Angelo Sifaleras
Fuhua Lin (Eds.)

LNCS 14799

Generative Intelligence and Intelligent Tutoring Systems

20th International Conference, ITS 2024
Thessaloniki, Greece, June 10–13, 2024
Proceedings, Part II

Part II

Springer

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Angelo Sifaleras · Fuhua Lin
Editors

Generative Intelligence and Intelligent Tutoring Systems

20th International Conference, ITS 2024
Thessaloniki, Greece, June 10–13, 2024
Proceedings, Part II

Springer

Editors
Angelo Sifaleras
University of Macedonia
Thessaloniki, Greece

Fuhua Lin
Athabasca University
Athabasca, AB, Canada

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-63030-9 ISBN 978-3-031-63031-6 (eBook)
https://doi.org/10.1007/978-3-031-63031-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 20th International Conference on Intelligent Tutoring Systems (ITS 2024) was held in Thessaloniki, Greece, from June 10 to June 13, 2024.

ITS 2024 is evolving to a new concept of Artificial Intelligence which can be found in various disciplines and can serve human education and well-being. This new concept is named: Generative Intelligence. Generative Intelligence concerns various AI systems, techniques, architectures, methods, or tools based on machine learning in particular deep learning which can be used to generate texts, images, solutions, and environments, able to enhance Human Intelligence, cognitive capacities, memory, and learning. Instead of trying to reproduce human behavior or recognize human traits using artificial intelligence techniques or learning analytics, Generative Intelligence provides means to increase human cognitive potential.

The conference featured eight distinct tracks in which the concepts mentioned above contribute to the increase of Generative Intelligence. Each track included (but was not limited to) a list of topics of interest which can be found below. Submitted papers refer to one of the tracks here below.

Conference tracks

T1: Generative Intelligence in Tutoring Systems

The goal of this track was to show how new techniques inspired by artificial intelligence and new methods in education can improve learning and teaching and generate the capacity for knowledge acquisition. The topics of this track include generative learning strategies, distance education, learning analytics for tutoring systems, deep learning and machine learning for tutoring systems, online and distance learning, generative learner models, emotion recognition, human-machine interaction, case-based reasoning, cognitive modeling, open learning, authoring systems, cultural learning, and adaptive learning.

T2: Generative Intelligence in Healthcare Informatics

The goal of this track was to show the progress of AI tools for increasing the propagation of healthcare techniques and their efficiency. Informatics provides means to improve the prediction, analysis, and treatment of disease and patients' control over their own care. The topics of this track were AI and telemedicine, medical image processing, virtual systems for healthcare, learning analytics in medicine, progress of AI for non-pharmacological Alzheimer's treatments, predictive modeling of healthcare, intelligent tutoring systems in medicine, machine learning and deep learning in healthcare, AI in medical education, AI in public health, home management of healthcare, neurofeedback techniques, games for healthcare, virtual reality (VR), and augmented reality (AR).

T3: Human Interaction, Games, and Virtual Reality

The goal of this track was to show the progress of interactive games using generative intelligence techniques. Intelligent games can adapt to the characteristics of the player and can be used to enhance learning, skills, memory, cognitive capacities, brain-computer interaction, and strategic decisions. They can be used in various applications (education, healthcare, group management, decision support systems, industry control). Multimedia allows an increase in the receptivity of sensors and reactions. The topics of this track included Brain-Computer Interaction (BCI), game design, intelligent immersive games, multi-agent systems, educational games, social games, generative simulations, theory of games, reinforcement learning in games, virtual and generative reality, simulation training, emotions recognition, neurofeedback games, generative scenario design, human interaction with games, multimedia technologies in games, fuzzy systems in games, artificial intelligence in games, and games content generation.

T4: Neural Networks and Data Mining

This track was a crucible for innovation, where the latest techniques in machine learning intersect with the rich, untapped data of educational environments, aiming to revolutionize the pedagogical landscape and pave the way for a future where intelligent tutoring systems are as nuanced and insightful as the educators they seek to augment. The topics of this track include supervised machine learning, genetic algorithms, Markovian regulation, smart sensor networks, determinate regulation, games and strategies, fuzzy systems, web information processing, applications of data mining in social sciences, data-driven reasoning, deep learning and statistical methods for data mining, big data mining, algorithms for data mining, ethical data analytics, and data mining for recommendation.

T5: Generative Intelligence and Metaverse

This track was dedicated to exploring the innovative synthesis of generative algorithms and the boundless educational landscapes within virtual environments. We delved into how generative AI is revolutionizing personalized learning experiences, creating dynamic content, and fostering engaging educational models that are as limitless as the Metaverse itself. Participants gained insights into the latest advancements, discussed the integration of AI-driven pedagogies in virtual spaces, and engaged with groundbreaking research that shapes the future of learning. Here, educators, technologists, and researchers came together to craft the nexus of next-generation learning platforms–where intelligence generation meets the expanse of the Metaverse, setting a new paradigm in digital education. The topics of this tracks included technology and creativity around Metaverse, gaming and interactivity, mixed reality and virtual world, social and digital identity, extended reality, digital art, social communication, applications of Metaverse in health, and global Metaverse.

T6: Security, Privacy, and Ethics in Generative Intelligence

As we step into an era where AI's capabilities to generate content are nearly indistinguishable from human output, we must also navigate the complex web of security challenges, privacy concerns, and ethical dilemmas that accompany these advancements. This track offered a multidisciplinary forum for examining the safeguarding of digital identities, the protection of intellectual property, and the moral imperatives guiding AI interactions in educational settings. Experts, scholars, and practitioners from around the globe converged to share their wisdom, debate best practices, and forge strategies to ensure that generative intelligence develops in a manner that is secure, respects privacy, and adheres to the highest ethical standards. Together, we will chart the course for responsible stewardship of AI technologies that enhance learning while honoring the trust placed in them by educators and learners alike. The topics of this track included commercial security, data privacy and security, web security, applied cryptography, authentication, identity management and biometrics, electronic payments, culture of ethics, business and human rights, diversity and inclusion in teaching and learning, environmental ethics, machine learning and security, cloud computing and data outsourcing security, mobile payments, security in games, security of peer-to-peer networks, security metrics, sustainability, language-based security, security and privacy for the Internet of Things, and socio-technical security.

T7: Generative Intelligence for Applied Natural Language Processing

This track was dedicated to unearthing and showcasing the transformative power of generative models that are reshaping the way we interact with language in computational settings. It served as a beacon for those who are leveraging these advancements to build sophisticated tutoring systems capable of understanding, generating, and personalizing language-based interactions. The topics of this track included language modeling, domain ontologies, computational linguistics, cognitive semantics, text mining, translation, question answering, dialogue systems, information retrieval, speech recognition and synthesis, discourse, machine translation, and lexical semantics.

T8: Generative Intelligence for Autonomous Robots and Learning

Generative intelligence with robots includes a variety of new criteria that provide more human characteristics to robots. Such elements concern emotions, mood, and facial expressions which give a more realistic interaction with humans. They transform robots into useful human-like companions. The topics of this track included but were not limited to emotional robots, voice recognition, intelligent agents, autonomous robots, planning and Goal reasoning, entertainments robotics, intelligent systems and robotics, applications of autonomous intelligent robots, sensors and vision systems for robots, generative exploration in hazardous situations, extraction of environment maps, robots in medicine, and teaching robots.

The call for scientific papers solicited works presenting substantive new research results in using generative artificial intelligence (GenAI), advanced computer technologies and interdisciplinary research for enabling, supporting, and enhancing human learning.

The international Program Committee consisted of 78 leading members (32 senior and 46 regular) of the Intelligent Tutoring Systems and AI communities, assisted by additional external reviewers.

Research papers came from 25 countries and were each reviewed by three reviewers through a double-blind process. ITS 2024 retained the strict and high standards in the review process that were established during the previous years, and which have rendered it a top-flight, rather selective, and high-quality conference. This year, 35 papers were accepted as full, whereas 28 papers were accepted as short. We believe that the selected full papers describe some very significant research and the short papers some very interesting new ideas.

The management of the review process and the preparation of the proceedings was handled through the EasyChair platform.

ITS 2024 had two outstanding Invited Speakers in the plenary sessions: Eftychios Protopapadakis (University of Macedonia, Greece), a renowned figure in Machine Learning, Optimization, and Computer Vision, and Gianna Martinengo, CEO of Didael Knowledge Technologies Services and President of Women &Tech ETS, specializing in Learning and AI (Milano, Italy). Both are leaders in different specialized areas of the ITS field. In addition, ITS 2024 had an excellent Tutorial Speaker: Panagiotis Fotaris (University of Brighton, UK), specializing in Digital Games and User Experience Design.

Moreover, ITS 2024 hosted two workshops, one under the title: Breaking Barriers with Generative Intelligence (BBGI'24) and another one under the title: Digital Transformation in Higher Education. Empowering Teachers and Students for Tomorrow's Challenges (Back2Basics). The program of ITS 2024 also included a panel on Potential Ways of Creative Use of Gen AI (e.g., ChatGPT) in Computer Science Departments.

In addition to the contributors mentioned above, we would like to thank all the authors, the members of the Program Committees of all tracks, the external reviewers, and the Steering Committee members as well as the Hosting Institution of the Conference, the University of Macedonia, in Thessaloniki, Greece.

We finally would like to acknowledge that ITS 2024 was held under the auspices of the Institute of Intelligent Systems (IIS) and was organized by Neoanalysis Ltd, under the guidance of Kitty Panourgia, the Organization Chair, and her team.

April 2024

Fuhua Lin
Angelo Sifaleras

Organization

Conference Committee

Program Committee Chairs

Oscar Lin	Athabasca University, Canada
Angelo Sifaleras	University of Macedonia, Greece

Program Advising Chairs

Maiga Chang	Athabasca University, Canada
Yugo Hayashi	Ritsumeikan University, Japan
Christos Troussas	University of West Attica, Greece
Seiji Isotani	Harvard University, USA

Organization Committee

Organization Chair

Kitty Panourgia	Neoanalysis Ltd., Greece

Organization Members

Maria Lafazani
Iias Petrodaskalakis
Kostantina Traka
Paula Dona
Isaak Tselepis

Program Committees

Senior Program Committee

Roger Azevedo	University of Central Florida, USA
Emmanuel Blanchard	Le Mans University, France

Benedict Du Boulay	University of Sussex, UK
Bert Bredeweg	University of Amsterdam, The Netherlands
Stefano A. Cerri	LIRMM, University of Montpellier, France
Maiga Chang	Athabasca University, Canada
Michel Desmarais	École Polytechnique de Montréal, Canada
Cyrille Desmoulins	Université Grenoble Alpes, France
Claude Frasson	University of Montreal, Canada
Sabine Graf	Athabasca University, Canada
Yugo Hayashi	Ritsumeikan University, Japan
Wu-Yuin Hwang	National Dong-Hwa University, Taiwan
Seiji Isotani	Harvard University, USA
Lewis Johnson	Alelo Inc., USA
Charalampos Karagiannidis	Aristotle University of Thessaloniki, Greece
Kinshuk	University of North Texas, USA
Vivekanandan Kumar	Athabasca University, Canada
Amruth Kumar	Ramapo College of New Jersey, USA
Susanne Lajoie	McGill University, Canada
Nian-Shing Chen	National Taiwan University, Taiwan
Alessandro Micarelli	Roma Tre University, Italy
Phivos Mylonas	Ionian University, Greece
Riichiro Mizoguchi	Japan Advanced Institute of Science and Technology, Japan
Roger Nkambou	Université du Québec à Montréal, Canada
Elvira Popescu	University of Craiova, Romania
Demetrios Sampson	University of Piraeus, Greece
Filippo Sciarrone	Roma Tre University, Italy
Marco Temperini	Sapienza University of Rome, Italy
Stefan Trausan-Matu	Politehnica University of Bucharest, Romania
Christos Troussas	University of West Attica, Greece
Julita Vassileva	University of Saskatchewan, Canada
Panagiotis Vlamos	Ionian University, Greece

Program Committee

Mohammed Abdel Razek	Al-Azhar University, Egypt
Galia Angelova	Bulgarian Academy of Sciences, Bulgaria
Renu Balyan	SUNY Old Westbury, USA
Maria Lucia Barron-Estrada	Instituto Tecnológico de Culiacán, Mexico
Azza Basiouni	Liwa College, UAE
Hamdi Ben Abdessalem	University of Chicoutimi, Canada
Jesus G. Boticario	UNED, Spain
Tharrenos Bratitsis	University of Western Macedonia, Greece

Ting-Wen Chang	Beijing Normal University, China
Chih-Yueh Chou	Yuan Ze University, Taiwan
Evandro Costa	Federal University of Alagoas, Brazil
Ali Akber Dewan	Athabasca University, Canada
Georgios Evangelidis	University of Macedonia, Greece
Nikolaos Fachantidis	University of Macedonia, Greece
Reva Freedman	Northern Illinois University, USA
Ashok Goel	Georgia Tech, USA
Xiaoqing Gu	East China Normal University, China
Ella Haig	University of Portsmouth, UK
Yusuke Hayashi	Hiroshima University, Japan
Dimitrios Hristu-Varsakelis	University of Macedonia, Greece
Qun Jin	Waseda University, Japan
Srecko Joksimovic	University of South Australia, Australia
Mizue Kayama	Shinshu University, Japan
Georgia Koloniari	University of Macedonia, Greece
Blair Lehman	Educational Testing Service, USA
Chao-Lin Liu	National Central University, Taiwan
Carla Limongelli	Roma Tre University, Italy
Pin-Chuan Lin	Mount Saint Vincent University, Canada
Frederick Li	University of Durham, UK
Mirko Marras	University of Cagliari, Italy
Wolfgang Müller	University of Education Weingarten, Germany
Silvia Margarita Baldiris Navarro	University of Cartagene, Colombia
Kuo-Liang Ou	National Hsin-Chu University of Education, Taiwan
Sasha Poquet	University of South Australia, Australia
Eftychios Protopapadakis	University of Macedonia, Greece
Valéry Psyché	TÉLUQ University, Canada
Ricardo Queiros	Agostinho Neto University, Portugal
Ioannis Refanidis	University of Macedonia, Greece
Flippo Sciarrone	Universitas Mercatorum, Italy
Riina Vuorikari	ALL DIGITAL, Belgium
Dunwei Wen	Athabasca University, Canada
Stylianos Xinogalos	University of Macedonia, Greece
Milos Kravcik	DFKI GmbH, Germany
Shaozi Li	Xiamen University, China
Theodore Kaskalis	University of Macedonia, Greece

Steering Committee

Chair

Claude Frasson	University of Montreal, Canada

Members

Stefano A. Cerri	LIRMM, University of Montpellier and CNRS, France
Maiga Chang	Athabasca University, Canada
Amruth Kumar	Ramapo College of New Jersey, USA
Yugo Hayashi	Ritsumeikan University, Japan
Isabel Fernandez-Castro	University of the Basque Country, Spain
Guy Gouardères	University of Pau and Pays de l'Adour, France
Alan Lesgold	University of Pittsburgh, USA
James Lester	North Carolina State University, USA
Alessandro Micarelli	Roma Tre University, Italy
Roger Nkambou	Université du Québec à Montréal, Canada
Giorgos Papadourakis	Hellenic Mediterranean University, Greece
Elliot Soloway	University of Michigan, USA
John Stamper	Carnegie Mellon University, USA
Daniel Suthers	University of Hawaii, USA
Stefan Trausan-Matu	University Politehnica of Bucharest, Romania
Christos Troussas	University of West Attica, Greece

Advisory Committee

Members

Luigia Carlucci Aiello	Sapienza University of Rome, Italy
Kinshuk	University of North Texas, USA
Demetrios Sampson	University of Piraeus, Greece

Workshop 1: Breaking Barriers with Generative Intelligence (BBGI)

Chair

Azza Basiouni	Liwa College Abu Dhabi, UAE

Senior Program Committee

Claude Frasson	University of Montreal, Canada
Oscar Lin	Athabasca University Canada
Kinshuk	University of North Texas, USA

Program Committee

Michael Lin	Mount Saint Vincent University, Canada
Xiaokun Zhang	Athabasca University, Canada
Khaled Shaalan	British University in Dubai, UAE
Ahmed Al-Gindy	Canadian University Dubai, UAE
Ray Al-Barazie	Liwa College Al Ain, UAE
Beenu Mago	Skyline University College, UAE
Rasha Hasan	Liwa College Abu Dhabi, UAE
Said Salloum	University of Salford, UK

Workshop 2: Back2Basics

Chairs

Anastasios Economides	University of Macedonia, Greece
Maria Perifanou	University of Macedonia, Greece

Program Committee

Alexandros Mikroyanidis	Open University, UK
Carlos Santos	University of Aveiro, Portugal
Tobias Ley	Danube University Krems, Austria
Maka Eradze	University of Aquila, Italy
Stavros Nikou	University of Strathclyde, UK
Rita Santos	University of Aveiro, Portugal
Luis Pedro	University of Aveiro, Portugal
Panagiotis Kosmas	University of Limassol, Cyprus

Invited Talks

Unleashing Potential: Harnessing the Power of Generative AI in Intelligent Tutoring Systems

Eftychios Protopapadakis

Department of Applied Informatics, University of Macedonia,
156 Egnatia Street, GR-546 36 Thessaloniki, Greece
eftprot@uom.edu.gr

Abstract. Intelligent Tutoring Systems (ITS) fueled by Generative AI (genAI) provide an exceptional opportunity to revolutionize education. Offering personalized learning experiences, tailored to individual needs and preferences, is not an easy task, yet it becomes easier by the year. In this keynote, we explore how genAI-powered ITS can transcend the limitations of traditional education models, paving the way for a new era of cognitive empowerment and skill development.

Through advanced algorithms and adaptive learning techniques, genAI-powered ITS can deliver targeted instruction, personalized practice problems, and immersive simulations that foster critical thinking and problem-solving skills. By leveraging the unique capabilities of generative AI, educators can further boost the human cognitive potential, moving beyond rote memorization to cultivate deep understanding and mastery.

However, alongside the promise of genAI-powered ITS come inherent challenges that must be addressed. Biases embedded within AI algorithms and concerns regarding data privacy and security demand careful consideration. Yet, by proactively addressing these challenges and fostering collaboration between AI developers, educators, and policymakers, we can ensure the responsible and ethical application of genAI in education.

The importance of collaboration and partnership between AI and human educators is a topic worthy of investigation. By combining the expertise of AI algorithms with the guidance and mentorship of teachers, we can create a symbiotic relationship that enhances the learning experience and maximizes student outcomes. Together, we can envision a future where genAI-powered ITS play a central role in democratizing education, empowering learners of all backgrounds to reach their full potential and contribute meaningfully to society.

Keywords: Intelligent Tutoring Systems · Generative AI · Personalized Learning · Cognitive Potential · Educational

Sharing from Experience: Competencies for "Intelligent Dialogues" with Emerging Technologies

Gianna Martinengo

CEO and Founder of Didael KTS and President of Women&Tech® ETS,
DKTS: Digital Knowledge Technologies Services, Via Ampère 61/A,
20131 Milan, Italy
gianna.martinengo@didaelkts.it

Abstract. We have witnessed the various phases of technologies supporting human learning and at the same time of artificial intelligence since 1982, mainly by means of business initiatives, within hundreds of projects of private and public interest including many EU pre-competitive ones. Lessons learned and tips for the future are at the core of our talk. The main issues we privilege are knowledge, competencies, and ethics for the future of modern, interactive AI including human learning as a first priority. As a side effect, we will shortly describe our efforts for women, by demonstrating the interest of a true complementarity between genders, as a source of innovation in society: future business as well as academic initiatives. Last, we briefly quote our contribution to the EU Parliament for the AI act recently approved. We are proud to see the initial implementation of an AI office, dedicated to the certification of AI products and services, as we proposed, motivated and documented in our written reports in 2021.

Keywords: digital transformation · AI · ethics · change management · evolving society · women's empowerment

Contents – Part II

Neural Networks and Data Mining

Generative Intelligence and Metaverse

Contents – Part I

Generative Intelligence and Healthcare Informatics

Elevating Medical Efficiency and Personalized Care Through the Integration of Artificial Intelligence and Distributed Web Systems

Marian Ileana(✉)

Interdisciplinary Doctoral School, National University of Science and Technology POLITEHNICA Bucharest, Pitesti University Center, Pitesti, Romania
marianileana95@gmail.com

Abstract. In the contemporary context of the medical field, the collaboration between technology and classical medicine becomes essential to improve the efficiency of the medical act and to be able to provide personalized care using the latest available technologies. The healthcare industry faces a number of challenges, including rising costs and a shortage of healthcare professionals. Effective and personalized medical treatment is one of the essential goals of healthcare systems spread across the globe. Distributed web infrastructure can help improve efficiency and reduce costs by simplifying communication and collaboration between healthcare providers. Distributed web systems offer a major new opportunity to improve the field through more effective interconnection and collaboration between healthcare providers and patients. Distributed web infrastructure improves data privacy and security by eliminating sensitive points. By allowing patients to share specific information with healthcare providers while maintaining privacy, they have more control over their medical data. This model of approach contributes to a model of patient-oriented medical care rather than distributing attention to the shortcomings of the information system. Decentralized data repositories and collaborative platforms can accelerate the pace of medical discovery and promote more efficient clinical trials. Integrating a distributed web infrastructure in healthcare promises to deliver personalized care and improve medical practice. This shift in direction towards decentralization will provide new opportunities for innovation, collaboration, and improving patient outcomes in the context of improved healthcare services.

Keywords: Distributed web systems · Healthcare informatics · Medical efficiency · Personalized care · Artificial intelligence

1 Introduction

Public health informatics encompasses the communication, surveillance, information, and learning systems relevant to public health, as well as their conceptualization, design, development, implementation, evaluation, and maintenance. Public health informatics requires a broad range of knowledge from a multitude of fields, including information and

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 3–11, 2024.
https://doi.org/10.1007/978-3-031-63031-6_1

communication technology, computer science, management, organizational theory, psychology, communications, political science, and law. In practice, these systems require knowledge in the fields of epidemiology, microbiology, toxicology, and statistics [1].

Public health informatics is distinguished from other types of informatics by the following characteristic [2]:

- The focus is on using the applications of science and information technology to improve the health of the population as a whole, not the individual.
- The priority is disease prevention rather than treatment.
- The focus is on prevention at all points that lead to illness, injury, or disability.
- The operation of such a system is typically done in a government setting, not a private setting.

As a result of increasing medical problems, hospitals have limited capacity, and there is an acute shortage of medical professionals. These problems enable the use of tools such as IoT, artificial intelligence, machine learning, and data analysis [3].

In the era of Industry 4.0, the application of IT technologies in the healthcare industry has become highly valued. Each segment has a significant impact on the growth and development of a nation's economy and plays a significant role in making a difference [4].

In terms of improving the accessibility, security, and personalization of health services, the distributed web infrastructure, which is characterized by the storage and processing of data in decentralized locations, plays an important role [5].

The adoption of advanced healthcare information systems and medical informatics requires an integrated approach highly sensitive to different social, economic, political, and cultural factors [6].

Healthcare reform, advances in information technology, and the emergence of big data [3]. Continued innovation represents changes that lead medical professionals to stay connected with the development of public health informatics systems.

2 Literature Review

2.1 Significant Works in the Field of Study

In an article by Yogesh et al. [7], in today's context with a significant amount of uncollected data, medical informatics is gaining ground. The article discusses trends and future directions in medical informatics for successful application in public health assurance. With the advancement of technology, healthcare facilities face new issues that need to be addressed by adopting appropriate policies and standards.

In a study by Sijm-Eeken et al. [8], the main objective was to develop a theory-based framework to be able to improve and accelerate the development, selection, and implementation of solutions that reduce the impact climate of medical organizations. In conclusion, it is suggested that research should assess its applicability in measuring and evaluating the impact of green health informatics solutions on environmental sustainability and climate resilience.

Research by Alanazi [9] emphasizes the importance of information technology in the context of the significant changes that the system has recently undergone. In order to properly prepare future health care providers, the author argues that it is essential to introduce information technology courses into health science programs. Experts identify seven key areas, stressing the need for introduction and the fact that they play a crucial role in producing competent graduates in the health sciences. Finally, the experts developed and validated the proposal.

2.2 Authors and Research Groups Analysis

By applying a dynamic analysis method to the subject of medicine, especially medical informatics, the specialized sites offer a diversity of publications signed by a large number of authors.

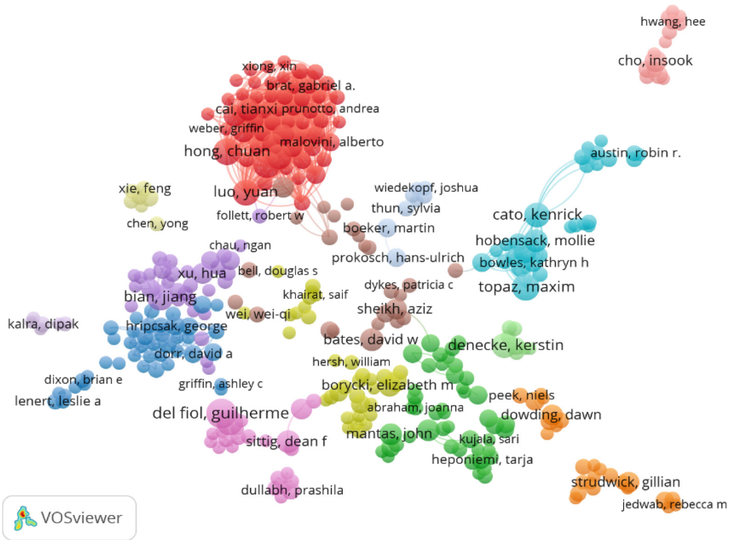


Fig. 1. A graphic map of authors for the search term "Healthcare Informatics"

In order to conduct this empirical study, the "Dimensions.ai" platform was used as a database, and its search functionality was used. The obtained results were interpreted using the VOSviewer application to easily identify the authors who published on this topic, with a graph as output. The group of identified authors represents distinct groups within an author. The dataset contains 12452 authors. An author's relevance is determined by appearing as an author in at least three papers, and that author also had to have at least two citations. A total of 497 people met these conditions (see Fig. 1).

Using the same data set, we also made a map of the countries from which these authors who write articles on medical informatics come from. For a country to be considered relevant, it had to have at least three papers, each of which had at least two citations. The total number of countries from which the authors originated is 97, after applying the previously stated conditions, 69 countries met the conditions (see Fig. 2).

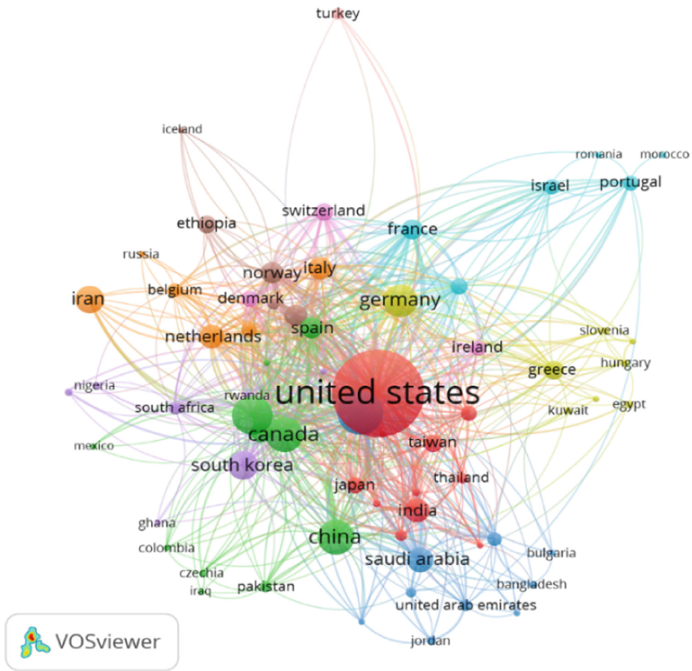


Fig. 2. A country map of authors for the search term "Healthcare Informatics"

3 Method

3.1 Distributed Web Systems Architecture

Distributed web systems are a type of computing architecture in which data processing (processing and storage) capabilities are spread across multiple nodes or locations, often geographically dispersed [10]. Features for which distributed web systems are used [11]: decentralization, scalable, data replication, load balancing, fault tolerance, interconnected nodes, parallel processing.

A Distributed Web Systems includes locally distributed server nodes that allow client applications to see multiple IP addresses [12]. A diagram of the architecture is available in Fig. 3. The user wants to access www.website.com. Initially, the Client makes a request to the Local DNS Server to decode the domain into an IP.

When the DNS Server connects to the Authoritative DNS Server, it responds with the IP 181.12.34.2, which is sent via the Local DNS Server back to the Client's laptop. Through an HTTPS request through the IP, it connects to Web Server 2, from where it retrieves the content desired by the user. Because DNS plays a very important role for Distributed Web Systems, authoritative DNS is included in the system box. The DNS mechanism assigns the Client's request to a target server during the address lookup process and the name of the website in the search phase [12].

A National Health System is a system of people, institutions, and resources that provide health care services to meet the health needs of a country's population [13]. It includes both primary health care services, such as those provided by family medicine, and secondary and tertiary care services, such as those provided by hospitals and private clinics [14].

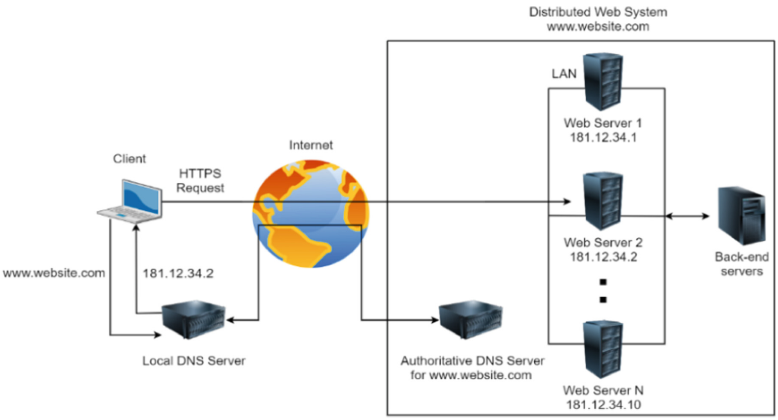


Fig. 3. Distributed Web Systems architecture diagram

Nowadays, the health system has become an essential component of national security, promoting the health of citizens and thereby contributing to the support of the population [15]. Consequently, each state must build a sanitary system that guarantees the health status of the population, because this is essential for improving the life and longevity of the population [16].

An efficient system developed on a distributed web system architecture can bring energy efficiency, resulting in a cost reduction on usage [17]. Through a properly designed IT system, it can help improve the quality of medical care by providing up-to-date medical information to support medical staff in making medical decisions and also to monitor the results of treatments and patients [18].

The proposed architecture is developed on the architecture realized in Fig. 3. It is developed around Data Centers: Data Center 1, Data Center 2, and Data Center 3. As can be seen in Fig. 4. The Data Centers are interconnected through a LAN network, which is connected to the Back-end Server. All data is accessed through the National Data Repository, and with the help of the DNS Server, the address of the Data Center that will receive the request is obtained.

In the upper left part of Fig. 4, you can see the medical area at the level of a county. Within it we have Public Hospitals, Family Medicine, and Emergency Medicine that exchange data with the help of a server dedicated to the County Health Department. The County Health Department server will connect to the National Data Repository after passing two FireWalls (one at the exit from the county network and the other before connecting to the National Data Repository). Between the two FireWalls, a connection will be made WAN.

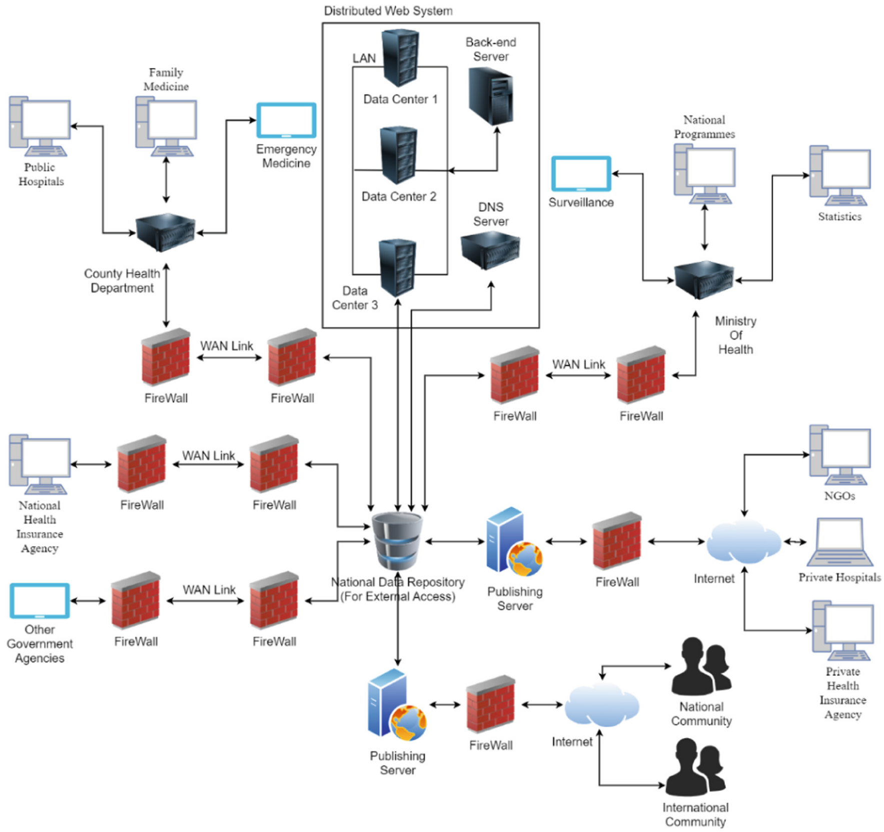


Fig. 4. Diagram of the national public health system's architecture

In the lower left part of Fig. 4, we find the National Health Insurance and Other Government Agencies connecting to the externally accessible server using the same method by going through two FireWalls.

In the upper right Fig. 4, we have a structure similar to that of the County Health Department, only made around the Ministry of Health, which has three connections: Surveillance, National Programs, and Statistics. The connection between the server at the Ministry and the server for external data uses the same connection procedure, passing through two FireWalls.

In the middle right part of Fig. 4, there are: NGOs, Private Hospitals, and Private Health Insurance, which connect via the Internet and will pass the security checks through a FireWall, and the data will reach a Publishing Server. The second layer of FireWall is not needed because the data stops on a server, only publishing the data in the system. Through a Publishing Server, they only have the right to read and write new data without deleting or altering existing data.

The National Community and the International Community can connect via the Internet, they will pass the checks of the FireWall, and the data will reach the Publishing Server.

The system is designed to be used to create an electronic health record. It will store medical information about patients, including medical history, test results and medical prescriptions. The registry will be located in the Data Center (Fig. 4). Such a registry helps to improve the coordination of medical care and reduce medical errors [16]. Its design is also oriented towards telemedicine support. Through this form of consultation, medical services can be offered at a distance via a smartphone, laptop, or tablet. The system can transmit medical images to support video consultations between patients and doctors (it can be seen in Fig. 4 at National Community). The system also supports the medical research area (via Statistics Fig. 4). The collected data is used to identify trends and develop new medical treatments.

3.2 Personal Care Using Artificial Intelligence

The previously integrated system is perfectly prepared to interface with AI technologies. In general, "Artificial Intelligence" (AI) refers to the capacity for a computer to imitate intelligent behaviors without the need for additional human intervention. Artificial intelligence is generally believed to have begun with the invention of robots [19]. The word "robota" is Czech in origin, which means biosynthetic machines used in forced labor. In this field, an important legacy is that left by Leonardo da Vinci regarding robot-assisted surgery, which bears his name. Da Vinci's sketches of robots helped set the stage for this innovation [20]. In 1495, Leonardo da Vinci's metal-clad warrior became the first robot to mimic human arm, jaw, and neck movements. His invention was a great source of inspiration for this trend.

Artificial intelligence, described as the science and engineering of making intelligent machines, was officially born in 1956. The phrase refers to a broad variety of aspects of medicine, such as medical diagnosis, medical surveillance, and medical statistics [19]. There are two primary areas of AI in medicine: virtual and physical. The virtual branch encompasses informatics approaches, from information management through generative learning to control health management systems, including electronic health records and the active guidance of physicians in their treatment decisions. The physical branch is used to assist surgeons during operations. Nanorobots are also included in this branch. The ethical and moral complications of these technologies require deeper attention, focusing on medical utility, economic value, and the need to develop interdisciplinary strategies for wider application. This paper will deal with the virtual branch.

With the help of the implemented system and through the opportunities offered by artificial intelligence, a number of aspects of medical care can be personalized [19]:

- Diagnosis: AI has the ability to identify abnormalities in medical images, such as X-rays and CT scans, to provide a more accurate diagnosis. In this way, conditions can be identified in time so that a short-term treatment is necessary without complications occurring.
- Treatment: AI can provide individualized treatments based on medical history, genetic data, and other factors. Thus, reducing the cost of treatments that do not work for the patient.
- Patient monitoring: AI can be used to continuously monitor patients' health and prevent complications. Monitoring can be done even through the patient's smart phone. It continuously sends data to the medical system about the user's lifestyle.

4 Results

By implementing the proposed architecture, it has the potential to bring major changes to the medical act. This is because the distribution enables access to medical information and resources in a fast and secure way. This will lead to a reduction in waiting time for patients, thus increasing accessibility to medical services.

Among the most important aspects of such a system [17]:

– Optimizing data flow: This type of infrastructure improves coordination and accelerates the exchange of medical information between system actors.
– Advanced security: Distributed infrastructure provides a high level of security.
– Administrative efficiency: Helps manage healthcare data and processes efficiently, reducing red tape and costs.
– Patient engagement: A secure web platform allows patients to access and participate in the management of their health information.

The proposed system is ready for integration with the latest technologies. Thus, the integration of healthcare personalization with AI into systems is a growing trend in healthcare [18]. Artificial intelligence has the potential to significantly improve the efficiency and quality of healthcare by providing healthcare tailored to the individual needs of each patient.

5 Conclusion

In conclusion, this paper has proposed an IT architecture for an efficient medical system. Such a system represents a significant evolution in improving the efficiency and quality of medical care [1]. This architecture will enable better collaboration between healthcare providers, increased accessibility to medical information, and personalized care. The latest AI innovations that help improve personalized healthcare were also discussed. This technology can be used in multiple ways, including analyzing large data sets to identify patterns that can be used to predict disease risk, customize treatment plans, and ultimately improve medical outcomes.

In today's medical context, the proposed IT architecture, together with the use of AI, has the potential to revolutionize the media system [19]. Through collaboration with specialists in the field of information and communication technology and the right investment. This would lead to efficient, affordable, and personalized healthcare that would benefit patients and healthcare providers alike.

References

1. Rahimi, B., Nadri, H., Afshar, H.L., Timpka, T.: A systematic review of the technology acceptance model in health Informatics. Appl. Clin. Inform. **09**, 604–634 (2018). https://doi.org/10.1055/s-0038-1668091
2. Ammenwerth, E., Nykänen, P., Talmon, J.L., Brender, J.: Factors influencing success and failure of health informatics systems. Methods Inf. Med. **45**, 125–136 (2006). https://doi.org/10.1055/s-0038-1634049

3. Pramanik, M., Lau, R.Y.K., Azad, M., Hossain, M., Chowdhury, M., Karmaker, B.K.: Healthcare informatics and analytics in big data. Expert Syst. Appl. **152**, 113388 (2020). https://doi.org/10.1016/j.eswa.2020.113388
4. Paul, S., et al.: Industry 4.0 applications for medical/healthcare services. J. Sens. Actuator Netw. **10**, 43 (2021). https://doi.org/10.3390/jsan10030043
5. Hassan, N.H., Maarop, N., Ismail, Z., Abidin, W.Z.: Information security culture in health informatics environment: a qualitative approach. In: International Conference on Research and Innovation in Information Systems (ICRIIS) (2017). https://doi.org/10.1109/icriis.2017.8002450
6. Balta, M., Valsecchi, et al.: Digitalization and co-creation of healthcare value: a case study in occupational health. Technol. Forecast. Soc. Change **168**, 120785 (2021). https://doi.org/10.1016/j.techfore.2021.120785
7. Yogesh, M., Karthikeyan, J.: Health informatics: engaging modern healthcare units: a brief overview. Front. Public Health **10** (2022). https://doi.org/10.3389/fpubh.2022.854688
8. Sijm-Eeken, M., Arkenaar, W., Jaspers, M.W.M., Peute, L.W.P.: Medical informatics and climate change: a framework for modeling green healthcare solutions. J. Am. Med. Inform. Assoc. **29**, 2083–2088 (2022). https://doi.org/10.1093/jamia/ocac182
9. Alanazi, A.: Interprofessional education in health informatics (IPEHI) for health sciences programs. Adv. Med. Educ. Pract. **14**, 1177–1182 (2023). https://doi.org/10.2147/amep.s422725
10. Cardellini, V., Casalicchio, E., Colajanni, M., Yu, P.S.: The state of the art in locally distributed Web-server systems. ACM Comput. Surv. **34**, 263–311 (2002). https://doi.org/10.1145/508352.508355
11. Viral, R., Khatod, D.K.: Optimal planning of distributed generation systems in distribution system: a review. Renew. Sustain. Energy Rev. **16**, 5146–5165 (2012). https://doi.org/10.1016/j.rser.2012.05.020
12. Sucurovic, S.: Implementing security in a distributed web-based EHCR. Int. J. Med. Inform. **76**, 491–496 (2007). https://doi.org/10.1016/j.ijmedinf.2006.09.017
13. Haldane, V., et al.: Health systems resilience in managing the COVID-19 pandemic: lessons from 28 countries. Nat. Med. **27**, 964–980 (2021). https://doi.org/10.1038/s41591-021-01381-y
14. Jarrett, J.P., et al.: Effect of increasing active travel in urban England and Wales on costs to the National Health Service. Lancet **379**, 2198–2205 (2012). https://doi.org/10.1016/s0140-6736(12)60766-1
15. Aldis, W.: Health security as a public health concept: a critical analysis. Health Policy Plan. **23**, 369–375 (2008). https://doi.org/10.1093/heapol/czn030
16. Wang, B.B., Wan, T.T.H., Burke, D., Bazzoli, G.J., Lin, B.Y.-J.: Factors influencing health information system adoption in American hospitals. Health Care Manag. Rev. **30**, 44–51 (2005). https://doi.org/10.1097/00004010-200501000-00007
17. Ileana, M.: Optimizing energy efficiency in distributed web systems. In: 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS) (2023). https://doi.org/10.1109/isas60782.2023.10391617
18. Tiwana, A., Konsynski, B.R.: Complementarities between organizational IT architecture and governance structure. Inf. Syst. Res. **21**, 288–304 (2010). https://doi.org/10.1287/isre.1080.0206
19. Johnson, K.B., et al.: Precision medicine, AI, and the future of personalized health care. Clin. Transl. Sci. **14**, 86–93 (2020). https://doi.org/10.1111/cts.12884
20. Yates, D.R., Vaessen, C., Rouprêt, M.: From Leonardo to da Vinci: the history of robot-assisted surgery in urology. BJU Int. **108**, 1708–1713 (2011). https://doi.org/10.1111/j.1464-410x.2011.10576.x

Human Interaction, Games and Virtual Reality

Cognitive Engagement Detection of Online Learners Using GloVe Embedding and Hybrid LSTM

Dharamjit Parmar, M. Ali Akber Dewan(✉), Dunwei Wen, and Fuhua Lin

School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University, Edmonton, Canada
dparmar1@learn.athabascau.ca, {adewan,dunweiw, oscarl}@athabascau.ca

Abstract. This paper presents a method for classifying discussion posts of online courses, aiming to improve students' cognitive engagement in online learning. This method utilizes deep learning models including a GloVe embedding and a hybrid long short-term memory (LSTM) network within an educational framework called interactive, constructive, active, and passives (ICAP), which classifies the students posts into interactive, constructive, active, and passives classes. These classes quantify the students' level of cognitive engagement in their online course discussion posts. We used textual attributes and label-specific characteristics (e.g., text length, sentiment polarity, and subjectivity) to gain comprehensive insights into the forum posts' emotional and cognitive depth. We further refined these features with a pre-trained GloVe embedding, which enhanced the classification accuracy. We interpreted the model's decision-making process using local interpretable model-agnostic explanations (LIME) for added transparency and interpretability. By employing deep learning models with ICAP and LIME, this study demonstrates an effective use of the proposed system for improving student cognitive engagement in online learning.

Keywords: Online learning environment · GloVe embedding · deep learning · cognitive engagement · ICAP framework · LSTM

1 Introduction

The evolution of online learning has reshaped the educational landscape of higher education [1]. The reach of online courses has become both geographically and culturally borderless. Students from different countries and cultural backgrounds are contributing their unique perspectives to these virtual learning environments. In online courses, course discussion forums play an important role in improving students' engagement with the instructors and their peers. Because of discussion posts, students do not only learn from course materials and lectures, but they also get the opportunity to learn from interacting with the others [2]. Online discussion helps improve students' critical thinking and problem-solving skills, decision making ability, communication skills and their

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 15–26, 2024.
https://doi.org/10.1007/978-3-031-63031-6_2

ability to organize and analyze information [3]. The asynchronous form of online course discussion posts helps to reveal students' behavioral and cognitive intelligence [4]. The form of discussion posts mimicking the traditional classroom interactions in the online courses while offering significant advantages by eliminating the limitation imposed by time and physical presence [5].

Course discussion forums enable educators to derive valuable pedagogical insights from students' posts, which reflect their cognitive capabilities. Discussion posts contain rich and multi-level cognition-related behavior patterns, which bring the potential for an in-depth investigation into the varying trend of a group and individual students' cognition processing through forum content analysis [6]. For example, posts showing concept application, critical analysis, or idea synthesis indicate higher cognitive engagement. Although discussion posts are useful to detect students' cognitive engagement, automating this detection process is still challenging. Employing coding schemes from educational frameworks (e.g., ICAP [7]) further open the opportunities for this type of analysis. The ICAP is grounded in the learning theory [8], which helps identifying engagement and interaction patterns and customizing learning. This framework informs on engaging course elements, areas needing support, impacting teaching methods on cognitive development, guiding curriculum, and refining instructional strategies.

In this study, we investigated the potential of using deep learning and natural language processing (NLP) techniques with the ICAP framework to identify the level of students' cognitive engagement for analyzing course discussion forum. We explored two main research questions, *RQ1: How can we automate cognitive engagement detection for analyzing course discussion forums? RQ2: What insights can we get from students' cognitive engagement detection?* We used GloVe embeddings, enhanced feature engineering and a hybrid LSTM model to classify discussion posts into ICAP's categories. We compared the performance of hybrid LSTM, standard LSTM, support vector machine (SVM), random forest (RF), K-nearest neighbor (KNN), logistic regression (LR), and gradient boosting (GB), where the hybrid LSTM with the enhanced features performed the best. This demonstrates the potential of using NLP and deep learning techniques in automating students' cognitive engagement detection for analyzing course discussion forum.

The rest of the paper is organized as follows. Section 2 discusses related works to this research study. Section 3 discusses the proposed framework. Section 4 presents the experiments and results. Finally, we conclude the paper with some future research directions in Sect. 5.

2 Related Work

Understanding students' cognitive engagement is critical in online learning for creating effective learning strategies. The ICAP framework, central to this study, analyzes learning behaviors to illuminate student engagement in digital settings, providing a detailed understanding of learning dynamics. The advancement in deep learning and NLP techniques facilitate educational researchers a deeper analysis of cognitive engagement and allows for personalized educational approaches to cater to individual student needs.

The ICAP framework suggests that interactive activities yield better learning outcomes than constructive or active ones [7]. This framework, based on research and case studies, provides a theoretical base to enhance learning activities. However, the study's focus on theory over practical application and the absence of empirical testing and tools suggest areas for further exploration and development. Gorgin et al. [9] assessed cognitive engagement in online forum using some traditional machine learning techniques with ICAP and Bloom's taxonomy. It explored linguistic and contextual features using decision tree, random forest, and SVM classifiers, highlighting the SVM's effectiveness. This study did not explore word embedding and deep learning techniques. Considering factors like course structure and facilitation method that affect engagement could further improve this study as well. Yogev et al. [10] created an automated system to classify cognitive engagement in the Nota Bene forum, showing higher cognitive engagement correlates with learning gains and helps improving course design and discussions. It visualizes cognitive engagement and highlights differences in different engagement levels. However, this system lacks detail on the classification process and emphasizes broad cognitive engagement categories over specific types. Hayati et al. [11] developed a system to categorize learners' cognitive engagement in forum threads, which helps instructors to reduce online learners disengagement. However, the classification of this system has not been backed up with any educational framework, such as ICAP [7] and CoI [12].

Several studies explored cognitive engagement detection using deep learning techniques. Liu et al. [13] examined the relationship between emotional and cognitive engagement in MOOC forums using a BERT-CNN model. It found a strong correlation between positive/confused emotions and high-level cognitive engagement, which is predictive of learning success. Wang et al. [14] investigated the relationship between forum discussion quality and learning outcomes using the ICAP framework and a LDA topic modeling on a psychology MOOC and found that the social topic discussions promoted higher-order thinking, unlike technical discussions, with off-topic conversations. Authors identified several challenges including limited availability of data for a detail engagement categorization and the inability to assess cognitive engagement outside forums, and recommended wider coding manual use and boosting of the forum participation to get for a deeper insight. Atapattu et al. [8] analyzed cognitive engagement in teacher PD MOOCs using word embeddings and the ICAP framework, with Doc2vec converting posts into vectors to assess engagement. It found three engagement types: active, constructive, and a blend, influenced by MOOC activities where reflection fostered constructive, and discussions encouraged active engagement. Liu et al. [15] automated cognitive engagement classification in online discussions using CoI and ICAP frameworks, excluding the "passive" category for a low engagement employing a semi-supervised learning with LIWC and BERT features. In this study, we investigated feature engineering, various machine learning models including deep learning, and NLP techniques, to improve the ICAP categorization and detection of the levels of students' cognitive engagement in course discussion forums.

3 System for Cognitive Engagement Detection

The proposed system uses deep learning models and NLP techniques with ICAP framework for students' cognitive engagement detection, analyzing posts in course discussion forum. Within the framework, the preprocessing steps, such as converting to string, and removing punctuation, numbers, and numbers have been applied for data cleaning and refinement. In the feature extraction step, GloVe embedding and a feature engineering technique have been applied to extract linguistic, contextual, and semantic features to improve student engagement detection with the ICAP categories. For the classification, the proposed system employs a hybrid LSTM within a modular architecture. Hybrid LSTM was chosen because of its ability to handle text sequence, context, and semantic comprehension. Model transparency and interpretability are described by using local interpretable model-agnostic explanations (LIME). The modular architecture of the framework emphasizes scalability and adaptability, making it suitable for various online learning platforms, which also makes it a valuable tool in the dynamic field of online education. The architecture of the proposed framework is shown in Fig. 1.

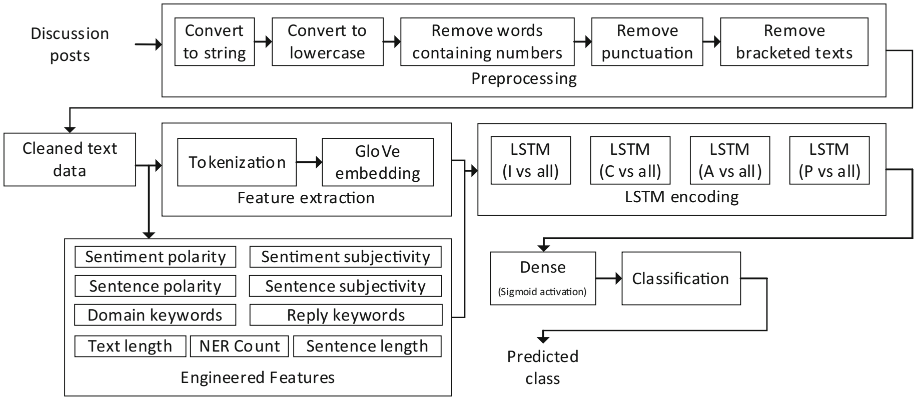


Fig. 1. Architecture of the proposed framework

3.1 Text Pre-processing and Feature Engineering

We applied standard text preprocessing steps, such as removing special characters, lowercasing, tokenization, and removing punctuation and stop words for the data cleaning. After pre-processing the texts, we applied a feature engineering step to extract key information from the non-tokenized text. This step analyzes text attributes that reflect broad textual properties instead of specific words or tokens.

The feature engineering step involves *domain-specific* and *discussion-oriented* feature extractions, each serving a distinct purpose in the text analysis. The *domain-specific* features (e.g., domain-keywords and reply-keywords) focus on capturing unique terminologies and elements relevant to a particular field, tailoring the analysis to the specific lexical context of the domain. The *discussion-oriented* features quantify the text's discourse structure and dynamics using different metrics to numerically capture the discussion's style and nature. The *discussion-oriented* features are further be categorized into *universal* features (e.g., sentence length, text length, named entity recognition (NER) count) and *label-specific* features (e.g., sentence polarity, sentence subjectivity, sentiment polarity, sentiment subjectivity). The *universal features* are suitable for detection of all the ICAP labels, whereas the *label-specific features* are suitable to enhance the detection of certain distinctive labels. For example, sentiment analysis plays an important role for subjectivity and polarity classification. The subjectivity classification classifies a given text into objective or subjective, i.e., separating facts from feelings, views, or beliefs, where as the polarity classification determines whether a text entails a positive or negative connotation [16].

3.2 GloVe Embedding

We extracted another set of features by applying GloVe embedding on the words after tokenization. Tokenization breaks text into meaningful elements (tokens). GloVe is an unsupervised learning technique for generating word embeddings [17, 18]. It involves creating a dictionary from sentences, with unique integers assigned to each word. By clustering semantically similar words, GloVe enhances model accuracy for related terms and offers robustness to new words through extensive dataset training [18]. It efficiently handles the higher-dimensional nature of texts by representing words in a lower-dimensional space, easing the training process. GloVe's unique weighting mechanism emphasizes word importance by maintaining semantic relationships and proximity, making it essential for complex text classification tasks and capturing the nuanced relationship between language and meaning.

In this study, we also use an autoencoder to fine-tune GloVe embeddings, adapting these pre-trained models to our dataset's unique linguistic need. This enhances the embeddings by improving feature relevance and minimizing noise. This fine-tuning also preserves GloVe's semantic integrity which boost both the efficiency and interpretability of language processing. After extracting features, applying GloVe embedding followed by fine tuning, and appended with the *domain-specific* and *discussion-oriented* features extracted in the previous step, we created a comprehensive feature set. These features are standardized to similar ranges, which facilitate to train the deep learning model of this proposed framework.

3.3 Model Architecture

For the classification task, we used a modular architecture of double-layer LSTM network [19], where each LSTM detects a single cognitive level as shown in Fig. 1. In this architecture, we used one-vs-all training approach for each module to classify the forum posts according to the ICAP framework. This modular architecture allows a balance between model complexity and generalization to achieve a high prediction accuracy. This also gives the flexibility of adding any new module with a new label or removing an existing one without interrupting the other modules in the network.

In this architecture, we used hybrid LSTM for the classification. LSTM has the form of a chain of repeating modules of neural networks. The modules have four neural network layers interacting in a very special way, which removes or adds information to the cell state by carefully regulating some structures called gates. Gates are a way to optionally let information go through. They are composed out of a sigmoid neural network layer and a pointwise multiplication operation [19]. The LSTM plays important roles in adopting complex sequential patterns and providing context-based interpretations of the input text. LSTM manages long-term dependencies through gating mechanisms and maintaining a 'cell state' for selective memory retention or deletion. This functionality grasps sentences' context and semantics.

To prevent overfitting, we used recurrent dropout and L2-regularization. The dropout randomly disables a fraction of input units during training to prevent dependency on certain features and improve generalization. The recurrent dropout applies this principle to the network's recurrent connections. L2-regularization adds a penalty based on the square of the weights to the loss function, which encourages smaller weights, reduces network's complexity, and mitigates overfitting. The output layer of the hybrid LSTM network features a dense layer with a single unit and a sigmoid activation function, ideal for binary classification. The sigmoid function outputs values between 0 and 1. We used a threshold 0.5, where the output value above 0.5 is represented as '1', and below 0.5 as '0'. For optimization, we used RMSprop optimizer with sparse data.

4 Experiment and Results

4.1 Dataset Preparation

We used Stanford's MOOC dataset [20] in our experiment. The labels in the original dataset are different than we need. We labeled the dataset manually by following the ICAP coding scheme presented in [7, 9]. This data annotation was done by three experts. Initially, two of the experts labeled the dataset individually without interacting with each other. Then a third expert compared the labels assigned by the two experts and discussed how to come into an agreement, where disagreement happened. If the two experts could not come to an agreement, the post was removed to avoid any further confusion. We considered four categories for the engagement labels: *social*, *active*, *constructive*, and *interactive*. While the original ICAP [7] included a category called, *passive* posts, we elected not to use this category, as this category could imply a lack of student involvement in generating discussion posts. We rather incorporated the category of *social* engagement as suggested by Gorgun et al. [9]. Tables 1 and 2 illustrate the coding scheme of the ICAP framework and the number of posts with different engagement levels, respectively.

Table 1. ICAP coding scheme [7, 9]

Category	Characteristics
Social (s)	Social engagement is characterized by being unrelated to the main course content or topics. It encompasses a variety of non-academic interactions, such as addressing administrative matters, sharing personal interests, exchanging greetings, introducing oneself, and expressing acknowledgments or gratitude
Active (a)	This engagement category is defined by activities involving comprehension and recall. It is marked by the use of or reference to previously covered course materials. In this category, students do not contribute new information or formulate novel arguments or products. Instead, it involves activities such as rephrasing existing content, organizing resources, or reiterating the same information
Constructive (c)	This engagement category focuses on analysis and application. It is characterized by the generation of new or supplementary ideas, arguments, or products. Activities within this category include the development of novel concepts, engaging in argumentation, making comparisons, exploring cause-and-effect relationships, and employing reasoning
Interactive (i)	This engagement category encompasses evaluation and creation. To qualify a post under this category, it must not only involve generating new ideas or outputs but also exhibit a dynamic of turn-taking. Activities indicative of this category include judging, building upon, challenging, and expressing agreement or disagreement with others' ideas and arguments. This also involves incorporating elements such as reasoning, cause-and-effect analysis, or comparisons

Table 2. Number of posts for each category.

Engagement Type	Number of posts from each category
Social	814
Active	656
Constructive	450
Interactive	420

4.2 Experiment Setup

As we mentioned earlier, we used GloVe embedding and feature engineering to extract features from the data. We used 3298 unlabeled posts to fine-tuning the GloVe embedding. The unlabeled data enhanced embedding precision and improved model's performance in academic discourse. We also used 2362 labeled posts to broaden the model's exposure and enhance the model's generalization capability in educational discussions. Our method treated all forum posts equally, whether they were initial threads or replies, aiming to assess cognitive engagement in each post independently. This allowed us to ignore variables like reply frequency or discussion recursion, focusing solely on the content's cognitive aspects. However, this approach also has some limitations. We potentially missed the interactive dynamics in extended threads where replies contribute to ongoing conversations. Future research could address this by developing methodologies that consider the context within threads, providing a deeper insight into the interactive and collective dimensions of cognitive engagement in online forums.

We used LSTM for the classification task. LSTM requires inputs of consistent shape and size, necessitating sequence padding for text data. We trained our model using an RMSprop optimizer with a learning rate of 0.001 and binary cross-entropy for loss, incorporating early stopping and a validation split to avoid overtraining and ensure efficacy on the unseen data. Each module was trained for each ICAP category, optimizing for precise feature learning relevant to distinct cognitive engagement levels. This enabled our classification models to classify forum posts within the ICAP framework with high accuracy. In classifying the posts with the cognitive engagement levels within the ICAP framework, we employed a one-versus-all strategy for its simplicity and interpretability in handling multi-class challenges, categorizing the posts into *social*, *active*, *interactive*, and *constructive*. One-versus-all breaks down the task into binary classifications and assigns posts to the class with the highest binary model confidence score, which can ease analysis and enhance model clarity.

4.3 Results

We tested the system performance using two LSTM variants (i.e., standard LSTM and hybrid LSTM) and five traditional machine learning models (i.e., SVM, RF, KNN, LR, and GB). We also tested the impact of feature engineering on the models. Overall, the hybrid LSTM performed best among the models we compared. Our experiment revealed a varied impact of feature engineering on the different models as shown in Fig. 2. Feature engineering significantly improved the hybrid LSTM model's accuracy, from 0.76 to 0.86, showcasing its enhanced ability to utilize additional features for extracting complex patterns from sequential data. In contrast, the standard LSTM model exhibited a slight decrease in performance post-feature engineering, with its accuracy dipping from 0.6502 to 0.6353. This outcome suggests a potential mismatch between the model's architecture and the complexity introduced by the new features. Feature engineering had mixed effects on the traditional machine learning models. SVM and RF achieved a modest gain, showing some capacity to utilize the new features. In contrast, KNN and GB performance declined, and LR was mostly unchanged. These outcomes underscore the necessity of customizing feature engineering to fit the specific attributes of each model.

The confusion matrices of predictions for the Hybrid LSTM without feature engineering and with feature engineering are shown in Fig. 3, where the Hybrid LSTM with feature engineering showed a clear improvement. Table 3 shows the performance of the models with and without enhancements of the features in terms of precision, recall and F1 score, where the hybrid LSTM significantly benefitted from the feature engineering leading to the highest F1-score.

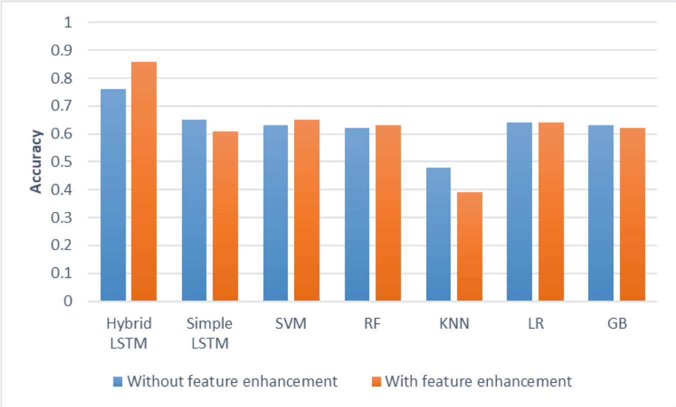
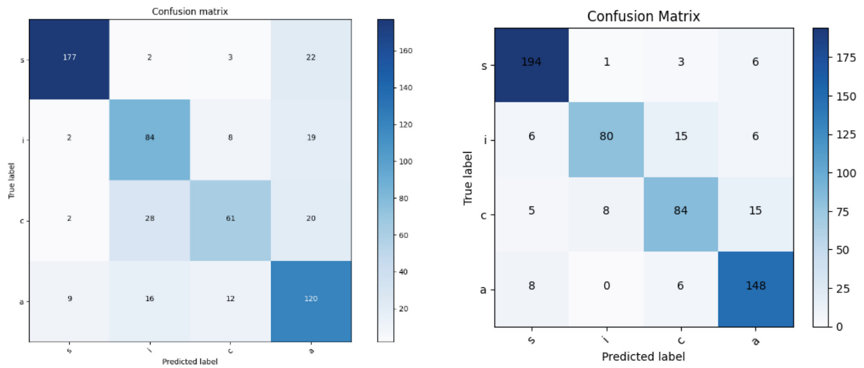


Fig. 2. Accuracy of different machine learning models with and without feature enhancement



(a) (b)

Fig. 3. Confusion matrix of prediction. (a) Prediction results of Hybrid LSTM without feature engineering. (b) Prediction results of Hybrid LSTM with feature engineering

4.4 Explaining the Hybrid LSTM Prediction with LIME

To explain the transparency and interpretability of the hybrid-LSTM model (with the feature engineering), we used LIME [21, 22]. LIME offers explanations by simplifying the model's predictions near a specific instance. In our study, LIME identified key words or phrases influencing predictions in forum post classifications, offering insights into their impact within the ICAP framework and improving the model's reliability.

Applying the LIME to hybrid LSTM model, we yielded some insightful information, which helped in demystifying the model's decision-making process by identifying key terms and their influence on the class labels. Figure 4 shows how certain words positively and certain words negatively impact the classification decision of the posts. In Fig. 4(a), the input post was "*in the meantime, you can click cc in the bottom right-hand corner, and it will show you the text that she's saying it will highlight and go along as the video plays*", which was classified as an *active* post. In this example, LIME identified words

Table 3. Performance of different machine learning models with feature enhancement

Model	With feature enhancement			Without feature enhancement		
	Prec	Rec	F1 Score	Prec	Rec	F1 Score
Hybrid LSTM	**0.87**	**0.86**	**0.86**	**0.77**	**0.76**	**0.76**
Standard LSTM	0.59	0.57	0.53	0.59	0.60	0.58
SVM	0.64	0.65	0.65	0.62	0.63	0.62
RF	0.62	0.63	0.62	0.62	0.62	0.62
KNN	0.38	0.38	0.38	0.49	0.48	0.48
LR	0.61	0.59	0.60	0.63	0.64	0.63
GB	0.60	0.59	0.59	0.62	0.63	0.62

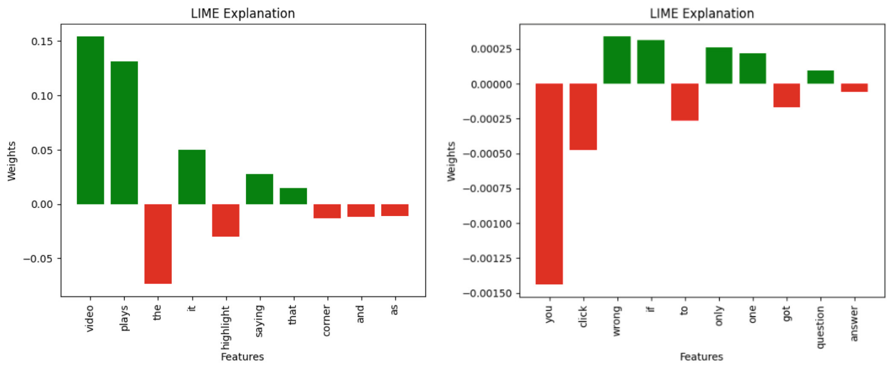


Fig. 4. LIME interpretation for the predicted labels: (a) active; (b) not constructive

like "videos" and "plays" impacted positively, while "the", "and", and "as" negatively impacted the 'active' classification. In Fig. 4(b), the input text was "*if you clicked one answer then you got the question wrong you need to click both*". In this example, the words like "you" and "click" negatively impacted and the words like "working" and "if" positively impacted the prediction of this post into not a *constructive* class. Since, *constructive* class does not include a dialog/debate between peers, thus "you" was a major negative word and had a lot of weight to classify this post as not a *constructive* post, while the word "question" plays positive role but not to a larger extent. This analysis sheds light on the effects of specific words on model outcomes, which enhances the understanding of the intricacies in language-based classification, as shown in Fig. 4.

5 Conclusion

This research study aimed to automate cognitive engagement detection within the *interactive*, *constructive*, *active*, and *passive* (ICAP) framework [7] through using deep learning models, particularly using hybrid LSTM, GloVe embeddings and enhanced feature engineering by analyzing the context and semantics of students' course discussion posts. The model was trained to classify posts into the ICAP's categories and achieved a notable success in quantifying cognitive engagement. Feature engineering played a crucial role, boosting our model's accuracy from 76% to 86%. Our method involved extensive text preprocessing and feature engineering, where we extracted domain-specific and discussion-specific features and appended the features with GloVe embedding. Our modular architecture of hybrid LSTM with regularization and dropout layers helped to avoid overfitting and increased the model's generalization capability. Finally, the use of LIME interpretation showed our model's transparency by revealing how it makes the decision of classification.

Our model's predictions revealed a dominance of *social* and *active* posts among the discussion posts in the dataset we used. This indicates a trend where many students passively consume content, with fewer actively engaging with course materials and peer discussions. Although our method made some progress in automating cognitive engagement detection, this study still has some limitations which could be addressed in further research. More focus should be given to the quality data annotation, which is crucial for any machine learning model's performance. Future research could explore additional or alternative feature engineering techniques, such as TF-IDF, word2vec, or BERT embeddings, to improve the model's prediction. Exploring cognitive engagement in other types of interactions, such as assignments, exams, and group projects, is a promising direction, which may offer a comprehensive perspective on student engagement beyond discussion posts. Integrating our model into learning management systems could automate engagement monitoring, which could help instructors identify at-risk students, tailor interventions, and enhance course outcomes. The model could be used for enhancing recommendation or question-answering systems by suggesting engaging discussion threads based on a student's past interactions or automating responses to frequent questions, thereby elevating online learning's efficiency. This can also be used for creating pedagogical agents and conversational agents to assist online learners.

Acknowledgements. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Innovates, Canada.

References

1. Haleem, A., Javaid, M., Qadri, M.A., Suman, R.: Understanding the role of digital technologies in education: a review. Sustain. Oper. Comput. **3**, 275–285 (2022)
2. Cheng, C.K., Pare, D.E., Collimore, L.-M., Joordens, S.: Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. Comput. Educ., 253–261 (2011)
3. Seethamraju, R., Hwang, G.-J.: Effectiveness of using online discussion forum for case study analysis. Educ. Res. Int. **2014** (2014)
4. Hew, K.F., Cheung, W.S.: Attracting student participation in asynchronous online discussions: a case study of peer facilitation. Comput. Educ. **51**, 1111–1124 (2008)
5. Parks-Stamm, E., Zafonte, M., Stephanie, P.: The effects of instructor participation and class size on student participation in an online class discussion forum. Br. J. Educ. Technol. **48** (2016)

6. Liu, Z., Kong, X., Liu, S., Yang, Z., Zhang, C.: Looking at MOOC discussion data to uncover the relationship between discussion pacings, learners' cognitive presence and learning achievements. Educ. Inf. Technol. **27**(6), 8265–8288 (2022)
7. Chi, M.: Active-constructive-interactive: a conceptual framework for differentiating learning activities. Top. Cogn. Sci. **1**(1), 73–105 (2009)
8. Atapattu, T., Thilakaratne, M., Rebecca, V., Katrina, F.: Detecting cognitive engagement using word embeddings within an online teacher professional development community. Comput. Educ. **140**, 103594 (2019)
9. Gorgun, G., Yildirim-Erbasli, S., Demmans, C.: Predicting cognitive engagement in online course discussion forums. In: International Conference on Educational Data Mining, Durham, UK (2022)
10. Yogev, E., Gal, K., Karger, D., Facciotti, M.T.: Classifying and visualizing students' cognitive engagement in course readings. In: Fifth Annual ACM Conference on Learning at Scale, NY, USA (2018)
11. Hayati, H., Idrissi, M.K., Bennani, S.: Classifying learners' cognitive engagement from online discussion using text mining. Int. J. Latest Trends Eng. Technol. **8**(2), 212–219 (2017)
12. Akyol, Z., et al.: A response to the review of the community of inquiry framework. J. Dist. Educ. **23**(2), 123–135 (2009)
13. Liu, S., Liu, S., Liu, Z., Peng, X., Yang, Z.: Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. Comput. Educ. **181**, 1–16 (2022)
14. Wang, X., Wen, M., Rose, C.P.: Towards triggering higher-order thinking behaviors in MOOCs. In: Proceedings of the Sixth International Conference on Learning and Analytics & Knowledge, New York (2016)
15. Liu, Z., et al.: Dual-feature-embeddings-based semi-supervised learning for cognitive engagement classification in online course discussions. Knowl.-Based Syst. **259**, 110053 (2023)
16. Carillo-de-Albornoz, J., Vidal, R.J., Plaza, L.: Feature engineering for sentiment analysis in e-health forums (2018)
17. Ni, R., Huan, C.: Sentiment analysis based on GloVe and LSTM-GRU. In: 2020 39th Chinese Control Conference (CCC, 2020) (2020)
18. Jeffery, P., Socher, R., Christopher, M.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (2014)
19. Olah, C.: Understanding LSTM networks, colah's blog, 27 August 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
20. Agrawal, A., Paepcke, A.: The Stanford MOOCPosts Data Set. https://datastage.stanford.edu/StanfordMoocPosts/. Accessed 12 Feb 2024
21. Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning (2016)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?": explaining the predictions of any classifier. In: Association for Computing Machinery (2016)

Assessing Cognitive Workload of Aircraft Pilots Through Face Temperature

Amin Bonyad Khalaj[1], Hamdi Ben Abdessalem[2](✉), and Claude Frasson[1]

[1] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Canada
amin.bonyad.khalaj@umontreal.ca, frasson@iro.umontreal.ca
[2] Département d'Informatique et de Mathématique, Université du Québec à Chicoutimi, Chicoutimi, Canada
hamdi_benabdessalem@uqac.ca

Abstract. This research delves into the intricate interplay between mental workload and facial temperature among participants undergoing simulated aircraft takeoff scenarios. Our objective was to discern how workload impacts facial temperature within the context of a flight simulator. Ten individuals were enlisted to serve as pilots in simulated A-320 flights, navigating through six distinct scenarios designed to mimic various levels of mental workload during takeoff, including both routine and emergency situations. We meticulously collected data, amassing a total of 120 takeoff instances and over 10 h of time-series data encompassing heart rate, workload assessments, and facial thermal images and temperatures. Through thorough comparative analysis of EEG data and different types of thermal images, compelling insights emerged. Notably, we observed a pronounced inverse relationship between workload and facial muscle temperatures, alongside facial landmark points. The implications of these findings extend beyond mere academic curiosity, offering valuable insights into the physiological repercussions of workload. Moreover, they hold promise for enhancing aviation safety protocols and optimizing pilot performance in demanding scenarios involving human interaction.

Keywords: Human Interaction · Mental Workload · EEG · Thermal Images · Flight Simulation · Aviation · Face Temperature · Brain Computer Interaction

1 Introduction

Understanding mental workload is crucial in cognitive psychology and human factors research, indicating the cognitive effort and attention required for tasks [1]. It encompasses various processes, including cognitive, perceptual, and motor functions. Managing mental workload is vital for job efficiency, task performance, and user experience, while reducing errors [2]. Professionals, especially in intensive care units, may suffer consequences from excessive workload [3]. Thermal imaging has garnered interest for its potential to correlate facial temperature with workload, offering real-time monitoring of physiological responses [4]. This correlation has implications for aviation safety,

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 27–37, 2024.
https://doi.org/10.1007/978-3-031-63031-6_3

particularly in understanding pilots' cognitive workload [5]. However, challenges such as standardization and feature extraction hinder its full potential [6]. Integrating thermal imaging into aircraft systems could help estimate cognitive stress [7]. Further research and methodological advancements are needed to fully utilize facial thermal imaging in workload assessment [8]. Combining EEG measures with thermal imaging offers a novel approach to studying mental workload [9]. Thermal imaging captures infrared radiation emitted by the body, providing insights into physiological reactions indicative of mental activity [10]. Methodological concerns in facial thermal imaging for workload assessment were noted, emphasizing accurate segmentation and thermal analysis reliability [11]. Researchers are exploring using facial skin temperature to categorize cognitive workload. Integration of thermal imaging into aircraft systems for estimating cognitive stress is suggested. However, challenges like standardization and temperature distribution representation need addressing. Further research and methodological breakthroughs are needed for full utilization in aviation workload assessment. Previous research [12], established a correlation between workload and EEG measures.

This paper aims to correlate EEG measures with thermal imaging technology. Exploring mental workload through thermal imaging technology presents an innovative method. Thermal images capture the infrared radiation emitted by the human body, providing unique insights into physiological responses that could signify mental activity. Understanding the connection between mental workload and facial thermal imaging has the potential to enhance our understanding of cognitive processes. Furthermore, it could contribute to the advancement of systems and environments that optimize human performance.

2 Related Work

Research has highlighted the efficacy of EEG analysis in categorizing exertion levels alongside other physiological indicators [13]. Integrating EEG with these signals offers a practical means of assessing workload in real-world scenarios [14]. Additionally, studies have shown the potential of EEG in evaluating cognitive effort by analyzing spectral powers at different cortical sites [15]. EEG has proven useful in quantifying brainwave activity during various tasks, shedding light on cognitive processes [16].

Understanding mental workload is crucial in transportation, with studies revealing its impact on operational performance, especially in multitasking situations [17–19]. Cognitive workload theory provides a framework for comprehending the strain on psychological capacity [20]. Additionally, research has explored the connection between facial expression and mental stress, underscoring the importance of managing workload to reduce errors [21].

Innovative approaches, such as using convolutional neural networks for EEG classification, offer insights into pilot workload during flight tasks [23]. Similarly, infrared facial thermography shows promise in assessing mental workload in different thermal environments [24]. Thermal imaging has also been employed to detect workload changes during flight simulation, with temperature variations in specific facial regions reflecting changes in mental effort [25]. In this study, workload was not precisely measured in real-time using EEG, and only the variation of face temperature in three facial regions, including the nose tip, nose area, and forehead, was recorded.

3 Experiments

The experiment aimed to gather real-time data on participants' workload and facial thermal images during takeoff procedures in an Airbus A320. A diverse range of flight scenarios was employed in the airplane simulator to replicate real-world flying conditions throughout the experiment's design and execution. It consisted of six scenarios varying in weather, time, and conditions, including potential failure during takeoff (Table 1). Scenarios one through three were designated for regular takeoff sessions, while scenarios four through six were for failure sessions. Participants, acting as pilots, were monitored by the experimenter.

Ten individuals, aged 25 to 35, formed a gender-balanced group for the study. Before participation, all received a briefing on the study's objectives and purpose and provided written consent following an ethics certificate. Half of the participants had flight simulation software experience, while the other half had none, ensuring balanced representation.

During the experiment, participants executed flight simulations using the X-plane flight simulator, while equipment documented physiological and cognitive data. Thorough calibration and setup procedures were conducted on all equipment before the experiment commenced. Heart rate monitoring was conducted using the Polar H10, while cognitive workload was measured using the EEG headset from BMU, with real-time data extraction facilitated by the NCO software.

Infrared cameras were employed to detect and record infrared light, transforming it into visual representations of temperature discrepancies. Thermal imaging visualized the temperature of participants' facial regions, revealing temperature variations across different facial areas. The experiment utilized an ICI-7640 infrared camera and IR Flash software to extract CSV files containing facial temperature data.

Table 1. An overview of the different scenarios [12].

Scenario	Detail		
	Time	Weather	Engine Failure
1	1:45 PM	No Wind, No Clouds	No
2	6:00 AM	Clouds at 2700ft, rain	No
3	9:00 PM	No wind, no clouds	No
4	5:30 AM	No wind, no clouds	Yes, EF at 80 knots
5	6:00 AM	15 knots crosswind	Yes, EF at 140 knots
6	6:00 AM	Low visibility, rain	Yes, EF at 80 knots

The experimental setup involved designating a participant as the pilot and the experimenter as the pilot monitor, configuring the X-plane simulator with an A320 model, and attaching EEG headsets and heart rate monitors to each participant. Comprehensive calibration and setup procedures were conducted on all equipment prior to commencing the experiment. Participants were randomly assigned to one of six scenarios to enhance the authenticity of their responses, as they were unaware of the specific scenarios beforehand. Additionally, A total of 120 takeoffs were recorded, resulting in over 10 h of time-series data. An Infrared camera captured facial thermal images and temperatures at a rate of five images per second, resulting in a total of 18,000 thermal images and temperature data points for each participant's face.

Throughout the entire flight, including different flight scenarios and participants' rest periods, all devices continuously and concurrently captured and recorded data. This comprehensive data collection approach aimed to capture nuanced variations in participant responses and provide valuable insights into the impact of different takeoff scenarios on pilot performance.

4 Results and Discussion

The subsequent crucial stage involves facial detection. For this purpose, the four types of processed images are utilized to improve the accuracy of detecting and identifying facial features and key facial points. Table 2 and Fig. 1, present the statistics of detected faces for each type of thermal camera face image. The results indicate that the Gray type exhibited the highest number of face detections, while the Rainbow type showed the lowest number.

Table 2. Comparison of participants average face detection for each type of face thermal image

Face Detection Status	In a Flight			
	Rainbow	Gray	Research	Iron
Detected (%)	**0.03**	**0.99**	**0.73**	**0.70**
Undetected (%)	**0.97**	**0.01**	**0.27**	**0.30**

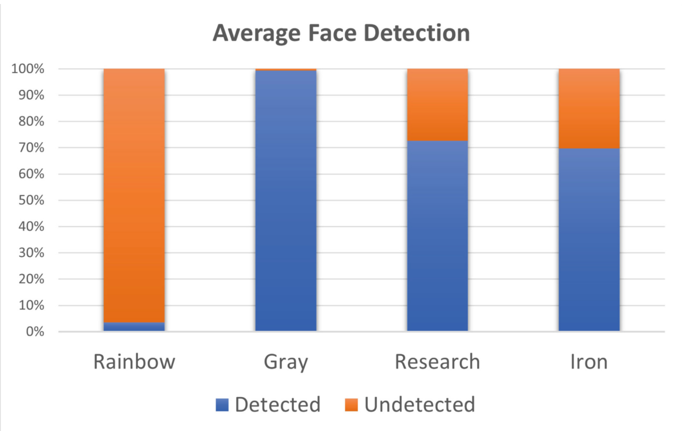


Fig. 1. Comparison of average face detection for each type of face thermal image.

The Fig. 2 illustrates a flowchart detailing a facial temperature detection subsystem crafted for precise temperature readings across various thermal image types. The subsystem comprises three pivotal stages: Face Detection, Facial Landmark Point Extraction, and Temperature Extraction from these Points. The objective is to furnish a robust solution for precise face detection and temperature measurement via thermal imaging.

4.1 Face Detection (Step 1)

The primary stage, Face Detection, sets the foundation for the entire process. To bolster accuracy, four distinct types of thermal images—Rainbow, Iron, Research, and Gray—are employed. Each image type offers a unique perspective for identifying facial features. The procedure unfolds as follows:

i. **Rainbow Thermal Image:**
   - If a face is successfully detected in the Rainbow thermal image, the system proceeds.
   - If no face is detected, the system proceeds to the next thermal image type.

ii. **Iron Thermal Image:**
   - If the Iron thermal image detects a face, the system moves forward.
   - If no face is detected, the system proceeds to the next thermal image type.

iii. **Research Thermal Image:**
   - Similar to prior steps, if the Research thermal image detects a face, the system advances.
   - If no face is detected, the system moves on to the final thermal image type.

iv. **Gray Thermal Image:**
   - If the Gray thermal image detects a face, the system progresses.
   - If no face is detected, the system disregards that image.

4.2 Extract Facial Landmark Points (Step 2)

Following successful face detection in Step 1, the system proceeds to extract facial landmark points. These points furnish a detailed map of facial features essential for precise temperature measurement. The human face is a complex structure comprising several different muscles that play a crucial role in various functions, including appearance, movement, and facial expressions. In Fig. 3, facial muscles are organized based on facial landmarks and indexes, accompanied by labels denoting the names of each muscle.

4.3 Extract Temperature from the Points (Step 3)

The concluding step involves extracting the temperature from the facial landmark points. Utilizing information garnered in the previous steps, the system accurately gauges the temperature at specific facial points, ensuring dependable and precise assessments.

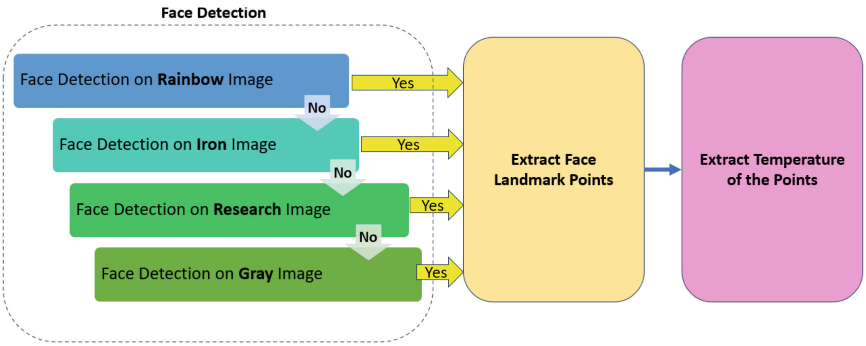


Fig. 2. Face detection cascade flowchart for face thermal image types

Figure 2 illustrates the cascade flowchart delineating a more accurate face detection approach and the process of measuring temperatures at facial landmark points.

We conducted an analysis on a randomly selected subset of participants' thermal face photographs. A landmark point was considered accurately recognized if the difference between the expected and actual positions fell within a range of five pixels. Based on investigations across various thermal camera image types, the accuracy of facial landmark point detection can be ranked as follows: Rainbow, Iron, Research, and Gray.

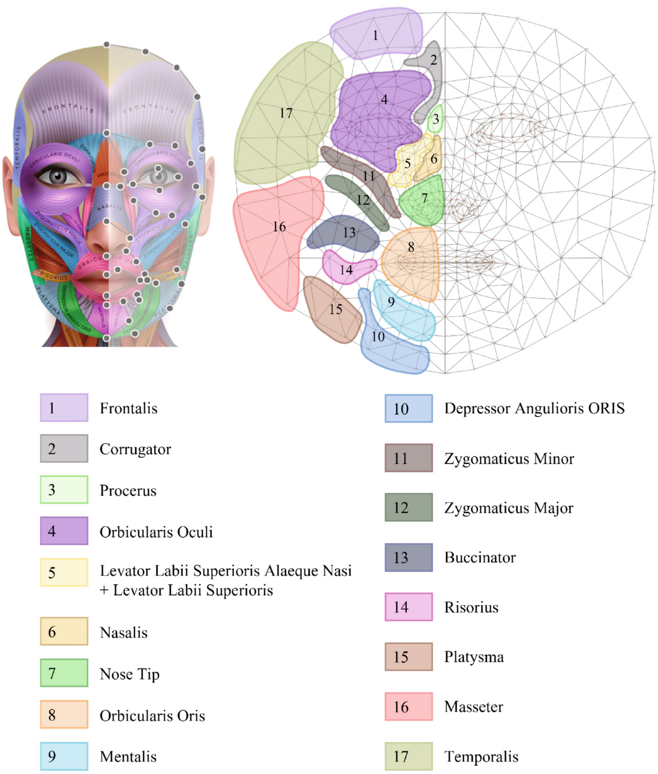


Fig. 3. Facial muscles grouped from facial landmarks and indexes and muscle names [26, 27]

Among these, Rainbow and Iron exhibit the highest accuracy, while Gray displays the lowest accuracy (see Fig. 4). When participants directly faced the camera, Rainbow and Iron thermal image types achieved accuracies of 98.2% and 98.1%, respectively, in detecting facial landmark points within a 5-pixel margin. On average, considering all head and face orientations, the overall accuracy is 95%. The framework attained a mean accuracy of 95%, indicating that 95% of all facial points were accurately recognized within a 5-pixel margin of error compared to the ground truth in the test image.

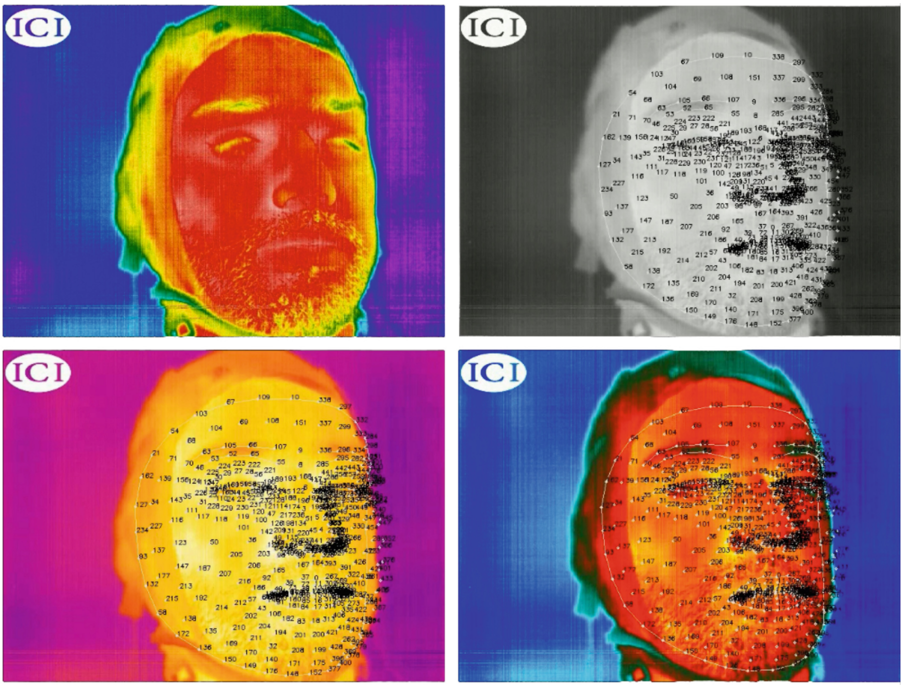


Fig. 4. Comparison of face detection for each type of face thermal image on extracted Face Landmark points.

The Fig. 5 presents a comprehensive depiction of workload and facial temperature changes during scenario 5 (Engine Failure after V1 speed). The red line denotes the average facial temperature, while the blue line represents workload. Vertical green and yellow dashed lines mark the scenario's start and end times, while pink dashed lines signify engine failure instances. Additionally, a dashed blue line indicates maximum workload. The figure illustrates workload escalation upon scenario initiation, followed by a decrease post-scenario, causing fluctuations in average facial temperature. A weak inverse correlation between average facial temperature and workload is evident, albeit with a time delay. Workload peaks one second after engine failure (Blue Point), indicating maximum workload, while average facial temperature responds with a noticeable delay, reaching its lowest value (Red Point) during this period. This delayed response underscores facial physiological dynamics and sensitivity to workload changes.

During scenario 5 flight, heart rate fluctuated but showed a gradual overall increase, peaking post-engine failure. The average heart rate after failure surpassed pre-failure levels. Multiple pre-failure heart rate increases occurred due to participants' anticipation of engine failure, rooted in prior experiences with scenario 4, where failure happened at 80 knots. This anticipation heightened stress levels, elevating heart rate and mental workload while decreasing facial temperature. Conversely, in scenario 5, engine failure occurred at a higher speed (V1 = 140 knots) when the aircraft was at high speed, lacking sufficient runway distance for braking. Participants needed adept control to execute a successful takeoff, making this scenario challenging. The Heart rate sharply rises post-engine failure, reflecting heightened stress.

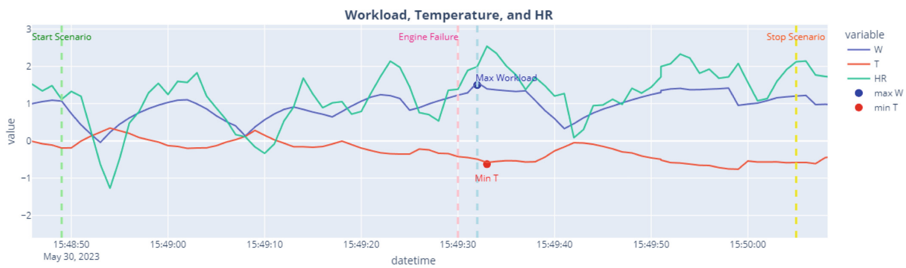


Fig. 5. Relation Between Workload (W) and Temperature (T) (Color figure online)

The Fig. 6 depicts variations in workload and nose area temperature. The red line denotes nose area temperature, while the blue line represents workload. Vertical green dashed lines indicate scenario start, pink dashed lines mark engine failure, and yellow dashed lines indicate scenario end. The dash blue line signifies maximum workload during the scenario. The purple line shows temperature changes of the Frontalis muscle (Face Muscle number 1), corresponding to the forehead, which remains relatively stable and shows no significant correlation with workload. In contrast, the green line represents temperature changes of face muscle 7 (Nose tip), displaying an inverse relationship with workload. As **workload increases**, **nose tip temperature decreases**, indicating a significant response in this area during high workload. Moreover, the red line illustrates temperature changes associated with the average temperature of facial muscles, including Procerus, Nasalis, and Nose Tip (corresponding to numbers 3, 6, and 7, respectively), highlighting the **most inverse relationship with workload**.

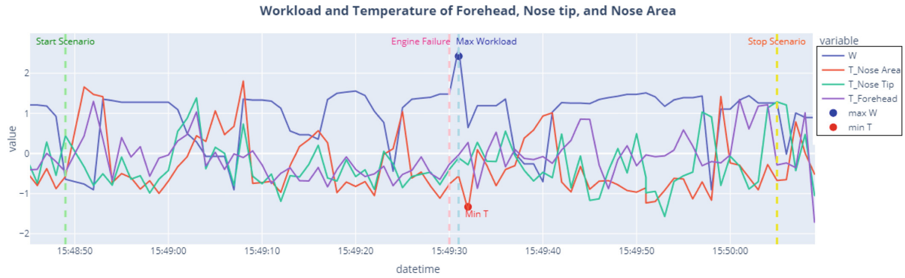


Fig. 6. Relation Between Workload (W) and Temperature of Nose Area (T_Nose Area), Nose tip (T_Nose Tip), and Forehead (T_Forehead) (Color figure online)

5 Conclusion

As air traffic increases, the management of pilot operations becomes increasingly critical. The rise in information flow can elevate pilots' workload, leading to potential confusion and errors. In a previous study, we highlighted the correlation between workload and brainwave activity measured via EEG. In line with aviation research, our current study aimed to explore the relationship between facial temperature and workload. The underlying concept is that an individual's workload can influence their body's thermoregulation, potentially reflected in facial thermal patterns. A heightened mental workload can induce increased stress and cognitive load, triggering physiological responses such as changes in facial temperature.

Our experiment involved 10 participants in six different aviation takeoff scenarios, encompassing normal and emergency situations. We collected continuous time-series data, including workload, heart rate, and facial thermal images and temperatures, totaling 120 h of takeoff and 9 h of data collection per participant. This allowed us to examine the relationship between workload and face muscle and landmark points temperature during takeoff. We observed that fluctuations in facial temperature correlated with participants' increasing workload during extended flight durations. Monitoring pilots' facial temperature during takeoff, landing, and long-haul flights could aid in identifying fatigue and stress, facilitating timely interventions to ensure aviation safety and pilot well-being.

Our ongoing project aims to integrate a machine learning model to calculate workload based on non-invasive physiological data. This model, leveraging EEG, heart rate, eye-tracking, and facial thermal imaging data, aims to provide a nuanced and real-time assessment of cognitive workload in aviation scenarios. By incorporating machine learning, we anticipate enhancing workload prediction accuracy and reliability, contributing to a deeper understanding of pilots' cognitive states.

The experimentation process is underway, involving continuous refinement of the machine learning model. This phase includes gathering additional data to expand the model's training dataset and evaluating its performance across various scenarios and individual differences. As participant numbers increase, we expect improved correlation discernment between mental workload and facial temperature. This iterative process is essential for fine-tuning the model's predictive capabilities and ensuring its applicability across diverse operational conditions in aviation.

Acknowledgments. We acknowledge the support from NSERC-Alliance, CRIAQ, CAE, Bombardier, and BMU for funding this work.

References

1. Devos, H., et al.: Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults. Brain Sci. **10**(12), 994 (2020). https://doi.org/10.3390/brainsci10120994
2. Hananingrum, P., Athqia, A.A., Wahyudiono, Y.D.A.: Relationship between age, gender, job placement, and social relationships with the mental workload of managers. Indones. J. Occup. Saf. Health **11**(3), 377–389 (2022). https://doi.org/10.20473/ijosh.v11i3.2022.377-389
3. Vidulich, M.A., Tsang, P.S.: Mental workload and situation awareness. In: Handbook of Human Factors and Ergonomics, pp. 243–273 (2012). https://doi.org/10.1002/9781118131350.ch8
4. Shah, M., Peikari, H.R.: Electronic prescribing usability: reduction of mental workload and prescribing errors among community physicians. Telemed. E-Health **22**(1), 36–44 (2016). https://doi.org/10.1089/tmj.2014.0246
5. Zahednezhad, H., Shokrollahi, N., Gheshlagh, R.G., Afshar, P.F.: Does heavy mental workload affect moral sensitivity among critical care unit nursing professionals? A cross-sectional study. BMC Nurs. **20**(1), 140 (2021). https://doi.org/10.1186/s12912-021-00662-8
6. Ganesh, K., Snekhalatha, U., Krishnan, P.T.: Deep learning techniques for automated detection of autism spectrum disorder based on thermal imaging. Proc. Inst. Mech. Eng. Part H J. Eng. Med. **235**(10), 1113–1127 (2021). https://doi.org/10.1177/09544119211024778
7. Stemberger, J., Allison, R.S., Schnell, T.: Thermal imaging as a way to classify cognitive workload. In: 2010 Canadian Conference on Computer and Robot Vision, pp. 231–238 (2010). https://doi.org/10.1109/crv.2010.37
8. Kosonogov, V., et al.: Facial thermal variations: a new marker of emotional arousal. PLoS ONE **12**(9), e0183592 (2017). https://doi.org/10.1371/journal.pone.0183592
9. Pollina, D.A., et al.: Facial skin surface temperature changes during a 'concealed information' test. Ann. Biomed. Eng. **34**(7), 1182–1189 (2006). https://doi.org/10.1007/s10439-006-9143-3
10. Wang, Z., et al.: A novel facial thermal feature extraction method for non-contact healthcare system. IEEE Access **8**, 86545–86553 (2020). https://doi.org/10.1109/access.2020.2992908
11. Liu, X., Wang, Y., Luan, J.: Facial paralysis detection in infrared thermal images using asymmetry analysis of temperature and texture features. Diagnostics **11**(12), 2309 (2021). https://doi.org/10.3390/diagnostics11122309
12. Antoine, M., Ben Abdessalem, H., Frasson, C.: Cognitive workload assessment of aircraft pilots. J. Behav. Brain Sci. **12**(10), 474–484 (2022). https://doi.org/10.4236/jbbs.2022.1210027
13. Liu, Y., Ayaz, H., Shewokis, P.A.: Multisubject 'learning' for mental workload classification using concurrent EEG, fNIRS, and physiological measures. Front. Hum. Neurosci. **11** (2017). https://doi.org/10.3389/fnhum.2017.00389
14. Hogervorst, M.A., Brouwer, A., Van Erp, J.: Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. Front. Neurosc. **8** (2014). https://doi.org/10.3389/fnins.2014.00322
15. Rebsamen, B., Kwok, K., Penney, T.B.: Evaluation of cognitive workload from EEG during a mental arithmetic task. Proc. Hum. Factors Ergon. Soc. Annu. Meet. **55**(1), 1342–1345 (2011). https://doi.org/10.1177/1071181311551279

16. Rojas, R.F., et al.: Electroencephalographic workload indicators during teleoperation of an unmanned aerial vehicle shepherding a swarm of unmanned ground vehicles in contested environments. Front. Neurosc. **14** (2020). https://doi.org/10.3389/fnins.2020.00040
17. Jafari, M., et al.: Assessment and monitoring of mental workload in subway train operations using physiological, subjective, and performance measures. Hum. Fact. Ergon. Manuf. Serv. Ind. **30**(3), 165–175 (2020). https://doi.org/10.1002/hfm.20831
18. Woods, B., Byrne, A.: The effect of multitasking on the communication skill and clinical skills of medical students. BMC Med. Educ. **18**(1), 76 (2018). https://doi.org/10.1186/s12909-018-1183-5
19. Lim, W.L., et al.: EEG-based mental workload recognition related to multitasking. In: 2015 10th International Conference on Information, Communications and Signal Processing (ICICS), pp. 1–4 (2015). https://doi.org/10.1109/icics.2015.7459834
20. Han, P., Xu, L., Lv, X.: Optimised analysis of community medical app user experience under cognitive load theory. In: E3S Web of Conferences, vol. 179, p. 02063 (2020). https://doi.org/10.1051/e3sconf/202017902063
21. Stone, R., Wei, C.: Exploring the linkage between facial expression and mental workload for arithmetic tasks. Proc. Hum. Factors Ergon. Soc. Annu. Meet. **55**(1), 616–619 (2011). https://doi.org/10.1177/1071181311551126
22. Deng, Z., et al.: An experimental study on web interface design optimization based on user cognitive load. Front. Artif. Intell. Appl. **365**, 506–518 (2023). https://doi.org/10.3233/faia220744
23. Hernàndez-Sabaté, A., et al.: Recognition of the mental workloads of pilots in the cockpit using EEG signals. Appl. Sci. **12**(5), 2298 (2022). https://doi.org/10.3390/app12052298
24. Wang, X., et al.: Can infrared facial thermography disclose mental workload in indoor thermal environments? In: Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization, pp. 87–96 (2019). https://doi.org/10.1145/3363459.3363528
25. Hassoumi, A., et al.: Thermal imaging of the face: mental workload detection in flight simulator. In: Neuroergonomics and Cognitive Engineering. AHFE (2022) International Conference, vol. 42. AHFE Open Access, USA (2022). https://doi.org/10.54941/ahfe1001822
26. Kim, J., et al.: Numerical approach to facial palsy using a novel registration method with 3D facial landmarks. Sensors **22**(17), 6636 (2022). https://doi.org/10.3390/s22176636
27. Reineg: Muscles of the Face, Colorful Anatomy Info Poster. https://stock.adobe.com/ca/images/muscles-of-the-face-colorful-anatomy-info-poster/309366859

Profiles of Performance: Game-Based Assessment of Reading Comprehension Skill

Katerina Christhilf(✉), Rod D. Roscoe, and Danielle S. McNamara

Learning Engineering Institute, Arizona State University, Tempe, AZ 85287, USA
{k.christhilf,rod.roscoe,danielle.mcnamara}@asu.edu

Abstract. The current study examines the use of literacy game performance in assessing reading comprehension skill. The goals for this study were (1) to examine the extent to which game performance can be used to predict reading comprehension skill, as estimated by standardized tests; and (2) to explore the development of student profiles using game performance, and how these profiles relate to reading test performance. The results indicate that performance on two games—games that involved vocabulary knowledge and inferencing—accounted for approximately 45% of variance in reading comprehension skill. These results suggest that game-based learning activities can serve as formative assessments of reading skill. Teachers can use such learning activities to provide their students with more frequent feedback, and themselves with a greater understanding of their students' current skills and areas for improvement.

Keywords: Game-Based Assessment · Reading Comprehension · Intelligent Tutoring Systems

1 Introduction

Effective reading assessments are crucial for understanding students' skills and tailoring instruction and feedback accordingly. Common reading assessment tasks, such as reading passages and responding to questions, are typically aimed at evaluating overall reading proficiency [19, 21]. However, traditional standardized assessments provide limited information to instructors, making it difficult for instructors to know why students received particular scores. For assessments to deliver more targeted feedback, they need to measure the moment-to-moment mental processes involved in comprehension, rather than just the comprehension product [1]. Adding to these challenges is the amount of valuable classroom time such assessments consume, which restricts the frequency of assessments. Infrequent testing further exacerbates the challenge of providing individualized feedback, hindering the iterative process of learning and improvement [22].

Recognizing the limitations of traditional assessment paradigms, researchers are increasingly exploring the integration of assessments *within* learning activities [23]. Learning activities can be used to capture reading processes such as paraphrasing

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 38–48, 2024.
https://doi.org/10.1007/978-3-031-63031-6_4

and inferencing, rather than overall reading proficiency [10]. Additionally, by merging learning and assessment, teachers gain more rapid insights into student progress. Such learning and assessment activities can be incorporated into games to further simulate real-world scenarios and increase engagement [20]. Game-based learning activities offer a unique opportunity for assessment, as they may increase the authenticity of the assessment, improve student engagement, and provide students with immediate feedback [2, 8, 15, 18]. However, there are persistent concerns regarding the validity of games as assessments [5]. The integration of game elements and assessment tasks necessitates meticulous attention to alignment and accuracy. Failure to harmonize these components may compromise reliability and validity, skewing the outcomes and undermining their diagnostic value. Further research is imperative to explore the effectiveness and validity of using games as assessments.

1.1 Assessing Reading Comprehension via Games

Within reading comprehension, embedding assessments in learning activities may help instructors better understand the specific skills, knowledge, and strategies students need to develop. For example, students may lack sufficient vocabulary knowledge to understand texts and may benefit from more knowledge-building activities. When reading a sentence, successful readers convert the explicit words in the sentence to a mental representation of the sentence's meaning. Without sufficient knowledge of the underlying words, one is unable to form this mental representation. Students with greater depth of vocabulary knowledge tend to receive higher scores on reading comprehension assessments [13]. Similarly, students who write better paraphrases are more likely to comprehend the text [7]. However, having vocabulary knowledge and being able to understand individual sentences is not enough for deep text comprehension. Successful readers can generate inferences as they read, connecting information across the text and with their prior knowledge [11, 17]. While overall reading scores may not inform teachers about the specific skills students are struggling with, students' performance in varied learning activities may provide greater insights into students' strengths and weaknesses.

The purpose of this research on game-based learning and assessment is to contribute to the validity and reliability of reading comprehension assessments while optimizing student engagement and learning outcomes. In service to this goal, the current study examined two research questions (RQs). First, *to what extent can performance on three games be used to predict reading comprehension skill, as estimated by a standardized test?* (RQ1). Second, *can students' performance on three reading games be used to form student profiles that relate to standardized test performance, while also providing insights into specific areas for improvement?* (RQ2). The study explored three games[1]: *Vocab Flash*, *Paraphrase Quest*, and *Map Conquest*. These games were taken from the *Interactive Strategy Training for Active Reading and Thinking* (*iSTART*) tutoring system,

[1] An additional two games were played by the participants but not included in analyses. *CON-Artist* was excluded due to difficulties during game administration: 29% of students did not correctly understand instructions and failed to navigate to the game. *Fix It* was excluded due to floor effects and low reliability. Approximately 64% of students performed with 50% accuracy or below, and split-half reliability was .14. For more information, see https://osf.io/sd7ny/.

further described in the Method section. They were selected to measure three aspects of reading comprehension skill: vocabulary knowledge, paraphrasing, and inferencing.

2 Method

2.1 Participants

Participants were Arizona State University undergraduate psychology students who were compensated with course credit. Students were required to have basic English proficiency to be eligible to participate. A total of 570 students enrolled in the study. Of these students, 46 were removed due to missing data and 119 were removed due to inattentive participation. Students were considered to be inattentive if they failed explicit attention check questions, selected the same answer for all questions within a measure, or provided nonsensical responses for open-ended questions. The final sample consisted of 405 participants, of which 184 participated in the fall semester and 221 participated in the spring semester. Participant ages ranged from 18.0 to 36.0 ($M = 19.0$, $SD = 1.4$). Among participants, 39.0% self-identified as a "woman", 58.5% as a "man", and 1.5% as "non-binary/third gender"; an additional 1.0% of participants selected "prefer not to say". Additionally, 54% of participants self-identified as White, 17% as Hispanic, 14% as Asian, 6% as Black or African American, 6% as multiracial, and 1% as American Indian or Alaska Native. Approximately 15% of participants reported learning English as a second (or later) language, and 22% self-identified as first-generation college students.

2.2 Materials

Demographics. Participants were asked to self-report sex, gender, age, race and ethnicity, education, native language, and use of English. See the OSF repository, Appendix A for a list of the questions.

Reading Skills Assessments
Gates-MacGinitie Reading Tests (GMRT). The Gates-MacGinitie Reading Test (4th ed.; Form T, level 10/12) [4] was used to assess reading comprehension and vocabulary via two distinct sub-scores. In the Reading Comprehension section, participants read passages that were 3 to 14 sentences long. After each passage, they answered two to six multiple choice questions that measured comprehension of shallow and deep level information. Participants had access to the passage as they answer questions, reducing the memory load of the test. The test was timed such that participants were given 20 min to complete the 48 questions. The GMRT Reading Comprehension section has been previously validated as a measure of reading skill ($\alpha = .85$–$.92$) [14]. In the Vocabulary section, students were shown underlined vocabulary words, each embedded in a sentence or phrase. They were asked to choose the closest synonym to the word from a list of five. Participants were timed and given 7 min to complete the questions. These assessments were both shortened versions of the test that have been used in previous studies [3, 9]. In the current study, the Reading Comprehension and Vocabulary sections both displayed high reliability ($\alpha = .90$ and $.89$, respectively).

User Experience

Enjoyment and Feedback Survey. This survey was given after the GMRT Reading Comprehension section, the GMRT Vocabulary section, the Self-Explanation Task, and each game. Participants were asked a few brief questions about their enjoyment of the measure, their perceived difficulty in completing the measure, and any ways it could be improved (see the OSF repository, Appendix B). These questions served to ask participants about their immediate reactions to the measures. The questions were used to compare enjoyment across tasks.

User Experience Survey. Participants were given this survey at the end of the study to obtain their overall impressions about the games (see the OSF repository, Appendix C). They were asked questions about their enjoyment and attention during the games. They were also asked to provide feedback on the usability of the games and any ways the games could be improved.

Games

iSTART. The games in the current study were selected from the Interactive Strategy Training for Active Reading and Thinking intelligent tutoring system (iSTART) [12]. iSTART is an interactive trainer that teaches students to use comprehension strategies through videos, animated agents, games, and self-explanation practice. The games provide practice in strategy identification, self-explanation, and synonym identification. iSTART games also have a strong potential to serve as stealth literacy assessments by measuring aspects of reading comprehension skill and vocabulary knowledge [3].

Vocab Flash. This game was included to measure vocabulary knowledge (see Fig. 1). Users were given five minutes to complete as many flashcards as possible. Each flashcard contained a word and four potential synonyms. Participants were instructed to select the word closest in meaning to the target word. There were seven flashcard levels of increasing difficulty. Participants progressed to the next level if they reached the score threshold for that level (e.g., 50 points for Level 1, 150 points for Level 2). Students received more points for responding correctly, and points were deducted for incorrect answers. If students dipped below the score threshold for a level, they were returned to the previous level. The dependent variable was the highest level achieved.

Paraphrase Quest. Paraphrase Quest was included as a measure of how well participants can understand and rephrase individual sentences (see Fig. 2). Participants read short texts and were asked to select the appropriate paraphrase of the bolded sentence out of three options. The game includes a narrative, in which participants assumed the role of a knight on a quest. When participants selected the correct paraphrase, the knight moved forward on the displayed map, and the hunger meter decreased. When participants selected an incorrect paraphrase, the knight failed to move forward, and the hunger meter increased. Participants won the game if they maintained a healthy hunger meter and reached the end of the map. The dependent variable was the proportion of correct responses.

Map Conquest. Map Conquest was included to measure how well participants can form inferences while they read (see Fig. 3). Participants were asked to provide self-explanations for bolded sentences in a text about aquatic biomes. Higher quality self-explanations earned more flags. Participants could strategically place the flags on the

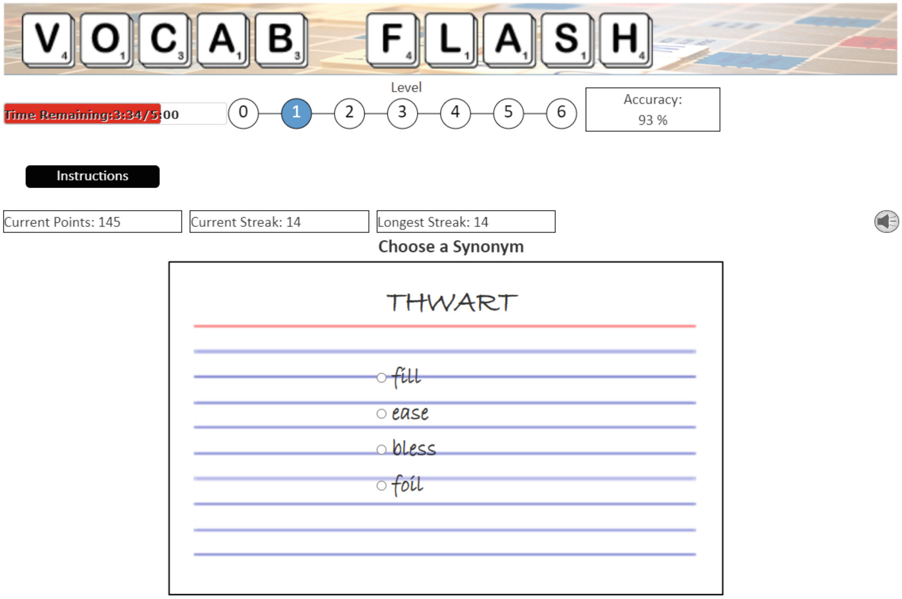


Fig. 1. Sample view in Vocab Flash game, containing time remaining, current level, average accuracy, current points, current streak, longest streak, target word ("thwart"), and four answer choices ("fill", "ease", "bless", and "foil")

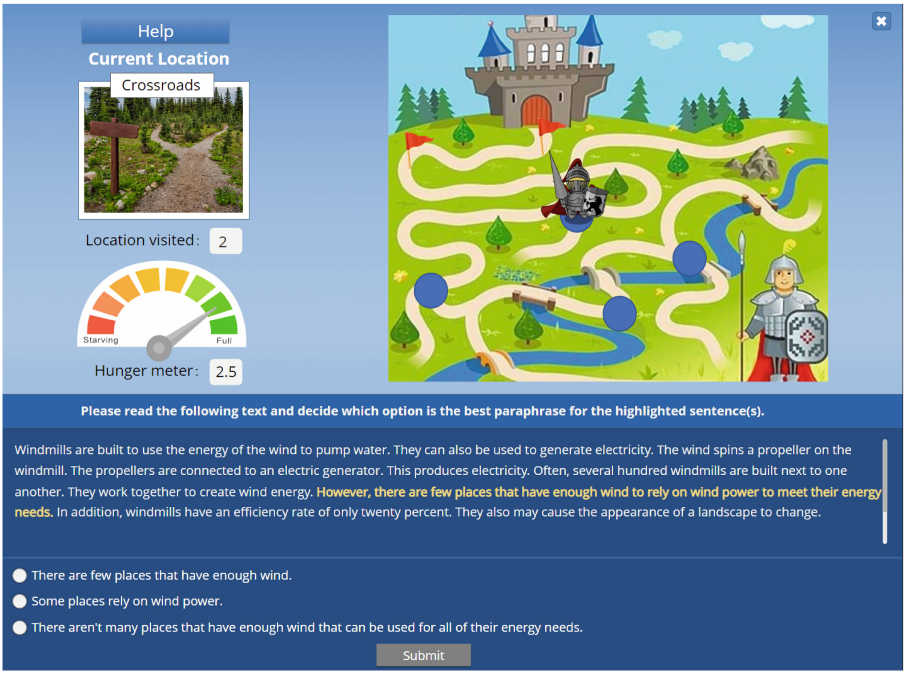


Fig. 2. Sample view in Paraphrase Quest game, containing player's current location, hunger meter, instructions, text, highlighted target sentence to be paraphrased, and three answer choices

map to conquer territories in a mini game. Self-explanation quality was calculated using natural language processing and ranged from 0–3. Responses with higher scores better integrated information across the text, while responses with lower scores primarily paraphrased the text or contained irrelevant information. The average of the scores served as the dependent variable.

2.3 Procedure

The study was conducted via Qualtrics. Participants completed the study in a single session and in the presence of research assistants and/or graduate students. Participation occurred in person in an auditorium setting to increase participant recruitment and data quality. Participants first completed a consent form and demographics questions. They next completed the GMRT Reading Comprehension and Vocabulary sections, and two socio-emotional assessments. Then, they completed the Self-Explanation Task, followed by five games. Students were randomly assigned to one of two game orders. Order A was CON-Artist, Paraphrase Quest, Fix It, Map Conquest, and Vocab Flash. Order B was the reverse order: Vocab Flash, Map Conquest, Fix It, Paraphrase Quest, and CON-Artist. Finally, participants completed the User Experience Survey. Students also completed short enjoyment questionnaires after both sections of the Gates-MacGinitie Reading Test, the Self-Explanation Task, and each of the games. The session length was 2 h, with a five-minute break before the Self-Explanation Task.

3 Results

3.1 Preliminary Analyses

Initial descriptive measures and correlations between variables are provided in Table 1. Split-half reliability (SHR) was calculated using the Spearman-Brown formula. Reading Comprehension, Vocabulary, Map Conquest, and Vocab Flash had good reliability, while Paraphrase Quest had poor reliability. Performance on the Reading Comprehension and Vocabulary sections of the GMRT was positively, moderately correlated with Vocab Flash performance and positively, weakly correlated with Paraphrase Quest and Map Conquest performance. Vocab Flash performance was positively, weakly correlated with Paraphrase Quest and Map Conquest performance.

An independent *t*-test was conducted for each game to determine whether performance differed by order of completion. Participants in Order A ($n = 207$) scored higher on Paraphrase Quest ($t = 3.319$, $p < .001$), which was the second game after CON-Artist (which was removed from analyses). Participants in Order B ($n = 198$) scored higher on Map Conquest ($t = -6.995$, $p < .001$) and Vocab Flash ($t = -2.019$, $p = .044$), which were the first two games that participants played. These results point toward the importance of limiting the number of games that participants play in any one session.

Table 1. Descriptive Statistics and Correlations of Measures

Variable	M	SD	Range	SHR	1	2	3	4
1. Reading Comprehension	27.50	8.83	3–47	.93	—			
2. Vocabulary	28.72	8.25	8–44	.92	.40**	—		
3. Vocab Flash	0.63	0.15	.09–.98	.80	.56**	.46**	—	
4. Paraphrase Quest	0.67	0.23	0–1	.41	.19**	.13**	.18**	—
5. Map Conquest	1.61	0.61	0–3	.89	.22**	.15**	.20**	.07

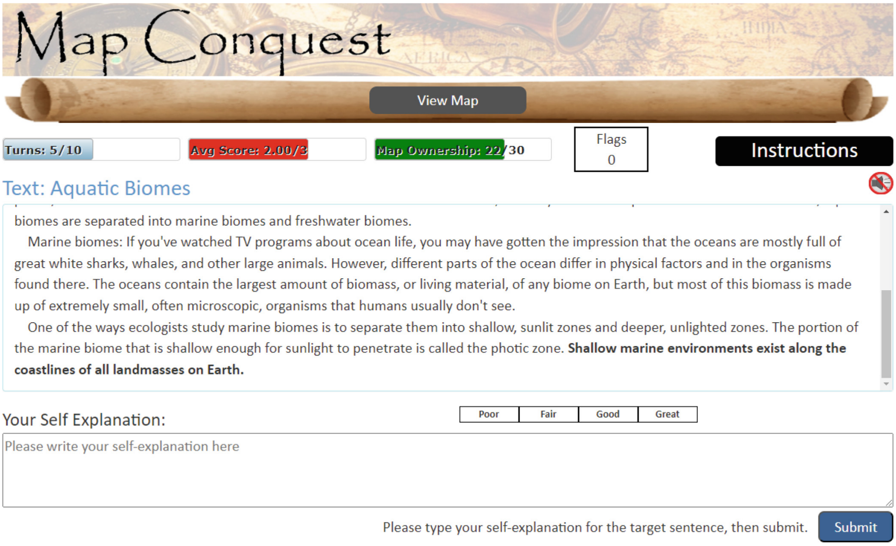


Fig. 3. Sample view in Map Conquest game, containing current turn, average self-explanation score, Map Conquest, current flags earned, text, bolded sentence to be self-explained, and text box in which the student types their self-explanation

3.2 RQ1: *Does Each Game Account for Unique Variance in GMRT Reading Comprehension Performance?*

A linear regression analysis was conducted to predict reading comprehension skill using the "lm" function in R [16]. The dependent variables were the highest level achieved in Vocab Flash, accuracy in Paraphrase Quest, and average score in Map Conquest. Cook's distance was used to check for influential observations. No distances greater than 1 were found, so all datapoints were retained in the analysis.

The model predicted 45.1% of variance in reading comprehension skill, $F(3, 401) = 112$, $p < .001$. Vocab Flash ($t = 16.7$, $p < .001$) and Map Conquest ($t = 2.32$, $p = .021$) were each significant predictors of reading comprehension skill. For every 1.0

standard deviation increase in Vocab Flash performance, reading comprehension performance increased by 0.640 standard deviations, holding all other predictors constant. For every 1.0 standard deviation increase in Map Conquest performance, reading comprehension performance increased by 0.087 standard deviations, holding all other predictors constant.

3.3 RQ2: *Can Students' Performance on Three Reading Games Be Used to Form Student Profiles that Relate to Standardized Test Performance, but also Provide Greater Insights into More Specific Areas for Improvement?*

To answer this question, we conducted an exploratory k-means cluster analysis to group students according to their game performance. Our model included three variables: (1) maximum level achieved in Vocab Flash, (2) accuracy in Paraphrase Quest, and (3) average score in Map Conquest. Each variable was z-scored prior to conducting the analysis. We used the elbow method to determine that the optimal number of clusters was three (see Fig. 4).

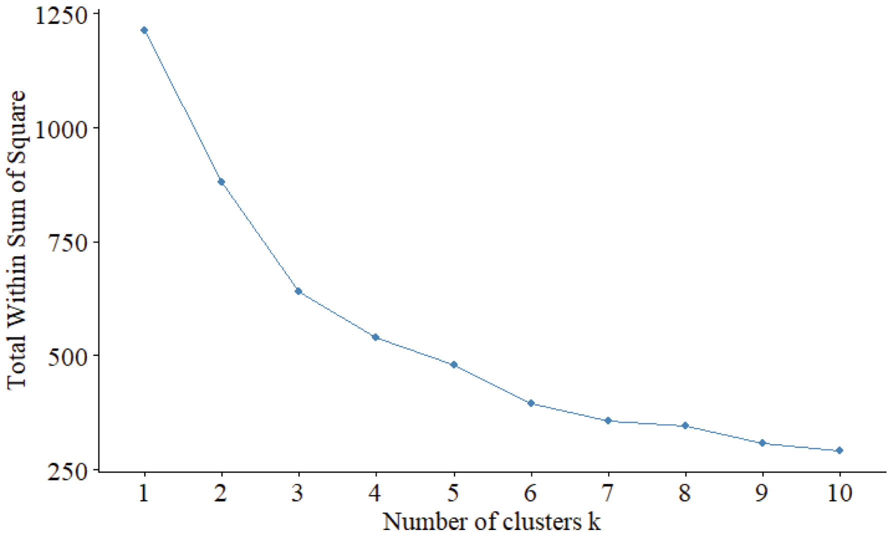


Fig. 4. Illustration of the relation of unexplained variance (total within sum of squares) to the number of clusters

Cluster 1 was labeled "Inferencers" ($n = 238$) and included students characterized by higher performance on Map Conquest and Vocab Flash, and above average performance on Paraphrase Quest (see Fig. 5). These students were notably excellent at self-explaining texts. Cluster 2, named "Paraphrasers" ($n = 220$), included students characterized by higher performance on Paraphrase Quest, with above average performance on Vocab Flash and lower performance on Map Conquest. These students mastered paraphrasing, but were less skilled at self-explaining texts. Finally, Cluster 3 was labeled as "Struggling Readers" ($n = 183$), which included students characterized by lower performance on all three games.

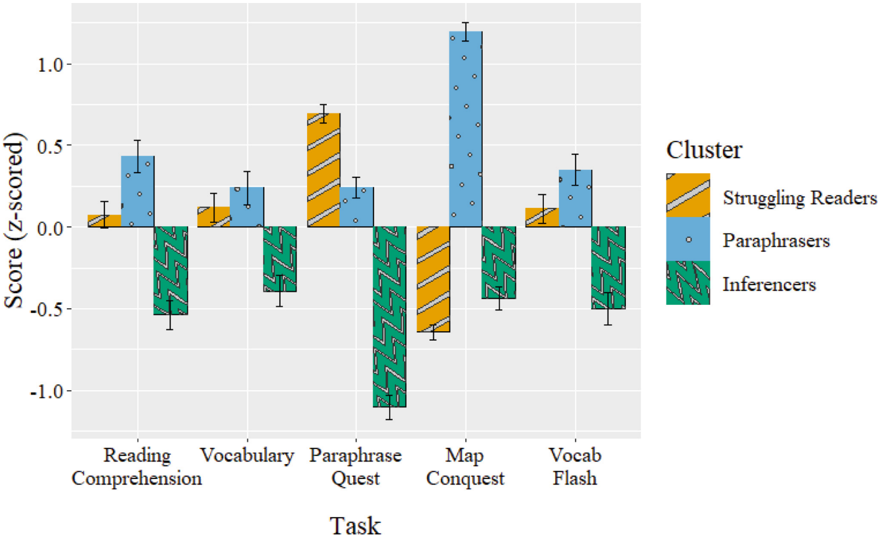


Fig. 5. Bar graph illustrating how each cluster differed, in terms of performance on the two validating variables (reading comprehension and vocabulary) and the three games used in the model (Paraphrase Quest, Map Conquest, and Vocab Flash)

We validated the three clusters using their reading test performance. An ANOVA was conducted to establish that the clusters significantly differed in their reading comprehension test performance, $F(2, 402) = 35.13, p < .001$. The clusters also significantly differed in their vocabulary test performance, $F(2, 402) = 15.28, \underline{p} < .001$. We conducted post-hoc Tukey *t*-test analyses to examine how the clusters differed from one another. "Struggling Readers" had lower performance than the other clusters on both the reading comprehension and vocabulary measures (*p*s < .001). "Paraphrasers" had lower reading comprehension performance than the "Inferencers" ($p < .004$), but similar vocabulary performance.

4 Discussion

The goals for this study were (1) to examine the extent to which game performance can be used to predict reading comprehension skill, as estimated by standardized tests; and (2) to explore the development of student profiles using game performance, and how these profiles relate to reading test performance. The results indicate that performance on two games—games that involved vocabulary knowledge and inferencing—accounted for approximately 45% of variance in reading comprehension skill. These results suggest that game-based learning activities can serve as formative assessments of reading skill. Teachers can use such learning activities to provide their students with more frequent feedback, and themselves with a greater understanding of their students' current skills and weaknesses.

The k-means cluster analysis demonstrated that we can develop student profiles based on game performance, and these profiles relate to reading comprehension skill and vocabulary knowledge. These results suggest that profiles can be used to determine which learning tasks might best help students improve their reading comprehension (e.g., vocabulary-building, knowledge-building, or practicing self-explaining). Students described as "Inferencers" performed well on all three games, and their high performance was reflected in their high reading comprehension scores. The "Paraphrasers" performed well on Vocab Flash and Paraphrase Quest, but they scored lower on Map Conquest. They had lower scores on the reading comprehension test than the Inferencers, while their vocabulary test performance was similar. They may need targeted instruction and practice on generating inferences and explaining texts while reading. Finally, the "Struggling Readers" did poorly on all three games. They may benefit from having more knowledge-building activities and from practice paraphrasing texts. These profiles can be formed within the context of intelligent tutoring systems to provide optimized selection of practice activities and to inform teachers about which skills students are struggling with. Students could receive practice for the skills they need most, getting continuous feedback that would further help them improve. Importantly, students' performance could be used to estimate their current reading skills.

One limitation of this study is that our analyses only included three games, one of which had low split-half reliability of .41. These games were designed primarily as learning activities rather than assessments. Future studies should examine a wider range of reading games that have been intentionally designed with both learning and assessment as equal goals. Additionally, the corpora for the games should be rigorously tested to ensure that students' reading skills can be assessed with sufficient reliability. Despite these limitations, the current study indicates the potential for games to be used as assessment of reading comprehension skill.

Acknowledgments. The research reported here was supported by The Office of Naval Research, through Grant N00014-20-1-2623 to Arizona State University. The opinions expressed are those of the authors and do not represent views of The Office of Naval Research. This study was preregistered at https://osf.io/sd7ny/.

References

1. Afflerbach, P., Cho, B.Y., Kim, J.Y.: Conceptualizing and assessing higher-order thinking in reading. Theory Pract. **54**(3), 203–212 (2015). https://doi.org/10.1080/00405841.2015.1044367
2. Crocco, F., Offenholley, K., Hernandez, C.: A proof-of-concept study of game-based learning in higher education. Simul. Gaming **47**(4), 403–422 (2016). https://doi.org/10.1177/1046878116632484
3. Fang, Y., Li, T., Huynh, L., Christhilf, K., Roscoe, R.D., McNamara, D.S.: Stealth literacy assessments via educational games. Computers **12**(7), 130 (2023). https://doi.org/10.3390/computers12070130
4. Gates-MacGinitie Reading Tests: Technical Report for Gates-MacGinitie Reading Tests Form S. Riverside Publishing (1989)
5. Gomez, M.J., Ruipérez-Valiente, J.A., Clemente, F.J.G.: A systematic literature review of game-based assessment studies: trends and challenges. IEEE Trans. Learn. Technol. (2022). https://doi.org/10.1109/TLT.2022.3226661

6. Hagaman, J.L., Casey, K.J., Reid, R.: Paraphrasing strategy instruction for struggling readers. Prev. Sch. Fail. **60**(1), 43–52 (2016). https://doi.org/10.1080/1045988X.2014.966802
7. Haynes, J.E., Fillmer, H.T.: Paraphrasing and reading comprehension. Read. World **24**(1), 76–79 (1984). https://doi.org/10.1080/19388078409557804
8. Howard-Jones, P.A., Jay, T., Mason, A., Jones, H.: Gamification of learning deactivates the default mode network. Front. Psychol. **6**, Article no. 1891 (2016). https://doi.org/10.3389/fpsyg.2015.01891
9. Kurby, C.A., Magliano, J.P., Dandotkar, S., Woehrle, J., Gilliam, S., McNamara, D.S.: Changing how students process and comprehend texts with computer-based self-explanation training. Fac. Res. Creat. Act. **25** (2012). https://doi.org/10.2190/EC.47.4.e
10. Magliano, J.P., Millis, K.K., RSAT Development Team, Levinstein, I., Boonthum, C.: Assessing comprehension during reading with the reading strategy assessment tool (RSAT). Metacognit. Learn. **6**, 131–154 (2011). https://doi.org/10.1007/s11409-010-9064-2
11. McMaster, K.L., et al.: Making the right connections: differential effects of reading intervention for subgroups of comprehenders. Learn. Individ. Differ. **22**(1), 100–111 (2012). https://doi.org/10.1016/j.lindif.2011.11.017
12. McNamara, D.S., Levinstein, I.B., Boonthum, C.: iSTART: interactive strategy training for active reading and thinking. Behav. Res. Methods Instrum. Comput. **36**(2), 222–233 (2004). https://doi.org/10.3758/BF03195567
13. Ouellette, G.P.: What's meaning got to do with it: the role of vocabulary in word reading and reading comprehension. J. Educ. Psychol. **98**(3), 554 (2006). https://doi.org/10.1037/0022-0663.98.3.554
14. Phillips, L.M., Norris, S.P., Osmond, W.C., Maynard, A.M.: Relative reading achievement: a longitudinal study of 187 children from first through sixth grades. J. Educ. Psychol. **94**(1), 3 (2002). https://doi.org/10.1037/0022-0663.94.1.3
15. Plass, J.L., Homer, B.D., Mayer, R.E., Kinzer, C.K.: Theoretical foundations of game-based and playful learning. Educ. Psychol. **50**(4), 258–283 (2020). https://doi.org/10.1080/00461520.2015.1122533
16. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2023). https://www.R-project.org/
17. Rapp, D.N., Broek, P.V.D., McMaster, K.L., Kendeou, P., Espin, C.A.: Higher-order comprehension processes in struggling readers: a perspective for research and intervention. Sci. Stud. Read. **11**(4), 289–312 (2007). https://doi.org/10.1080/10888430701530417
18. Sailer, M., Homner, L.: The gamification of learning: a meta-analysis. Educ. Psychol. Rev. **32**(1), 77–112 (2020). https://doi.org/10.1007/s10648-019-09498-w
19. Schrank, F.A., Wendling, B.J.: The Woodcock–Johnson IV: tests of cognitive abilities, tests of oral language, tests of achievement. In: Flanagan, D.P., McDonough, E.M. (eds.) Contemporary Intellectual Assessment: Theories, Tests, and Issues, pp. 383–451. The Guilford Press (2018)
20. Shute, V., et al.: Maximizing learning without sacrificing the fun: stealth assessment, adaptivity and learning supports in educational games. J. Comput. Assist. Learn. **37**(1), 127–141 (2021). https://doi.org/10.1111/jcal.12473
21. Smolkowski, K., Cummings, K.D.: Evaluation of the DIBELS (sixth edition) diagnostic system for the selection of native and proficient English speakers at risk of reading difficulties. J. Psychoeduc. Assess. **34**(2), 103–118 (2016). https://doi.org/10.1177/0734282915589017
22. Swart, E.K., Nielen, T.M., Sikkema-de Jong, M.T.: Supporting learning from text: a meta-analysis on the timing and content of effective feedback. Educ. Res. Rev. **28**, Article no. 100296 (2019). https://doi.org/10.1016/j.edurev.2019.100296
23. Wang, L., Shute, V., Moore, G.R.: Lessons learned and best practices of stealth assessment. Int. J. Gaming Comput. Med. Simul. (IJGCMS) **7**(4), 66–87 (2015). https://doi.org/10.4018/IJGCMS.2015100104

Towards Neuro-Enhanced Education: A Systematic Review of BCI-Assisted Development for Non-academic Skills and Abilities

Ziqi Pan(✉) and Alexandra I. Cristea

Department of Computer Science, Durham University, Durham, UK
{ziqi.pan2,alexandra.i.cristea}@durham.ac.uk

Abstract. Students' success in the 21st century demands not only strong academic skills but also well-developed *Non-academic Skills and Abilities* (NaSAs) such as critical thinking, concentration, and emotion regulation. The emerging field of *Brain-Computer Interfaces* (BCIs) offers intriguing possibilities for enhancing student learning and development, by providing real-time neurofeedback that could inform personalised strategies. While research on the impact of BCIs on academic skills is growing, there is limited research regarding their potential to promote NaSAs. This Systematic Literature Review (SLR) aims to address this gap, by investigating and analysing the existing research on how BCIs can be used to assist the development of these crucial skills and abilities. This SLR provides a *comprehensive analysis of 46 empirical studies between 2013–2023* selected via the well-accepted PRISMA method from 922 candidate studies. This analysis explores the diverse ways BCIs can facilitate or enhance the development of cognitive, motor, and inter-/intra- personal NaSAs, either passively or actively. The findings of this SLR offer valuable insights into the potential of BCIs to revolutionise education towards a *neuro-enhanced future that promotes not only academic achievement, but also holistic student growth.*

Keywords: Brain-Computer Interface (BCI) · 21st Century Skills · Student Development · Non-academic Skills and Abilities

1 Introduction

Characterised by the dynamic interplay of learning and teaching, education aims to equip students with essential knowledge and skills to succeed in their later life [49]. Traditionally, formal education has primarily focused on the development of students' *academic skills* that directly relate to their academic performance, such as literacy and numeracy [77]. However, recent research suggests that academic performance are not the sole determinants of student success. It has become increasingly recognised that a range of *Non-academic Skills and Abilities (NaSAs)* also play a significant role [58]. Consequently, contemporary educational frameworks are demonstrating a growing emphasis on developing

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 49–66, 2024.
https://doi.org/10.1007/978-3-031-63031-6_5

these skills and abilities in students, encompassing areas such as being able to pay attention and concentrate, confidence, problem-solving, logical ability, and emotional regulation.

Unlike academic skills, which can arguably be easily and accurately measured by quantitative measurements, such as exam scores, NaSAs are much more difficult to be assessed, due to their abstract nature [67]. In addition, compared to academic skills, which are primarily honed through formal education, NaSAs are not solely affected by school-based education, but also depend on various factors, such as personalities and family environment [32]. Thus, it is more difficult to help students to develop NaSAs, based on each individual's needs. Furthermore, the development of some NaSAs is inherently context-dependent, often requiring immediate and context-specific feedback [73]. However, providing such real-time feedback presents a significant challenge in traditional educational settings. It is not feasible for educators to monitor every student's progress and identify the optimal moment for individual feedback, due to the constraints of class sizes and the diversity of student needs. Additionally, NaSAs manifest primarily through internal cognitive processes, making it difficult to identify the optimal feedback moment, solely relying on external behavioural observations [24].

In this context, Brain-Computer Interfaces (BCIs) emerge as promising tools. A BCI can potentially provide a more intuitive and objective assessment of NaSAs, by measuring brain activity associated with specific skills or abilities [87]. In addition, BCIs empower educators to offer students real-time interventions, based on real-time neurofeedback [45]. BCI-based learning systems also have the advantages on providing a more personalised and context-specific learning experiences [94], enhancing engagement [4] and motivation [5], and improving learning effectiveness [78]. This can complement traditional methods, offering deeper insights into students' learning potential and progress.

1.1 Research Rationales and Objectives

Within the burgeoning field of BCI-assisted education, two distinct approaches emerge. The first approach utilises the output data from the BCI equipment solely as a measurement, gauging participant behaviour, like attention and concentration. Conversely, the second approach delves deeper, further leveraging output data as input to develop personalised interventions. This SLR aims to review researches using these two approaches by defining:

- **BCI-monitored education**: This approach utilises BCI technology to non-invasively monitor brain activity. Here, the BCI acts as a passive data collection tool, providing insights into brain activity, without actively influencing users' behaviour and experience. However, educators or researchers can use the data collected to deliver interventions later on.
- **BCI-based education**: This approach goes beyond mere monitoring, actively utilising real-time brain data as input to shape and adapt the educational environment or learning content. Here, the BCI itself becomes an interactive tool, generating personalised real-time interventions based on individual brain activity patterns.

While numerous literature reviews have explored the application of BCIs in education (e.g., [47,65,84,89,90]), a critical gap exists in clearly differentiating between BCI-monitored and BCI-based approaches. Most existing reviews fail to categorise studies based on this distinction, leading to ambiguity in comprehending the actual use of BCIs within the educational context. Furthermore, certain prior literature reviews have been limited in scope, solely examining one or two individual skills or abilities rather than providing a comprehensive overview (e.g., [47,89]). Therefore, this SLR aims to bridge this gap by systematically investigating and analysing research papers from 2013 to 2023 on BCIs in education through the lens of both BCI-monitored and BCI-based perspectives, with a special focus on a range of non-academic skills and abilities. We posit that this SLR is crucial for gaining a clear understanding of the current state and future potential of BCIs in education.

To address these rationales, we formulate the following research questions:

- *RQ1:* ***Which non-academic skills or abilities (NaSAs) can be facilitated or enhanced through the application of Brain-Computer Interface (BCI) technology?***
- *RQ2:* ***In what ways can BCI technology be leveraged to support students' development of non-academic skills or abilities (NaSAs)?***

2 Methodology

2.1 Searching Strategy

This SLR adhered to the reporting standards outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA 2020) guidelines [63]. A search for relevant literature was conducted through the following reputable online databases: ScienceDirect, IEEEXplore, and Scopus. The literature search spanned a decade (2013–2023) to capture the latest advancements in BCI applications within the educational context. Additionally, the search was narrowed by focusing solely on document titles, optimising the search results and ensuring their relevance to the topic. The primary search strategy utilised for locating pertinent publications was: ("brain computer interface" OR "BCI" OR "EEG") AND ("skill" OR "ability" OR "educat*" OR "student" OR "learn*"). As BCIs are relatively new, it is reasonable to expect their mention in the title; similarly, EEG are currently the most common signal type used in BCI, thus these are also considered relevant (e.g., [26,55]); the rest of the keywords represent the application domain.

2.2 Selection Process

Following title-based identification of potentially relevant studies, data from multiple databases were aggregated and duplicate publications were excluded. Subsequently, abstracts were screened, resulting in the exclusion of studies not aligned with the field of education. For instance, some papers with the word "learn" or "learning" in the title are actually about machine learning or deep learning, rather than the learning process of students. Thus, these papers were excluded, due to incompatible research field. Full texts of the remaining eligible studies were then retrieved and assessed against other pre-defined inclusion and exclusion criteria (see Table 1). Following the final selection of papers, a supplementary citation search was conducted, to identify and include additional relevant studies not indexed in the three databases initially used. To ensure selection reliability, this systematic literature review employed a cross-checking method between authors. The full identification and selection process is explained in Fig. 1.

Table 1. Inclusion and Exclusion Criteria

	Inclusion Criteria	Exclusion Criteria
Publication Year	*IC1:* Published between January 2013 and December 2023	*EC1:* Published outside this year rage
Source Reliability and Validity	*IC2:* Reviewed or peer-reviewed empirical studies	*EC2:* Not reviewed or peer-reviewed empirical studies (e.g., reviews, blog)
Language	*IC3:* English	*EC3:* Non-English
Participants	*IC4:* Students or children at different stages, from kindergarten to university	*EC4:* Not students or children
Topic	*IC5:* Investigate the use of BCI on students' or children's non-academic skills and abilities	*EC5.1:* -Did not use BCI or related technology *EC5.2:* -Not in the field of education or child development *EC5.3:* -Not related to non-academic skills or abilities

2.3 Data Extraction

The following information of each selected article was extracted: the characteristics of the participants, the type(s) of NaSAs addressed, experimental design, tasks performed, and the findings. We categorised the NaSAs investigated into three types:

1. **Cognitive NaSAs**: This category refers to mental skills or abilities associated with cognitive processes, such as attention, concentration, and critical thinking.
2. **Motor NaSAs**: This category refers to any skills and abilities related to physical movement, coordination, and interaction with the environment. It is important to note that studies in this category do not necessarily involve actual physical activities. As long as their target brain activity related to these functions, they will be included.
3. **Interpersonal and Intrapersonal NaSAs**: This category refers to skills and abilities related to interact with other individuals or related to one's self-management, such as communication skills and emotion regulation.

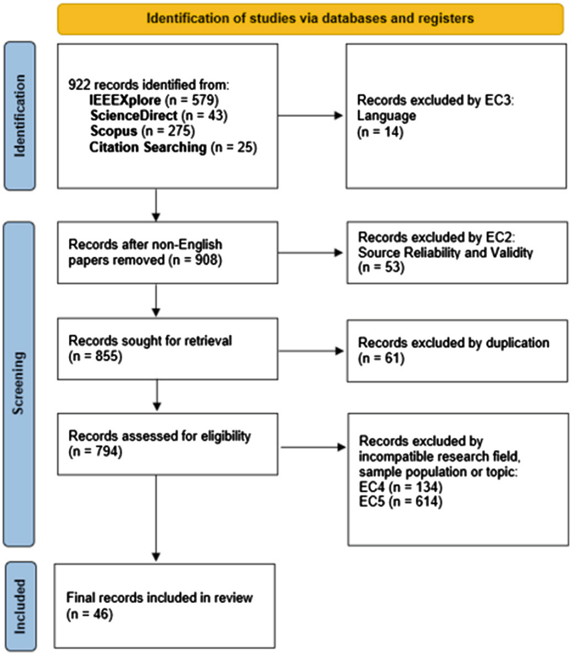


Fig. 1. Publication Selection Process

3 Results

This SLR thus includes 46 papers out of 922 candidate studies. EC 1,2 and 3 were automatically applied during the string search in the digital databases. For EC 4 and EC 5, papers that were not eligible were excluded manually (Table 1).

3.1 RQ1: Which Non-academic Skills or Abilities (NaSAs) Can Be Facilitated or Enhanced Through the Application of Brain-Computer Interface (BCI) Technology?

To answer this and the next research question, Table 2 offers a detailed breakdown of the NaSA types and participant types included in each study, as well as whether the study is BCI-monitored or BCI-based. Our analysis of the identified studies revealed that 47.83% employed BCI technology for monitoring purposes, while a surprisingly higher proportion of 52.17% actively utilised BCI for interventions.

Figure 2 visually summarises the distribution of specific NaSA types and participant types. For studies involved participants under 18, all of them claimed that they followed the ethical guidelines provided by their research institute and received consent from both the participants and their guardians. According to our analyses, 32.61% of the studies involved primary school students, 15.22%

Table 2. BCI-monitored and BCI-based Studies Categorised by Participant Type and NaSAs Type

	Participant	Non-academic Skills and Abilities (NaSAs)		
		Cognitive	*Motor*	*Inter/Intra-personal*
BCI-Monitored	*Primary School Students*	[2]	[92]	[44,53]
	High School Students	[26]	[62]	[96]
	University Students	[16,29,39,72,86] [1,7,36,56,83]		[7,40,55,56]
	Other Types of Students	[34,42]	[69]	[42]
BCI-Based	*Primary School Students*	[13,41,54,66,79] [35,50,59]	[46]	[6,74]
	High School Students	[15]	[10,19]	[95]
	University Students	[76]	[93]	[31]
	Other	[52,60]	[12,23,37]	[70]

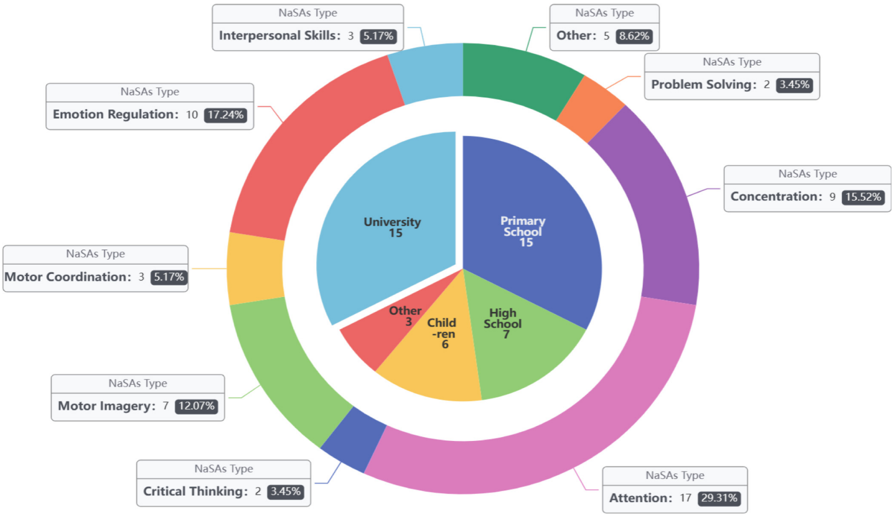


Fig. 2. An Overview of NaSAs Type and Participants Type *Note. Some research investigated more than one NaSAs*

involved high school students, 32.61% involved university students. Within the category of "other", 13.04% involved children under 18, whose educational backgrounds varied, or were not mentioned in the paper (e.g., children with cerebral palsy age from 6 to 16 [37]). 6.52% of studies involved participants whose educational background fell outside the predefined categories, such as tourism students [34,42]. In terms of NaSAs investigated in the study, there is a predominant focus on cognitive NaSAs, which comprised 56.52% of the analysed studies. Within this category, attention emerged as the most frequently investigated cognitive NaSA (17 studies), followed by concentration (9 studies). Research exploring motor NaSAs constituted only 21.74% of the studies, encompassing areas like motor imagery (7 studies) and motor coordination (3 studies). Finally, 28.26% of the studies belonged to the interpersonal and intrapersonal NaSAs category, investigating the use of BCI for the development of skills such as social interaction (3 studies) and emotion regulation (10 studies). Notably, some studies assessed multiple NaSAs. For example, [21] investigated memory, attention, and logic within a single study. [42] monitored both the brain activities reflecting concentration and anxiety in their study. Some subcategories in Table 2 remained empty, due to the absence of studies meeting all the defined inclusion criteria.

3.2 RQ2: In What Ways Can BCI Technology Be Facilitated or Leveraged to Support Students' Development of Non-academic Skills or Abilities (NaSAs)?

As can be seen from Fig. 3, there are various ways BCIs can be used or leveraged to assist students' development of NaSAs. To start with, BCIs have the potential to monitor students' brain activity and provide an indication of their current NaSAs performance. As evidenced by the literature reviewed in this SLR, different types of NaSAs are associated with distinct frequency bands in brain activity. For example, studies investigating cognitive workload typically focus on the beta wave and the gamma wave which related to concentration. (e.g., [76]), while research on motor NaSAs often examines the sensorimotor rhythm (12–15 Hz) and mid-beta band (15–20 Hz) [46]. Similarly, studies exploring interpersonal and intrapersonal NaSAs tend to focus on the alpha and theta bands, and the ratio between them [6,31].

Leveraging its monitoring capabilities, BCIs can offer students real-time feedback, fostering their self-awareness of their current NaSAs level which assist in their NaSAs development. This empowers students to adapt their learning behaviours strategically to achieve optimal outcomes. One example of this use of BCIs is the study conducted by Cai et al. (2022) [13], in which students perform various tasks in augmented reality like bending a spoon, blowing up a balloon, and launching rockets, using their attention. This innovative design provides students with real-time feedback on their current attention state, enabling them to learn and develop their ability to focus more effectively. Similarly, BCIs can also assist children in recognising their current anxiety levels and subsequently working towards reducing them [74].

Furthermore, BCIs present the potential to actively promote the development of NaSAs, by influencing the brain's internal structure. Emerging evidence suggests that BCI-based interventions can induce neuroplastic changes, potentially leading to enhanced cognitive and social capabilities. For instance, Qian et al. (2018) [70] investigated the effects of an 8-week BCI-based attention intervention on children with ADHD. Their findings revealed that the intervention modified brain network connectivity, specifically by reducing connectivity within specific networks and between task-positive networks and subcortical regions. These changes coincided with improvements in ADHD symptoms and enhanced social abilities. Additionally, BCIs can potentially facilitate targeted modulation of specific brain activity, fostering the development of various aspects of NaSAs. As an example, Gruzelier et al. (2014) [31] demonstrated improved creative thinking in participants, following BCI-based neurofeedback training. This study employed auditory feedback contingent upon the participants' brainwave activity (theta and alpha waves) with the aim of enhancing theta activity while maintaining wakefulness.

Beyond its capabilities modulating brain internal structure and activity, BCIs also hold potential for enabling individuals to control external objects and interact with the surrounding environment through their brain activity. This approach is particularly relevant for enhancing motor NaSAs, as evidenced by studies exploring muscle activation (e.g., [46]) and motor imagery (e.g., [92]).

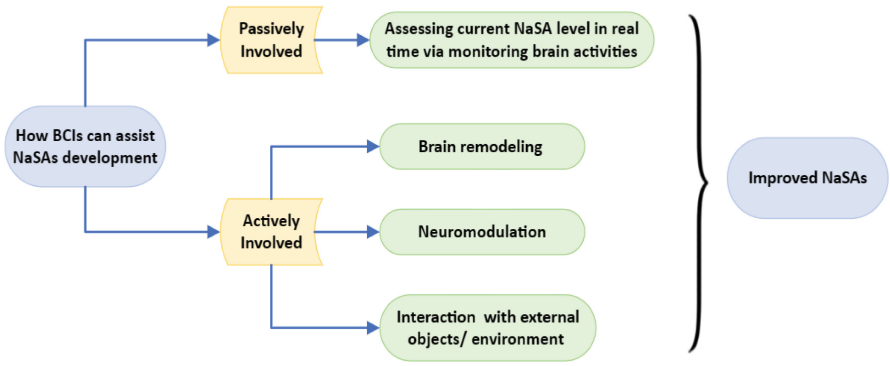


Fig. 3. BCI Taxonomy for NaSAs development

4 Discussion

Our findings have culminated in the establishment of the first-ever taxonomy for Non-Academic Skills and Abilities (NaSAs) development, differentiated based on whether the use of BCIs within the study is passive or active (see Fig. 3). Furthermore, this SLR contributes to the field by investigating which types of NaSAs can be developed with the assistance of BCIs and elucidating the specific mechanisms by which BCIs can facilitate or enhance this developmental process.

4.1 Brain Computer Interface (BCI) and Its Application in Education

Our SLR exhibits that BCIs have gained considerable attention within education, particularly concerning the development of NaSAs across various age groups. BCIs have shown advantages in both passively monitoring and actively being involved in the development of cognitive, motor, as well as interpersonal and intrapersonal skills and abilities. Our analysis reveals that BCI achieved such purposes via various methods, ranging from change the internal brain structure and activities of users to improve the self-awareness of one's current NaSAs level. Furthermore, there is a growing trend toward integrating BCIs with other cutting-edge technologies like VR [68], AR [13], and electromyographic biofeedback [59] to enhance the processes and experiences of learning and overall development. The potential benefits of BCIs extend across diverse student populations, encompassing both individuals with typical development and those with mental, physical, or learning disabilities.

Although the majority of studies included in this SLR collected and analysed the entirety of the data captured by the BCI equipment, rather than targeting specific brain regions known to be associated with the measured NaSAs, we can still observe some differences in the brain areas investigated. For example, motor NaSA development appears to be most closely linked to the motor cortex which is in charge of movement, and parietal cortex which is related to perception and sensory. For instance, [92] developed a motor imagery (MI)-based training system to aid motor rehabilitation in children with cerebral palsy. During this process, they monitored the children's sensorimotor cortex, premotor and supplementary motor cortices, and the parietal cortex. Interpersonal and intrapersonal NaSAs, such as emotion regulation, mostly correlates with temporal lobe and orbitofrontal cortex (e.g., [53]). And studies investigating cognitive NaSAs development mostly monitored the prefrontal cortex, which is associated with the attention and concentration (e.g., [36]).

This observed correlation between brain activity and specific skills strengthens the argument for the potential of BCIs in education. We argue that BCIs hold a great promise in facilitating the overall development of students. Our SLR reveals that BCIs can offer a more comprehensive understanding of and greatly assist in the development of various NaSAs. Additionally, existing research demonstrates the widespread application of BCIs in enhancing students' or children's academic development. For instance, by offering an objective, continuous, and intuitive measure of brain activity, BCIs can effectively reflect students' learning processes as well as cognitive and affective states, which may assist in identifying potential difficulties faced in learning and enhancing the learning experience and outcome [38]. In addition, BCIs also show advantages in delivering personalised learning, which aims to tailor learning experiences to individual student needs in real-time. By monitoring brain activity and identifying areas of difficulty, the system can adjust the pace, complexity, and teaching methods to optimise learning outcomes and experience [27]. Moreover, BCI can be used to create interactive learning environments that respond to students' brain activity in real-time, which can increase engagement and motivation [4]. Therefore, the use of BCIs in education can complement traditional methods, better assisting students' overall development.

4.2 Non-academic Skills and Abilities (NaSAs) in Education

One of the main purposes of education is to equip students with knowledge and fostering their ability to learn and adapt in a ever-changing world. For students, successfully mastering knowledge involves two processes: acquiring and retaining information, followed by understanding and applying that knowledge to solve problems [80]. These two processes are affected by various factors, ranging from external factors, such as the classroom environment [33] and teaching tools [20] to internal factors, such as the student's personality [18] and cognitive skills [14]. Successful knowledge acquisition and retention rely on NaSAs such as being able to pay attention and concentrate, while knowledge comprehension and application of knowledge involve other NaSAs, such as critical thinking and problem-solving. Therefore, although NaSAs are not directly related to academic content itself, they still have impact on students' academic success [64].

Beyond academic development, NaSAs also correlate with other crucial aspects of student development, such as physical and mental well-being. For instance, studies have found that NaSAs such as motor skills are associated with students' physical development [3], while emotional regulation [30] and stress management [43] contribute to students' mental well-being. Thus, by cultivating these NaSAs, students are empowered to excel not just in the academic realm, but also in their overall development.

Well-developed NaSAs continue to benefit students long after they finished their school years. One example of this is that certain NaSAs can help students excel in their professional careers. As stated by The World Economic Forum (2023), critical thinking is one of the most crucial employability skills [91]. Other NaSAs, such as communication [48], teamwork [82], and problem-solving [17] are also highly desired abilities in employment. Beyond professional success, well-developed NaSAs can also empower individuals to become resilient to life's changes and adapt swiftly [75]. What is more, individuals with strong NaSAs, such as communication skills, are better at building healthy relationships in their lives [71].

Therefore, understanding how NaSAs can be fostered and how educators can better equip students with these capabilities is crucial, which demonstrates the importance of our SLR.

4.3 Current Limitations and Future Avenues

While this SLR offers valuable insights into the current use of BCIs for developing NaSAs, there are some limitations that cannot be ignored. One example of this is the lack of specific BCI equipment analysis in this SLR, which might limit comparability of results across studies as differences in equipment can impact measurement and data presentation. We also did not discuss how limitations inherent to the BCI equipment employed in some studies restricted the scope of their analysis. For instance, due to the limited number of electrodes, these studies could only measure activity in a specific, confined portion of the brain (e.g., [76] used Neurosky mindwave which only has one electrode). This restricted approach impedes a comprehensive understanding of brain activity changes during NaSA development. In a addition, our keyword search in document titles may have missed relevant studies without those terms in their titles.

Beyond the limitations of this specific SLR, a broader concern regarding research in the field of BCIs is a potential bias in the limited types of NaSAs investigated. Notably, limited research has been conducted on the impact of BCIs on interpersonal skills and abilities like communication and teamwork. While it is true that interpersonal skills and abilities often involve interacting with the external environment and other individuals which could potentially introducing noise into brain activity data and complicating interpretation [25,61], they have a crucial role in student success and well-being [81]. BCIs could be a good option for assessing such skills and abilities as current assessment methods for these interpersonal skills or abilities often lack objectivity [85], while BCIs can provide more reliable and quantifiable measures.

In addition, a further limitation lies in the imbalanced participants sample. We observed a prevalent use of university students in this field. While these individuals represent a readily accessible participant pool, some of their non-academic skills may be already well-developed [8,22], potentially hindering investigations into the development process itself. Consequently, including a broader range of younger participants in future studies is crucial to gain a deeper understanding of how NaSAs develop across different age groups and to explore how BCIs can facilitate this development in younger populations. However, conducting research with younger age groups presents challenges, including the ethics complexities of obtaining informed consent from both children and their guardians [11]. Additionally, most BCIs are designed for adults, potentially compromising accuracy and reliability due to size incompatibility with children's heads. To gain a more comprehensive understanding of BCIs' role in fostering children's NaSAs development, future research should consider exploring adaptations to existing BCI technology to address potential size issues. Moreover, the field exhibits a bias towards studies related to motor NaSAs involving participants with physical impairments, while the potential benefits for healthy individuals remain underexplored. Investigating the use of BCIs by healthy students for learning physical activities or by young athletes to optimise training could offer valuable insights and broaden the scope of research in this area.

What is more, we observe a notable lack of research exploring the impact of individual differences within the use of BCIs on students' NaSAs development. While a number of studies included in this SLR incorporated participants' demographic information (e.g., gender) and individual characteristics (e.g., personality), very few studies have specifically investigated the moderating effects of these factors on the efficacy of BCI-assisted interventions in developing NaSAs. However, the uniqueness of each individual's brain should not be ignored. Previous studies demonstrate significant individual variation in brain activity [88], such as performing cognitive tasks [57]. In addition, Gong et al. (2011) [28] argued that researchers must consider gender when designing or interpreting studies of brain activity, as male and female brains exhibit differences in network topology, indicating the organisational patterns of whole-brain connectivity. Similarly, Bell et al. [9] also found that males and females differ in brain activation during cognitive tasks. Moreover, differences in personality can also affect brain responses and patterns [51]. Therefore, it is worthy to take these individual differences into account in future research.

5 Conclusion

By exploring the use of BCI technology to facilitate and enhance the development of NaSAs in students and children, this SLR provides a comprehensive understanding of BCIs' potential to revolutionise educational practices and promote holistic student development. This review analysed 46 empirical studies and revealed the promising use of BCI across the development of a diverse range of NaSAs, encompassing cognitive (e.g., attention, concentration), motor (e.g., motor imagery, coordination), and interpersonal and intrapersonal (e.g., emotion regulation, social interaction) skills and abilities. This review shows that BCIs offer a unique approach, by monitoring brain activity and providing real-time feedback on it, enabling students to become more self-aware of their current NaSAs state and adjust their learning behaviours, accordingly. Furthermore, BCI-based interventions hold the potential to induce brain structure and activity changes, potentially leading to enhanced NaSAs. These findings warrant further exploration and emphasise the need for educators and researchers to collaboratively explore the potential of BCIs in developing personalised learning strategies that effectively promote NaSAs development in students.

For future research, a comprehensive BCI equipment analysis would be beneficial. We also suggest that future research to include more children at the crucial development age. Additionally, future research is needed to investigate the use of BCIs on the development of interpersonal NaSAs such as teamwork, and explore the moderating effects of individual differences (e.g., gender, personality) on the efficacy of BCI-assisted interventions.

References

1. Aggarwal, S., Lamba, M., Verma, K., Khuttan, S., Gautam, H.: A preliminary investigation for assessing attention levels for massive online open courses learning environment using EEG signals: an experimental study. Hum. Behav. Emerg. Technol. **3**(5), 933–941 (2021)
2. Aleksandrova, N.A., Khramova, M.V., Hramov, A.E.: Study of the elementary school students' attention functioning on the basis of the brain-computer interface. In: 2019 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS), pp. 448–451. IEEE (2019)
3. Anderson, D.I.: Motor development: far more than just the development of motor skills. Kinesiol. Rev. **7**(2), 99–114 (2018)
4. Andujar, M., Gilbert, J.E.: Let's learn! enhancing user's engagement levels through passive brain-computer interfaces. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 703–708 (2013)
5. Angeloni, C., Salter, D., Corbit, V., Lorence, T., Yu, Y., Gabel, L.: P300-based brain-computer interface memory game to improve motivation and performance. In: 2012 38th Annual Northeast Bioengineering Conference (NEBEC), pp. 35–36. IEEE (2012)
6. Antle, A.N., Chesick, L., Sridharan, S.K., Cramer, E.: East meets west: a mobile brain-computer system that helps children living in poverty learn to self-regulate. Pers. Ubiquit. Comput. **22**, 839–866 (2018)

7. Apicella, A., Arpaia, P., Frosolone, M., Improta, G., Moccaldi, N., Pollastro, A.: EEG-based measurement system for monitoring student engagement in learning 4.0. Sci. Rep. **12**(1), 5857 (2022)
8. Beauchamp, M.H., Anderson, V.: Social: an integrative framework for the development of social skills. Psychol. Bull. **136**(1), 39 (2010)
9. Bell, E.C., Willson, M.C., Wilman, A.H., Dave, S., Silverstone, P.H.: Males and females differ in brain activation during cognitive tasks. Neuroimage **30**(2), 529–538 (2006)
10. Beveridge, R., Wilson, S., Coyle, D.: Can teenagers control a 3D racing game using motion-onset visual evoked potentials? Brain Comput. Interfaces **4**(1–2), 102–113 (2017)
11. Birbeck, D.J., Drummond, M.J.: Research with young children: Contemplating methods and ethics. J. Educ. Enquiry **7**(2) (2007)
12. Breshears, J.D., et al.: Decoding motor signals from the pediatric cortex: implications for brain-computer interfaces in children. Pediatrics **128**(1), e160–e168 (2011)
13. Cai, S., Liu, Z., Liu, C., Zhou, H., Li, J.: Effects of a BCI-based AR inquiring tool on primary students' science learning: a quasi-experimental field study. J. Sci. Educ. Technol. **31**(6), 767–782 (2022)
14. Carneiro, P., Crawford, C., Goodman, A.: The impact of early cognitive and non-cognitive skills on later outcomes (2007)
15. Chen, C.M., Huang, S.H.: Web-based reading annotation system with an attention-based self-regulated learning mechanism for promoting reading performance. Br. J. Edu. Technol. **45**(5), 959–980 (2014)
16. Chen, C.M., Wang, J.Y., Yu, C.M.: Assessing the attention levels of students by using a novel attention aware system based on brainwave signals. Br. J. Edu. Technol. **48**(2), 348–369 (2017)
17. De Fruyt, F., Wille, B., John, O.P.: Employability in the 21st century: complex (interactive) problem solving and other essential skills. Ind. Organ. Psychol. **8**(2), 276–281 (2015)
18. De Raad, B., Schouwenburg, H.C.: Personality in learning and education: a review. Eur. J. Pers. **10**(5), 303–336 (1996)
19. Dion Kelly, M., Kirton, A.: Pediatric brain-computer interface competency: a pilot study. In: 8th Graz Brain-Computer Interface Conference, vol. 2019, pp. 1–6 (2019)
20. Eady, M., Lockyer, L.: Tools for learning: technology and teaching. In: Learning to Teach in the Primary School, pp. 71–89 (2013)
21. Esquicha-Tejada, J.D., Pari-Larico, S., Llerena-Urday, B., Álvaro, del Carpio, F., Rosas-Paredes, K.: Interactive toy to strengthen the memory, attention and logic of primary education students using sphero, arduino and neurosky mindwave EEG (2020), https://api.semanticscholar.org/CorpusID:235433556
22. Van der Fels, I.M., Te Wierike, S.C., Hartman, E., Elferink-Gemser, M.T., Smith, J., Visscher, C.: The relationship between motor skills and cognitive skills in 4–16 year old typically developing children: a systematic review. J. Sci. Med. Sport **18**(6), 697–703 (2015)
23. Floreani, E.D., Rowley, D., Kelly, D., Kinney-Lang, E., Kirton, A.: On the feasibility of simple brain-computer interface systems for enabling children with severe physical disabilities to explore independent movement. Front. Hum. Neurosci. **16**, 1007199 (2022)
24. Floyd, R.: Assessment of cognitive abilities and cognitive processes. Practical Handbook of School Psychology: Effective Practices for the 21st Century, pp. 48–66 (2010)

25. Frith, C.D.: The social brain? Philos. Trans. R. Soc. B Biol. Sci. **362**(1480), 671–678 (2007)
26. Fuentes-Martinez, V.J., Romero, S., Lopez-Gordo, M.A., Minguillon, J., Rodríguez-Álvarez, M.: Low-cost EEG multi-subject recording platform for the assessment of students' attention and the estimation of academic performance in secondary school. Sensors **23**(23), 9361 (2023)
27. Geeta, N., Gavas, R.D.: Enhanced learning with abacus and its analysis using BCI technology. Int. J. Mod. Educ. Comput. Sci. **6**(9), 22 (2014)
28. Gong, G., He, Y., Evans, A.C.: Brain connectivity: gender makes a difference. Neuroscientist **17**(5), 575–591 (2011)
29. González-Hernández, H.G., Medina-Pozos, J.M., Cantú-González, V., Amozurrutia-Elizalde, A., Flores-Amado, A., Mora-Salinas, R.J.: Looking for experimental evidence of critical thinking through EEG. Int. J. Interact. Des. Manuf. **15**(2–3), 333–351 (2021)
30. Gross, J.J., Muñoz, R.F.: Emotion regulation and mental health. Clin. Psychol. Sci. Pract. **2**(2), 151 (1995)
31. Gruzelier, J.H., Thompson, T., Redding, E., Brandt, R., Steffert, T.: Application of alpha/theta neurofeedback and heart rate variability training to young contemporary dancers: state anxiety and creativity. Int. J. Psychophysiol. **93**(1), 105–111 (2014)
32. Gutman, L.M., Schoon, I.: The impact of non-cognitive skills on outcomes for young people. A literature review (2013)
33. Hannah, R.: The Effect of Classroom Environment on Student Learning (2013)
34. Hsu, L.: A tale of two classes: tourism students' cognitive loads and learning outcomes in face-to-face and online classes. J. Hospitality Leisure Sport Tourism Educ. **29**, 100342 (2021)
35. Huang, J., et al.: Focus: enhancing children's engagement in reading by using contextual BCI training sessions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1905–1908 (2014)
36. Ishtiaque, F., Motahar, E.B., Sajal, M.S.R., Amin, M.A., Sarker, F., Mamun, K.A.: Students' attention classification during class lecture using BCI and machine learning: a pathway towards neurofeedback-based learning. In: 2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pp. 1–6. IEEE (2023)
37. Jadavji, Z., Zewdie, E., Kelly, D., Kinney-Lang, E., Robu, I., Kirton, A.: Establishing a clinical brain-computer interface program for children with severe neurological disabilities. Cureus **14**(6), e26215 (2022)
38. Jamil, N., Belkacem, A.N., Ouhbi, S., Guger, C.: Cognitive and affective brain-computer interfaces for improving learning strategies and enhancing student capabilities: a systematic literature review. IEEE Access **9**, 134122–134147 (2021)
39. Jamil, N., Lakas, A., Belkacem, A.N.: Improving students' cognitive abilities in online environment based on neurofeedback. In: 2023 IEEE Global Engineering Education Conference (EDUCON), pp. 1–3. IEEE (2023)
40. Jiang, J., et al.: Leader emergence through interpersonal neural synchronization. Proc. Natl. Acad. Sci. **112**(14), 4274–4279 (2015)
41. Johnstone, S.J., Roodenrys, S.J., Johnson, K., Bonfield, R., Bennett, S.J.: Game-based combined cognitive and neurofeedback training using focus pocus reduces symptom severity in children with diagnosed AD/HD and subclinical AD/HD. Int. J. Psychophysiol. **116**, 32–44 (2017)

42. Juan, Y.K., Chen, Y.: The influence of indoor environmental factors on learning: an experiment combining physiological and psychological measurements. Build. Environ. **221**, 109299 (2022)
43. Keogh, E., Bond, F.W., Flaxman, P.E.: Improving academic performance and mental health through a stress management intervention: outcomes and mediators of change. Behav. Res. Ther. **44**(3), 339–357 (2006)
44. Khng, K.H., Mane, R.: Beyond BCI-validating a wireless, consumer-grade EEG headset against a medical-grade system for evaluating EEG effects of a test anxiety intervention in school. Adv. Eng. Inform. **45**, 101106 (2020)
45. Khorshidtalab, A., Salami, M.J.E.: EEG signal classification for real-time brain-computer interface applications: a review. In: 2011 4th International Conference on Mechatronics (ICOM), pp. 1–7. IEEE (2011)
46. Kim, T.W., Lee, B.H.: Clinical usefulness of brain-computer interface-controlled functional electrical stimulation for improving brain activity in children with spastic cerebral palsy: a pilot randomized controlled trial. J. Phys. Ther. Sci. **28**(9), 2491–2494 (2016)
47. Kinney-Lang, E., Auyeung, B., Escudero, J.: Expanding the (kaleido) scope: exploring current literature trends for translating electroencephalography (EEG) based brain-computer interfaces for motor rehabilitation in children. J. Neural Eng. **13**(6), 061002 (2016)
48. Kleckner, M.J., Marshall, C.: Critical communication skills: developing course competencies to meet workforce needs. J. Res. Bus. Educ. **56**(2), 59–81 (2014)
49. Kohlberg, L., Mayer, R.: Development as the aim of education. Harv. Educ. Rev. **42**(4), 449–496 (1972)
50. Kosmyna, N., Soetaert, N., Scheirer, C.: A pilot study of using brain-computer interfaces in classrooms for promoting formal educational activities. In: Arai, K. (ed.) FTC 2021. LNNS, vol. 360, pp. 590–609. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-89912-7_46
51. Kumari, V., Williams, S.C., Gray, J.A., et al.: Personality predicts brain responses to cognitive demands. J. Neurosci. **24**(47), 10636–10641 (2004)
52. Lee, P.J., Chin, S.W.: Early childhood educator assistant with brain computer interface. In: International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014, pp. 52–57. IET (2014)
53. Leventon, J.S., Stevens, J.S., Bauer, P.J.: Development in the neurophysiology of emotion processing and memory in school-age children. Dev. Cogn. Neurosci. **10**, 21–33 (2014)
54. Lim, C.G., et al.: A randomized controlled trial of a brain-computer interface based attention training program for ADHD. PLoS ONE **14**(5), e0216225 (2019)
55. Liu, L., Ji, Y., Gao, Y., Li, T., Xu, W.: A novel stress state assessment method for college students based on EEG. Comput. Intell. Neurosci. **2022**, 4565968 (2022)
56. Liu, N., Zhang, Y., Mark, G., Li, Z., Rau, P.-L.P.: Mindfulness meditation: investigating immediate effects in an information multitasking environment. In: Rau, P.-L.P. (ed.) HCII 2019. LNCS, vol. 11576, pp. 531–542. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22577-3_38
57. Locke, H.S., Braver, T.S.: Motivational influences on cognitive control: behavior, brain activation, and individual differences. Cogn. Affect. Behav. Neurosci. **8**(1), 99–112 (2008)
58. Lotkowski, V.A., Robbins, S.B., Noeth, R.J.: The role of academic and non-academic factors in improving college retention. Act policy report. American College Testing ACT Inc. (2004)

59. Maurizio, S., et al.: Comparing tomographic EEG neurofeedback and EMG biofeedback in children with attention-deficit/hyperactivity disorder. Biol. Psychol. **95**, 31–44 (2014)
60. Mercado, J., Escobedo, L., Tentori, M.: A BCI video game using neurofeedback improves the attention of children with autism. J. Multimodal User Interfaces **15**, 273–281 (2021)
61. Noordzij, M.L., Newman-Norlund, S.E., De Ruiter, J.P., Hagoort, P., Levinson, S.C., Toni, I.: Brain mechanisms underlying human communication. Front. Hum. Neurosci. **3**, 478 (2009)
62. Ose Askvik, E., Van der Weel, F., van der Meer, A.L.: The importance of cursive handwriting over typewriting for learning in the classroom: a high-density EEG study of 12-year-old children and young adults. Front. Psychol. **11**, 1810 (2020)
63. Page, M.J., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. Int. J. Surg. **88**, 105906 (2021)
64. Palardy, G.J.: School peer non-academic skills and academic performance in high school. In: Frontiers in Education, vol. 4, p. 57. Frontiers (2019)
65. Papanastasiou, G., Drigas, A., Skianis, C., Lytras, M.: Brain computer interface based applications for training and rehabilitation of students with neurodevelopmental disorders. A literature review. Heliyon **6**(9), e04250 (2020)
66. Pari-Larico, J.E.T.S., Llerena-Urday, B., Carpio, A., Rosas-Paredes, K.: Interactive toy to strengthen the memory attention and logic of primary education students using sphero arduino and neurosky mindwave EEG. In: Proceedings of the 6th Iberoamerican Conference of Computer Human Interaction, vol. 2747, pp. 245–254 (2020)
67. Patterson, F., Galbraith, K., Flaxman, C., Kirkpatrick, C.M.: Evaluation of a situational judgement test to develop non-academic skills in pharmacy students. Am. J. Pharm. Educ. **83**(10), 7074 (2019)
68. Pietroszek, K., Tahai, L.: Scalebridge VR: immersive proportional reasoning game for children with brain-computer interface for difficulty scaling. In: Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology, pp. 1–1 (2019)
69. Pineda, J.A., Carrasco, K., Datko, M., Pillen, S., Schalles, M.: Neurofeedback training produces normalization in behavioural and electrophysiological measures of high-functioning autism. Philos. Trans. R. Soc. B Biol. Sci. **369**(1644), 20130183 (2014)
70. Qian, X., et al.: Brain-computer-interface-based intervention re-normalizes brain functional network topology in children with attention deficit/hyperactivity disorder. Transl. Psychiatry **8**(1), 149 (2018)
71. Rosenberg, M.B., Chopra, D.: Nonviolent Communication: A Language of Life: Life-Changing Tools for Healthy Relationships. PuddleDancer Press (2015)
72. Roslan, N.S., Izhar, L.I., Saad, M.N.M., Sivapalan, S.: Innovations for teaching and learning: an exploration of the potential of EEG for individual and group learning activities. In: 2015 IEEE Student Conference on Research and Development (SCOReD), pp. 384–388. IEEE (2015)
73. Ruitenberg, M.F., De Kleine, E., Van der Lubbe, R.H., Verwey, W.B., Abrahamse, E.L.: Context-dependent motor skill and the role of practice. Psychol. Res. **76**, 812–820 (2012)
74. Schoneveld, E.A., Malmberg, M., Lichtwarck-Aschoff, A., Verheijen, G.P., Engels, R.C., Granic, I.: A neurofeedback video game (mindlight) to prevent anxiety in children: a randomized controlled trial. Comput. Hum. Behav. **63**, 321–333 (2016)

75. Schoon, I.: Risk and Resilience: Adaptations in Changing Times. Cambridge University Press (2006)
76. Sethi, C., Dabas, H., Dua, C., Dalawat, M., Sethia, D.: EEG-based attention feedback to improve focus in e-learning. In: Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, pp. 321–326 (2018)
77. Shapiro, E.S., Keller, M.A.: Academic skills problems. In: Clinician's Handbook of Child Behavioral Assessment, pp. 605–630. Elsevier (2006)
78. Shen, Y.T., Chen, X.M., Lu, P.W., Wu, J.C.: Use BCI to generate attention-based metadata for the assessment of effective learning duration. In: Zaphiris, P., Ioannou, A. (eds.) LCT 2018. LNCS, vol. 10925, pp. 407–417. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91152-6_31
79. Shereena, E., Gupta, R., Bennett, C., Sagar, K., Rajeswaran, J.: EEG neurofeedback training in children with attention deficit/hyperactivity disorder: a cognitive and behavioral outcome study. Clin. EEG Neurosci. **50**(4), 242–255 (2019)
80. Spiro, R.J., Vispoel, W.P., Schmitz, J.G., Samarapungavan, A., Boerger, A.: Knowledge acquisition for application: cognitive flexibility and transfer in complex content domains. In: Executive control processes in reading, pp. 177–199. Routledge (2013)
81. Spitzberg, B.H., Cupach, W.R.: Interpersonal skills. The Sage Handbook of Interpersonal Communication, vol. 4, p. 481527 (2011)
82. Suarta, I.M., Suwintana, I.K., Sudhana, I.F.P., Hariyanti, N.K.D.: Employability skills required by the 21st century workplace: a literature review of labor market demand. In: International Conference on Technology and Vocational Teachers (ICTVT 2017), pp. 337–342. Atlantis Press (2017)
83. Tang, H., Dai, M., Du, X., Hung, J.L., Li, H.: An EEG study on college students' attention levels in a blended computer science class. Innov. Educ. Teach. Int. 1–13 (2023)
84. Tangermann, M., et al.: Review of the BCI competition IV. Front. Neurosci. **6**, 55 (2012)
85. Tanner, K.: Increasing objectivity in the assessment of interpersonal skills and attitude. J. Paramedic Pract. **6**(11), 566–571 (2014)
86. Udayana, I.P.A.E.D., Sudarma, M., Putra, I.K.G.D., Sukarsa, I.M.: Effect of different online learning screen sizes during the COVID-19 pandemic: an EEG study. In: 2022 International Conference on Data and Software Engineering (ICoDSE), pp. 47–52. IEEE (2022)
87. Vaid, S., Singh, P., Kaur, C.: EEG signal analysis for BCI interface: a review. In: 2015 fifth International Conference on Advanced Computing & Communication Technologies, pp. 143–147. IEEE (2015)
88. Van Horn, J.D., Grafton, S.T., Miller, M.B.: Individual variability in brain activity: a nuisance or an opportunity? Brain Imaging Behav. **2**, 327–334 (2008)
89. Wang, J.W., Zhang, D.W., Johnstone, S.J.: Portable EEG for assessing attention in educational settings: A scoping review (2023)
90. Wegemer, C.: Brain-computer interfaces and education: the state of technology and imperatives for the future. Int. J. Learn. Technol. **14**(2), 141–161 (2019)
91. World Economic Forum: The future of jobs: Employment, skills, and workforce strategy for the fourth industrial revolution (2023). https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf
92. Xie, J., et al.: Rehabilitation of motor function in children with cerebral palsy based on motor imagery. Cogn. Neurodyn. **15**, 939–948 (2021)
93. Yang, C., Ye, Y., Li, X., Wang, R.: Development of a neuro-feedback game based on motor imagery EEG. Multimedia Tools Appl. **77**, 15929–15949 (2018)

94. Zaharija, G., Bogunović, P., Mladenović, S.: Brain computer interface in enhanced learning system. In: INTED2018 Proceedings, pp. 198–205. IATED (2018)
95. Zhang, J., Jadavji, Z., Zewdie, E., Kirton, A.: Evaluating if children can use simple brain computer interfaces. Front. Hum. Neurosci. **13**, 24 (2019)
96. Zhang, W., et al.: Event-related synchronization of delta and beta oscillations reflects developmental changes in the processing of affective pictures during adolescence. Int. J. Psychophysiol. **90**(3), 334–340 (2013)

From Novice to Expert: Unraveling the Impact of Experience on Cognitive Load and Physiological Responses in Aviation Pilots

Amin Bonyad Khalaj[1], Hamdi Ben Abdessalem[2](✉), and Claude Frasson[1]

[1] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Canada
amin.bonyad.khalaj@umontreal.ca, frasson@iro.umontreal.ca

[2] Département d'Informatique et de Mathématique, Université du Québec à Chicoutimi, Chicoutimi, Canada
hamdi_benabdessalem@uqac.ca

Abstract. This study explores the intricate correlation between participants' experiences and their cognitive and physiological responses in aviation. Using workload, heart rate, and facial temperature as focal points, two experiments involved a variety of participants, including real pilots, pilot engineers, and individuals with varying flight simulation backgrounds. The findings highlight distinct patterns, emphasizing how experience influences responses during flight simulation. Novices face higher heart rate, cognitive workload and heightened physiological arousal, while experienced participants exhibit controlled responses. The study underscores the non-linear relationship between experience and physiological responses, influenced by individual differences and coping mechanisms. Over time, repeated exposure leads to a learning curve, showcasing the adaptive nature of human responses in aviation.

Keywords: Pilot Experience · Mental Workload · Thermal Image · Heart Rate · EEG · Flight Safety · Human Computer Interaction

1 Introduction

Modern aviation requires a deep understanding of human cognitive and physiological performance in complex environments. Flight simulation is crucial for studying these factors. This research focuses on workload, heart rate, and facial temperature as indicators of cognitive and physiological responses during flight simulation experiments. In aviation, pilots need extensive knowledge and experience for safe operations. The impact of experience on pilot cognitive load and physiological responses has been extensively studied. Paper [1] explored the effect of familiarity on associative memory by comparing experienced pilots, novices, and non-pilots, revealing differences in cognitive processes among these groups.

Paper [2] underscored the role of pilot experience in cognitive performance during flight operations. Meanwhile, paper [3] investigated the visual scanning strategies of

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 67–79, 2024.
https://doi.org/10.1007/978-3-031-63031-6_6

professional and novice pilots during manual landings, revealing superior perceptual efficiency and attention distribution among experts. Beyond cognitive load, managing physiological responses and stress is crucial for pilot performance, as noted in paper [4], which highlighted the impact of sleep restriction on emergencies. Similarly, [5] stressed the importance of fatigue management in aviation, emphasizing the physiological and performance consequences of sleep loss and heavy workload on pilots. In aviation, communication and cooperation play crucial roles, as evidenced in the literature. Paper [6] showed that experienced pilots recall multiple flight scenarios during collaboration, emphasizing familiarity's impact on collaborative memory. Paper [7] discussed the challenge of excessive voice communication in air traffic control, which can affect pilot memory and cognitive performance. Additionally, the physiological challenges of modern high-performance aircraft are significant. Paper [8] stressed understanding physiological responses to challenging flight conditions, while [9] though highlighted mild hypoxia's impact on pilot performance, underlining the importance of physiology and mental conditions for flight safety and performance.

In this paper, we aim to investigate how pilots' experience influences their cognitive load and physiological responses. Understanding this relationship is crucial for improving pilot training, performance, and safety in aviation. Our objective is to assess the impact of pilot experience on flight performance using physiological sensors. We will measure **workload** variation using data from **electroencephalograms** (EEG) and **thermal measures** of the face across different simulated flight conditions.

2 Related Work

Evaluating pilot experience through cognitive performance, heart rate, and facial temperature is crucial for flight safety and pilot well-being. Various studies have investigated the relationship between physiological symptoms and cognitive performance in pilots. For instance, paper [10] discussed using pilot physiological assessments like EEG signals and ocular parameters to assess cognitive performance. Meanwhile, paper [11] revealed significant modulations in pilots' behavior and cognitive demand by evaluating the influence of dynamic workload and incorporating heart rate variability analysis. Moreover, in [12], the importance of facial boundary image in pilot cognitive performance was demonstrated. Research has also shown that facial expressions can elicit emotion-specific autonomic nervous system activity, complementing physiological cues like heart rate and skin temperature to assess pilot emotions and cognitive labor [13]. Machine learning and EEG were employed to differentiate pilot cognitive performance during flight, highlighting various components including environmental factors and pilot experience [2]. Paper [14] conducted a comprehensive review of employing machine learning models for heart rate estimation from facial videos. Study [15] found fixation duration to be a valuable metric for attention allocation, distinguishing between expert pilots and novices, guiding the development of training programs to improve aviation safety. Previous research [16] established a correlation between workload and EEG measures, while paper [17] identified relationships between workload, facial temperature, and heart rate, highlighting stress indicators. These studies stress integrating heart rate, facial temperature, and cognitive workload to assess pilot experience accurately. Technological advancements can improve pilot well-being and safety by understanding how physiological and psychological factors interact during challenging tasks.

3 Experiments

We conducted a study on cognitive workload, heart rate, and facial thermal responses during Airbus A320 takeoff procedures. The research involved two distinct experiments with 23 participants, including real pilots, pilot engineering without a license, individuals with limited flight simulation experience, and those with no flight experience. All participants provided written consent following ethical guidelines, including members from CAE, Bombardier, and other volunteers.

3.1 Experiment 1

In the first experiment, 13 pilots from CAE and Bombardier participated in a study that involved six different **scenarios**. The scenarios included standard takeoff sessions (scenarios 1–3), and failure sessions (takeoff with emergency situations, scenarios 4–6). Participants operated the Airbus A320 as a pilot, while the experimenter served as a pilot monitor. Data were collected in real time, including **cognitive workload and heart rate,** resulting in 136 takeoffs over a span of more than 9 h.

The study included 13 male subjects (all between 24 and 49 years of age) with an average age of 36 years. The 13 participants included 7 pilots with piloting experience, piloting license and A320 piloting experience. The other 6 participants were engineers at Bombardier and CAE who were familiar with most aircraft procedures but without holding a piloting license. Participants were assigned to two **groups**, with **group one** initiating 20-min *regular* takeoff session followed by 20-min *failure* session, and **group two** following the reverse session order. Table 1 provides a detailed breakdown of each group's sessions. This approach was adopted to compare how the sequential difficulty levels of scenarios impact both workload and heart rate.

Table 1. An overview of the groups and their respective sessions.

Group	Session 1	Session 2
1	Normal	Failure
2	Failure	Normal

3.2 Experiment 2

The second experiment focused on 10 participants and utilized similar scenarios as the first experiment. Participants, acting as pilots, operated the Airbus A320, while the experimenter served as a pilot monitor. Real-time data collection included **cognitive**

workload, heart rate, and **facial thermal images** and **temperatures**. A total of 120 instances of takeoffs were recorded, generating over 10 h of time-series data.

In this second experiment, 10 individuals were recruited, ensuring a gender-balanced group with ages ranging between 25 and 35 years old. Among the ten participants, five of them had prior experience and knowledge in piloting and aviation simulation. The remaining five participants did not have any previous exposure to aviation simulation. This division allowed us to have a balanced representation of both experienced and inexperienced individuals, which can help in drawing more comprehensive conclusions from the study. Participants were briefed on the study's objectives and gave informed consent before the experiments. Combining the results of both experiments data including cognitive workload, heart rate, and facial thermal responses provided a thorough understanding of pilots' cognitive and physiological responses during takeoff, offering insights into the challenges they face during critical flight phases.

3.3 Flight Scenarios

To simulate realistic flying circumstances, a variety of flight scenarios were utilized in the airplane simulator during the design and implementation of this experiment [16]. The scenarios varied in time, weather conditions, and whether a failure will occur. We will use scenarios one through three for the regular takeoff sessions and four through six for the failure sessions. Table 2 shows the details of the different scenarios.

Table 2. An overview of the different scenarios [16].

Scenario	Detail		
	Time	Weather	Engine Failure
1	1:45 PM	No Wind, No Clouds	No
2	6:00 AM	Clouds at 2700ft, rain	No
3	9:00 PM	No wind, no clouds	No
4	5:30 AM	No wind, no clouds	Yes, EF at 80 knots
5	6:00 AM	15 knots crosswind	Yes, EF at 140 knots
6	6:00 AM	Low visibility, rain	Yes, EF at 80 knots

3.4 Materials and Measures

We monitored heart rate using the Polar H10 chest strap, ensuring accuracy by normalizing data to accommodate individual variations, facilitating a fair assessment of physiological responses. For measuring cognitive workload, we utilized the EEG headset from BMU enterprise, by OpenBCI, with NCO software extracting real-time data. Infrared cameras, like the ICI-7640 used in our experiment, captured facial temperature variations, converting thermal energy into visual representations, aiding analysis of temperature differences across facial areas.

3.5 Procedure

In a series of experiments, we investigated cognitive workload, heart rate, and facial temperature during simulated A320 takeoff scenarios using the X-plane flight simulator. Ethical approval and informed consent were obtained, ensuring adherence to ethical standards. Participants received a thorough briefing on the A320 takeoff procedure a week before the experiment, covering visual displays and aircraft handling. Specific scenario details were withheld to increase cognitive workload and unpredictability. Participants were designated as pilots, with experimenters monitoring. EEG headsets and heart rate monitors were attached, and equipment calibrated before the experiment. Six scenarios were randomly assigned to participants to ensure authentic responses. In Experiment 2, a thermal camera captured facial thermal images and temperatures at five images per second, totaling 18,000 data points per participant. Each scenario was repeated within a one-hour session to provide diverse cognitive demands. Participants remained unaware of scenarios to maintain authenticity. Continuous data collection aimed to capture nuanced responses and understand the impact of takeoff scenarios on pilot performance.

4 Results and Discussion

In this section, we thoroughly analyze data from Experiment 1 and Experiment 2 to validate our initial hypotheses. Experiment 1's results are examined in two subsections. The first explores the performance of pilots and pilot engineers in groups with different starting scenario difficulties, while the second focuses on licensed pilots and pilot engineers without licenses. We scrutinize variations in Heart Rate (HR) and Workload to uncover patterns within these subgroups. Experiment 2 investigates participants with and without prior flight simulator experience, revealing the impact on HR, Workload, and Face Temperature. Synthesizing the results from both experiments, we conduct a comprehensive comparison across diverse scenarios and two flights of the same scenario.

4.1 Results of Experiment 1

Scenario Difficulty Order Effects. The comparison between the two pilot groups illustrates **how the sequential difficulty levels of scenario experiences** influence both heart rate and workload. Figure 1 illustrates normalized heart rate (HR) variations among two distinct groups of pilots, represented by blue and red box-and-whisker plots. Group 1, depicted in blue, exhibits lower mean, maximum, and range values for HR compared to Group 2, emphasizing a significant difference in physiological responses between the two pilot groups. It shows that the HR range and mean of most participants in Group 1 are lower than the HR range and mean of Group 2, indicating that participants who start the experiment with the easier scenario experienced less stress than those who faced the harder scenario at the beginning.

Figure 2 illustrates normalized heart rate variations between the two pilot groups across the six different scenarios using box and whisker plots. The blue plot denotes the first group, and the red plot signifies the second group. Generally, the first group shows lower mean, maximum, and range of heart rates than the second group in all scenarios. This demonstrates that participants who initiated the experiment with the easier scenarios showed consistently lower mean and range values of HR across all six scenarios compared to those who started with the harder scenarios, indicating that Group 1 (7 pilots) experienced lower stress levels in all six scenarios than Group 2 (6 pilots).

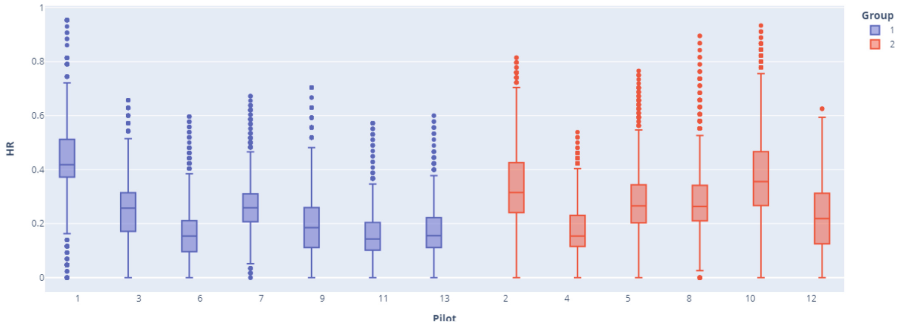


Fig. 1. Normalized HR variations comparison between the two pilot groups

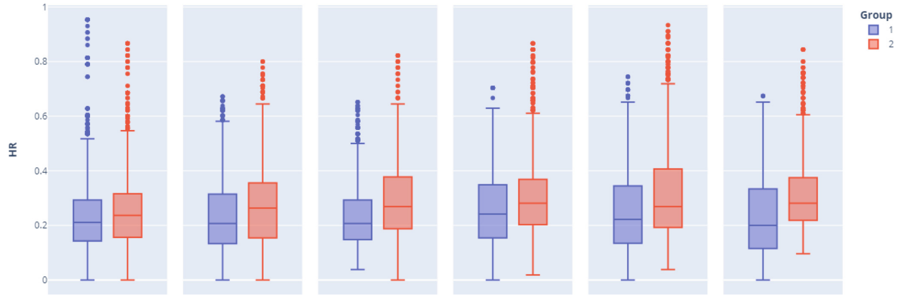


Fig. 2. Normalized HR variations for the two different Pilot groups in different scenarios

Figure 3 depicts **workload** variations for two distinct pilot groups in different scenarios, with the blue box and whisker plot representing the first group and the red plot representing the second group. Across all scenarios, the maximum of the first group is slightly lower, and the mean workload for the first group consistently exhibits lower values compared to the corresponding metrics for the second group of pilots. This demonstrates that participants who begin with easier scenarios and then face difficult scenarios experience a lower overall workload in the entire experiment compared to participants who start with difficult scenarios.

Comparing Experimented vs. Unexperimented Pilots. This subsection contrasts experimented and unexperimented pilots to demonstrate the impact of experience on heart rate and workload. Figure 4 shows average heart rate (HR) variations for experienced (blue) and inexperienced (red) pilots across six scenarios. In the normalized data, experimented pilots exhibit lower mean, maximum, and range of heart rates compared to unexperimented pilots, suggesting lower stress levels. Additionally, Fig. 5 indicates that experimental pilots (red) have significantly lower average workload than pilot engineers without experience(blue), particularly evident in challenging scenarios like engine failure (scenarios 4, 5, and 6). Lower heart rate and workload suggest reduced stress and potentially better performance under pressure.

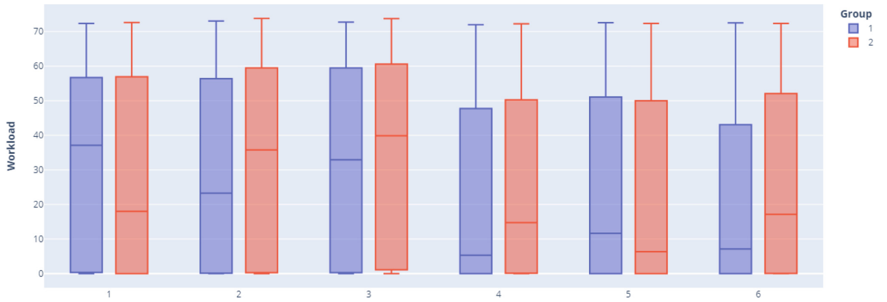


Fig. 3. Workload variations for the two different Pilot groups in different scenarios

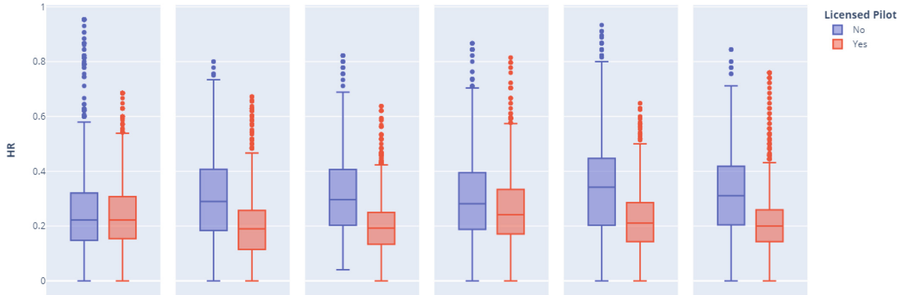


Fig. 4. Normalized HR variations for the pilots with/without the License in different scenarios

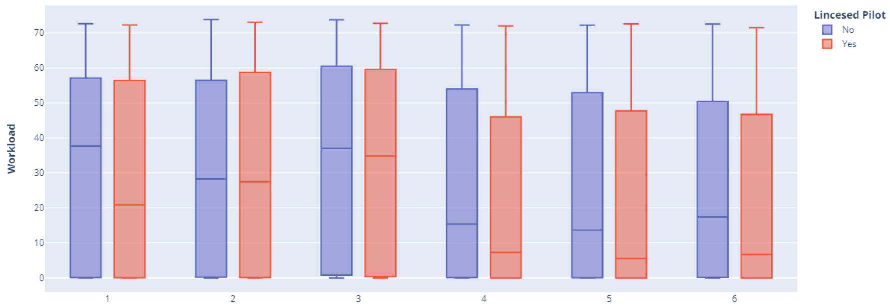


Fig. 5. Workload variations for the pilots with/without the License in different scenarios

4.2 Results of Experiment 2

In Experiment 2, a comparison between the normalized Heart Rate (HR) and Workload variations of 12 flights (two flights for each scenario) is illustrated in Fig. 6 and Fig. 7. The key observations from the figures indicate that, in contrast to the first flight of the scenario, the second flight shows a decrease in both Heart Rate and Workload. These plots facilitate a more meaningful comparison between the two flights, across all six different scenarios. These findings suggest that, with the repetition of the same scenario in the second flight, participants experienced a reduction in both physiological stress (reflected in the Heart Rate) and perceived workload. This could imply that participants adapted or became more accustomed to the scenario during the second iteration, resulting in a more relaxed physiological and psychological response compared to their initial exposure in the first flight.

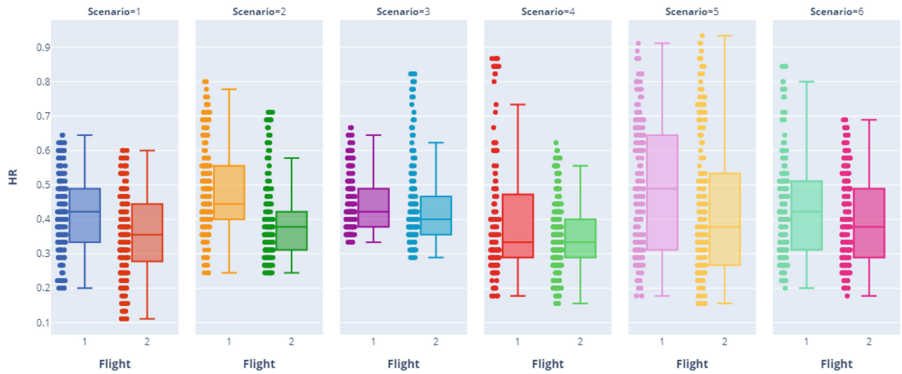


Fig. 6. Normalized HR variations for two Flights of the 10 Participants

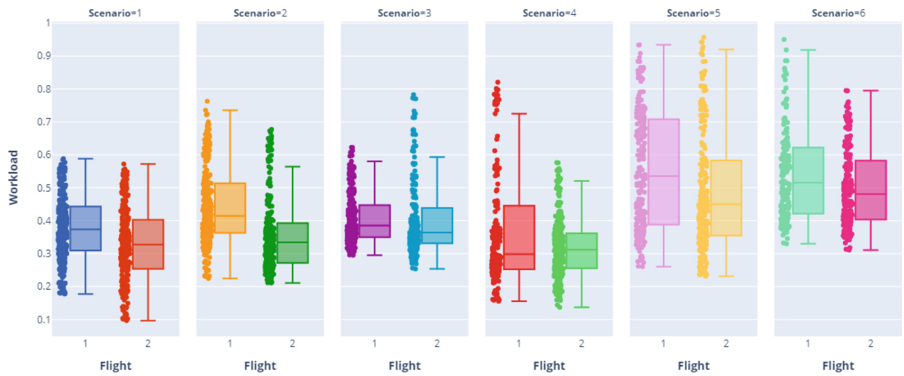


Fig. 7. Normalized Workload variations for two Flights of the 10 Participants

Figure 8 presents a comparative analysis of two participants across 12 flights, encompassing all six distinct scenarios (with 2 flights for each scenario). Participant 1 has prior experience, while Participant 2 lacks such familiarity. The results indicate that Participant 1 consistently exhibits higher values for delta Heart Rate (ΔHR), delta Workload (ΔW), and the absolute value of delta face temperature (ΔT) compared to Participant 2. This suggests that the experienced participant, when confronted with various scenarios, undergoes more pronounced physiological responses and thermal changes. These findings underscore the potential impact of prior experience on individuals' reactions during flight scenarios, as reflected in the measured parameters.

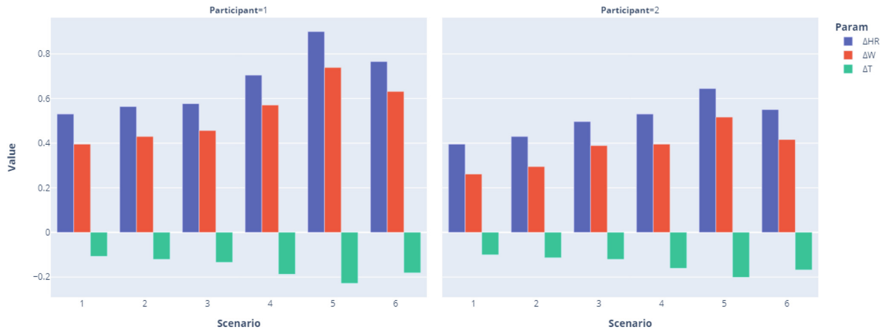


Fig. 8. Comparative Analysis of two Participants with/without experience in Two Flights for each Scenario (Participant 1 is with experience, Participant 2 is without prior experience)

The overall results of the average changes in heart rate, workload, and facial temperature in the first and second flights across all scenarios for both experiments are depicted in Fig. 9. Moving from left to right on the x-axis indicates an increase in the flight experience of participants. Table 3 and Fig. 9 provide a comprehensive comparison of Heart Rate (HR), Workload, and Face Temperature average variations across four distinct participant types in both the first and second flights, conducted under identical scenarios and conditions. The participants are categorized as 1) Without Experience, 2) With Experience, 3) Pilot Engineering (without license), and 4) Real Pilot (with license).

During the initial (first) flight of each scenario, individuals **lacking experience** demonstrate the highest delta workload, heart rate (HR), and facial temperature, whereas **Real Pilots** exhibit the **lowest values**. Participants **with prior experience** and those in **Pilot Engineering** show comparatively lower values, with Pilot Engineering having the lower one. Increased variations signify heightened stress levels and decreased control stability.

In the second flight under the same scenario, the participant Without Experience experiences a slight decrease in delta workload and HR but a slight increase in delta face temperature, suggesting elevated heart rate and workload persist across both flights, accompanied by a lower temperature in the nose area. Participants with Experience show a higher decrease in delta workload and HR but a notable increase in delta nose area temperature. Pilot Engineering participants also experience a significant decrease, though less than those with Experience. Real Pilots demonstrate the lowest decrease in delta workload and HR, indicating a greater level of stability and adaptation to the repeated scenario (Table 4).

Table 3. Comparison of HR, Workload, and Nose Area Temperature Average variations

Participant	In a Flight			In two Flight with Same Scenario		
	ΔHR	ΔW	ΔT	ΔHR	ΔW	ΔT
Without Experience	**High**	**High**	**High**	**-Low**	**-Low**	**+Low**
With Experience	**Low-Mid**	**Low-Mid**	**Low-Mid**	**-High**	**-High**	**+High**
Pilot Engineering	**Low-Mid**	**Low-Mid**	**N/A**	**-High**	**-High**	**N/A**
Real Pilot	**Low**	**Low**	**N/A**	**-Low**	**-Low**	**N/A**

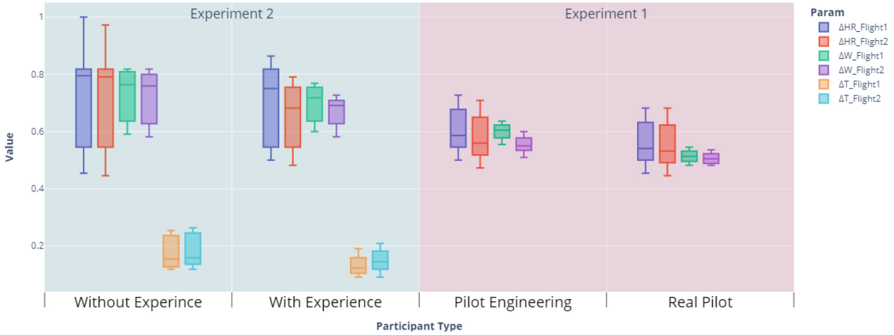


Fig. 9. Comparison of HR, Workload, and Nose Area Temperature Average variations for Participants

Table 4. Workload, HR, and Age of participants.

Participant Type	Workload				HR				Age			
	Min	Mean	Median	Max	Min	Mean	Median	Max	Min	Mean	Median	Max
Group 1	5.39E−16	27.01	19.78	73.05	49	72.56	74	107	28	35.28	32	47
Group 2	4.84E−17	27.40	19.64	73.79	56	74.03	73	113	24	37	36	49
With License	4.84E−17	25.79	15.12	73.05	60	76.70	77	107	28	32	31	43
Without License	3.26E−15	28.78	25.52	73.7	49	69.92	69	113	24	40.83	46	49

Through an analysis of figures in the results section, specifically Fig. 9 and Table 3, discernible results emerge, providing substantial evidence to support our hypothesis. The following assessment parameters and results show the **impact of experience** on cognitive load and physiological measures in aviation pilots.

In the study on cognitive workload in flight simulation, novice participants, lacking experience, faced higher cognitive demands due to the novelty of aviation tasks, while experienced individuals, like real pilots, exhibited lower workload thanks to their familiarity with procedures. Novices often showed elevated heart rates, reflecting stress from unfamiliar tasks, whereas experienced participants demonstrated better stress management. Facial temperature patterns also varied, with novices exhibiting lower temperatures in nose area, indicative of stress, while experienced individuals maintained more stable readings. The relationship between experience and physiological responses was complex, influenced by individual differences and situational factors. With repeated exposure, participants showed a learning curve, leading to reduced workload, stabilized heart rates, and normalized facial temperatures, highlighting the adaptive nature of human responses to simulated environments.

5 Conclusion

Our study reveals how participants with different aviation experience levels respond during simulated flight scenarios. Inexperienced participants showed elevated workload, heart rate, stress, and decreased nasal area temperature in both initial and subsequent flights, indicating sustained challenges. In contrast, licensed pilots exhibited lower workload, heart rate, and stress levels in both flights, highlighting their stability and adaptability. Their ability to perform the second flight scenarios with even less stress suggests enhanced performance. An intriguing revelation from this research is the notable performance of participants with limited familiarity with the flight simulator and flight engineers lacking a pilot's license but possessing a foundational understanding of flight principles. This cohort displayed the most significant changes between the first and second flights in various scenarios, suggesting a learning curve and adaptability. The observed decrease in workload, heart rate, and stress in their second flight performances indicates a practical application of gained experience from the initial flights, highlighting the potential for skill acquisition and improvement over time. These findings not only contribute valuable insights to aviation psychology but also underscore the importance of experience and training in enhancing performance under challenging conditions.

In conclusion, this study delves into the intricate dynamics among workload, heart rate, and facial temperature in diverse participant groups engaged in flight simulation experiments. The obtained results provide a nuanced understanding of both cognitive and physiological aspects of aviation performance, offering potential improvements for training practices and pilot well-being. The multi-faceted relationship identified between participants' experience and workload, heart rate, and facial temperature emphasizes the pivotal role of experience in shaping cognitive and physiological responses during flight simulation. Designing training programs that consider the learning curve and stress adaptation is crucial. Gradually exposing novice participants to increasingly complex tasks can aid in building resilience and diminishing physiological responses over time. Additionally, incorporating stress management techniques and interventions during training positively influences participants' capacity to manage cognitive workload, subsequently impacting heart rate and facial temperature responses. These discoveries not only propel the field of aviation psychology forward but also emphasize the importance of experience and training in optimizing performance under challenging conditions. This contributes to the overall enhancement of aviation capabilities and well-being.

Acknowledgments. We acknowledge the support from NSERC-Alliance, CRIAQ, CAE, Bombardier, and BMU for funding this work.

References

1. Meade, M.L., Nokes, T.J., Morrow, D.G.: Expertise promotes facilitation on a collaborative memory task. Memory (2009). https://doi.org/10.1080/09658210802524240
2. Gorji, H.T., Wilson, N., VanBree, J., Hoffmann, B., Petros, T.V., Tavakolian, K.: Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight. Sci. Rep. (2023). https://doi.org/10.1038/s41598-023-29647-0
3. Lounis, C., Peysakhovich, V., Causse, M.: Visual scanning strategies in the cockpit are modulated by pilots' expertise: a flight simulator study. PLoS ONE (2021). https://doi.org/10.1371/journal.pone.0247061
4. Wolkow, A., Ferguson, S.A., Aisbett, B., Main, L.C.: Effects of work-related sleep restriction on acute physiological and psychological stress responses and their interactions: a review among emergency service personnel. Int. J. Occup. Med. Environ. Health (2015). https://doi.org/10.13075/ijomeh.1896.00227
5. Choi, J.-K.: The improvement of pilot fatigue management. Korean J. Aerosp. Environ. Med. (2023). https://doi.org/10.46246/kjasem.230006
6. Barnier, A.J., Klein, L., Harris, C.B.: Transactive memory in small, intimate groups: more than the sum of their parts. Small Group Res. (2017). https://doi.org/10.1177/1046496417712439
7. Prinzo, O.V., Morrow, D.G.: Improving pilot/air traffic control voice communication in general aviation. Int. J. Aviat. Psychol. (2002). https://doi.org/10.1207/s15327108ijap1204_2
8. Summerfield, D.T., Raslau, D., Johnson, B., Steinkraus, L.: Physiologic challenges to pilots of modern high performance aircraft (2018). https://doi.org/10.5772/intechopen.75982
9. Holt, T.B., Luedtke, J.R., Perry, J., Hight, M., Schindler, C., Ward, P.: General aviation hypoxia and reporting statistics. J. Aviat. Technol. Eng. (2019). https://doi.org/10.7771/2159-6670.1176
10. Hebbar, P.A., Bhattacharya, K., Prabhakar, G., Pashilkar, A.A., Biswas, P.: Correlation between physiological and performance-based metrics to estimate pilots' cognitive workload. Front. Psychol. (2021). https://doi.org/10.3389/fpsyg.2021.555446
11. Mohanavelu, K., et al.: Cognitive workload analysis of fighter aircraft pilots in flight simulator environment. Def. Sci. J. (2020). https://doi.org/10.14429/dsj.70.14539
12. Wang, H., Pan, T., Si, H., Li, Y., Jiang, N.: Research on influencing factor selection of pilot's intention. Int. J. Aerosp. Eng. (2020). https://doi.org/10.1155/2020/4294538
13. Levenson, R.W., Ekman, P., Friesen, W.V.: Voluntary facial action generates emotion-specific autonomic nervous system activity. Psychophysiology (1990). https://doi.org/10.1111/j.1469-8986.1990.tb02330.x
14. Pagano, T.P., et al.: Machine learning models and videos of facial regions for estimating heart rate: a review on patents, datasets, and literature. Electronics (2022). https://doi.org/10.3390/electronics11091473
15. Ghaderi, M., Khalaj, A.B., Ben Abdessalem, H., Frasson, C.: Attention assessment of aircraft pilots using eye tracking. In: Frasson, C., Mylonas, P., Troussas, C. (eds.) ITS 2023. LNCS, vol. 13891, pp. 209–219. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-32883-1_18

16. Antoine, M., Ben Abdessalem, H., Frasson, C.: Cognitive workload assessment of aircraft pilots. J. Behav. Brain Sci. **12**(10), 474–484 (2022)
17. Bonyad, A., Abdessalem, H., Frasson, C.: The relation between mental workload and face temperature in flight simulation. J. Behav. Brain Sci. **14**(02), 64–92 (2024). https://doi.org/10.4236/jbbs.2024.142006

Kahoot! as a Tool to Maintain Students' Attention and Increase Retention Rates: An Experience Report with Computer Science Students

Mariana Madalina Nastase(✉) and Elvira Popescu

Computers and Information Technology Department, University of Craiova, Craiova, Romania
nastase.mariana.e8b@student.ucv.ro, elvira.popescu@edu.ucv.ro

Abstract. In the present study, we explore the current research on Kahoot! and address some concerns related to the effectiveness of the tool. We report on an experiment that we conducted with 39 computer science students in an introductory programming course, which aimed to investigate the use of Kahoot! in two-hour-long practice sessions. Results highlight some interesting aspects revealed from the students' feedback: Kahoot! proved to be an exciting and motivational tool that enabled learners to focus better. The students also perceived the tool as helpful in remembering previously taught concepts but not so much for newly taught concepts. Their overall feedback was that Kahoot! is a very useful add-on to the lesson but cannot replace traditional teaching methods. Some of the limitations of Kahoot! are related to teacher involvement (as they need to choose the right questions and allocated time) and the necessity of using a mobile device in the classroom, which might cause distractions.

Keywords: Kahoot! · Microlearning · Gamification · Computer Science education · Introductory programming · Student motivation

1 Introduction

A widely discussed topic in educational research is related to the attention span of young people, which has faced a visible decrease in recent years. Most research found a correlation between the tendency of young people to get distracted and their mobile phone use [1,2]. As for the reasons the students might use their phones during lessons, the researchers highlighted the so-called Fear of Missing Out, which is characterized as the fear of not being up to date with the latest news or not being actively engaged on social media [3]. However, studies also found that disengagement might show up because of boring lectures, thus shifting the perspective related to the cause [3].

In the context of an educational system where students' attention is dropping every year, teachers, researchers, and tech companies struggle to find new ways of

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 80–87, 2024.
https://doi.org/10.1007/978-3-031-63031-6_7

maintaining students' attention for as long as possible. Combining gamification elements with microlearning activities and collaborative engagement, they strive to find the best tools for fighting against young students' current tendency to get easily distracted during lectures, thus not retaining enough helpful information from what they are presented. One of the tools that have been intensively analysed in recent years is Kahoot!. According to the company, Kahoot! is a globally used learning platform that uses gamification elements such as characters, points, leaderboards and visually pleasing animations to keep participants engaged during lessons [4].

Although the body of research on Kahoot! is quite large, there are still some conflicting findings [1]. The literature review conducted in [1] highlighted some interesting insights and showed that Kahoot! mainly had positive results. However, few studies focused on teachers' perceptions and how class dynamics are influenced. Even though the number of studies mentioning students' anxiety is relatively low, their analysis revealed that students feel less stressed when playing Kahoot! due to the anonymity factor and because it was fun and engaging. Also, the students did not feel the pressure of judgement for their answers [1].

Most studies highlight that Kahoot! is an excellent tool for motivating students to learn since it creates a fun and competitive environment [5–8]. However, other researchers raise some worries about creating a competitive learning space, stating that students' competitiveness might act as a distractor. Since the students might be focused on earning more points and pressured by time constraints, they might tend to guess the answer quicker rather than analyse it thoroughly [9,10].

Other conflicting opinions appear regarding using Kahoot! as a testing tool, and the current research explores the possibility from students' perspectives. Paper [11] investigated the feasibility of using Kahoot! as a testing tool on medical students, and their perception indicated some potential. However, the analysis conducted in [12] showed that the students perceived Kahoot! as an unfair testing tool due to time pressure. They also mentioned that Kahoot! might generate a noisy environment prone to distractions.

From the teachers' perspective, Kahoot! might prove tedious since it requires a strong internet connection, and the time allocation must be carefully chosen [13]. Other comments on Kahoot! are related to not having the question and answers on the student's devices, which might make it hard for some students to see them correctly [12,14].

Over the years, Kahoot! was tested in different disciplines, such as language learning [15,16], social sciences [17], psychology [18], computer science [20] and medicine [11], and in different educational levels, from middle school to higher education. However, there is little to no research on Kahoot! done in an environment other than a formal educational system - so more studies could be performed in corporate and non-formal education [1].

In our review, we also noticed that Kahoot! is primarily used in short sessions. Thus, in the current paper, we attempt to evaluate Kahoot! in multiple longer sessions. We therefore conducted an experiment across two sessions of two hours each; the study was aimed to help us understand the efficiency of Kahoot! during longer sessions with computer science students. The investigation was done from the students' point of view and targeted their engagement and preference for digital tools instead of traditional methods when reviewing previously learned concepts.

The next section describes the context of our experiment and the research questions that guided it. Section 3 summarises the student survey results and discusses how they relate to the current body of research we presented in the introduction. The final section concludes the paper, outlining the limitations of our study and the future research directions.

2 Methodology

The experiment targeted a cohort of first-year computer science students taking a course in introductory programming using C++. The sample size was 39 students: 17 with some programming skills and 22 beginners.

We attempted to evaluate whether or not the effectiveness of Kahoot! is also apparent when used in longer sessions. Thus, the experiment took place in the last two laboratory sessions of the semester, having allocated two hours per meeting. Each Kahoot! session had 20 questions that targeted the students' basic understanding of concepts practised during the semester, such as recursion, memory allocation, function parameters, and strings; in addition, a few new concepts were also introduced during the Kahoot! sessions. An example can be seen in Fig. 1.

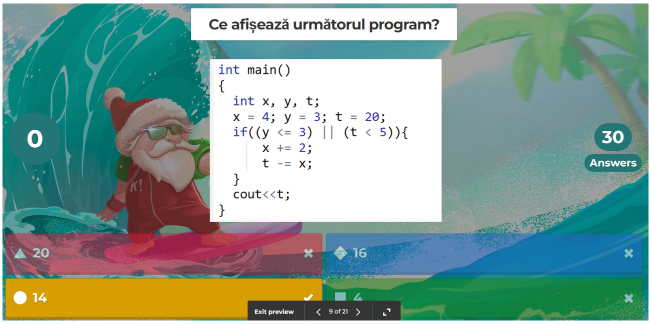


Fig. 1. Example of Kahoot! question

Our research focused on evaluating the effectiveness of Kahoot! in maintaining the attention of students during longer sessions and in increasing the retention of information. Thus, the following questions guided the experiment:

1. Do the students perceive Kahoot! as an effective tool for helping them retain newly taught concepts in longer sessions?

2. Do the students perceive Kahoot! as an effective tool for helping them remember previously taught concepts in longer sessions?
3. Do the students perceive Kahoot! as an effective tool for keeping students focused and motivated in longer sessions?
4. Do the students perceive digital tools like Kahoot! as replacing traditional methods, such as whiteboards and notebooks, for reviewing previously taught concepts?

The students who attended the Kahoot! sessions could give feedback on their experience of using Kahoot! in class by completing a dedicated survey. The questionnaire contained five demographic and background questions and five questions that evaluated aspects related to students' perception of progress, information retention, the pleasure of using Kahoot! and maintaining attention (assessed on a 5-point Likert scale). Each multiple-choice question was accompanied by an open-ended question where the students could motivate the answer or give additional comments. Furthermore, the students were asked to what degree they considered the mobile phone a distractor during the lesson and they were also offered the chance to give any additional feedback or suggestions. Twelve of the 39 students chose to fill in the questionnaire and provided insights on how they perceived the learning experience.

3 Results

The demographic data shows that the students had ages ranging between 19 and 22 years old (75%), 22 and 25 years old (8.3%), and 25 to 30 years old (16.7%). The males and females were split evenly and related to their high school background: 66.7% had a high school degree in mathematics and informatics, 25% in humanities, and 8.3% in the food industry.

The quantitative analysis was done based on the 5-point Likert scale questions, where one stands for total disagreement and five for total agreement.

As shown in Table 1, 91.7% of the students perceived the Kahoot! sessions as an excellent method of remembering and reinforcing previously learned concepts. Opinions were slightly less positive regarding the impact of Kahoot! on understanding newly taught concepts, with only 58.3% of students strongly agreeing that the session helped. Similar results were also obtained by Licorish et al. [19], where only 9 out of 14 students felt that Kahoot! is useful for learning new concepts, while 12 out of 14 agreed it is a great tool for revising previously learned concepts. Overall, these results provide moderate support for the first research question and strong support for the second one.

The qualitative analysis for the open-ended questions used an inductive thematic analysis approach, which implies multiple steps. Firstly, this process requires familiarisation with the data and highlighting the main keywords or phrases that describe the content. This approach was the most suitable, considering the small amount of data. After highlighting the codes in our data, we identified recurring themes.

Table 1. Student survey results

Item	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Kahoot! sessions helped me remember previously learned concepts	0%	8.3%	0%	0%	91.7%
Kahoot! sessions helped me understand newly learned concepts	0%	8.3%	8.3%	25%	58.3%
I enjoyed the Kahoot! session	0%	0%	0%	8.3%	91.7%
I consider digital tools, such as Kahoot!, more helpful than traditional methods in remembering information	0%	0%	16.7%	33.3%	50%
I consider digital tools, such as Kahoot!, more helpful than traditional methods in maintaining focus during the session	0%	8.3%	25%	16.7%	50%

Most students agreed that Kahoot! was fun to play as it created a competitive environment where they were motivated to answer the questions to outperform their peers. They also appreciated that it was a novel experience, which they would like to see more often. All these findings provide strong support for the third research question.

Moreover, students appreciated that Kahoot! encouraged active participation, as opposed to the traditional way of teaching, which is still primarily passive. The students also liked that the questions were short and practical, focusing on the central concepts of the course and guiding them through all the learned topics. The students also mentioned that they finished the session with a sense of accomplishment regardless of whether or not they answered most questions correctly, which is in line with the findings in [19].

However, when asked to compare the Kahoot! sessions with the traditional ones in reviewing previously learned concepts, they stated that they still prefer to have their notes written and appreciated that after each question, the teacher explained the answer. Other studies [9,19–21] also highlighted the preference for an additional explanation from the teachers. Most students agreed that these two methods should be used complementary as both play a role in improving the learning experience. Coupled with the quantitative data from Table 1, we can conclude that digital tools are perceived mostly as an add-on to traditional teaching methods and not a complete replacement, which provides an answer for the fourth research question.

Finally, since Kahoot! requires a device, there is the worry of encouraging students to use smartphones during lectures [3]. Students who participated in our experiment also offered some insights regarding the distraction caused by mobile phones. They admitted that they might tend to get distracted if they receive messages and notifications. Still, they do not see the mobile phone as a distractor in itself, especially when used for educational purposes. A study conducted in this regard is reported in [23], which analyses the students' tendency to multitask during lectures when encouraged to use digital devices, such as laptops and smartphones. The study confirmed that using multiple apps and attempting multitasking during lectures negatively impacted academic performance. However, they dismissed banning devices in academic environments and proposed a shift in the teaching practices. To address the students' engagement challenges, they proposed solutions such as shortening lecture duration and adding additional breaks, but more importantly, including various resources and activities during class to avoid social media use, online chatting and mind wandering. These observations are also confirmed by previous research on multitasking [24,25] and distractions [26].

4 Conclusions and Perspectives

The limitations of our study consist in the small sample size and the fact that it only involved computer science students. Nevertheless, the results provide interesting insights into the use of Kahoot! in longer study sessions, a context which has been less explored in the literature. The majority of the students expressed their desire to use Kahoot! also in the future, highlighting their appreciation for instant feedback, active participation, playful and competitive environment.

A few students also mentioned in the additional comments question that their study domain might influence their preference for digital tools. In the same vein, the study conducted in [22] also concluded that computer science students experienced a higher level of satisfaction with using Kahoot! than social science students. Thus, the experiment could be extended to students from different study domains. Furthermore, research in more diverse environments, such as corporate training and self-paced non-formal training, could bring beneficial insights, as suggested in [1].

Students also mentioned the time allocated for each question, which was sometimes deemed too short. This type of limitation was also reported in other studies [12,13]. The time allocation and choosing suitable activities that serve the lesson purpose are issues that are difficult to manage, partially because each student has an individual set of needs, as discussed also in [27].

Our current interest is to explore different tools that promise to reduce mind wandering and maintain students' attention during lectures. As future work, on the one hand, we will continue to explore such tools and analyse their strengths and weaknesses; on the other hand, we will work on developing a tool focused on microlearning activities and continuous engagement from the learners while also providing an environment where teachers can easily find and create their desired activities to enrich their lessons.

References

1. Wang, A.I., Tahir, R.: The effect of using Kahoot! for learning - a literature review. Comput. Educ. **149**, 103818 (2020). https://doi.org/10.1016/j.compedu.2020.103818
2. Mendoza, J., Pody, B., Lee, S., Kim, M., McDonough, I.: The effect of cellphones on attention and learning: The influences of time, distraction, and nomophobia. Comput. Hum. Behav. **86**, 52–60 (2018). https://doi.org/10.1016/j.chb.2018.04.027
3. Al-Furaih, S.A.A., Al-Awidi, H.M.: Fear of missing out (FoMO) among undergraduate students in relation to attention distraction and learning disengagement in lectures. Educ. Inf. Technol. **26**, 2355–2373 (2021). https://doi.org/10.1007/s10639-020-10361-7
4. About Kahoot! (2024). https://Kahoot!.com/company/
5. Wirani, Y., Nabarian, T., Romadhon, M.S.: Evaluation of continued use on Kahoot! as a gamification-based learning platform from the perspective of Indonesia students. Procedia Comput. Sci. **197**, 545–556 (2022). https://doi.org/10.1016/j.compedu.2020.103818
6. Anugerah, Y.B., Efendi, A., Maryono, D.: The effectiveness of using Kahoot! in basic graphic design subject to students' attention levels and learning achievement. J. Inf. Vocat. Educ. **2**(2), 88–92 (2019). https://doi.org/10.20961/joive.v2i2.35766
7. Martin-Somer, M., Moreira, J., Casado, C.: Use of Kahoot! to keep students' motivation during online classes in the lockdown period caused by COVID-19. Educ. Chem. Eng. **36**, 154–159 (2021). https://doi.org/10.1016/j.ece.2021.05.005
8. Martínez-Jiménez, R., Pedrosa-Ortega, C., Licerán-Gutiérrez, A., Ruiz-Jiménez,M.C., García-Martí, E.: Kahoot! as a tool to improve student academic performance in business management subjects. Sustainability **13**, 2969 (2021). https://doi.org/10.3390/su13052969
9. Owen, H.E., Licorish, S.A.: Game-based student response system: the effectiveness of Kahoot! on junior and senior information science students' learning. J. Inf. Technol. Educ. Res. **19**, 511–553 (2020). https://doi.org/10.28945/4608
10. Toma, F., Diaconu, D.C., Popescu, C.M.: The use of the Kahoot! Learning platform as a type of formative assessment in the context of pre-university education during the COVID-19 pandemic period. Educ. Sci. **11**, 649 (2021). https://doi.org/10.3390/educsci11100649
11. Neureiter, D., Klieser, E., Neumayer, B., Winkelmann, P., Urbas, R., Kiesslich, T.: Feasibility of Kahoot! as a real-time assessmenttool in (Histo-)pathology classroom teaching. Adv. Med. Educ. Pract. **11**, 695–705 (2020). https://doi.org/10.2147/AMEP.S264821
12. Chiang, H.H.: Kahoot! in an EFL reading class. J. Lang. Teach. Res. **11**(1), 33–44 (2020). https://doi.org/10.17507/jltr.1101.05
13. Adnyani, K.E.K., Adnyana, I.W., Murniasih, N.N., Suwastini, N.K.A.: Implementing Kahoot! for Japanese language learning in Indonesian high school. J. Educ. Technol. **6**(2), 217–225 (2022). https://doi.org/10.23887/jet.v6i2.44674
14. Suharsono, A.: The use of quizizz and Kahoot! in the training for millennial generation. Int. J. Indonesian Educ. Teach. **4**(2), 332–342 (2020). https://doi.org/10.24071/ijiet.v4i2.2399
15. Marsa, C.S., Kuspiyah, H.R., Agustina, E.: The effect of Kahoot! game to teaching reading comprehension achievement. J. English Teach. **7**(2), 133–149 (2021). https://doi.org/10.33541/jet.v7i2.2738

16. Yürük, N.: Using Kahoot! as a skill improvement technique in pronunciation. J. Lang. Linguist. Stud. **16**(1), 137–153 (2020). https://doi.org/10.17263/jlls.71266
17. Tetep, Arista, Y.: Students' perception towards Kahoot! learning media and its influence on students' motivation in learning social studies and civic education amid pandemic in SMKN 9 garut. Innov. Soc. Stud. J. **4**(1), 99–108 (2022). https://doi.org/10.20527/iis
18. Warsihna, J., Ramdani, Z., Prakoso, B.H.: Using Kahoot! to improve students' achevement and critical thinking in undergraduate of psychology students. In: IADIS International Conference Cognition and Exploratory Learning in Digital Age (2019). https://doi.org/10.33965/celda2019_201911L018
19. Licorish, S.A., Owen, H.E., Daniel, B., George, J.L.: Students' perception of Kahoot!'s influence on teaching and learning. Res. Pract. Technol. Enhanced Learn. **13**(9), 1–23 (2018). https://doi.org/10.1186/s41039-018-0078-8
20. Fuster-Guilló A., Pertegal-Felices M.L., Jimeno-Morenilla A., Azorín-López J., Rico-Soliveres M.L., Restrepo-Calle F.: Evaluating impact on motivation and academic performance of a game-based learning experience using Kahoot!. Front. Psychol. **10**, 2843 (2019). https://doi.org/10.3389/fpsyg.2019.02843
21. Korkmaz, S., Öz, H.: Using Kahoot! to improve reading comprehension of English as a foreign language learners. Int. Online J. Educ. Teach. **8**(2), 1138–1150 (2021). https://www.iojet.org/index.php/IOJET/article/view/1002
22. Pertegal-Felices M.L., Jimeno-Morenilla A., Sánchez-Romero J.L., Mora-Mora H.: Comparison of the effects of the Kahoot! tool on teacher training and computer engineering students for sustainable education. Sustainability **12**(11), 4778 (2020). https://doi.org/10.3390/su12114778
23. Limniou, M.: The effect of digital device usage on student academic performance: a case study. Educ. Sci. **11**, 121 (2021). https://doi.org/10.3390/educsci11030121
24. Waite, B., Lindberg, R., Ernst, B., Bowman, L., Levine, L.: Off-task multitasking, note-taking and lower- and higher-order classroom learning. Comput. Educ. 120, 98–111 (2018). https://doi.org/10.1016/j.compedu.2018.01.007
25. Bowman, L., Waite, B.: Multitasking and Attention. The Wiley Handbook of Psychology, Technology, and Society (2015). https://doi.org/10.1002/9781118771952.ch22
26. Lee, S., et al.: Cellphone addiction explains how cellphones impair learning for lecture materials. Appl. Cogn. Psychol. **36**(3), 1–13 (2020). https://doi.org/10.1002/acp.3745
27. Nastase, M.M., Popescu, E.: Towards Integrating learnersourcing, microlearning and gamification in Moodle. In: Augmented Intelligence and Intelligent Tutoring Systems. ITS 2023. Lecture Notes in Computer Science, vol. 13891, pp. 352–363 (2023). https://doi.org/10.1007/978-3-031-32883-1_32

Adoption of Digital Games as Pedagogical Aids for Teachers and Pupils in Primary Schools to Overcome Learning Problems in Arithmetic

Karima Boussaha[1](✉), Abdelhafid Rahab[2], Khelifa Khelifi Touhami[2], and Charef Eddine Abdelaziz Tayeb Boudraa[2]

[1] Department of Computer Science, LIAOA Laboratory, University of Oum El Bouaghi, Oum El Bouaghi, Algeria
karima.boussaha@univ-oeb.dz

[2] Department of Computer Science, University of Oum El Bouaghi, Oum El Bouaghi, Algeria

Abstract. Arithmetic is a fundamental component of mathematics education that forms the basis for many other disciplines, including science, engineering, and finance. However, for many schoolchildren, learning arithmetic can be a difficult and often tedious task. Traditional methods of teaching arithmetic, such as rote memorization and repetitive practice, may not always be effective in engaging pupils and promoting a deep understanding of the material.

To address these challenges, educators have explored alternative methods of teaching arithmetic that can engage pupils and promote a deeper understanding of the subject. A promising approach is the use of digital games.

The objective of this research paper is to create a digital game in the form of a "Game & learn" web application that can be used to teach arithmetic to primary school children, the application is accessible via a web browser and includes a user-friendly interface suitable for children aged 6 to 11. Our game use interactive challenges and feedback to help children learn at their own pace and track their progress.

Keywords: Education · arithmetic operations · game-based learning · primary school pupils · pupil performance · digital games

1 Introduction

Game-based learning is an innovative and effective approach that uses games to teach and engage primary school pupils in a variety of subjects, including mathematics. Mathematics is one of the most important subjects taught in schools, but it can also be the most difficult subject to learn for pupils [1], with using games to teach mathematics and arithmetic, pupils can deepen their understanding of mathematical concepts and arithmetic. Furthermore, students showed greater motivation and greater interest in mathematics when engaged in gaming activities. The benefits of game-based learning in teaching arithmetic are not limited to academic success and motivation. Game-based learning can also promote important skills, such as critical thinking, problem-solving, and collaboration. Games that force children to work together and solve problems can help develop those skills, which are essential for success in school and beyond [2–5].

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 88–95, 2024.
https://doi.org/10.1007/978-3-031-63031-6_8

Today, numerous studies prove the effectiveness of game-based learning in improving school performance and highlight many benefits that help solve traditional learning problems at different ages. Therefore, our research question is as follows:

- How can we help pupils overcome learning difficulties in arithmetic?
- What is the pedagogical method that can be used to overcome the difficulties of pupils in arithmetic?

Our contribution is to suggest studying via a selection of computer games that we created to simultaneously learn arithmetic in an approachable and enjoyable manner.

The rest of this paper is organized as follows: Sect. 2 presents the learning problems of primary school children. In Sect. 3, we discuss the subjects in which pupils face the most difficulties in learning. Section 4 is reserved for the definition of both concepts of game-based learning and gamification. In Sect. 5 we describe the general architecture of the proposed game -***game & Learn***- and its main modules. Finally, the general conclusion and future works are presented in Sect. 6.

2 Learning Problems for Primary School Children

The most common and important challenges facing educators are:

- Cognitive development
- Individual differences (Learning Styles)
- Motivation and commitment

2.1 Cognitive Development

Cognitive development is the gradual growth of cognitive abilities from birth to adulthood, involving acquisition, retention, and use of knowledge, perception, memory, language, problem-solving, reasoning, and decision-making. These skills are crucial for learning and retaining new information. Children may face cognitive developmental problems [6, 7, 8] such as:

- Attention Deficit Disorder with Hyperactivity (ADHD):
- and Dyslexia:
- Disorders of the Autism Spectrum (ASD):

2.2 Individual Differences (Learning Styles)

Individual differences refer to variations in abilities, talents, and characteristics between individuals [9]. In the field of education, individual differences may refer to the different ways pupils learn and process information, their cognitive abilities, and their motivation level. Understanding individual differences is crucial in education as it enables teachers to develop teaching strategies and techniques that meet each pupil's unique needs [9].

Individual differences pose a major challenge for teachers when it comes to teaching primary school children. Each pupil comes into the classroom with a unique set of experiences, abilities, and backgrounds that can impact their learning. These differences can manifest in various ways, including in:

- The learning styles,
- levels of competence,
- Attention and Behavior Problems
- Motivation and commitment.

2.3 Motivation and Commitment

The internal urge to accomplish a task, including learning new things and overcoming obstacles, is known as motivation. Motivation in the learning environment refers to the effort, focus, and fervor put forth in reaching a specific goal [2–5]. Contrarily, engagement entails investing in and actively participating in the learning process, which includes behavioral and emotional components. For a child to learn and succeed in school, they need to be motivated and dedicated. Promoting motivation and commitment is a common task for primary school instructors, who must contend with obstacles like lack of interest, difficulty, unfavorable views, and lack of autonomy [3, 4].

3 Primary School Children Are Faced with These Problems

Primary school students may have learning difficulties in a variety of topics. Reading, writing, and mathematics are a few popular disciplines.

These three topics are often regarded as foundational knowledge for elementary school pupils since they provide the foundation for the majority of other courses that pupils will study. Even while other disciplines like science, social studies, and foreign languages are as important, success in many other academic fields depends on pupils being proficient in arithmetic, reading, and writing [10, 11].

The use of games or game components, which may enhance learning outcomes and involve pupils in the learning process, is one approach that educators investigate to solve these difficulties.

Game-based learning and gamification are two techniques that have become more popular recently. These methods make use of the immersive and interactive learning opportunities offered by games and technology to increase pupil motivation and engagement while promoting cognitive growth and customized learning. The learning environment as a consequence is more effective and fruitful [11].

4 Game-Based Learning vs Gamification

Game-based learning is a methodology that involves using games as a tool to teach specific content or skills. In this approach, the game is the main educational tool and pupils learn by playing the game. Games can be designed specifically for educational purposes (serious games) or they can be adapted from entertainment games available in commerce [11]. Whereas, Gamification is an approach that uses game elements (such as points, challenges, comments, and rewards) in non-game contexts. It is effective in increasing engagement and motivation [11].

By suggesting a selection of digital games, we want to shed light on game-based learning in the area of mathematics for school-aged children in this article. The activities are specially made to benefit both pupils and teachers. They serve as both an educational tool to assist the pupil in understanding arithmetic operations and a pedagogical tool to help the teacher explain his/her lesson.

5 The General Architecture of the Proposed Digital Game: Game and Learn

The general architecture of our game contains four main modules. Three actors can use the proposed game

- The learning module: used by pupils in their learning process
- The assessment module: used by pupils in their assessment process
- The classification module: used the teacher in his/her task of his/her pupil's classification
- The administration module: used by the director of the school in his/her management process.

6 Methodology

6.1 Case Study

In Algerian schools, primary education is free and compulsory for all children over the age of six.

Education in Algeria passes through four stages: the primary stage, the intermediate stage, the secondary stage, and the university stage. In this research, we are interested in the primary stage, which is considered compulsory for every student up to six years of age, in which the student graduates through five years from the first year, where he studies the process of addition in mathematics and then moves on. Until the second year,

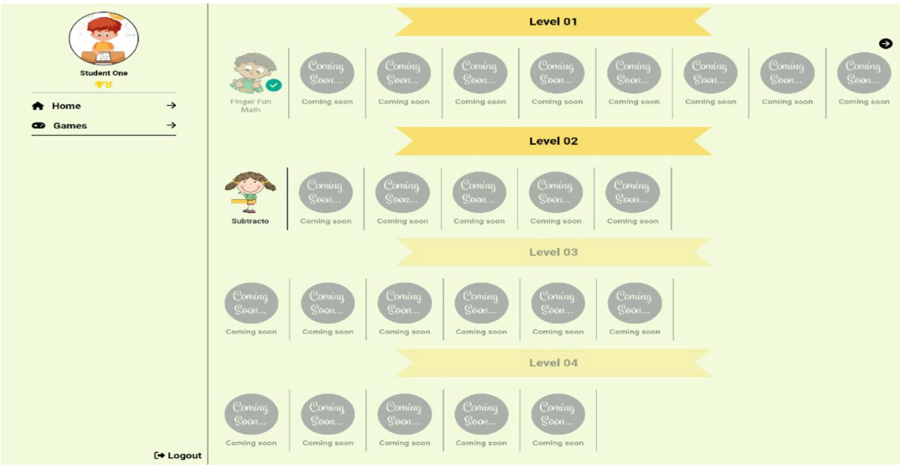


Fig. 1. The "Games" section in the pupil's space after moving to the second year

when he studies subtraction. In the third year, he studies the multiplication process, and in the fourth year, he studies the division process and other more difficult arithmetic operations. In the fifth year, the student then moves to the intermediate stage.

In this section, we present our game. We chose the children who studied at the second-year level. In Algerian schools. In the second year, the pupil must learn the subtraction operation after inevitably learning the addition operations in the first year. So the main page of our Game & learn is illustrated in the following Fig. 1.

The pupil must play the game "substracto" in order to learn the subtraction operation. Figure 2 demonstrates the home page of the "substracto" game.

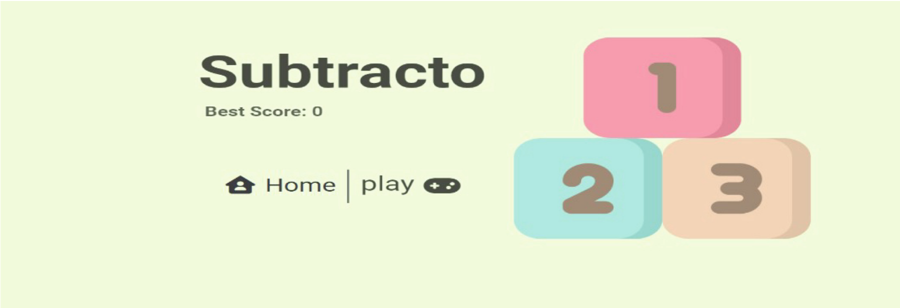


Fig. 2. Substracto game home page

To start playing, the pupil must press the "play" heading, the game starts with his background music. In the beginning, the pupil has 10 s will appear in the "Timer" to choose the right answer from the three possible answers. As it is shown in Fig. 3.

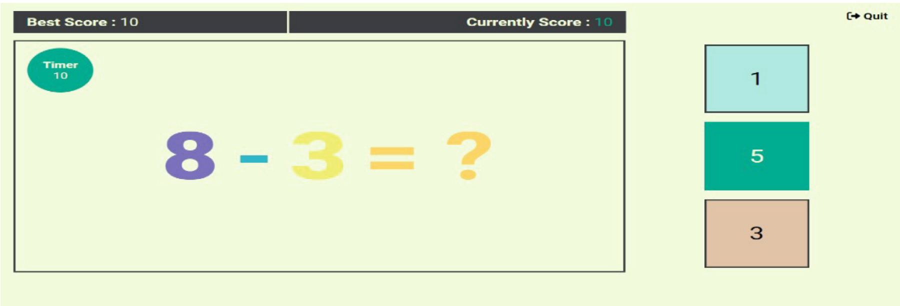


Fig. 3. The game is started

6.2 Different Scenarios Offered by the "Substracto" Game

The pupil starts to play the game. Different scenarios are given to him/her:

The Case When the Pupil Gives the Correct Answer
If the pupil chooses the right answer:

- It will become green.
- The sound of the correct answer is triggered.
- 10 points are added to your current “Currently Score” score.
- If the current score exceeds the best “Best Score” score, the best score will take the actual score value at each change.
- 5 s added to “Timer” if the current score is less than 500, Otherwise only 2 s are added.
- Go to the next operation.

If he/she continues to play and chooses a wrong answer

The Case When the Pupil Gives the Wrong Answer

- It will become red.
- The sound of the wrong answer is triggered.
- The correct answer is marked with green.
- If the current score is less than 500, 5 points will be deducted from the present score. Otherwise, 10 points will be deducted.
- If the current score is less than 500, 5 s will be deducted from the “Timer”. Otherwise, 10 s will be deducted.
- Go to the next operation. Figure 4 shows the details

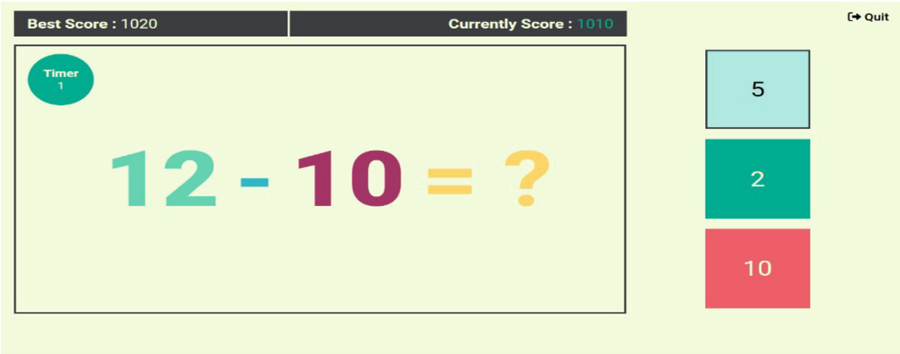


Fig. 4. The game when the pupil gives the wrong answer

The pupil continues to play

1. if the time passed before the current score reaches the specified value to finish the game (1500):
   - The round ends and the score is displayed.
   - The sound of time that has passed is triggered.

- Then it is transferred to the main page of the game.

2. without time ending, until he/she reaches the specified score to finish the game (1500):

- The sounds of applause are triggered.
- The falls of colored confetti.
- The trophy will be given.

7 Conclusion

Game-based learning is an innovative and effective teaching approach that uses games to deepen children's understanding of concepts, improve motivation, and develop key skills.

Recent studies have proven that children learn more by playing, that's why we have presented in this research paper, an educational game whose main objective is to teach children arithmetic operations in mathematics, we have chosen the field of mathematics because it is considered one of the most important subjects in the educational path for children and they encounter many difficulties to learn this on the one hand, and the other hand, as teachers are faced in the primary school a set of problems to teach children this subject which introduces abstract concepts.

Therefore, we have developed the web application "Game & learn", an innovative solution for primary school children, aimed at facilitating the different arithmetic operations by solving the learning problems that its children face in traditional learning methods, adopting a fun and effective approach based on games. As perspectives, we plan to:

- Add other games.
- Expand the types of games used such as strategy games and simulation games.
- Added tools to help the teacher analyze student data, allowing them to act with precision.

References

1. Alt, D.: Assessing the benefits of gamification in mathematics for student gameful experience and gaming motivation. Comput. Educ. **200**, 104806 (2023)
2. Videnovik, M., Bogdanova, A.M., Trajkovik, V.: Game-based learning approach in computer science in primary education: a systematic review. Entertain. Comput., 100616 (2023)
3. Zheng, Y., et al.: Effects of digital game-based learning on students' digital etiquette literacy, learning motivations, and engagement. Heliyon **10**(1) (2024)
4. Acquah, E.O., Katz, H.T.: Digital game-based L2 learning outcomes for primary through high-school students: a systematic literature review. Comput. Educ. **143**, 103667 (2020)
5. Brezovszky, B., et al.: Effects of a mathematics game-based learning environment on primary school students' adaptive number knowledge. Comput. Educ. **128**, 63–74 (2019)
6. Cherry, K.: Piaget's 4 stages of cognitive development explained (2018)
7. Latif, S.: What Is Cognitive Development? 3 Psychology Theories. PositivePsychology.com (2023). https://positivepsychology.com/cognitive-development/
8. Cht, A.B.P.D.: What is cognitive development & why is it important? Verywell health (2022)

9. Kubat, U.: Identifying the individual differences among students during learning and teaching process by science teachers. Int. J. Res. Educ. Sci. **4**(1), 30–38 (2018)
10. Liu, Y., Ng, J.T.D., Hu, X., Ma, Z., Lai, X.: Adopt or abandon: facilitators and barriers of in-service teachers' integration of game learning analytics in K–12 classrooms? Comput. Educ. **209**, 104951 (2024)
11. Tapingkae, P., Panjaburee, P., Hwang, G.J., Srisawasdi, N.: Effects of a formative assessment-based contextual gaming approach on students' digital citizenship behaviours, learning motivations, and perceptions. Comput. Educ. **159**, 103998 (2020)

Educational Games for Computational Thinking: Evaluation of the Scaffolded aMazeD Game

Christina Tikva[1](✉), Maria Mousiou[2], and Efthimios Tambouris[1]

[1] University of Macedonia, Thessaloniki, Greece
ch.tikva@uom.edu.gr
[2] Hellenic Open University, Patra, Greece

Abstract. Computational Thinking (CT) has recently gained increased interest from educational and academic world. The intense interest in CT has led to the utilization and design of a variety of learning tools, with particular interest given to digital games. Several studies are investigating their effectiveness in learning CT, however more research is needed on the specific features of these tools, such as scaffolding features. This study evaluates a scaffolding educational game and investigates how middle school students perceive the effectiveness of its scaffolding features. For this purpose, 28 middle school students participated in an educational intervention. The study adopts a survey research approach. The results regarding ease of use, usefulness, attitude, accessibility and overall experience of the scaffolded game are promising. Specifically, students found scaffolding features of the game, useful for solving the game and effective in learning CT.

Keywords: Educational games · Computational Thinking · Scaffolding

1 Introduction

Computational Thinking (CT) is a process that "involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science" [1]. According to Brennan and Resnick framework [2], CT includes the following three dimensions: **CT concepts**: Sequences, Loops, Parallelism, Events, Conditionals, Operators, Data; **CT practices:** Being incremental and iterative, Testing and debugging, Reusing and remixing, Abstraction and modularity and **CT perspectives:** Expressing, Connecting, Questioning.

The intense interest in CT has led to the utilization and design of a variety of tools. Furthermore, several research efforts (e.g., [3–10]) which review CT studies, include in their scope the analysis of these tools. This contributes to the argument that CT tools play an important role in the process of learning and teaching CT.

Particular interest has been paid on digital games for learning CT [11]. Several studies [12–16] explore the development of CT in the context of digital games and game-based learning environments, suggesting that games could be utilized for teaching and learning

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 96–103, 2024.
https://doi.org/10.1007/978-3-031-63031-6_9

CT. Other studies [11], explore some important features of digital games for learning CT, such as the provision of scaffolding and feedback to students.

This study aims to evaluate a CT game and to investigate how students perceive the effectiveness of its scaffolding features.

2 Literature Review on CT Educational Games

In this section we present educational games widely proposed for learning CT, providing a categorization in terms of CT concepts covered (Table 1) and feedback and scaffolding features (Table 2).

In terms of CT Concepts, sequences, conditionals and loops are the most covered concepts. This could be explained as these are considered the fundamental programming and CT concepts that students could deal with. Variables, expressions and events are also covered, but to a lesser extent.

All selected games incorporate some level of feedback or scaffolding features. Integrated help and integrated tutorials are the features with the greatest presence with 9 games incorporating them. This is followed by highlighting the command being executed (6 games) and run the program in different speeds (6 games), giving the time needed to understand what the problem was. Other features included are the provision of feedback after level failure (5 games), the provision of feedback before each level (5 games), translation of the solution into a text-programming language (3 games) and the ability to restart a level (3 games).

Table 1. CT educational games' general characteristics and CT concepts

Game	CT Concepts
Alice	Sequences, Conditionals, Loops, Variables, Expressions, Events
Blockly Games	Sequences, Conditionals, Loops, Variables, Expressions
CodeCombat	Sequences, Conditionals, Loops, Variables, Expressions
CodeMonkey	Sequences, Loops
Kodable	Sequences, Conditionals, Loops
Kodu	Sequences, Variables, Expressions
Lightbot 2.0	Sequences
Minecraft	Sequences, Conditionals, Loops, Variables
Program Your Robot	Sequences, Conditionals, Loops
Run Marco	Sequences, Conditionals, Loops
Scratch	Sequences, Conditionals, Loops, Variables, Expressions

Table 2. CT educational games' feedback and scaffolding features

Feature	
Integrated Tutorial	Alice, Blockly Games, CodeCombat, CodeMonkey, Kodable, Lightbot 2.0, Minecraft, Run Marco
Integrated Help	Alice, Blockly Games, CodeCombat, CodeMonkey, Kodable, Kodu, Minecraft, Run Marco, Scratch
Feedback before each level	Blockly Games, Kodable, Program Your Robot, Run Marco
Feedback after level failed	CodeCombat, CodeMonkey, Kodable, Program Your Robot, Run Marco
Solution translated in text programming language	Blockly Games, CodeCombat, CodeMonkey
Highlighting	Blockly Games, CodeCombat, CodeMonkey, Kodable, Lightbot 2.0, Run Marco
Run the program at different speeds	Blockly Games, CodeCombat, CodeMonkey, Kodable, Lightbot 2.0, Run Marco
Restart Level	Blockly Games, CodeCombat, CodeMonkey

3 aMazeD CT Educational Game

The aMazeD (see Fig. 1) CT game [17] consists of closed levels and covers the following CT concepts: Sequences, Conditionals and Loops. The game incorporates some of the scaffolding features presented in the previous section (Table 3).

The player uses programming blocks to give appropriate instructions to complete the levels. He/She is not allowed to move to any level of his/her choice. The player starts at level 1 and by submitting his/her answer moves to each subsequent level.

The game environment consists of the following parts: the navigation bar, the instruction bar, the main game frame, the results box, the Blockly toolbox and the workspace. The Blockly toolbox contains the available blocks for each level. Below the main game frame are the three buttons "Play", "Reset" and "Submit". By clicking the Play button, the player can see the visual execution of the code entered in the workspace. When the code is executed, the executed blocks are highlighted. No level output is displayed after execution. The Play button allows students to see the execution of their designed solutions and debug their code. By clicking the Submit button, the code is executed and a message is displayed with the level output. The submitted instruction is translated into JavaScript and displayed on the screen, while the game moves to the next level. The player goes to the next level, regardless of whether the current level is successfully completed. The result box displays a Pass or Fail message, in addition to the current level's score and the player's total score up to that level. Level output and level score are displayed after player submission. In addition, some additional information is displayed, such as the time taken to complete and the times the paly button was pressed.

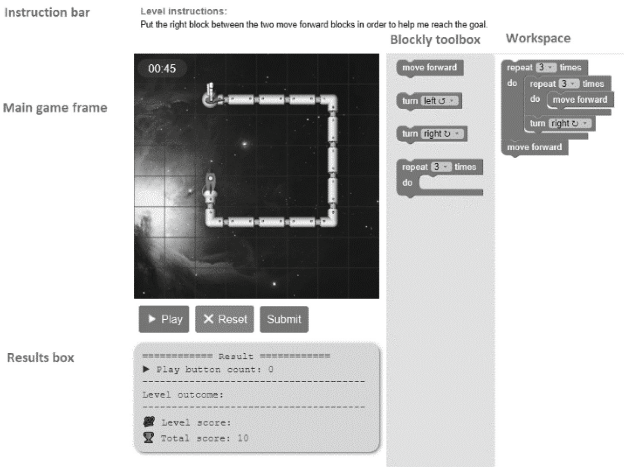


Fig. 1. The aMazeD game environment

Table 3. aMazeD feedback and scaffolding features

Feature
Feedback before each level: a) Explanations of CT concepts required for the solution of each level; b) Support regarding the login behind the solution design
Feedback after level failed: Failure message
Translate solution in text programming: Submitted instruction is translated into JavaScript and displayed on the screen
Highlighting the command being executed
Run the program at different speeds: a) Low speed when running the solution; b) High speed after submitting the solution
Solution execution: Running the solution multiple times before submitting
Restart Level: Reset and initialize the layer by clicking the restart button

4 Methodology

4.1 Research Questions

The research questions of the study are:

Q1. Do students perceive the aMazeD game as ease to use?

Q2. Do students perceive the aMazeD as effective on learning CT?

Q3. Do students perceive the scaffolding features of the aMazeD Game as effective in learning CT?

Q4. Do students have a positive attitude towards learning CT through games?

4.2 Research Design and Participants

To answer the research questions of the study, we adopted a survey research approach. For this purpose, we designed and implemented an educational intervention in a Greek school for students from grades 7 to 9 (ages 13 to 15) that has been approved by the Ethics Committee of the university of the authors. The intervention took place during formal teaching hours and lasted one and a half hours (two teaching hours) for each grade. Students played the aMazeD game for one hour and subsequently were asked to complete a questionnaire for about 30 min. Only students whose parents gave their written consent participated in the intervention. A total of 28 students were finally participated in the study.

4.3 Instrument

We adapted the instrument developed by Park [18]. The questionnaire is divided in the following sections: Perceived ease of use (PE), Perceived usefulness (PU), Attitude (AT), Accessibility (AC). A 5-point Likert scale from 1 to 5 was used for, where 1 equal "Strongly Disagree", 2 equals "Disagree", 3 equals "Indifferent", 4 equals "Agree" and 5 equals "Strongly Agree". In addition to the above sections, an open-ended question was included about the overall experience. The scale had a good level of internal consistency, as determined by a Cronbach's alpha of 0.761. The following paragraphs present the results of each section of the scale.

5 Findings

5.1 Evaluation of the aMazeD Game

Perceived Ease of Use (PE)
The following items was used to investigate how students perceive the ease of use of the aMazeD game, answering the first (Q1) research question: 1) PE1. I find the aMazeD programming and CT game easy to use; 2) PE2. Learning how to use a programing and CT game is easy for me. 64,3% of the students perceived the game as easy/very easy to use which is a slightly higher than the 60,7% who answered that they find easy/ very easy to learn how to use a CT game. While only 3,6% answered that disagrees that the game is easy to use.

Perceived Usefulness (PU)
The following items was used to investigate how students perceive the effectiveness of the aMazeD game on their CT learning, answering the second (Q2) research question: 1) PU1. The aMazeD game would improve my understanding of the concepts and practices of programming and CT; 2) PU2. The aMazeD game could make it easier to study the concepts and practices of programming and CT. A high percentage of 92.9% of the students answered that the aMazeD game would improve their understanding of CT practices. 57.1% consider that the game could make it easier for them to study CT concepts and practices, while 7.2% of students answered that they disagree/strongly disagree.

The following items was used to investigate how students perceive the effectiveness of the scaffolding features of the game, answering the third (Q3) research question: 1) PU3. The prompts the game provide me were enough to help me solve the levels; 2) PU4. The prompts the game provide me were useful to help me solve the levels; 3) PU5. The prompts the game helped me understand the basic concepts of programming and CT. 75% of students answered that the prompts provided were enough to help them solved the levels. 67.9% found them useful and 53.6% found that the prompts helped them understand CT concepts and practices. While only 17.1% stated that they disagree/strongly disagree that the game helped to understand the basic programming and CT concepts.

Attitude (AT)

The following items was used to investigate students' attitude towards learning CT through games, answering the fourth (Q4) research question: 1) AT1. Studying CT and programming through games such as aMazeD is a good idea; 2) AT2. I'm positive about programming and CT games. 82.1% has a positive attitude towards learning CT through games such as aMazeD. While only 3.6% express a negative attitude. 92.9% has a positive attitude towards CT games, while 3.6% express a negative attitude.

Overall Experience

Students were asked to answer the following open-ended question: "Write a few words about your experience of playing aMazeD. What did you like or dislike? What impressed you?". 25 students answered this open-ended question while three left it blank.

Regarding how students perceived the game, students generally found the game nice, interesting and fun. 11 students stated that the game was "nice"/ "very nice" / "interesting" / "fun" / "challenging".

Three students focused on the ease of use of the game. For example, one student stated that "The game is very well designed and easy to use."

Three students focused on the prompts. For example, St1: "I loved playing this game because of its ease of use. I was impressed by how helpful the tips were."; St2: "This is my second time doing programming, and the instructions given to us helped me to solve them [the levels] more easily."

The majority of the students also perceived the game as effective on learning CT and programming. This is supported by the following quotes: St3: "It was a really nice experience. The game helps in thinking and creativity."; St4: "I liked that it helped me understand CT a little bit."; St5: "The game was interesting to get acquainted with the programming."; St6: "I quite liked it because it is a fun way to learn things about programming". St7: "The thought process helps you understand CT concepts."

Finally, only two students express moderate or negative statements about the game. St10: "Although I did not find it very useful it was quite interesting"; St11: "I didn't like it."

6 Discussion and Conclusions

In this paper we evaluate a CT educational game (aMazeD) and investigate how students perceive the effectiveness of its scaffolding features. The results of the evaluation regarding ease of use, usefulness, attitude and overall experience are promising. Specifically, students seem to consider aMazeD and similar games as easy to use and accessible. What is also important is that students are in general positive to CT games. The results in questions regarding how students perceive usefulness of the game indicate that CT and programming games could help students develop CT. This is constant with prior research (e.g., [14, 16, 19]) that found that programming games could be effectively utilized to help students develop their CT. It is characteristic that a high percentage of 92% believe that the game could improve their CT. Students also found scaffolding features and specifically prompts useful for solving the game and effective in learning CT. This is reflected in their answers to the open-ended question where they evaluate the game and their experience as a whole. Almost all the comments are extremely positive, focusing on both the ease of use of the game and the effectiveness of its scaffolding features.

7 Limitations

Although this study is a first step in evaluating the aMazeD CT game and its scaffolding features, it has significant drawbacks and further research is needed to address them. First, the intervention is designed to include only one research group. And most importantly, the results are based on a self-report measure and capture student's opinions and perceptions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wing, J.M.: Computational thinking. Commun. ACM **49**(3), 33–35 (2006). https://doi.org/10.1145/1118178.1118215
2. Brennan, K., Resnick, M.: New frameworks for studying and assessing the development of computational thinking. In: Annual American Educational Research Association Meeting, Vancouver, pp. 1–25. http://web.media.mit.edu/~kbrennan/files/Brennan_Resnick_AERA2012_CT.pdf
3. Grover, S., Pea, R.: Computational thinking in K-12: a review of the state of the field. Educ. Research. **42**(1), 38–43 (2013). https://doi.org/10.3102/0013189X12463051
4. Lye, S.Y., Koh, J.H.L.: Review on teaching and learning of computational thinking through programming: what is next for K-12? Comput. Hum. Behav. **41**, 51–61 (2014). https://doi.org/10.1016/j.chb.2014.09.012
5. Kalelioglu, F., Gulbahar, Y., Kukul, V.: A framework for computational thinking based on a systematic research review. **4**(3), 583–596 (2016)
6. Buitrago Flórez, F., et al.: 'Changing a generation's way of thinking: teaching computational thinking through programming. Rev. Educ. Res. **87**(4), 834–860 (2017). https://doi.org/10.3102/0034654317710096

7. Hsu, T.-C., Chang, S.-C., Hung, Y.-T.: How to learn and how to teach computational thinking: suggestions based on a review of the literature. Comput. Educ. **126**, 296–310 (2018). https://doi.org/10.1016/j.compedu.2018.07.004
8. Ching, Y.-H., Hsu, Y.-C., Baldwin, S.: Developing computational thinking with educational technologies for young learners. TechTrends **62**(6), 563–573 (2018). https://doi.org/10.1007/s11528-018-0292-7
9. Papadakis, S.: The impact of coding apps to support young children in computational thinking and computational fluency. A literature review. Front. Educ. (2021). https://doi.org/10.3389/feduc.2021.657895
10. Trakosas, D., Tikva, C., Tambouris, E.: Visual programming and computational thinking environments for K-9 education: a systematic literature review. Int. J. Learn. Technol. **18**(1), 94–121 (2023). https://doi.org/10.1504/IJLT.2023.131313
11. Eguíluz, A., Garaizar, P., Guenaga, M.: An evaluation of open digital gaming platforms for developing computational thinking skills. Simul. Gaming (2017). https://doi.org/10.5772/INTECHOPEN.71339
12. Hershkovitz, A., et al.: Creativity in the acquisition of computational thinking. Interact. Learn. Environm. **27**(5–6), 628–644 (2019). https://doi.org/10.1080/10494820.2019.1610451
13. Garneli, V., Chorianopoulos, K.: Programming video games and simulations in science education: exploring computational thinking through code analysis. Interact. Learn. Environ. **26**(3), 386–401 (2018). https://doi.org/10.1080/10494820.2017.1337036
14. Zhao, W., Shute, V.J.: Can playing a video game foster computational thinking skills? Comput. Educ. (2019). https://doi.org/10.1016/j.compedu.2019.103633
15. Giannakoulas, A., Xinogalos, S.: A Review of Educational Games for Teaching Programming to Primary School Students, pp. 1–30 (2020). https://doi.org/10.4018/978-1-7998-4576-8.ch001
16. Karakasis, C., Xinogalos, S.:BlocklyScript: design and pilot evaluation of an RPG platform game for cultivating computational thinking skills to young students. Inform. Educ. **19**(4), 641–668 (2020). https://doi.org/10.15388/INFEDU.2020.28
17. Tikva, C., Tambouris, E.: The effect of scaffolding programming games and attitudes towards programming on the development of Computational Thinking. Educ. Inf. Technol. **28**, 6845–6867 (2023). https://doi.org/10.1007/s10639-022-11465-y
18. Park, S.Y.: An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. J. Educ. Technol. Soc. **12**(3), 150–161 (2009)
19. Giannakoulas, A., Xinogalos, S.: A pilot study on the effectiveness and acceptance of an educational game for teaching programming concepts to primary school students. Educ. Inf. Technol. **23**(5), 2029–2052 (2018). https://doi.org/10.1007/S10639-018-9702-X

Neural Networks and Data Mining

MonaCoBERT: Monotonic Attention Based ConvBERT for Knowledge Tracing

Unggi Lee[1], Yonghyun Park[2], Yujin Kim[3], Seongyune Choi[1], and Hyeoncheol Kim[1(✉)]

[1] Korea University, Seoul, Republic of Korea
[2] Seoul National University, Seoul, Republic of Korea
[3] KB Bank, Seoul, Republic of Korea
codingchild@korea.ac.kr, enkeejunior1@snu.ac.kr
{csyun213,harrykim}@korea.ac.kr

Abstract. Knowledge tracing (KT) is a research area of predicting students' knowledge states using their interaction data, such as concepts, questions, and responses. Most deep learning-based KT models have suffered from attributions of KT datasets such as the data sparsity, changeability of the knowledge state, and educational domain. Recently, most KT models use attention mechanisms to solve these problems. However, few studies tried to redesign the attention mechanism to restrict coverage of the local receptive field, for which the model can optimize to find the latent representation locally and globally in the students' interaction. In this study, we propose MonaCoBERT, with a redesigned attention mechanism by combining monotonic attention (MA) and span-based dynamic convolution (SDC), in order to represent global and local features together and to apply students' forgetting. As a result, MonaCoBERT achieves remarkable performance on most benchmark datasets. In addition, we used a classical test-theory-based embedding strategy to reflect the difficulty degree of knowledge concepts. We conducted ablation studies and further analysis to explain the remarkable performance of our model quantitatively. The analysis results demonstrate that SDC and MA complement one another. We also demonstrate that our model represents the relationship between concepts.

Keywords: knowledge tracing · educational data mining · intelligent tutoring system · personalized learning · attention

1 Introduction

Following the outbreak of COVID-19, a swift transition to online-based learning took place, profoundly impacting both students and educators [16]. This sudden change exposed widening learning gaps among students based on individual factors like socioeconomic status and race [1,3,14]. To bridge these disparities, there has been a growing interest in 'learning with AI'-an approach that leverages artificial intelligence to assist students [6,13,17,18,30].

Knowledge tracing (KT) has emerged as a pivotal research area within AI-enhanced learning. Its primary objective is to predict a student's knowledge

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 107–123, 2024.
https://doi.org/10.1007/978-3-031-63031-6_10

state, assessing if they have grasped the underlying concepts [33]. Intelligent tutoring systems (ITS) utilize KT to suggest appropriate questions tailored to individual students adaptively. By facilitating adaptive learning, ITS armed with KT tools could enhance educational opportunities for students of diverse backgrounds. Previous studies have further highlighted the efficacy of KT in streamlining curriculum mastery across subjects like mathematics and English. By doing so, KT conserves resources and reallocates saved time to other educational pursuits, thereby enriching students' learning experiences [4,34,36]. Such potential benefits have driven the expansion of KT research.

However, challenges arise when considering the nature of KT datasets. Their structure resembles natural language processing (NLP) datasets, given that both comprise sequences. Firstly, KT datasets are often sparser; each sequence represents one student's learning history, and the high attrition rates in online education render many of these sequences relatively brief and incomplete. [8,11,20] As such attempts to apply NLP models like BERT directly to KT have yielded subpar results, possibly because the Transformer architecture needs to capture locality [24,38,39].

Secondly, unlike static NLP datasets, a student's knowledge state in KT datasets evolves continually. Therefore, KT models must recognize global trends and specific moments in a student's learning journey. Moreover, KT is inherently educational, requiring insights from pedagogy, such as educational psychology and learning sciences. [9] For instance, as students may forget earlier topics more quickly than recent ones, considering factors like question difficulty-often calculated using item-based theory (IRT), becomes crucial in predicting student performance.

Given the attention mechanisms and transformers' success in various fields [40], KT models have incorporated similar techniques to address data sparsity. Examples include SAKT [32], MF-DAKT [47], and CL4KT [22], which amalgamate attention with methodologies like graph neural networks and the contrastive learning framework. However, there remains ample scope to optimize the attention mechanism for KT. Existing implementations often rely on standard attention [40], which primarily captures global sequences [25]. Given the unique characteristics of KT datasets, a more balanced representation that encompasses both global and local aspects of sequences is imperative.

We have re-envisioned the attention mechanism by drawing inspiration from studies like ConvBERT [15]. Our model captures both global and local representations while integrating educational insights. We employ span-based dynamic convolution (SDC) to represent local sequence features. Complementing this, monotonic attention, aligned with educational perspectives, captures the phenomenon of student forgetfulness [9]. Furthermore, we use an embedding based on the classical test theory (CTT) derived from student interactions in training datasets to denote concept difficulty.

Consequently, we introduce MonaCoBERT, a model employing a hybrid attention mechanism that fuses span-based dynamic attention with monotonic attention. In benchmark tests, MonaCoBERT consistently excels in metrics like

AUC and RMSE. Through ablation studies and supplementary analyses using the element-wise version of Grad-CAM [10,35], we ascertain that our methods adeptly capture both global and local information while elucidating inter-concept relationships.

2 Related Work

2.1 Knowledge Tracing

Knowledge tracing (KT) is a research area of predicting the knowledge states of the students using their interaction data, such as concepts, questions and responses. Since the first introduction of DKT, significant research in this area using deep neural networks has been conducted. DKT used recurrent neural network (RNN) architecture which designed for handling sequential data. [33]. For example, DKVMN [46] uses transformed memory-augmented neural network architecture, DKT+ [44] uses effective regularization for enhancing the performance. Researchers have recently focused on attention architectures. SAKT [32], SAINT+ [36], which uses attention, achieves a better performance than previous models. Moreover, AKT [9] was presented with self-attention, and a new architecture for retrieving latent knowledge representations was suggested. For AKT, a new embedding method that considers the educational perspective has also been suggested. PEBG [28], MF-DAKT [47] uses graph based embedding, achieves great performance. CL4KT [22] uses contrastive learning framework with data augmentation techniques and achieves a better performance than AKT. However, there has been less concerned about redesigning the attention mechanism. In this study, we focus on refining the attention mechanism which can represent global and local both for KT.

2.2 Locality Issues of Transformer

The Transformer architecture is widely used in most research areas and has achieved remarkable performance [19]. Transformer was used self-attention architecture which is considering the relationship between tokens, which model understand the context and the relationships naturally [40]. Many research reports that Transformers are effective in modeling long-distance dependencies of sequences [25]. Contrastively, some research revealed that Transformer architecture, especially self-attention, had trouble with locality-agnostics which is insensitive to the local context [24]. In Computer Vision (CV) research, Swin Transformer [29] used a hierarchical Transformer with shifted local windows and achieved both efficiencies and better performance in the same CV tasks. Meanwhile, some research tried to inject the locality inductive bias to solve this problem. Lee and colleagues [21] used locality self-attention to solve the lack of locality on small-size datasets. Li and colleagues [23] used lightweight CNN-based locality guidance for Vision Transformer (ViT) on tiny datasets. In NLP research, most of research focus on the global or long-range dependencies of models [2,26,42,45]. Whereas

ConvBERT [15] used a mixed attention mechanism that combined span-based dynamic convolution and attention to capture local dependencies. By using the mixed attention, ConvBERT can achieve better performance than the BERT [7], and ELECTRA [5].

Unfortunately, in KT research, no research investigated the locality problem of Transformer, and a few research [9] only redesigned the attention architecture to the best of our knowledge. Considering that the attribution of KT datasets is sparsity, solving the locality issue can be a solution for enhancing the model's performance.

3 Methodology

3.1 Problem Statement

Knowledge tracing aims at predicting the probabilities of students' correctness through sequences of interaction data gathered by an LMS or ITS. Student interactions can be expressed as $x_1, ..., x_t$. Interaction data in KT is consist of three things; question id, question's educational concept and student response. The t-th interaction can be denoted as $x_t = (c_t, q_t, r_t)$. Here, the c_t is t-*th* question's educational concept. The q_t is t-*th* question id. The $r_t$is t-th student's response, where $r_t \in \{0, 1\}$, in which 0 indicates an incorrect response and 1 is a correct answer.

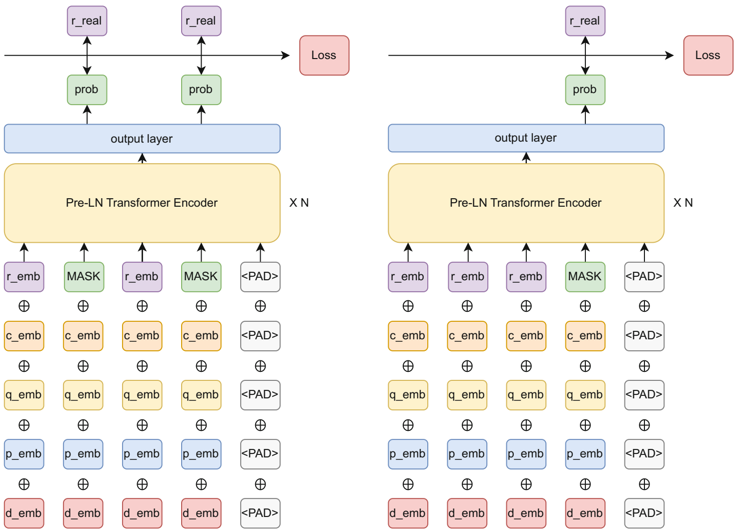


Fig. 1. Different strategies of MonaCoBERT. *left* is training and *right* is testing session.

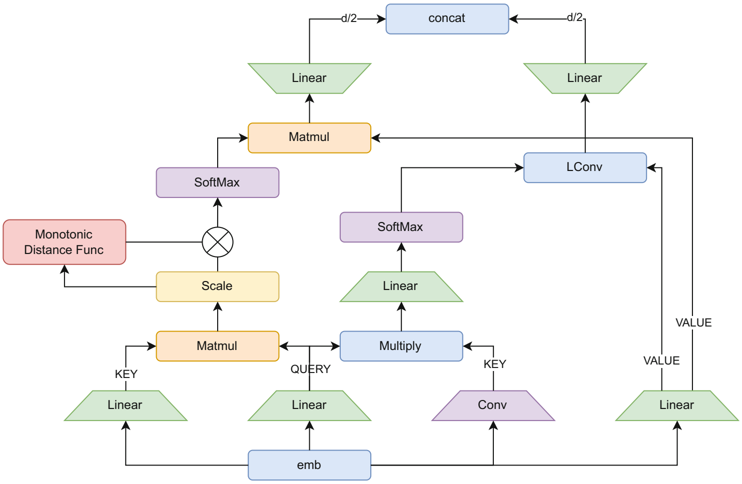


Fig. 2. Architecture of monotonic convolutional multi-head attention, combined with monotonic attention and span-based dynamic convolution (SDC).

3.2 Proposed Model Architecture

BERT Based Architecture for Knowledge Tracing. To create our model baseline, we mainly referenced BERT [7], BiDKT [38], BEKT [39], and BERT4Rec [37]. To optimize our research into KT, we changed some of the BERT architecture. First, we used a pre-layer normalization (pre-LN) Transformer in our model. Previous research [27] has suggested that Transformer is difficult to train without a training strategy, such as a warm-up start. By contrast, the pre-LN Transformer can be trained without a warm-up start and converges much faster than the original Transformer [41]. Second, we used a different strategy for the training and testing processes. During the training process, the proposed model predicted the masking position. The masking ratio used in the training process was the same as with the original BERT, which used 15% embedding, 80% of which was actual masking, 10% was a reversal, and 10% did not change. During the testing process, masking was applied to the last position of each sequence. Referring to the previous BERT-based studies on KT [38] or recommendation systems [37], the model predicts the correctness of the students using their previous history of interaction sequences. Figure 1 describes the different training and testing strategies of our model.

Embedding Strategy. Most KT models use concepts, questions, and correctness as the input vectors for training. Previous studies have explored new input features. For example, AKT created Rasch embedding vectors by using concepts, items, and responses [9]. However, an item response theory (IRT), such as Rasch, can be applied to the dataset collected from tests or examinations because IRT assumes that the ability of a student does not change during the trial. In KT, the states of student knowledge change during learning [43]. Therefore, we used the classical test theory (CTT) for handling the difficulty features.

Note that we extracted the correctness of each question from the training set and made the questions difficulty. If the question in the validation or test set were not contained in training set, we replaced that question difficulty as a arbitrarily number like 75. Subsequently, we added the difficulty to the embedding blocks. In a previous study, BEKT [39] used five difficulty ranges in its embedding blocks. Nevertheless, we used a difficulty range of 100. Similar to BERT embedding layers, MonaCoBERT uses embedding vectors E_{input}, which is element-wise of each embedding; learnable positional embedding E_{pos}, concept embedding E_c, item embedding E_q, correctness embedding E_r, and CTT embedding E_{ctt}, where $E_{input}, E_{pos}, E_c, E_q, E_r, E_{ctt} \in R^{m \times h}$. Embedding layers E_{input} are formulated as follows:

$$E_{input} = E_{pos} + E_c + E_q + E_r + E_{ctt} \tag{1}$$

Pre-LN Transformer-Based Encoder Architecture. The encoder blocks used the pre-LN Transformer architecture [41]. In this study, 12 encoder layers were used. First, the embedding vectors E_{input} are normalized through the pre-LN LN_{pre}

$$z = LN_{pre}(E_{input}) \tag{2}$$

Second, the normalized value z was changed to the query, key, and value of monotonic convolutional multihead attention A_{mc}. The results were passed through dropout layer D and added to the embedding vectors as a residual connection.

$$a = x + D(A_{mc}(z, z, z)) \tag{3}$$

Third, the results were normalized with layer normalization LN. Normalized results passed through fully connected layers fc. The results were also normalized through the dropout layer D. The second result was added as a residual connection.

$$l = a + D(fc(LN(a))) \tag{4}$$

The fully connected layers fc is consist of two layer blocks W_{fc1}, W_{fc2} and activation function. The activation function is LeakyReLU σ. The fc are formulated as follows.

$$fc = W_{fc2}(\sigma(W_{fc1})), \tag{5}$$

where $W_{fc1} \in R^{h \times (h*n)}$, $W_{fc2} \in R^{(h*n) \times h}$.

Monotonic Convolutional Multihead Attention. We suggest the use of monotonic convolutional multihead attention. This architecture is combined with ConvBERT's [15] span-based dynamic convolution and AKT [9] monotonic attention. In previous research, attention with the span-based dynamic convolution achieved a higher performance than normal attention with BERT. Meanwhile, the sequence data in KT contain latent information about the forgetting of the students. To represent such forgetting, we used the exponential decay mechanism of monotonic attention. Figure 2 shows the monotonic convolutional multihead attention architecture.

The monotonic convolutional multihead attention A_{mc} consists of the concatenation ($[;]$) of monotonic multihead attention A_m and span-based dynamic convolution SDC. Here, A_{mc} can be formulated as follows:

$$A_{mc}(Q,K,V) = [A_m(Q,K,V); SDC(Q,K,V)]. \tag{6}$$

First, Monotonic multihead attention A_m has an exponential decay mechanism for measuring the distance between sequences. The exponential decay mechanism is a dot product with query linear W_Q and key linear W_K. The learnable parameter δ is multiplied by these values. In addition, A_m can be formulated as follows:

$$Am = softmax(\frac{exp(-\delta \cdot d(t,\tau)) \cdot W_Q^T \cdot W_K}{\sqrt{D_k}}), \delta > 0. \tag{7}$$

Here, $d(t,\tau)$ is the distance function, where t is the present time step, and τ is the previous time step. Details can be found in AKT paper [9]. $d(t,\tau)$ and $\gamma_{t,t'}$ can be formulated as

$$d(t,\tau) = |t-\tau| \cdot \sum_{t'=\tau+1}^{t} \gamma_{t,t'}. \tag{8}$$

$$\gamma_{t,t'} = \frac{exp(\frac{W_t^T \cdot W_{t'}^T}{\sqrt{D_k}})}{\sum_{1 \leq \tau' \leq t} exp(\frac{W_t^T \cdot W_{t'}^T}{\sqrt{D_k}})}, t' \leq t. \tag{9}$$

Second, the span dynamic convolution SDC can be formulated below, where W is a linear layer, and $\otimes$ can be denoted as a point-wise multiplication (Details can be found in ConvBERT paper [15]).

$$SDC(Q,K,V) = LConv(V, softmax(W(Q \otimes K))), \tag{10}$$

$$LConv(X,W) = \sum_{j=1}^{k} W_j \dot{X}_{i+j-[\frac{[k+1]}{2}]} \tag{11}$$

Table 1. Overall performance of KT models based on five benchmark datasets. The best performance is denoted in bold underline, the second in bold, and the third in underline. MCB-C indicates that MonaCoBERT used classical test theory (CTT), whereas MCB-NC indicates that it did not. We can see that MCB-C achieved the best results, and MCB-NC was second for most of the benchmark datasets.

Dataset	Metrics	DKT	DKVMN	SAKT	AKT	CL4KT	MCB-NC	MCB-C
assist09	AUC	0.7285	0.7271	0.7179	0.7449	<u>0.7600</u>	**0.8002**	**<u>0.8059</u>**
	RMSE	<u>0.4328</u>	0.4348	0.4381	0.4413	0.4337	**<u>0.4029</u>**	**0.4063**
assist12	AUC	0.7006	0.7011	0.6998	<u>0.7505</u>	0.7314	**0.8065**	**<u>0.8130</u>**
	RMSE	0.4348	0.4355	0.4360	<u>0.4250</u>	0.4284	**0.3976**	**<u>0.3935</u>**
assist17	AUC	**<u>0.7220</u>**	<u>0.7095</u>	0.6792	0.6803	0.6738	0.6700	**0.7141**
	RMSE	**<u>0.4469</u>**	**0.4516**	<u>0.4591</u>	0.4722	0.4713	0.4727	0.4630
algebra05	AUC	0.8088	0.8146	<u>0.8162</u>	0.7673	0.7871	**0.8190**	**<u>0.8201</u>**
	RMSE	0.3703	<u>0.3687</u>	**0.3685**	0.3918	0.3824	0.3940	**<u>0.3584</u>**
algebra06	AUC	0.7939	<u>0.7961</u>	0.7927	0.7505	0.7789	**0.7997**	**<u>0.8064</u>**
	RMSE	**0.3666**	**<u>0.3661</u>**	0.3675	0.3986	0.3863	0.3835	<u>0.3672</u>
EdNet	AUC	0.6609	0.6602	0.6506	<u>0.6687</u>	0.6651	**0.7221**	**<u>0.7336</u>**
	RMSE	0.4598	<u>0.4597</u>	0.4629	0.4783	0.4750	**0.4572**	**<u>0.4516</u>**

3.3 Experiment Setting

Datasets. We validated the effectiveness of our model using six benchmark datasets widely used in KT research. These benchmark datasets were obtained from realistic online-based educational environments. All students' Personal Identifiable Information (PII) was removed to address privacy concerns. In our research, we excluded students with at most five interactions. When a dataset contained multiple concepts within a single interaction, we treated the combination of those concepts as unique.

- **Assitment09, 12, 17**: The ASSISTment datasets were sourced from the ASSISTment ITS, primarily from American middle schools. Participants in these datasets were randomly assigned [12]. We used assist09, assist12, and assist17 but excluded assist15 due to its lack of question information[1]
- **Algebra05, 06**: These algebra datasets, which were provided by the KDD Cup 2010 Educational Data Mining Challenge, originated from the Cognitive Tutor. This ITS, developed by Carnegie Learning, targets middle school students [34].[2]
- **EdNet**: The EdNet dataset was provided by a South Korean EdTech company named Santa, primarily emphasizing the English test TOEIC administered by ETS. It boasts 131,441,538 interactions from 784,309 students collected since 2017. This dataset primarily caters to adult learners aiming to

[1] retrieved from https://sites.google.com/site/assistmentsdata/home.
[2] retrieved from https://pslcdatashop.web.cmu.edu/KDDCup.

certify their English proficiency [4][3] Due to the large size of the dataset and in order to expedite our experiments, we extracted a subset of 5,000 interactions from the original dataset for our research. Table 2 enumerates the number of features in each benchmark dataset.

Table 2. Six benchmark datasets used our work. Benchmark dataset ignored student data with less than five interactions. #Concepts are the same as the skills.

Dataset	#Students	#Concepts	#Questions	#interactions
assist09	3,695	149	17,728	282,071
assist12	24,429	264	51,632	1,968,737
assist17	1,708	411	3,162	934,638
algebra05	571	271	173,113	607,014
algebra06	1,318	1,575	549,821	1,808,533
EdNet	5,000	1,472	11,957	641,712

Evaluation Metrics and Validation. By referencing CL4KT, we used both AUC and the RMSE as the performance metrics. We also used a five-fold cross-validation for the evaluation.

Baseline Models. We compared MonaCoBERT to the baseline models, such as DKT [33], DKVMN [46], SAKT [32], and the latest models, such as AKT [9] and CL4KT. [22].

Hyperparameters for Experiments. To compare each model, we used the same parameters for the model training.

- **batch size**: The batch size was 512. Owing to a limitation of resources, we also used a gradient accumulation.
- **early stop**: The early stop parameter was 10. For example, if the validation score was not successively increased during the ten iterations, the training session was stopped.
- **training, validation, test ratio**: The training ratio was 80% of the entire dataset, and the test ratio was 20%. The valid ratio was 10% of the training ratio.
- **learning rate and optimizer**: The learning rate was 0.001, and Adam was used as the optimizer.
- **embedding size**: The embedding size was 512.

[3] retrieved from https://github.com/riiid/ednet.

– **others**: We used eight attention heads for MonaCoBERT. The Max sequence length was 100, and the encoder number was 12. Other models such as AKT[4] and CL4KT[5] used the default settings.

4 Result and Discussion

4.1 Overall Performance

Figure 1 illustrates the overall performance of each model. Every model used a five-fold cross-validation for the estimation. MonaCoBERT-C, which was trained using CTT, was the best model in most benchmark datasets and was a new state-of-the-art model in assist09, assist12, algebra05, and ednet. MonaCoBERT-NC was the second-best model for most of the datasets. This result indicates that CTT embedding affects the performance of the model. For all datasets, MonaCoBERT-C performed better than MonaCoBERT-NC. This result indicates that it was difficult for MonaCoBERT-NC to learn the latent representations of the item difficulty from the dataset.

Our estimation differs from that of previous research. Except for MonaCoBERT-NC and MonaCoBERT-C, the best model was modified for each dataset. For instance, the AUC and RSME of assist17, and the RMSE, DKT, and DKVMN of algebra06 showed that these were the best and second-best models, respectively. This indicates that DKT and DKVMN are still helpful in predicting certain cases. These results may stem from pre-processing methods or the training of the hyperparameter settings.

4.2 Ablation Studies

In this section, we explore why MonaCoBERT performed better than the other models and which parts of the model affected the increase in performance.

Impact of Attention Mechanisms. In Table 3, we compare the attention mechanisms. For comparison, we used the assist09 and assist09-CTT datasets. The assist09 dataset is a normal dataset that contains concepts, questions, and correctness; however, assist09-CTT contains the concepts, questions, correctness, and CTT-based difficulty.

We detached each part of the monotonic convolutional multi-head attention and created four attention mechanisms: normal multi-head attention, monotonic multi-head attention, convolutional multi-head attention, and monotonic convolutional multi-head attention. We also used a five-fold cross-validation and an early stop 10 times. The other hyperparameters used to determine the overall performance were the same.

As a result, monotonic convolutional multihead attention exhibited the best performance for both comparisons. Convolutional multihead attention and

[4] https://github.com/arghosh/AKT.
[5] https://github.com/UpstageAI/cl4kt.

monotonic multihead attention achieved the second-best performance under each setting. The increments differed for each setting and were approximately 2% for assist09 and 1-2% for assist09-CTT.

Table 3. AUC performances of each attention mechanism using the assist09 and assist09-CTT datasets. The increments were written based on normal attention.

Dataset	Attn	MonoAttn	ConvAttn	MonoCoAttn
assist09	0.7736	**0.7993**	<u>0.7959</u>	<u>**0.8002**</u>
increment	0	+ 0.026	+ 0.022	+ 0.027
assist09-CTT	0.7858	<u>0.8039</u>	**0.8054**	<u>**0.8059**</u>
increment	0	+ 0.018	+ 0.020	+ 0.021

Impacts of Embedding Strategy. In Table 4, we compare each embedding strategy. The first embedding strategy emb_{cq} is an element-wise sum of the concept embedding emb_c, question embedding emb_q, and correctness embedding emb_r.

$$emb_{cq} = emb_c + emb_q + emb_r \tag{12}$$

Moreover, the second embedding strategy emb_{rasch} is an element-wise sum of concept and Rasch embedding, as suggested by AKT. Rasch embedding uses concept embedding emb_c and learnable question scalar emb_q or a combination of concepts and answer embedding emb_{cr} to calculate the difficulty, where $emb_c, emb_{cr} \in R^{n\times h}$ and $emb_q \in R^{n\times 1}$. Note that IRT Rasch embedding differs from AKT Rasch embedding because the condition of IRT assumes that the knowledge state of the student is fixed and does not change when estimated.

$$emb_{rasch-c} = emb_c + emb_q * emb_c \tag{13}$$

$$emb_{rasch-cr} = emb_{cr} + emb_q * emb_{cr} \tag{14}$$

The last embedding strategy, emb_{CTT}, is an element-wise sum of concept embedding, question embedding, correctness embedding, and CTT embedding, emb_{ctt}, which was suggested in this study. We set emb_{ctt} as the probability of the difficulty and the integer type, where $0 \leq emb_{ctt} \leq 100$.

$$emb_{CTT} = emb_c + emb_q + emb_r + emb_{ctt} \tag{15}$$

As a result, in Table 4, emb_{CTT} generally showed a better performance than the other embedding strategies. DKVMN, AKT, and MonaCoBERT performed

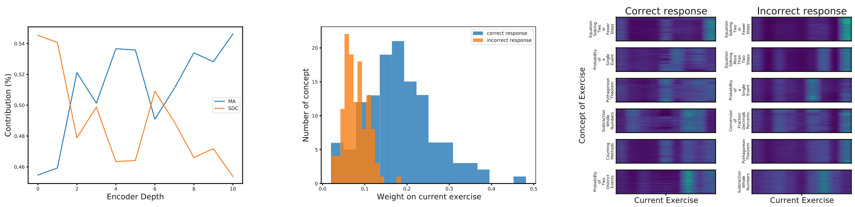


Fig. 3. Analysis of underlying behaviors of SDC. The figure on the *left* illustrates the proportion of the importance of each module. SDC showed an importance competitive to that of MA in most layers. In particular, the SDC showed the most significant contribution in the first layer. The histogram in the *center* figure represents the current input weight of the concept. When the response of the student was correct, the SDC allocated more weight to the interaction. In addition, even if the response was the same, the weight varied considerably based on the concept. The figure on the *right* shows examples of SDC filters arranged based on the correctness and concept.

well when using emb_{CTT}. This result indicates that the models did not learn the difficulty representation during training. Meanwhile, CL4KT and SAKT showed slightly better performances when using emb_{rasch}. DKT was not affected by the embedding strategy.

Table 4. Comparison of each embedding strategy with KT models in the assist09 dataset.

Embedding Strategy	emb_{cq}	emb_{rasch}	emb_{CTT}
DKT	0.7263	**0.7274**	0.7239
DKVMN	0.7188	0.7255	**0.7313**
SAKT	0.6822	**0.6941**	0.6693
AKT	0.7440	0.7449	**0.7632**
CL4KT	0.7600	**0.7601**	0.7461
MCB	0.8002	0.7736	**0.8059**

4.3 In-Depth Analysis of Attention and Embedding

In this subsection, we analyze the attention and embedding in depth. We used Grad-CAM and t-SNE for the analysis and visualization.

Analysis of SDC. First of all, we observed that SDC was more critical than MA in the early layer. Figure 3-*Left* shows the relative importance ratios of SDC and MA. The contribution of SDC was especially greater than that of MA in the first layer. To define the importance of each module, we used an element-wise version of Grad-CAM as a metric [10,35]. We also found that SDC in the first layer extracted useful information regarding the properties of the current input. Specifically, SDC focused on the current input when the student answered correctly. In Fig. 3-*Center, Right*, we can see that SDC assigned higher weights to the current inputs when the student responded correctly. Moreover, the large variance of weights given correct responses implies that SDC considers not only the correctness of responses but also the importance of the concept. (Figure 3-*Center*, blue) This result shows us that MonaCoBERT implicitly learned what concepts or questions were essential for inferring the ability of the students. This indicates the possibility of using MonaCoBERT to automatically find the problem essential to estimating the student's ability, which can be used to support the estimation and assessment.

CTT Based Embedding. We showed that CTT-based embedding helps the model represent the difficulty of the problem. Figure 4 shows a visualization using t-SNE [31]. Figure 4-*Left* shows the visualization of the CTT-based embedding vector, and Fig. 4-*Right* shows the visualization of the No-CTT-based embedding. Unlike No-CTT-based embedding, where different difficulties are mixed in each cluster, CTT-based embedding (i.e., emb_{CTT}) showed that the difficulty of the information was smoothly distributed globally.

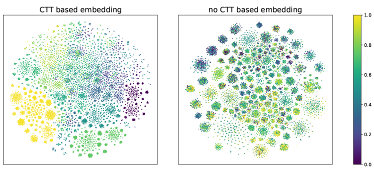


Fig. 4. Visualization of the embedding vector. The figure on the *left* shows the results with CTT-based embedding. The figure on the *right* shows the results of No-CTT-based embedding. We can see that the results of CTT-based embedding not only represent the difficulty information globally, they also help avoid a difficulty in the mixing in each cluster.

4.4 Discovery of Relationships Between Concepts

To determine whether our model understood the relevance between concepts, we analyzed the monotonic attention weights of the last encoder layer after passing through the softmax function. The results are shown in Fig. 5-*left*. We averaged the attention scores of the questions using the same concepts to obtain the relevance between concepts. We created a directed graph, as shown in Fig 5-*Center*, by selecting only those concepts with attention weights of higher than 0.1.

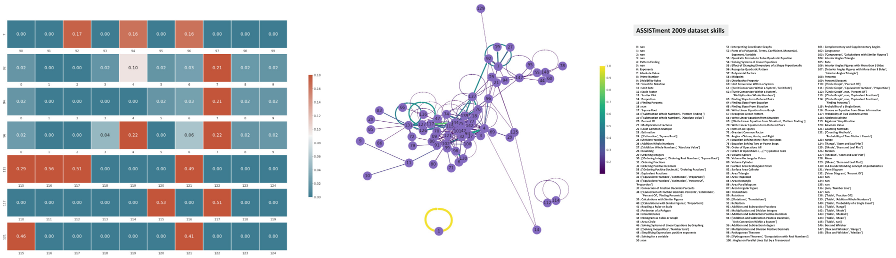


Fig. 5. Analysis results of the relevance between knowledge concepts, exploiting attention weights of the monotonic attention part after the model was trained using monotonic convolutional multi-head attention. The figure on the *left* shows a heatmap of the attention weights between each pair of concepts. It shows how much attention each concept on the y-axis (e.g., 7th, 92nd, 94th, 96th) assigns attention to some selected concept on the other x-axis. The *center* figure shows a directed graph of the relevance between concepts. It shows how the concepts of assist09 influence one another. The source concept nodes are assigned a high attention weight to the destination concept nodes, and the concept nodes can be connected in both directions. We set the threshold to 0.1 and ignored edges lower than the threshold. When the threshold was decreased, more skill nodes were connected, and vice versa. The concept information of the assist09 dataset can be found on the *right*. 'nan' means concepts that are not defined in the original dataset.

According to the concept network shown in Fig. 5-*center*, we can see that the model learns the relevance between skills. For example, as shown in Fig. 5-*left*, the 7th concept (Absolute Value) was connected with some concepts of subtraction, such as 92 (Addition and Subtraction Fractions), 94 (Addition and Subtraction Positive Decimals), and 96 (Addition and Subtraction Integers). This means that you need to be good at subtraction to calculate the correct absolute value. Accordingly, the 117th concept (Probability of a Single Event) and 115th concept (Probability of Two Distinct Events) assigned high attention weights to each other, since concept 117 is a prerequisite for concept 115. 121st concept (Counting Methods) is also connected with 115 and 117. However, the concept network shown in Fig. 5 is not perfect because some concepts did not connect to each other despite their similarities. This result may be due to the monotonic attention decreasing the attention weight according to the time step. Nevertheless, observing the attention weights can help uncover new connections between previously inconceivable concepts.

5 Conclusion

This study presents MonaCoBERT, a cutting-edge Knowledge Tracing model combining monotonic attention with span-based dynamic convolution. The model Effectively addresses students' forgetting curves and integrates global and local representations. Across benchmark datasets, MonaCoBERT has showcased superior performance, highlighting its immense potential in real-world educational settings. The model adeptly captures learning sequences from granular and overarching viewpoints by integrating monotonic convolutional multi-head attention. Our distinctive approach, grounded in Classic Test Theory, has been instrumental in elevating the model's efficacy. MonaCoBERT offers a pioneering solution to the challenges presented by data sparsity and intricacies in the education domain. Its practical uses, such as forecasting student performance and customizing learning experiences, leverage its profound understanding of student learning patterns. Such features emphasize MonaCoBERT's capability to instigate notable educational advancements, spurring further investigation and deployment. However, further in-depth research is required to fine-tune its effectiveness for different teaching scenarios and age groups.

References

1. An, Y., Kaplan-Rakowski, R., Yang, J., Conan, J., Kinard, W., Daughrity, L.: Examining k-12 teachers' feelings, experiences, and perspectives regarding online teaching during the early stage of the COVID-19 pandemic. Educ. Tech. Res. Dev. **69**(5), 2589–2613 (2021)
2. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: the long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
3. Bonal, X., González, S.: The impact of lockdown on the learning gap: family and school divisions in times of crisis. Int. Rev. Educ. **66**(5), 635–655 (2020)
4. Choi, Y., et al.: EdNet: a large-scale hierarchical dataset in education. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 69–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7_13
5. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
6. Cukurova, M., Luckin, R., Kent, C.: Impact of an artificial intelligence research frame on the perceived credibility of educational research evidence. Int. J. Artif. Intell. Educ. **30**(2), 205–235 (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Feng, W., Tang, J., Liu, T.X.: Understanding dropouts in MOOCs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 517–524 (2019)
9. Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2330–2339 (2020)
10. Gildenblat, J., contributors: Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam (2021)
11. He, J., Bailey, J., Rubinstein, B., Zhang, R.: Identifying at-risk students in massive open online courses. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)
12. Heffernan, N.T., Heffernan, C.L.: The assistments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Int. J. Artif. Intell. Educ. **24**, 470–497 (2014)

13. Holmes, W., Bialik, M., Fadel, C.: Artificial Intelligence in Education. (2020)
14. Hu, Y., Spiro, R.J.: Design for now, but with the future in mind: a "cognitive flexibility theory" perspective on online learning through the lens of MOOCs. Educ. Tech. Res. Dev. **69**(1), 373–378 (2021)
15. Jiang, Z.H., Yu, W., Zhou, D., Chen, Y., Feng, J., Yan, S.: ConvBERT: improving BERT with span-based dynamic convolution. Adv. Neural. Inf. Process. Syst. **33**, 12837–12848 (2020)
16. Karakaya, K.: Design considerations in emergency remote teaching during the COVID-19 pandemic: a human-centered approach. Educ. Tech. Res. Dev. **69**(1), 295–299 (2021)
17. Kim, S., Kim, W., Jang, Y., Choi, S., Jung, H., Kim, H.: Student knowledge prediction for teacher-student interaction. In: EAAI, pp. 15560–15568 (2021)
18. Kim, S., Kim, W., Jung, H., Kim, H.: DiKT: dichotomous knowledge tracing. In: Cristea, A.I., Troussas, C. (eds.) International Conference on Intelligent Tutoring Systems, pp. 41–51. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80421-3_5
19. Kitaev, N., Kaiser, Ł., Levskaya, A.: Reformer: the efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
20. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 170–179 (2013)
21. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021)
22. Lee, W., Chun, J., Lee, Y., Park, K., Park, S.: Contrastive learning for knowledge tracing. In: Proceedings of the ACM Web Conference 2022, pp. 2330–2338 (2022)
23. Li, K., Yu, R., Wang, Z., Yuan, L., Song, G., Chen, J.: Locality guidance for improving vision transformers on tiny datasets. arXiv preprint arXiv:2207.10026 (2022)
24. Li, S., et al.: Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
25. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: LocalViT: bringing locality to vision transformers. arXiv preprint arXiv:2104.05707 (2021)
26. Lin, Y., et al.: BertGCN: transductive text classification by combining GCN and BERT. arXiv preprint arXiv:2105.05727 (2021)
27. Liu, L., Liu, X., Gao, J., Chen, W., Han, J.: Understanding the difficulty of training transformers. arXiv preprint arXiv:2004.08249 (2020)
28. Liu, Y., Yang, Y., Chen, X., Shen, J., Zhang, H., Yu, Y.: Improving knowledge tracing via pre-training question embeddings. arXiv preprint arXiv:2012.05031 (2020)
29. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
30. Luckin, R., Holmes, W., Griffiths, M., Forcier, L.B.: Intelligence Unleashed: An Argument for AI in Education (2016)
31. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)
32. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. arXiv preprint arXiv:1907.06837 (2019)
33. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, vol. 28 (2015)

34. Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.: Cognitive tutor: applied research in mathematics education. Psychon. Bull. Rev. **14**, 249–255 (2007)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
36. Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., Choi, Y.: Saint+: integrating temporal features for EdNet correctness prediction. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 490–496 (2021)
37. Sun, F., et al.: BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1441–1450 (2019)
38. Tan, W., Jin, Y., Liu, M., Zhang, H.: BiDKT: deep knowledge tracing with BERT. In: Bao, W., Yuan, X., Gao, L., Luan, T.H., Choi, D.B.J. (eds.) ADHOC-NETS/TridentCom -2021. LNICST, vol. 428, pp. 260–278. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98005-4_19
39. Tiana, Z., Zhengc, G., Flanaganb, B., Mic, J., Ogatab, H.: BEKT: deep knowledge tracing with bidirectional encoder representations from transformers. In: Proceedings of the 29th International Conference on Computers in Education (2021)
40. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
41. Xiong, R., et al.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning, pp. 10524–10533. PMLR (2020)
42. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
43. Yeung, C.K.: Deep-IRT: make deep learning based knowledge tracing explainable using item response theory. arXiv preprint arXiv:1904.11738 (2019)
44. Yeung, C.K., Yeung, D.Y.: Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale, pp. 1–10 (2018)
45. Zaheer, M., et al.: Big bird: transformers for longer sequences. Adv. Neural. Inf. Process. Syst. **33**, 17283–17297 (2020)
46. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th International Conference on World Wide Web, pp. 765–774 (2017)
47. Zhang, M., Zhu, X., Zhang, C., Ji, Y., Pan, F., Yin, C.: Multi-factors aware dual-attentional knowledge tracing. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2588–2597 (2021)

Detection of Pre-error States in Aircraft Pilots Through Machine Learning

Massimo Pietracupa[1], Hamdi Ben Abdessalem[2](✉), and Claude Frasson[1]

[1] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal H3C 3J7, Canada
massimo.pietracupa@umontreal.ca, frasson@iro.umontreal.ca

[2] Département d'Informatique et de Mathématique, Université du Québec à Chicoutimi, Chicoutimi G7H 2B1, Canada
hamdi_benabdessalem@uqac.ca

Abstract. This study explores the feasibility of training a machine learning model to recognize pre-error signals in the anterior cingulate cortex (ACC) using Flanker test data from the COG-BCI dataset, and subsequently employing this model to detect pre-error states in aviation pilots. To address this issue, we applied various machine learning models to the dataset, including Support Vector Machines (SVM), and Random Forests, double Convolutional Neural Network (CNN) model, and a Transformer model, renowned for handling sequential data efficiently. Pilot experiments were conducted in an Airbus A320 simulator to assess real-time cognitive activity during takeoff, involving seven pilots and six engineers. Cognitive workload (CW), heart rate (HR), and pupil diameter (PD) were measured using an EEG headset, Polar H10 heart rate monitor strap, and Gazepoint GP3 eye tracker. Results from the analysis of the FLANKER dataset using various models revealed the superiority of the transformer model, with notable reductions in false negatives and a final F1 score of 0.610. Moving beyond typical study conclusions, our objective extends to assessing model applicability in a secondary domain—evaluating the classifiers' discriminative power during takeoff procedures for aviation pilots. Despite a slight reduction in performance, the transformer model still outperforms other models in classification with an F1 of 0.578. Although there's room for improvement in erroneous state detection, these results indicate trends in electrical brain activity that correlate with decreased behavioral performance. The transformer's real-time performance, with an inference speed of 0.01 s, positions it favorably for Brain-Computer Interface (BCI) applications. As we anticipate increases in classifier performance with more training data and extended polling bands, this study lays the groundwork for further research in erroneous state prediction and machine learning optimization models for BCI and real-world applications.

Keywords: Brain Controller Interfaces · EEG · Transformers · Aviation · Pilot · Error Prediction

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 124–136, 2024.
https://doi.org/10.1007/978-3-031-63031-6_11

1 Introduction

Pilots frequently encounter scenarios requiring swift processing of significant information within tight timeframes, leading to the potential for piloting errors and, unfortunately, fatal outcomes. To investigate and address this issue, a 15-year observational methodology study was undertaken, employing a Line Operations Safety Audit (LOSA). Expert observers were strategically placed in the cockpit during routine flights to document potential threats. The study's reports highlighted a range of common errors, with one notable example being procedural lapses resulting in incorrect executions, such as entering erroneous data into the flight management computer [1]. In response to these challenges, deviation detection systems have been developed to assist pilots in verifying task adherence to established guidelines and facilitating quick adjustments [2]. However, it's crucial to note that, despite these advancements, certain mistakes remain irrecoverable [3–5]. This underscores the significance of prioritizing error prevention measures, and catching these potential errors earlier, over relying solely on error correction systems.

Brain Computer Interfaces (BCI) have been growing in popularity with regards to scientific research in neurotechnology [6]. These BCI devices can come in multiple forms, from non-invasive, and partially invasive to invasive, based on how close electrodes get to brain tissue [8]. Electroencephalograms (EEG), a form of non-invasive BCI devices, can read the user's brain electrical activity through a mesh of electrodes. This electrical activity can then be interpreted to understand subject brain states along with their brain connectivity [7]. One such study mapped connectivity of cognitive monitoring and executive control of strategy adjustment to the anterior cingulate cortex (ACC) through neuroimaging and brain potential studies [9]. These monitoring processes are heavily intertwined with cognitive task performance and can be indicative of performance degradation. It is believed that the ACC can detect potential errors or conflicting responses prior to being consciously aware of them [10] and then begins to signal the need to regulate future uncertainty and risk minimization. An example of this could be a student answering a question on an exam for which he is uncertain of the response. The ACC will detect potential errors or conflicting responses, such as misinterpretations of questions and then activate adaptive control processes, prompting the student to adjust their approach, clarify uncertainties, and optimize their performance to minimize the risk of errors to enhance overall exam success. Ridderinkhof was able to identify these markers using the Flanker test and an EEG device. They concluded that errors were observed to be presaged by a distinct pattern of electrophysiological brain activity on the trial preceding the error and that the ACC serves to indicate the need to engage control processes (including response inhibition) to minimize the risk of errors [11].

As a result, the hypothesis that will be tested in this article is as follows:

Hypothesis: Is it possible to train a machine learning model to identify these ACC pre-error signals using flanker test data, and apply this model to detect pre-error states in aviation pilots?

First, we will present a section for related works in the domain of predicting perceptual decision-making errors. Next, we will examine the methods implemented in this paper, and finally Sect. 4 will dive into the results.

2 Related Works

Predicting perceptual errors in BCIs is still in a novel state with Batmanova et al. claiming to have taken the first steps towards predicting these error states. They setup a perceptual decision-making task to collect behavioral data and brain activity signals [12]. This experiment required participants to perceive 400 stimuli in the form of neckar cubes continuously with a brief pause interval, attempting to replicate real-life situations where decision making would be done in a stressful environment. Their machine learning stack involved using two one-dimensional convolutional neural networks (CNN) in series. The first CNN performed an EEG channel wise convolution, while the second performed a time-point-wise convolution, with the size output being a hyperparameter, K, that was selected based on model performance. Their EEG data was taken two seconds before the stimulus and two seconds after the stimulus, filtered by performing z-score normalization and thresholding at a value of 1. Despite receiving exceptional results, an F1-score of 88% and accuracy of 92.6%, there is no mention of transferability of the model beyond the dataset. They also mention a limitation of not addressing participant leakage, where a single participant's data may be found in both training and validation set. The error and success trials of a participant should be locked in either the test or validation set, if not, results can be skewed by providing exceptional F1 and accuracy scores, as the paper demonstrates.

A secondary study explores the use of EEGs to measure human decision confidence levels [13]. A visual perceptual decision confidence experiment involves 14 participants performing a task while EEG data is recorded. The task involves showing blurred images of animals to the participants, while they must decide for the animal group, then report their confidence levels about the selection from one through five. The study uses two classifiers (Support vector machine with RBF kernel and a deep neural network ANN with shortcut connections to retain original information) trained with EEG features. Their results show EEG signals can assess decision confidence, achieving peak accuracy of 49.14% and F1-score of 45.07% when validating all five confidence levels, and 91.28% accuracy with an average F1-score of 88.92% for extreme confidence levels (confidence of 1 and 5, while ignoring confidence levels in between). Though model generality is important for real time applications such as flight where cases of confidence between 2 and 4 are quite common in real world applications, which may be problematic for model performance.

Another study claiming to be the first online study in real car decoding driver's error-related brain activity introduces an EEG-based BCI designed to decode error-related brain activity for use in driving assistance systems [14]. Conducted in both a car simulator ($N = 22$) and a real car ($N = 8$), participants received directional cues before approaching intersections. The study classified error-related potentials from EEG using Linear Discriminant Analysis as a supervised learning algorithm to predict whether the cued direction aligned with the driver's intention. Offline experiments achieved an average classification accuracy of 0.698 ± 0.065 in the simulator and 0.682 ± 0.059 in the real car, both significantly above chance level. Online tests showed equivalent performance in simulated and real car driving, supporting the BCI's feasibility for decoding signals and estimating driver intention in real-world driving scenarios.

3 Methods

3.1 Flanker Dataset

The COG-BCI dataset was used to train the machine learning models, as it is a standardized and highly regulated dataset containing flanker EEG and behavioral data [15]. The Flanker task is a choice reaction task designed to induce errors and conflict during binary decisions. In its arrowhead version, participants are presented with stimuli consisting of 5 horizontal arrows, where they must respond to the middle arrow while disregarding flanking arrows. Flanker stimuli can be congruent (flanking arrows point in the same direction) or incongruent (flanking arrows point in the opposite direction to the central arrow). Each trial involves a 2000 ms inter-stimulus interval (ISI) followed by a 16 ms display of the stimulus. Stimuli are presented equally frequently in a pseudo-random order, and participants respond by indicating the target direction with keyboard keys. Feedback about trial outcomes is provided, and the task involves 120 trials (30 for each stimulus type), lasting about 10 min. The structure of the test can be seen in Fig. 1. Participant responses, error rates, and reaction times are recorded throughout the task. Instructions are given before the run begins. The dataset was compiled over three sessions, each separated by one week.

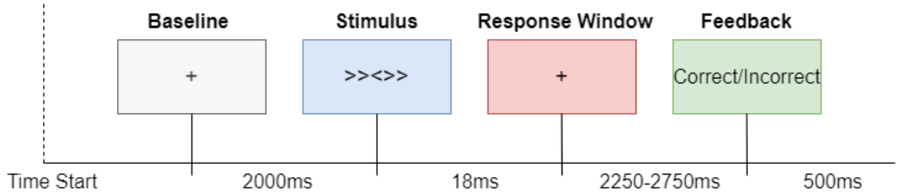


Fig. 1. Flanker test procedure and timing windows.

3.2 EEG Preprocessing and Dataset Creation

The EEG signals underwent a comprehensive preprocessing pipeline to enhance their quality and prepare them for subsequent analysis. Initially, we applied a finite impulse response filter (FIR) with an automatic filter length set at 1 Hz, leveraging the capabilities of the Python MNE toolbox. Following this, a notch filter at 50 Hz was employed to effectively eliminate any line noise that might have been present in the signals. Subsequently, we down sampled the frequency from the original 500 Hz to 250 Hz to optimize computational efficiency without sacrificing critical information. To address potential artifacts such as heartbeat and eye blinks, we performed Independent Component Analysis (ICA). An epoch, representing a time sequence of 1 s before and 0.5 s after the stimulus signal, was extracted for each trial. Every individual epoch underwent visual inspection, and any instances with lingering large amplitude artifacts were removed. Setting a z-score threshold of 5 during the rejection procedure was a decision aimed at maximizing the number of samples and throwing away egregiously bad EEG spikes without being too critical. The dataset creation overview can be seen below in Fig. 2.

After completing the rejection procedure, our dataset comprised of 1153 erroneous samples and 8521 non-erroneous samples, ensuring a robust and refined dataset for further analysis and model development. It should be noted that training and validation sets were created with no participant leakage (Single participant data would not be found in both training and validation set).

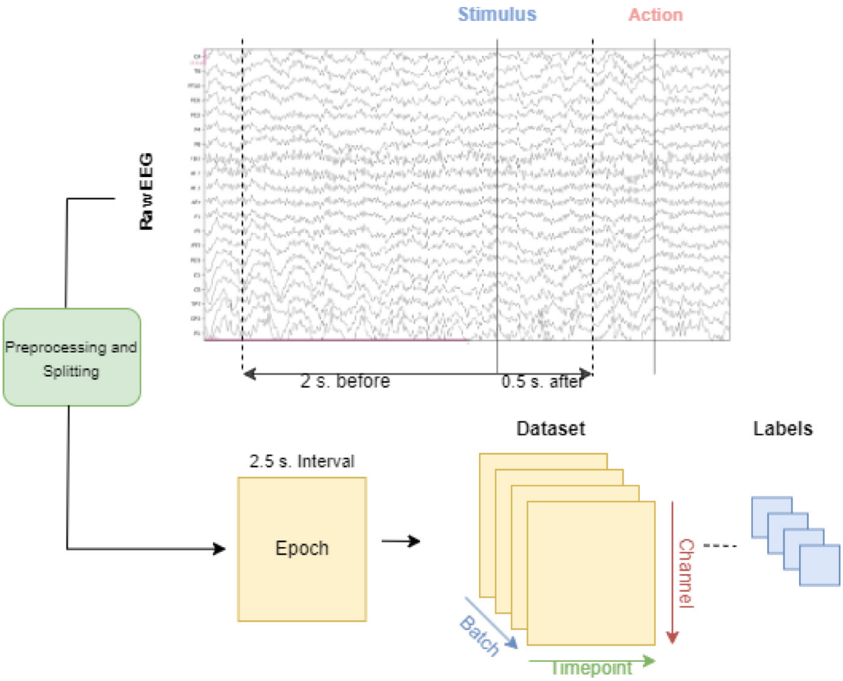


Fig. 2. EEG processing pipeline to create labeled data.

3.3 Adjusting Imbalanced Datasets

Imbalanced datasets refer to situations where the distribution of classes is uneven, with one or more classes having significantly fewer instances than others. In our COG-BCI dataset, the occurrence of erroneous data points is significantly lower compared to the correct ones. Addressing imbalanced datasets is crucial as it can lead to biased model training and poor generalization. Given the limited size of the dataset, undersampling is not a feasible choice. As a result, we explored two oversampling techniques: Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-sampler, with the latter demonstrating superior performance.

3.4 Models

Several machine learning models were used on the dataset to compare performance. Hyperparameter tuning techniques such as Cross Validation Grid Search from SKLearn and open-source library Optuna were used to find optimal configurations for each model.

SVM and Random Forest Models. These models are supervised machine learning models described as classification algorithms. They are both relatively easy to implement and have been proven to provide great results on EEG data [16, 17]. These models serve as a great starting point in terms of identifying trends in our EEG dataset.

Support Vector Machines (SVM) work by finding the optimal hyperplane that maximally separates data points of different classes in a high-dimensional space. They can handle both linear and non-linear data depending on the kernel function selected. We decided to use an RBF kernel function, which is commonly used with EEG data. The parameters chosen for the SVM model can be found in Table 1.

Table 1. SVM Parameters.

Parameter	Value
C	1.0
Kernel	RBF
Gamma	Scale
Probability	True
Shrinking	True
Tol	0.001

Random Forest models operate by constructing multiple decision trees during training and outputting the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree is built using a random subset of the training data

Table 2. Random Forests parameters

Parameter	Value
Bootstrap	True
Ccp_alpha	0.0
Criterion	Gini
Max_depth	50
Max_features	0.7
Min_samples_leaf	1
Min_samples_split	10
N_estimators	100

and a random subset of features, reducing overfitting and improving generalization. The combination of diverse trees results in a robust and accurate model. The parameters chosen for the Random Forests can be found in Table 2.

Double Convolution. Convolutional Neural Networks (CNNs) are a type of deep learning architecture, specifically designed for tasks involving structured grid-like data, such as images. To process the dataset through CNNs we follow similar architecture to Batmanova et al. with a few modifications by performing the following steps (also illustrated in Fig. 3):

1. We first take our Channel by Time-point matrix from our batch and perform a 1-dimensional convolution across the channel axis. This convolution slides along the dimension of the EEG channels, while keeping the time-point dimension fixed, performing an averaging of EEG amplitude at every moment.
2. A second 1 dimensional convolution is performed across the remaining 1-dimensional time dimension (where each point is an average across all channels) resulting in a 1-dimensional vector with k features. This value k is then treated as a hyperparameter that can be tuned to improve model performance. Naturally, with too few k values, the model will not efficiently learn, while with too many k values, the model will have redundant features.
3. We then perform a 1-dimensional batch normalization to improve training stability and convergence speed. The batch normalization is also useful to mitigate the vanishing and exploding gradient problems during backpropagation.
4. A dropout layer is added with a dropout rate of 0.3. This regularization technique helps to break tight coupling between neurons, reducing the risk of overfitting and improving the model's ability to generalize to unseen data.
5. Finally, two fully connected layers with hidden dimensions of size 1024 is used to perform the classification. A leaky-Relu activation is performed in between these two layers with a sigmoid activation function for the binary classification output.

Double Convolution Hyperparameters. We experimented with Xavier Uniform and Random Uniform initializers, finding that the former yielded superior outcomes. Employing a learning rate of 0.00671, we conducted training over 27 epochs with a batch size of 314. Leaky Relu activation functions were utilized, demonstrating the most favorable results. Additionally, we utilized Adaptive Gradients (Adagrad) optimizers and found that a value of $K = 79$ produced the optimal results.

Transformer. Transformers are a type of deep learning architecture that has gained significant popularity, especially in natural language processing tasks. Transformers are known for their self-attention mechanism, allowing them to capture relationships between different elements in a sequence, making them particularly powerful for handling long-range dependencies. The transformer architecture was initially proposed for natural language processing but has been successfully applied to various other domains, including computer vision, speech processing, and EEG [18]. The architecture that will be used for this task was inspired by the Conformer architecture conceived by Song et al. The architecture, can be split up into three main parts as follows:

- Part 1 - Convolution Module: The convolution module is designed by decomposing the two-dimensional convolution operator into two one-dimensional layers, separating temporal and spatial convolutions, much like the double convolution used previously. The first layer employs k kernels performing convolutions along the time dimension. The second layer consists of k kernels acting as a spatial filter to capture interactions between electrode channels. Batch normalization is applied for enhanced training and reduced overfitting, while exponential linear units (ELUs) serve as the activation function for nonlinearity. The third layer performs average pooling along the time dimension, smoothing temporal features to prevent overfitting and decrease computational complexity. The resulting feature maps are rearranged, with the electrode channel dimension squeezed and the convolution channel dimension transposed, enabling the feeding of all feature channels of each temporal point as tokens into the next module.
- Part 2 - Self-Attention Module: The Self-Attention Module is introduced to capture global temporal dependencies in EEG features, utilizing self-attention to enhance the decoding of context-dependent representations within low-level temporal-spatial features. Tokens from the previous module are linearly transformed into query (Q), key (K), and value (V), and their correlations are evaluated using dot product, with a scaling factor to prevent vanishing gradients. The resulting attention score is obtained through a Softmax function and applied to V with a dot product. This process is repeated N times in the self-attention module, incorporating a multi-head strategy to

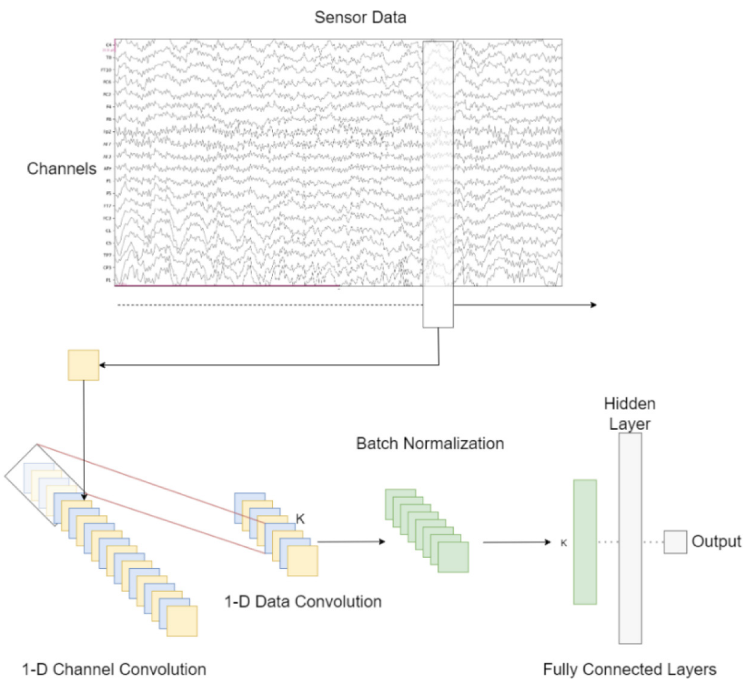


Fig. 3. Illustration of the double convolution neural network structure from raw EEG data.

enhance representation diversity. The multi-head attention results are concatenated, and the entire attention computation is performed N times.

- Part 3 - Classifier Module: Finally, the Classifier Module employs two fully connected layers to output an M-dimensional vector after a Softmax function, using binary cross-entropy as the loss function for the entire framework.

As a result, we have a model that processes EEG data through temporal and spatial convolution layers, arranges them into tokens with a pooling layer, applies self-attention layers, and uses fully connected layers for classification results.

Transformer Hyperparameters. We utilized Xavier Uniform initializers, a learning rate of 0.0072, and conducted training for 30 epochs with a batch size of 75. Employing a CyclicLR optimizer scheduler cycling between 0.001 and 0.01 yielded the most optimal results. Additionally, we fine-tuned the Adam optimizer, achieving optimal outcomes with beta-1 set to 0.285 and beta-2 set to 0.927. Exploring values of K between 10 and 100, we found that K = 20 yielded the best results.

4 Experiments

Experiments were conducted to assess real-time cognitive activity in pilots during takeoff in an Airbus A320 simulator, employing the previously outlined methodology. The participants, consisting of seven pilots, including five A320 pilots, and six engineers with expertise in aircraft maneuvers, were divided into two groups experiencing different takeoff scenarios to eliminate learning bias. The participants had to release the parking brake, do a takeoff procedure, and climb until 3000 ft without using autopilot. Cognitive workload (CW), heart rate (HR), and pupil diameter (PD) were measured using an EEG headset, Polar H10 heart rate monitor strap, and Gazepoint GP3 eye tracker. A total of 136 takeoffs were performed by the 13 participants, who had an average age of 36 years, 604 flight hours, and 8.5 years of piloting experience [19].

5 Results

The results from the analysis of the Flanker dataset using the models are presented in Table 3. It should be noted that accuracy values are misleading to the dataset imbalance. With an error rate close to the 13% mark, a model that has learned to simply pick the most abundant class without actually classifying between the two will achieve a high accuracy. For this reason, we have accumulated Macro-Averaged Precision, Recall, and F1 scores – which essentially factor in the class imbalance and give us a better understanding in the model's performance. A Macro-Averaged F1 score above 0.6 would indicate a classifier with moderate to good performance, while above 0.8 indicating excellent performance. Additionally, we have included the area under the receiver operating characteristic (AUC ROC), which serves as a metric that quantifies the overall performance of a binary classification model by measuring the area under the curve formed by plotting the true positive rate against the false positive rate across various classification thresholds. A value of 0.5 for the AUC ROC is indicative of a random classification with no discriminative power, with values closer to 1 presenting a strong discriminative power.

We can see that the simpler machine learning algorithms, namely SVM and Random Forests, receive high accuracy values but underperform in all other metrics. Their classification often seems random or biased towards one class, indicating that the complexities in the EEG data might not be suited for these particular classifiers. The double convolution model begins to discern between the two classes; however, its performance still remains quite low and contains a lot of false negatives. The transformer model manages to outperform the other models by a significant margin. We see a reduction in the number of false negatives and a final F1 of 0.610, which serves as promising signs of its classification power for erroneous states. It is interesting to note that a channel reduction from 62 to 16 exhibited no loss of information and increased the performance of the classifier. The 16 chosen electrodes followed the standard 10–20 system distribution and not simply a random subset of the original 62 channels.

Table 3. Results of the different machine learning models on the Flanker dataset

DATASET	MODEL	ACC.(%)	AUC ROC	*PRECISION	*RECALL	*F1
FLANKER	SVM (Linear)	70.86	0.559	0.519	0.533	0.512
	SVM (RBF)	80.48	0.523	0.493	0.493	0.491
	Random Forests	**86.30**	0.484	0.439	0.500	0.468
	Double Convolution	83.88	0.591	0.583	0.555	0.563
	Conformer (62 Chanel)	84.08	0.633	0.599	0.600	0.600
	Conformer (16 Chanel)	82.40	**0.652**	**0.611**	**0.610**	**0.610**

While many studies conclude their experiments at this juncture, our objective extends beyond, aiming to assess the model's applicability in a secondary domain - substantiating the transferability of the classifiers. To achieve this, we have collected EEG data from aviation pilots and intend to evaluate the classifiers' discriminative power specifically during takeoff procedures. We can examine the performance of these models in Table 4, where we can see that the performance reduces slightly, but performs relatively as expected.

Table 4. Results of the different machine learning models on the Pilot dataset

DATASET	MODEL	ACC.(%)	AUC ROC	*PRECISION	*RECALL	*F1
FLANKER	SVM (Linear)	78.81	0.499	0.498	0.499	0.499
	SVM (RBF)	82.68	0.486	0.423	0.486	0.452
	Random Forests	**88.42**	0.498	0.443	0.498	0.469
	Double Convolution	77.38	0.532	0.537	0.532	0.534
	Conformer (62 Chanel)	76.55	0.567	0.560	0.567	0.565
	Conformer (16 Chanel)	77.06	**0.584**	**0.574**	**0.584**	**0.578**

The transformer model still outperforms the other models in terms of classification with a macro-averaged F1 above random classification, which is a promising indicator. Although there was a decrease in F1 score, these results highlight the existence of electrical brain activity trends that could foreshadow a decrease in behavioral performance. The flanker dataset that was used was quite small in contrast to common transformer training datasets, yet the transformer was still able to receive an adequate score. With an increase in training data, we believe that there could be a substantial increase in classifier performance.

Furthermore, our polling began 1 s prior to the stimulus due to constraints in Flanker testing procedures. We also believe that increasing the polling band prior to the stimulus will also increase the classifier performance, and transferability.

The transformer model works exceptionally well in real time, which would be its main use case when dealing with BCI devices. It has an inference speed of 0.01 s, resulting in the bulk of computation time being associated to input processing.

6 Conclusion

With such small margins for error in real-time systems, such as aviation, the need for predictive error detention is crucial. In this study, we have demonstrated the ability of different machine learning models to predict erroneous states within a perceptual-decision-making flanker and piloting task, using EEG. These signals were recorded prior to the behavioral response and achieved a maximum macro averaged F1 score of 0.601 with a transformer-based model. Furthermore, testing the transferability of this model to an aircraft piloting task yielded promising results with a maximum F1 score of 0.578, suggesting that the signals in behavioral responses may be similar across certain domains and can be used to substitute the gaps in data required – consequently accelerating the creation of error prevention systems that can revolutionize transportation safety. While this study highlights the potential to predict errors, further research and experimentation with larger datasets is essential to construct a more robust machine learning model for this task.

Acknowledgments. We acknowledge the support from NSERC- Alliance, CRIAQ, CAE, Bombardier, and BMU for funding this work.

References

1. Helmreich, R.L.: On error management: lessons from aviation. BMJ **320**(7237), 781–785 (2000). https://doi.org/10.1136/bmj.320.7237.781.PMID:10720367;PMCID:PMC1117774
2. Pietracupa, M., Ben Abdessalem, H., Frasson, C.: An approach to automatic flight deviation detection. In: Frasson, C., Mylonas, P., Troussas, C. (eds.) Augmented Intelligence and Intelligent Tutoring Systems. ITS 2023. Lecture Notes in Computer Science, vol. 13891. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-32883-1_47
3. BBC News: No survivors as Pakistan plane crash kills 152. BBC News (2010)
4. Lamb, K. (2015, December 13). AirAsia crash: crew lost control of plane after apparent misunderstanding. The Guardian
5. Cloudberg, A.: Levers of Power: The crash of Yeti Airlines flight 691. Medium (2024)
6. Hramov, A.E., Maksimenko, V.A., Pisarchik, A.N.: Physical principles of brain–computer interfaces and their applications for rehabilitation, robotics and control of human brain states. Phys. Rep. **918**, 1–133 (2021). https://doi.org/10.1016/j.physrep.2021.03.002
7. Koenig, T., Studer, D., Hubl, D., Melie, L., Strik, W.K.: Brain connectivity at different time-scales measured with EEG. Philos. Trans. R. Soc. B Biol. Sci. **360**(1457), 1015–1024 (2005). https://doi.org/10.1098/rstb.2005.1649
8. Martini, M.L., Oermann, E.K., Opie, N.L., Panov, F., Oxley, T., Yaeger, K.: Sensor modalities for brain-computer interface technology: a comprehensive literature review. Neurosurgery **86**(2), E108–E117 (2020). https://doi.org/10.1093/neuros/nyz286
9. Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., Cohen, J.D.: Anterior cingulate cortex, error detection, and the online monitoring of performance. Science **280**(5364), 747–749 (1998). https://doi.org/10.1126/science.280.5364.747. PMID: 9563953
10. Holroyd, C.B., Coles, M.G.H.: The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. **109**(4), 679–709 (2002). https://doi.org/10.1037/0033-295X.109.4.679. PMID: 12374324
11. Ridderinkhof, K.R., Nieuwenhuis, S., Bashore, T.R.: Errors are foreshadowed in brain potentials associated with action monitoring in cingulate cortex in humans. Neurosci. Lett. **348**(1), 1–4 (2003). https://doi.org/10.1016/s0304-3940(03)00566-4. PMID: 12893411
12. Batmanova, A., et al.: Predicting perceptual decision-making errors using EEG and machine learning. Mathematics. **10**(17), 3153 (2022)
13. Li, R., Liu, L.D., Lu, B.L.: Discrimination of decision confidence levels from EEG signals. In: 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), Italy, pp. 946–949 (2021). https://doi.org/10.1109/NER49283.2021.9441086
14. Zhang, H., Chavarriaga, R., Khaliliardali, Z., Gheorghe, L., Iturrate, I., Millán, J.D.R.: EEG-based decoding of error-related brain activity in a real-world driving task. J. Neural Eng. **12**(6), 066028 (2015). https://doi.org/10.1088/1741-2560/12/6/066028
15. Hinss, M.F., Jahanpour, E.S., Somon, B., et al.: Open multi-session and multi-task EEG cognitive dataset for passive brain-computer interface applications. Sci Data **10**, 85 (2023)

16. Richhariya, B., Tanveer, M.: EEG signal classification using universum support vector machine. Expert Syst. Appl. **106**, 169–182 (2018). https://doi.org/10.1016/j.eswa.2018.03.053
17. Edla, D.R., Mangalorekar, K., Dhavalikar, G., Dodia, S.: Classification of EEG data for human mental state analysis using random forest classifier. Procedia Comput. Sci. **132**, 1523–1532 (2018). https://doi.org/10.1016/j.procs.2018.05.116
18. Song, Y., Zheng, Q., Liu, B., Gao, X.: EEG conformer: convolutional transformer for EEG decoding and visualization. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 710–719 (2023). https://doi.org/10.1109/TNSRE.2022.3230250
19. Antoine, M., Abdessalem, H., Frasson, C.: Cognitive workload assessment of aircraft pilots. J. Behav. Brain Sci. **12**, 474–484 (2022). https://doi.org/10.4236/jbbs.2022.1210027

Mining Discriminative Sequential Patterns of Self-regulated Learners

Amine Boulahmel[1(✉)], Fahima Djelil[1], Jean-Marie Gilliot[1], Philippe Leray[2], and Gregory Smits[1]

[1] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France
{amine.boulahmel,fahima.djelil,jm.gilliot,gregory.smits}@imt-atlantique.fr

[2] Nantes Université, CNRS, LS2N UMR 6004, 44000 Nantes, France
philippe.leray@univ-nantes.fr

Abstract. This research explores the links between self-regulation behaviors and indicators of learning performance. A data mining approach coupled with appropriate qualitative measures is proposed to extract behavioral sequences that are representative of learning success. Applied on an online programming platform, obtained results allowed to highlight important self-regulation behaviors during the planning and engagement phases. It e.g. appears that successful self-regulated learners are those who analyze their tasks before working on them. This work brings methodological contributions in the field of self-regulation learning measurement and is a first step towards the design of intelligent tutoring systems.

Keywords: Self-Regulated Learning · Discriminative Sequential Patterns Mining · Learning Performance · Programming Tasks · Learning Analytics

1 Introduction

Although Self-Regulated Learning (SRL) has a long existence in the education literature [17], it is still a key challenge in modern Online Learning Environments (OLE) [6]. The measurement of SRL has gained increased research interest aimed at a more refined understanding of SRL behaviors and strategies [7].

OLEs make it possible to capture and store students' interactions with the platform as trace data, on which fine-grained analyses can be applied to *in fine* provide learners with accurate feedback on their self-regulation skills.

In this paper, the relationship between SRL behaviors and learner performance is studied. It is assumed that learning sessions are discretized in pedagogical activities ending with a success or failure indicator of performance. The objective of this work is to discover discriminative behavioral sequential patterns that more likely lead to learning success.

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 137–149, 2024.
https://doi.org/10.1007/978-3-031-63031-6_12

The following main question is addressed: How to leverage trace data to point out the link between learners' SRL behavior and learning performance? Answering this question is a first step towards the conception of an Intelligent Tutoring System (ITS) to support learning. Provided contributions are: 1) a data mining pipeline including appropriate measures to extract sequential patterns of actions that are characteristic of learning performance, and 2) an implementation on real data collected from an OLE, demonstrating the relevance of this approach.

The paper is structured as follows. Section 2 gives the principles of the proposed approach. Section 3 describes a positioning wrt. existing works in SRL measurement, from one side, and discriminative sequential pattern mining, from the other side. The proposed sequence mining method is detailed in Sect. 4, its implementation and data collection are described in Sect. 5. Results are discussed in Sect. 6, conclusions and implications are drawn in Sect. 7.

2 Principle and Notations

2.1 Assumptions

Our main objective is to quantify the relevance of sequences of action sequences wrt. learning performance.

Collected sequences of actions end up with an indicator of performance which is a success or a failure in an exercise. Sequences can thus be divided into two sets, those that lead to success (positive set) and oppositely those that lead to failure (negative set). The more a sequence of actions is observed in the positive set and the less in the negative set, the more characteristic of success is the sequence. A sequence is also all the more characteristic of success, in that its actions have to appear in a specific order.

Moreover, this work relies on the SRL model of Siadaty et al. [13] that suggests three common phases of learning behaviors, from planning through engagement, to self-reflection and evaluation. It maps trace data to micro-level SRL processes that are already categorized under macro-level processes. In our case, phases that can be observed from the data are planning and engagement. Reflection and evaluation are not instrumented in the concerned OLE. Planning through *task analysis* includes action events such as clicking on exercise assignments and documentation. Engagement through *working on the task* and *applying strategy changes*, include respectively action events such as keystrokes, click-streams, mouse movements and reviewing activities.

Finally, the extent to which a sequence is characteristic of success depends on its **discrimination degree** with the set of negative sequences, but also on the **order dependency of its actions**. Although the two phases of SRL planning and engagement can be observed and micro-level processes that affect learning performance can be examined from trace data, it is very likely that there is no unique behavior, reflected in a sequence of actions, that characterizes the performance of self-regulated learners.

Behavioral sequences must therefore be analyzed at different levels: platform, exercise, and learner. Defining the two sets of positive and negative sequences is crucial to the analysis.

2.2 Preliminaries

From trace data, sequences of SRL actions performed by each learner can be reconstructed. A set of task actions is denoted by $\mathcal{A}$, and a sequence, often denoted by s hereafter, is an ordered list of actions $s = \langle a_1, \ldots a_{|s|} \rangle$ where $|s|$ is the length of s and $a_i, i = 1..|s| \in \mathcal{A}$. A sequence s' is a sub-sequence of another sequence s, denoted by $s' \subseteq s$, if it exists a one-to-one mapping from s' to s that preserves the order of s and s'. One denotes by $supp(s, \mathcal{S})$ the support of the sequence s in a set $\mathcal{S}$ of sequences and is computed as follows:

$$supp(s, \mathcal{S}) = \frac{|\{s' \in \mathcal{S}, s \subseteq s'\}|}{|\mathcal{S}|} \tag{1}$$

Definition 1. *A sequence s is said to be* **frequent** *if $supp(s, \mathcal{S}) \geq \alpha$, α being a predefined frequency threshold.*

Sequences that are characteristic of learning success are identified through a differential analysis between two sets of sequences denoted $\mathcal{S}^+$ and $\mathcal{S}^-$ that gather the sequences leading to success and failure respectively. In addition to being frequent, sequences of interest in a differential analysis have to be discriminative of the positive set $\mathcal{S}^+$ compared to the negative one $\mathcal{S}^-$, which means that the sequence has to cover a larger ratio of $\mathcal{S}^+$ than $\mathcal{S}^-$.

3 Related Works

3.1 Sequence Mining for SRL Measurement

The use of sequence mining to measure SRL has emerged as a promising method, as it appears in some recent works. It is mainly applied to identify SRL behaviors and student profiles, as for [14] to reveal learning sequences on trace coded behaviors and strategies, and to analyze behavior patterns and how they affect the course performance [4]. Describing students' SRL patterns is an important analysis task performed with sequence mining. For example, it allows to explore coded sequences of learning actions to investigate the effects of personalized scaffolding on students' learning activities [11]. Similarly, it enables the exploration of behavioral patterns of SRL and the examination of the significance of self-regulation scales among different groups of students [15], as well as the exploration of sequences of learners' activities to understand how learners utilize the SRL supports [16]. To our knowledge, no work has focused on the mining of discriminative sequential patterns in the measurement of SRL behaviors. In this work, we focus on the most relevant qualitative measures to identify sequential patterns expressing SRL behaviors that affect learners' performance.

3.2 Discriminative Sequential Pattern Mining

A key constraint that a sequential pattern must satisfy to be considered interesting is related to its frequency of occurrence in the analyzed dataset [9].

When quantifying the informativeness of a pattern in a differential setting materialized by two subsets of sequences, an additional criterion to its frequency is often considered. A pattern is indeed interesting if it is both frequent in its subset of assignment and discriminative wrt. other subsets. In [3], the quantification of the discriminative degree of a given sequence is studied in a general setting when multiple classes of sequences exist. But, in most cases, sequences are split into two sets and a sequence is all the more discriminative as it frequently occurs in its set and not in the other one.

A common way to quantify the extent to which a sequence s is characteristic of a set $\mathcal{S}^+$ wrt. another one $\mathcal{S}^-$ is to calculate the Support Difference (SD) [12]:

$$SD(s) = supp(s, \mathcal{S}^+) - supp(s, \mathcal{S}^-) \tag{2}$$

Compared to other existing measures of discrimination (see [12] for more details), the SD measure (Eq. 2) ranges in $[-1, 1]$, thus making its output easier to interpret. Another important property to consider when providing users with the most discriminative sequences is to check that the discriminative degree of a sequence s is not due to any of its sub-sequences, which would lead to redundancy in the provided list of discriminative sequences [10].

4 Mining Successful Behavioral Sequences

The first step is to determine the properties, and subsequently the appropriate measures, of the sequences of behavioral actions that are characteristic of learning success. Candidate sequences to characterize a learning success are obviously taken from $\mathcal{S}^+$ and have first to be frequent, this set of frequent sequences is denoted by $\mathcal{S}_F^+$ and built as follows:

$$\mathcal{S}_F^+ = \{s \in \mathcal{S}^+, \text{ st. } supp(s, \mathcal{S}^+) \geq \alpha\} \tag{3}$$

where α is a predefined frequency threshold that can be adjusted to control the number of candidate sequences to take into account in the next steps. To be characteristic of success, a sequence also has to be differential from sequences that lead to failure. This is where the measure SD (Eq. 2) comes into play.

Definition 2. *A sequence s is* **discriminative** *of $\mathcal{S}^+$ wrt. $\mathcal{S}^-$ if $SD(s) > 0$.*

A sequence s may be discriminative but also redundant, as s may contain a sub-sequence that is itself discriminative. For instance, let us consider a frequent and discriminative sequence $s \in \mathcal{S}_F^+ : \langle a_1, a_2, a_3 \rangle$ and one of its sub-sequences $s' : \langle a_1, a_2 \rangle$ that is necessarily frequent too but also more discriminative $(SD(s') > SD(s))$. Then, it is useless to present the user both s and s' as s' is by itself frequent and discriminative. The measure $\overline{red}(s, \mathcal{S}_F^+)$ quantifies the redundancy of a sequence s according to its discriminative sub-sequences:

$$\overline{red}(s) = SD(s) - \min(SD(s), \max_{s' \subset s} SD(s')). \tag{4}$$

The measure $\overline{red}(s)$ is a revision of the measure introduced in [10], so as to obtain a non-redundancy degree normalized in the unit interval. The sequence s being compared to its discriminative sub-sequences, then $SD(s) \in]0, 1]$ and $\max_{s' \subset s} SD(s') \in]0, 1]$.

The measures of discrimination and non-redundancy are classically used in existing approaches to extract discriminative sequential patterns. In order to identify sequences of SRL actions characterizing success, an additional property has to be considered. A sequence of actions constitutes an SRL behavior if the order in which the actions appear in the sequence is important. Indeed, in the SRL model of learning [13], planning behaviors precede engagement behaviors, otherwise, it is the set of actions that matters and not the sequence itself. Thus, we quantify the extent to which the order in which the actions appear in a frequent sequence is decisive. To do so an order dependence measure, denoted $ordDep(s)$, is introduced that compares the discriminative degree of s with its permutations. It is based on the entropy measure of data series [1]:

$$ordDep(s) = SD(s) - \min(SD(s), \max_{s' \in \mathfrak{s}(s)} SD(s')), \tag{5}$$

where $\mathfrak{s}(s)$ denotes the set of permutations of s.

Definition 3. *A sequence expresses an SRL behavior that is characteristic of the learner's performance iff. it is both* **frequent***,* **discriminative***,* **not redundant** *and contains* **order-dependent actions**. *The overall qualitative degree attached to each sequence s to express the extent to which it constitutes an SRL behavior of $\mathcal{S}^+$ is denoted by $\mu(s)$ and computed as follows:*

$$\mu(s) = \min(SD(s), \overline{red}(s), ordDep(s)). \tag{6}$$

Lemma 1. *Based on the discriminative degree computed using SD, sequences having a strictly positive μ degree are the most characteristic sequences of $\mathcal{S}^+$.*

Proof. It is straightforward to show that a sequence s having a strictly positive $\mu(s)$ is a representative sequence of $\mathcal{S}^+$. As $s \in \mathcal{S}_F^+$ then s is frequent. Let $s' \subseteq s$ be a frequent sub-sequence (resp. permutation) of s. If s' is more discriminative than s then $\min(SD(s), SD(s')) = SD(s)$ and $\overline{red}(s) = 0$ (resp. $ordDep(s) = 0$) leading to $\mu(s) = 0$.

In summary, the first step of our approach leverages a sequential pattern mining algorithm [8] to extract all the patterns in $\mathcal{S}^+$ that occur frequently. Then, to keep discriminative patterns only, the support of the patterns that frequently occur in $\mathcal{S}^+$ has to be computed on $\mathcal{S}^-$. Moreover, using only the set of somewhat discriminative patterns (i.e. $SD > 0$), redundant and non order-dependant sequences are discarded. Especially to check the non-redundancy with their sub-sequences, discriminative sequences are processed in an increasing order of the size.

Table 1. Learning actions library for Quick-Pi trace data.

Learning Action	Code	Description
READ_TASK	RDT	The learner reads the task set by the exercise
NAVIGATION	NAV	The learner navigates through platform exercises
PROGRAMMING	PRG	The learner is programming, therefore attempting to solve the exercise
SUBMISSION_FAIL	SFL	The learner submits a code but the solution is invalid
SUBMISSION_SUCCESS	SSC	The learner submits a code and passes
CODE_DEBUG	CDG	The learner is debugging a code
CODE_TEST	CDT	The learner experiments with code before submitting
HOVER_TASK	HTK	The learner hovers over the task panel

5 Implementation and Data Collection

As an implementation of the proposed approach, a study on real data is conducted addressing the following questions: 1) What are the SRL behaviors that impact learner performance? 2) Is it meaningful to search for these behaviors at a *holistic* platform level or at a more individual level, i.e. learning task or learner level?

Data are collected from the platform Quick-Pi[1], that provides learners with pedagogical content for learning programming along with a range of activities related to the Internet of Things for school or home use. The recorded data consists of timestamped raw traces describing the low-level interactions of the learner with the platform interface (i.e. clickstream data). The data are first cleaned: 1) Noisy captures correction, to correct noisy events captured by the platform, 2) Event overlap correction, to correct the time-stamped sampling errors, and 3) Learning actions fusion, to merge successive chunks of the same learning actions into single blocks. Then, raw traces are translated into interpretable learning actions. Events on the use of the platform are aggregated into interpreted learning actions leading to a learning action library (Table 1).

The participants are a total of N = 506 learners who connected online to the platform, within a period of 40 d. Data privacy is ensured as the platform

[1] https://quick-pi.org/.

does not collect any demographic or personal information about learners. All 8 exercises of the first module of Quick-Pi were attempted by the learners.

The data show that the number of attempters and the number of successful attempters tend to decrease as the difficulty of the exercise increases.

The dataset has been segmented into two sets, one containing sequential learning actions leading to a success $\mathcal{S}^+$ (sequences ending with SUBMISSION_SUCCESS), and the other containing sequential learning actions leading to failure $\mathcal{S}^-$ (sequences ending with SUBMISSION_FAIL).

At the platform level, all the observed data sequences are split into two groups ($|\mathcal{S}^+| = 2139$ and $|\mathcal{S}^-| = 12927$). At the exercise level (24 exercises available), there are averages of 89.12 positive sequences and 538.62 negative sequences. At the learner level (506 learners), the average number of positive (resp. negative) sequences per learner is 4.22 (resp. 25.54). The prefixspan[2] algorithm has been used to mine frequent patterns, as it has the advantage of reducing processing time and memory compared with other apriori-like algorithms.

Table 2. Statistics on the frequent sequences computed at platform, exercise, and learner levels using SD, $\overline{red}$, ordDep and μ.

	Platform				Exercise				Learner			
	min	max	avg	std	min	max	avg	std	min	max	avg	std
SD	−0.31	0.006	−0.10	0.09	−0.38	0.008	−0.23	0.03	−0.2	0.05	−0.10	0.03
$\overline{red}$	−0.10	0.06	−0.0004	0.02	−0.27	0.03	−0.001	0.02	−0.2	0.02	0.0003	0.01
ordDep	−0.27	0.02	−0.002	0.009	−0.27	0.02	−0.002	0.02	−0.2	0.002	−0.001	0.01
μ	−0.31	0.06	−0.10	0.09	−0.41	−0.01	−0.23	0.03	−0.25	−0.04	−0.16	0.01

6 Results and Discussion

The discriminative sequential patterns mining at the three levels, platform, exercise, and learner, revealed that there are more frequent patterns that are characteristic of failure than of success (negative values of *min* and *avg SD*, and very low values of *max SD*) (see Table 2). This can be explained by the nature of the data collected from the Quick-Pi platform, which is accessible online to a very wide audience and used in contexts that are not only school-based but also outside the classroom. These data therefore do not necessarily come from teacher-supervised sessions where self-regulation is supported.

6.1 SRL Behaviors that Impact the Learner Performance

Platform Level. Based on the discriminative power computed using SD, $\overline{red}$, *ordDep*, sequences having a strictly positive μ value are the two atomic sequences

[2] https://pypi.org/project/prefixspan/.

$READ\_TASK$ ($SD = \overline{red} =$ ordDep $= \mu = 0.69$), and $CODE\_DEBUG$ ($SD = \overline{red} =$ OrdDep $= \mu = 0.08$). This suggests that $READ\_TASK$ is the most prevalent action within the set $\mathcal{S}^+$, and thus the most indicative of success across the platform. This indicates the importance of task analysis (refereed by reading the task) within the SRL planning phase for learner performance. Although $CODE\_DEBUG$ has low discriminative power compared to $READ\_TASK$, it appears to be important for success in the programming tasks. Debugging code allows errors to be identified and helps the code to run well. For learners, this is very characteristic of reviewing tasks during their SRL engagement with the programming tasks. This second result shows the importance of reviewing tasks for learner performance.

Table 3. Discriminative sequential patterns at the exercise level ($SD > 0.1$).

Exercise Code	Difficulty	Sequential pattern	Length	SD
MEL	Medium	PRG → RDT → (PRG → HTK)^3→ CDG	9	0.33
	Hard	(PRG → HTK)^3 → CDG	7	0.23
INS	Easy	RDT → PRG → HTK → CDG	4	0.22
	Medium	PRG → HTK → CDG → RDT	4	0.40
	Hard	(PRG → HTK)^3 → PRG → RDT → (PRG → HTK)^2→ PRG → CDG	14	0.26
AVT	Easy	(HTK → PRG)^2→ RDT → CDG	6	0.36
	Medium	(PRG → HTK)^3→ PRG → CDG	8	0.44
	Hard	RDT → PRG → NAV → PRG → HTK → PRG → RDT → NAV → PRG → (RDT → PRG → (HTK → PRG)^n)^m → CDG	9+m $(2 + 2n)$	0.16
SRV	Easy	RDT → PRG → HTK →(PRG → RDT)^2 → CDG	8	0.47
	Medium	RDT → (PRG → HTK)^2 → CDG	6	0.21
	Hard	RDT → CDG → RDT	3	0.23

Exercise Level. A total of 45 sequences appear to be representative of success ($\mu(s) > 0$). Table 3 lists the most discriminative sequences (based on $SD > 0.1$ although $\overline{red} = ordDep = \mu = 0$). These most discriminative sequences begin with the read task action (RDT), implying for learners a conscientious effort to ensure alignment with the task instructions, which is consistent with subsequent success (Table 3). Some sequences begin with the programming action (PRG) followed by reading (RDT) or hovering (HTK) the task. These behaviors are characteristic of task analysis in the SRL planning phase. In addition, these sequences end with either code debugging (CDB) or occasional reading of the task (RDT). These results indicate that during learners' SRL engagement phase, they typically review their code at the end of the task to ensure that it works correctly. Moreover, many exercises exhibit a learning strategy where learners iteratively program and check the completion of the task, resulting in sequence loops $(PRG \rightarrow HTK)^k$, where $k \in \mathbb{N}^*$ is the loop length. For instance, for the hard version of the exercise AVT, an extended sequence with a nested loop was observed $s =$(RDT $\rightarrow$ PRG $\rightarrow$ (HTK $\rightarrow$ PRG)^n)^m, where $n, m \in \mathbb{N}^*$. This suggests that learners consistently apply strategy changes by reviewing their tasks during their SRL engagement phase, which may lead to prolonged task completion. Indeed, the length of the pattern in this case indicates a greater effort in problem-solving due to the difficulty of the exercise. For the exercise SRV, no loops were found, regardless of the difficulty of its versions, due to the limited number of learners who successfully completed it.

Table 4. Most discriminative sequential patterns characteristic of success at the learner level ($\mu > 0$).

Learner	Sequential pattern	Length	SD	$\overline{red}$	$ordDep$	μ
L1	(PRG $\rightarrow$ HTK)^2 $\rightarrow$ (PRG $\rightarrow$ HTK)^2 $\rightarrow$ PRG	9	0.79	0.03	0.005	0.005
L2	(PRG $\rightarrow$ HTK)^2 $\rightarrow$ PRG $\rightarrow$ RDT	6	0.85	0.03	0.05	0.05
L3	(HTK $\rightarrow$ PRG)^4 $\rightarrow$ RDT	9	0.90	0.10	0.005	0.005
L4	RDT $\rightarrow$ (PRG $\rightarrow$ HTK)^3	8	0.85	0.02	0.003	0.003
L5	RDT $\rightarrow$ HTK	2	0.55	0.19	0.01	0.01
L6	PRG $\rightarrow$ CDT	2	0.38	0.04	0.015	0.015
L7	PRG $\rightarrow$ RDT	2	0.2	0.05	0.05	0.05
L8	CDG $\rightarrow$ HTK $\rightarrow$ PRG $\rightarrow$ RDT	4	0.92	0.12	0.005	0.005
L9	CDG $\rightarrow$ PRG	2	0.25	0.29	0.05	0.05
L10	CDG $\rightarrow$ HTK	2	0.43	0.46	0.015	0.015
L11	CDG $\rightarrow$ PRG $\rightarrow$ RDT $\rightarrow$ HTK $\rightarrow$ PRG	5	0.95	0.01	0.01	0.01
L12	(RDT $\rightarrow$ PRG)^2 $\rightarrow$ RDT	5	0.68	0.05	0.005	0.005
L13	CDG $\rightarrow$ (PRG $\rightarrow$ HTK)^2 $\rightarrow$ (PRG $\rightarrow$ RDT)^2 $\rightarrow$ CDG $\rightarrow$ CDT $\rightarrow$ (RDT $\rightarrow$ PRG)^2	17	0.91	0.07	0.07	0.07

Learner Level. A total of 488 sequences based on $\mu > 0$ were found at the learner level (i.e. 18 learners and one sequence per learner). As shown in Table 4, most of the learner sequences exhibit loops, mainly combining the actions programming and hovering over the task (i.e. (PRG $\rightarrow$ HTK)^k) with $k \in \mathbb{N}$. This behavior is characteristic of reviewing tasks during the engagement SRL phase.

Different learners' patterns were identified. For instance, learner L13 shows a reviewing task tendency where he/she proceeds through code debugging (CDG), programming and reading or hovering the task, code testing (CDT), and again task reading and programming. The appearance of a CDT preceded by (RDT $\rightarrow$ PRG)^2 potentially displays a state where learners have tested their program before submission, and made adjustments before proceeding with a submission. Learners L1, L2, L4, L8, L11, L12 showed quiet similar behaviors. Moreover, we observe that some sequential patterns start with code debugging. We can assume that these patterns occurred when learners had already attempted the tasks and possibly failed, before reviewing these tasks and changing their strategies to succeed in their tasks. Few patterns include code testing which is only displayed at the learner level, showing that most learners do not test their code before submission, which may explain the imbalance of occurrences of successful vs. failure sequences, skewed in favor of failure.

6.2 Discriminative Sequential Patterns at the Platform, Task and Learner Level

The discriminative sequential patterns mining at the different levels, platform, exercise, and learner shows that the proportion of discriminative sequences based on $\mu > 0$ tends to increase as we move from a holistic to a more individual level (i.e. exercise level and learner-level). At the platform level, only two atomic patterns were identified (reading the task and debugging the code). This highlights the importance of these actions for learner's success at a *holistic* level. At the exercise level, very low variability of the discriminative measures values was observed ($\overline{red}$, $OrdDep$ and μ), see Fig. 1. These measures have more variability at the learner level, as shown in Fig. 2, where one can observe that these measures discriminate better at the individual level. This leads us to say that more discriminative behaviors may be observed at a user level than at a platform level.

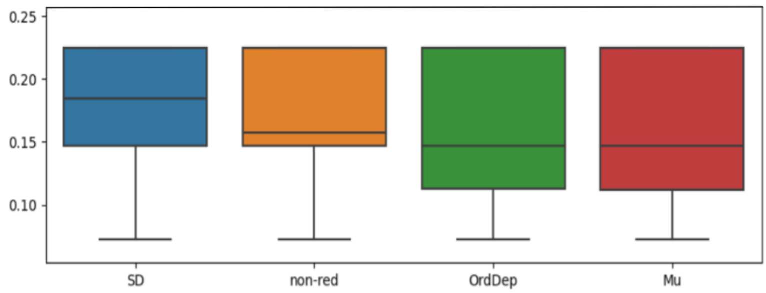


Fig. 1. Comparing the discriminative degree measures of sequences computed at the exercise level (SD, $\overline{red}$, $OrdDep$ and μ >0). (Color figure online)

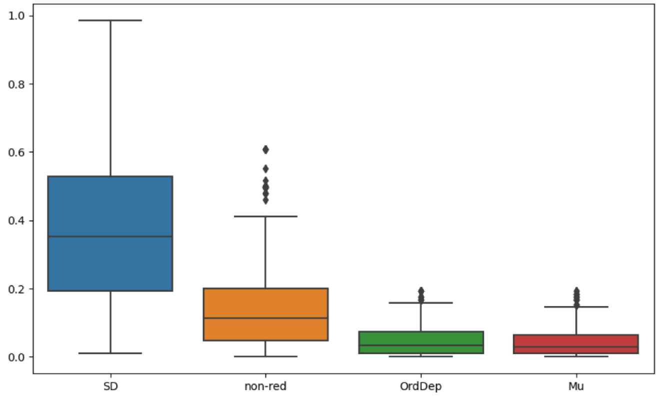


Fig. 2. Comparing the discriminative degree measures of sequences computed at the learner level (SD, $\overline{red}$, $ordDep$ and μ >0). (Color figure online)

7 Conclusion and Implications

In this work, we contribute to research in SRL with a new data mining approach that provides sequential patterns of SRL behaviors that may explain learning performance. A first contribution is the formalization of a generic data mining pipeline that includes appropriate measures for identifying self-regulated learner behaviors that lead to learning success. These measures allow us to determine the frequency, discriminative degree, non-redundancy, and order dependency of sequential learner actions, leading to a more fine-grained analysis of SRL behaviors. A second contribution is the implementation of the proposed approach on real trace data from an online programming platform. The trace data is translated into learning actions, resulting in two sets of sequence data, on the one hand, sequences that lead to learning success, and on the other hand, sequences that lead to learning failure.

Results showed that self-regulated learners who demonstrated high performance were those who showed a planning SRL phase before an engagement phase. During the planning phase, the most successful learners were those who read their tasks before working on the programming tasks. During the SRL engagement phase, learners showed the behavior of reviewing tasks. This behavior appears to be essential for the completion of high-difficulty tasks, where learners constantly make strategy changes and invest more effort in problem-solving. We mine the data at three different levels, platform, exercise, and learner resulting with a different discriminative degree expressed by the proposed measures. This revealed that the more we analyze the patterns at an individual level, the more we identify the most characteristic behaviors impacting self-regulated learners' performance. This work has scholarly and practical implications. It provides interesting insights into the design of intelligent tutoring systems by providing behavioral actions to recommend as SRL scaffolds and strengthen learners' reflection. This work is not without limitations. Our approach uses only the success or failure of an exercise to explain the learning performance. It would be relevant to consider other modeling methods that estimate learner skill acquisition. Future work is motivated to design an approach that allows the exploration of behavioral patterns that contribute to learners' skill acquisition, based on a Bayesian Knowledge Tracing (BKT) model [2,5]. Finally, it would be worth investigating how to provide learners with useful SRL scaffolds based on this data mining approach.

Acknowledgements. This work is part of the ANR xCALE project, funded by the French National Research Agency (ANR-20-CE38-0010).

References

1. Bandt, C., Pompe, B.: Permutation entropy: a natural complexity measure for time series. Phys. Rev. Lett. **88**(17), 174102 (2002)
2. Boulahmel, A., Djelil, F., Gilliot, J.-M., Smits, G.: Towards a skill-based self-regulated learning recommendation system. In: Doctoral Consortium of the 18th European Conference on Technology Enhanced Learning (2023)
3. Cheng, H., Yan, X., Han, J., Philip, S.Y.: Direct discriminative pattern mining for effective classification. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 169–178. IEEE, (2008)
4. Cicchinelli, A., et al.: Finding traces of self-regulated learning in activity streams. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 191–200 (2018)
5. Couland, Q., Leray, P., Boulahmel, A.: Un modèle générique avec structuration des compétences et facteurs externes pour le bayesian knowledge tracing. In: Proceedings of the 11èmes Journée Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes (2023)
6. Djelil, F., Gilliot, J.-M., Garlatti, S., Leray, P.: Supporting self-regulation learning using a Bayesian approach. Some preliminary insights. In: International Joint Conference on Artificial Intelligence IJCAI-21, Workshop Artificial Intelligence for Education (2021)
7. Fan, Y., et al.: Improving the measurement of self-regulated learning using multi-channel data. Metacogn. Learn. **17**, 1025–1055 (2022)
8. Fournier-Viger, P., Lin, J.C.-W., Kiran, R.U., Koh, Y.S., Thomas, R.: A survey of sequential pattern mining. Data Sci. Pattern Recogn. **1**(1), 54–77 (2017)
9. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2000)
10. He, Z., Zhang, S., Gu, F., Wu, J.: Mining conditional discriminative sequential patterns. Inf. Sci. **478**, 524–539 (2019)
11. Lim, L., et al.: Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. Comput. Hum. Behav. **139**, 107547 (2023)
12. Pham, H.S., Virlet, G., Lavenier, D., Termier, A.: Statistically significant discriminative patterns searching. In: Ordonez, C., Song, I.-Y., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DaWaK 2019. LNCS, vol. 11708, pp. 105–115. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27520-4_8
13. Siadaty, M., Gasevic, D., Hatala, M.: Trace-based micro-analytic measurement of self-regulated learning processes. J. Learn. Analytics **3**(1), 183–214 (2016)

14. Srivasta, N., et al.: Effects of internal and external conditions on strategies of self-regulated learning: a learning analytics study. In: Proceedings of the 12th International Learning Analytics and Knowledge Conference, pp. 392–403 (2022)
15. Sun, J.C.-Y., Tsai, H.-E., Cheng, W.K.R.: Effects of integrating an open learner model with AI-enabled visualization on students' self-regulation strategies usage and behavioral patterns in an online research ethics course. Comput. Educ. Artif. Intell. **4**, 100120 (2023)
16. Wong, J., Khalil, M., Baars, M., de Koning, B., Paas, F.: Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course. Comput. Educ. **140**, 103595 (2019)
17. Zimmerman, B.J.: A social cognitive view of self-regulated academic learning. J. Educ. Psychol. **81**(3), 329–339 (1989)

Analysis of Machine Learning Models for Academic Performance Prediction

Andres Benitez Amaya(✉), Harold Castro Barrera, and Ruben Manrique

Universidad de los Andes, Cra 1 18a-12, Bogotá, Colombia
{af.benitez,hcastro,rf.manrique}@uniandes.edu.co

Abstract. The prevalent issue of increased student dropouts, shared by universities worldwide, often culminates in decreased academic performance and prolonged completion times for degree programs. Prompt detection of those students facing a likely chance of failing a course could allow universities to intervene with sufficient support and guidance, facilitating an improvement in their performances. Numerous studies have explored the problem of performance prediction from various perspectives using different representations, algorithms, and data sets. The diversity in research strategies, however, complicates comparisons. In this study, we present a thorough evaluation of various predictive algorithms, representations, and predictive targets for the task of predicting student performance across 77 different courses in three distinct programs at the Universidad de los Andes: Systems and Computer Engineering, Industrial Engineering, and Economics. The results show that representing data in windows of time spanning 3 previous semesters, in conjunction with the LSTM-based algorithm for binary classification, yields the best results, achieving a precision of 0.838.

Keywords: Academic performance · Machine learning models · Grade prediction

1 Introduction

The present-day challenges in higher education, such as academic performance and student success-defined as students completing their programs, are significant [9]. A preliminary data analysis of some undergraduate programs at Universidad de los Andes in Colombia revealed a rising trend in program durations by almost 34% for the ISIS and IIND programs and 22.5% for the ECOM program. Concurrently, there was an almost 18% decline in academic performance as assessed by Grade Point Average (GPA). These trends present a challenge for universities, necessitating the creation of more efficient mechanisms and systems for detecting academic risk, and the development of intervention programs that provide support for students identified as at risk. This study proposes to examine various machine learning models to predict a student's performance in a specific course and determine the likelihood of the student passing or failing. With these

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 150–161, 2024.
https://doi.org/10.1007/978-3-031-63031-6_13

predictions, the university can intervene early, providing students with support, advice, suggestions, and other aids to boost their academic performance and mitigate student attrition [10,13].

This study, therefore, presents a process of constructing a machine learning supported system that predicts performance in a specific course. A literature review revealed a multitude of strategies and approaches for this task. However, the vast number of possibilities and the absence of comparative work on a data corpus at a university scale make the selection process challenging. Therefore, this study endeavors to fill this void by presenting a comparative analysis of different representations, machine learning algorithms, performance prediction targets, and hyperparameter selection processes on a corpus of more than 77 courses across three different university programs. The findings derived from such a broad corpus can guide future research.

To achieve this, our methodology includes the following macro-phases:

i. Compare different machine learning algorithms with different combinations of hyperparameters.
ii. In each algorithm-hyperparameter configuration, evaluate them using different feature representations.
iii. Repeat the process for each course in the curriculum of each of the study programs: *Computing Engineering (ISIS), Industrial Engineering (IIND), and Economics (ECOM)* at the Universidad de los Andes.
iv. Identify the model(s) configuration(s) that, in most cases during the iterative process, allows for the best performance in predicting if a student is at risk of failing a course.

In the rest of this paper, we introduce our methodology in Sect. 3. The results are presented in Sect. 4, with conclusions drawn from both the process and the results in Sect. 5.

2 Related Work

Student performance prediction has been addressed using various approaches, with machine learning models being the most common in the literature. The effectiveness of these models hinges on the relevant features they are trained on, such as numerical/alphabetical grades and demographic information, among others [13]. Grades emerge as the most frequently used feature in the literature due to the nature of the problem [2,13]. Other explored features include grades aggregated as the Grade Point Average (GPA) [3], pre-university grades [5], and demographic characteristics like gender, age, socio-economic stratum, place of residence, among others [15]. Despite efforts to incorporate new features to enhance their performance, most studies lack a consensus regarding the benefits of such inclusions.

The definition of the task can also play a crucial role in the outcomes of the models. For instance, if the objective is to predict the student's grade, it may involve a regression problem [11]. However, it is different if the task is a classification problem, where we might encounter binary outcomes (pass or fail a course) [6], or when the goal is to predict whether a grade falls within one or more grade ranges [8]

For binary classification problems, logistic regression (LR) algorithms demonstrate good performance [17]. Decision trees (DT) are also effective, especially when there is enough historical student data for training [8]. Neural networks (NN) also yield promising results compared to algorithms such as Support Vector Classifier (SVC) or Random Forest (RF) [14]. If there is written feedback available, such as teacher reviews or exam responses, Naive Bayes models show good classification results due to their effectiveness in text analysis [16].

Neural networks have also been utilized for solving academic performance prediction problems. However, despite their often noted efficiency in various tasks, it's not guaranteed that they will consistently outperform traditional models. Therefore, it's crucial to exercise critical judgment when selecting the most suitable model for a given problem. For instance, LSTMs tend to perform better when the feature set is not extensive, and there is a sufficient number of timesteps in the series [12]. Conversely, if the number of features is substantial and the timesteps are limited, they may exhibit subpar performance, and superior results might be achieved with ensembled models or Gated Recurrent Units (GRU) [1]. In this study, we will evaluate which machine learning model performs best based on the relevant features extracted from the data in our study. We will assess the most frequent models in the literature to solve the student performance prediction problem. Additionally, we will analyze their performance in different tasks: grade prediction [Sect. 3.5], binary classification [Sect. 3.5] and multinomial classification [Sect. 3.5]. We will vary the representation Sect. 3.3 of the training data for the models and the historical data from different academic programs at the Universidad de los Andes.

3 Research Methodology

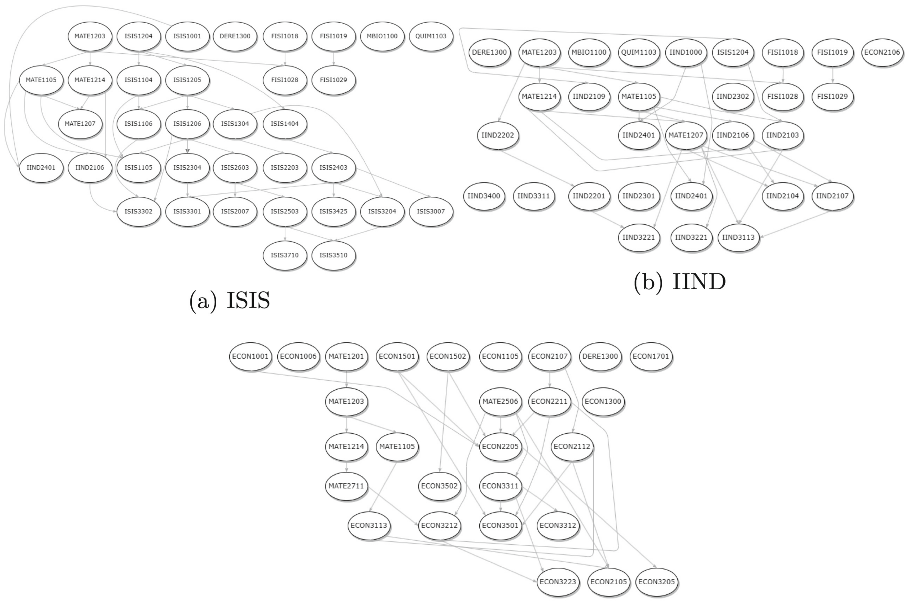
3.1 Dataset

The research dataset comprises the academic histories of 5,387 students enrolled in the Universidad de los Andes between 2001 and 2019. This spans three academic programs: Computing Engineering (ISIS), Industrial Engineering (IIND), and Economics (ECOM), with student numbers at 533, 3,156, and 1,698 respectively. Key features for model training, such as anonymized student IDs, academic semester, course ID, semester during which the course was taken, academic credits of the course, and obtained grades, were extracted.

3.2 Preprocessing

Data from 2001 to 2003 was not included in the analysis due to differing curriculum structures in comparison to the rest of the dates covered by the study. Post-anonymization, the data underwent a feature extraction process targeting crucial characteristics discovered in the previous stage. Moreover, this excluded first-semester courses from the target course set due to an insufficient basis for performance prediction. Nevertheless, such courses were included if they were prerequisites for other courses.

Curriculum-specific courses are the focus of this study, hence professional electives, free-choice courses, basic cycle courses specific to Universidad de los Andes, and other professional electives were dismissed. The analysis covered 29 courses for ISIS, 24 for IIND, and 24 for ECON. Finally, a curriculum representation (depicted in Fig. 1) based on department recommendations was devised for each program, serving as input for creating different features representations (Sect. 3.3).



(a) ISIS

(b) IIND

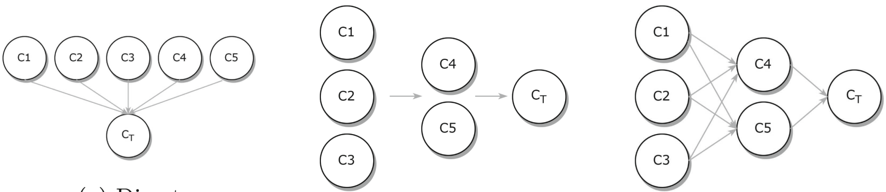
(c) ECON

Fig. 1. Courses prerequisites graphs

The data analysis from the 77 courses studied reveals a significant imbalance in class distribution. On average, there exists six times more records of positive class 1 (students passing the course) compared to the negative class 0 (students failing the course). This might be attributable to the Universidad of los Andes' policy that allows students to withdraw from courses until a certain period in the semester. The majority of students tend to opt for this alternative over failing the course. This dataset solely considers final course grades and disregards the characteristics of courses from which students withdrew, leading to significantly fewer instances of failure. Synthetic Minority Over-sampling Technique (SMOTE) was employed to rectify the issue of class imbalance. Past research indicates that SMOTE significantly improves classification tasks in machine learning models [7].

3.3 Representations

A representation consists of the data extracted for model training. In other words, for a target course C_T, a different subset of data D_{CT} is created for each of the representations. This D_{CT} will be the data used as input to train the various machine learning models. These representations are crafted to analyze which relationships are more relevant for determining a student's academic performance in the course C_T with greater precision, as each representation captures different information about the students. Some representations are direct (Fig. 2a), others are sequential over time (Fig. 2b), and other capture prerequisite relationships (Fig. 2c). The methodology involved the development of various representations. Firstly, two direct representations were established: a weighted average of course grades up to the first semester (R6), and course grades from the previous semester (R5). Sequential representations were also employed, which involved historical student grades up to the first semester (R4), student grades from the prior semester (R3), and grades from preceding semesters within specific time windows (R7). Additionally, prerequisite-based representations were constructed which included the grades for prerequisite courses (R1), and the sequence of prerequisite course grades leading up to the first semester (R2).



(a) Direct (b) Sequential (c) Prerequisites

Fig. 2. Feature representations

3.4 Models

Predicting a student's performance poses a significant challenge due to inherent data dispersion and relevant characteristics [5]. Moreover, each machine learning model's performance is contingent on the dataset utilized for training since no universally optimal model exists for predicting performance [14]. Consequently, we will train various models to compare and identify the most suitable one for our data, particularly concerning the relevant features extracted from it. The literature review revealed that the commonly used models for this problem include:

Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), Gradient Boosting (GBoosting), XGBoost (XGB), and Long Short-Term Memory (LSTM).

3.5 Prediction Objectives

Prediction Grade Problem. The initial stage involved training six different models; Linear Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), Gradient Boosting (GBoosting), and Extreme Gradient Boosting (XGBoost), on six preliminary course representations. This was done for 29 Computer Engineering, 24 Industrial Engineering, and 24 Economics targeted courses. The models underwent an 80% training, 20% validation split, which was consistent with the subsequent stages. Evaluation was carried out utilizing two metrics: R^2-score and Mean Absolute Error (MAE). The results were computed as an average across all the courses examined for each metric and the findings have been documented in Table 1.

Table 1. R^2-score and MAE for model.

Representation	LR		SVM		MLP		RF		Gboosting		XGBoost	
	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
R1	0.43	0.49	0.35	0.38	0.29	0.43	**0.23**	**0.33**	0.26	0.38	0.25	0.31
R2	0.39	0.41	0.32	0.37	0.37	0.41	0.25	0.37	0.24	0.38	**0.21**	**0.34**
R3	0.41	0.46	0.37	0.43	0.26	0.37	0.27	0.36	**0.23**	**0.33**	0.25	0.39
R4	0.39	0.43	0.37	0.46	0.29	0.34	0.34	0.41	0.31	0.37	**0.21**	**0.31**
R5	0.43	0.59	0.39	0.52	0.35	0.46	0.31	0.35	0.29	0.34	**0.24**	**0.32**
R6	0.37	0.51	0.33	0.38	0.32	0.44	0.26	0.35	**0.24**	**0.33**	0.26	0.36

The efficacy of predictive models can be evaluated through the R^2-score and MAE metrics, ranging from 0 to 1, where lower values denote higher effectiveness. Tree-based ensemble methods have exhibited superior prediction performance, offering more accurate results in comparison to other models [3]. Hence, for future iterations, models such as RF, GBoosting, and XGBoost are preferred over the others.

Binary Classification Problem. The obtained prediction metrics revealed a degree of inaccuracy, with the error rate proving significant in some instances based on the relationship between the target course C_T and the remaining courses in each curriculum. To address this, the problem was converted into a binary classification task where '1' signifies course pass, and '0' indicates otherwise. Established models such as Random Forest (RF), Gradient Boosting (GBoosting), and XGBoost (XGB) were then adjusted to their corresponding classification models.

We also introduce the use of the Long Short-Term Memory (LSTM) model, previously demonstrated to produce desirable results for academic performance prediction tasks [12]. Although this model is computationally expensive, it is deemed beneficial for a problem that is not overly extensive and does not demand substantial resources. To employ this model, a time window representation was required as the current representations are incompatible with this model. Consequently, a new representation was established ($R7$), based on the T previous semesters before the target course. Therefore, if the chosen window T equals 5, data from the five semesters preceding the target course would be collected and used to train this model.

Table 2. Number of courses with best Macro F1-Score per time window

Time window										
T	1	2	3	4	5	6	7	8	9	10
	3	12	37	16	4	1	3	1	–	–

Past studies have shown that the time window used can impact the classification results as it determines the more relevant data and features over time needed to predict a student's performance [6]. It was therefore necessary to determine the optimal time window for our data. An experiment was conducted on the 77 courses of the analysis belonging to the three academic programs to identify the best-performing time window within a range from $T = 1$ to $T = 10$. Notably, the average duration for completing a degree among the students in the analysis was 11 semesters. The results in Table 2 demonstrate that the optimal value for T is 3, meaning taking the data from the 3 semesters preceding the target subject C_T that we want to predict. Additionally, some isolated cases of higher time windows, such as 7 and 8, are shown, which may also depend on the course being tested and the execution process itself. With this clarity, the LSTM with a window of $T = 3$ will be used along with the previously identified models RF, GBoosting, and XGBoost in the following experiments.

Multinomial Classification Problem. The previous binary model was efficient in predicting if a student would pass a particular course. However, the aim is to develop an early warning system for vulnerable students, thereby offering guidance and intervention strategies. Hence, it's crucial to not only identify if a student is at risk, but also to segregate them into finer categories. This approach allows students to comprehend the potential scope of their performance, assisting them in making decisions about undertaking the intended course.

To structure this problem is divided into five distinct classes that represent the range of possible grades for the target course. These ranges are as follows: 1 = [0,1), 2 = [1,2), 3 = [2,3), 4 = [3,4), and 5 = [4,5].

4 Results

The research examines the performance of models across various academic program courses by tailoring training conditions for each, owing to their unique interrelations with other subjects in the curriculum. Factors such as prerequisite relationships or GPA distinctively impact each subject. Each course, C_T, from the 77-course pool, C, has been individually represented in seven different feature representations, serving as unique training data for the model set, M. These model-representations combinations, R_M, iterate over all combinations of the hyperparameters set, H. This means that each model is trained using different representations and various hyperparameter combinations for each course. This approach leads to a multitude of results, denoted as $C_T \times R \times M \times (H \times H)$.

Binary Classification Problem. When executing the process outlined in the Sect. 3, for each representation-model pair R_M, we identify the best set of hyperparameters that optimizes the model for that specific representation. Subsequently, the results are grouped to evaluate and compare which representation and model combination consistently yields the best metrics across most courses. It is important to note that the set of hyperparameters may vary for each element, depending on the course, representation, and model. A count is performed over all 77 executions of the courses in each program. The results are presented in Tables 3, 4 and 5. Note that in these tables the number of courses with best performance in terms of macro F1-score for a model-representation combination is shown.

Table 3. Computing Engineering

Model	RF	Gboosting	XGBoost	LSTM
R1	1	–	–	–
R2	–	4	4	–
R3	2	–	–	–
R4	–	2	3	–
R5	–	–	–	–
R6	–	1	–	–
R7	–	–	–	12

Table 4. Industrial Engineering

Model	RF	Gboosting	XGBoost	LSTM
R1	–	2	–	–
R2	-	2	4	–
R3	1	–	–	–
R4	–	1	3	–
R5	–	–	–	–
R6	–	–	2	–
R7	–	–	–	9

Table 5. Economy

Model	RF	Gboosting	XGBoost	LSTM
R1	–	1	2	–
R2	–	3	3	–
R3	1	–	–	–
R4	–	2	3	–
R5	–	–	–	–
R6	1	1	–	–
R7	–	–	–	7

The findings reveal that representation 7, in combination with the LSTM model in a time window of $T = 3$, is the most consistently successful model across the three academic programmes investigated. This model outperformed others in the binary classification task for 28 courses, accounting for slightly over a third of the total courses analyzed. Subsequent analyses focused on identifying the hyperparameter combination which delivered the optimal results. Considered metrics included precision, recall, Micro F1-Score and Macro F1-score. The most effective performance was noted with a total of 150 units, a batch size of 32, 20 epochs, and the use of a relu activation function. This combination of hyperparameters was identified via a grid search. The study notes that the inclusion of additional parameters in the grid search (dropout, learning rate, layers, etc.) may deliver marginally improved model performance. However, this would greatly enhance the computational cost given the process is executed separately for each model, representation, and course assessed in the analysis.

The results obtained from applying the LSTM on all courses per program in a training and test partition of 70:30 are presented in Table 6.

Table 6. LSTM with $R7$ average results per program

ISIS		IIND		ECON	
Metric	Value	Metric	Value	Metric	Value
Precision	0.8386	Precision	0.7673	Precision	0.8369
Recall	0.7872	Recall	0.7906	Recall	0.7687
Micro F1-score	0.7987	Micro F1-score	0.7702	Micro F1-score	0.7923
Macro F1-score	0.8498	Macro F1-score	0.8412	Macro F1-score	0.8350

Multinomial Classification Problem. For this problem, we build upon the previously obtained results where the LSTM model shows better performance in a significant number of courses for each academic program. We assume that although the behavior may vary, it will maintain similar results. We run the process again for all courses in each program, but this time treating it as a multinomial problem, where the labels are each of the specified ranges and using LSTM model with $R7$. The results are presented in Table 7 as the average of the outcomes for the courses in each program across the different ranges. Divided in this way: 1 = [0,1), 2 = [1,2), 3 = [2,3), 4 = [3,4), 5 = [4,5].

Table 7. Multinomial Classification Results

ISIS					IIND				
Range	Precision	Recall	Micro F1	Macro F1	Range	Precision	Recall	Micro F1	Macro F1
1	0.4714	0.3232	0.4012	0.5007	1	0.5017	0.3055	0.3824	0.5162
2	0.5060	0.3444	0.4128	0.5167	2	0.5214	0.3944	0.4501	0.5574
3	0.5796	0.3777	0.4033	0.5791	3	0.5560	0.3942	0.4628	0.5667
4	0.7421	0.7571	0.7495	0.7406	4	0.7083	0.3777	0.4657	0.5945
5	0.7703	0.7392	0.7545	0.7504	5	0.7586	0.6470	0.6984	0.7862

ECON				
Range	Precision	Recall	Micro F1	Macro F1
1	0.4468	0.6176	0.5185	0.6126
2	0.4583	0.6470	0.5365	0.6243
3	0.4634	0.5588	0.5066	0.6202
4	0.6721	0.7321	0.7008	0.6702
5	0.7431	0.7814	0.7570	0.7435

5 Conclusions

The generation of various representations aids in evaluating which data is deemed more relevant for the prediction task. The results indicate that employing a representation based on time windows ($R7$) better captures a student's academic behavior, even when compared to prerequisite relationships or cumulative GPA. Furthermore, by generating representations, we develop small datasets for each student's course, assisting machine learning models in identifying similar patterns in academic history and effectively grouping students with comparable outcomes. This approach has shown to enhance the performance of tree-based ensemble models [4], which predominantly yielded the best results in the initial phase.

The models for the binary classification problem (i.e., pass or fail) demonstrate strong performance, achieving over 90% accuracy in certain specific courses, particularly those that are highly interconnected with others and are pivotal to the academic curriculum. However, the reported data does not consistently exceed 90% precision because some courses are isolated cases (without prerequisites or early semesters) where neither the representation nor the model captures the relevant characteristics for accurate prediction, leading to a decrease in overall performance.

The Systems Engineering and Computer Science program shows better results based on the metrics in the LSTM model tested with a window frame equal to 3 semesters, which could be attributed to the need for recent concepts or courses to determine success in a subject. In contrast, disciplines like economics, with their curricular structure, may require concepts beyond the time window.

Although they do not show poor performance, it tends to be lower than that of engineering programs.

Multinomial classification offers an advantage in the generation of early warnings by providing a clearer overview, as it classifies the student's performance in a more specific range, and is not only limited to determining only pass or fail results, not only limited to determining pass or fail outcomes alone. This significantly impacts the advice that can be offered and the subsequent steps a counselor may take. Knowing that a student will pass a course differs from classifying them into a lower risk category (for example, range 4=[3,4), which includes the passing grade limit of 3). In the latter scenario, specific recommendations for reinforcing topics or considering alternatives can be provided. Although the model's performance is lower in the multinomial case compared to the binary case, it still produces good results, which is expected given the complexity of the task.

The obtained results are promising both in the binary and multiclass problems. These results are tailored to our current data and can be replicated for other programs at the Universidad de los Andes. However, it is recognized that exploring numerous options such as new representations, models, architectures, or incorporating additional data could further enhance results. Nonetheless, considering the scale of the problem, These findings significantly contribute to early at-risk student detection, enabling timely interventions to reduce dropout rates and extend academic timelines.

References

1. Al-Sulami, A., Al-Masre, M., Al-Malki, N.: Predicting at-risk students' performance based on LMS activity using deep learning. Int. J. Adv. Comput. Sci. Appl. **14**(6) (2023). https://doi.org/10.14569/IJACSA.2023.01406129
2. Alyahyan, E., Düştegör, D.: Predicting academic success in higher education: literature review and best practices. Int. J. Educ. Technol. High. Educ. **17**(1), 3 (2020). https://doi.org/10.1186/s41239-020-0177-7
3. Ayulani, I.D., Yunawan, A.M., Prihutaminingsih, T., Sarwinda, D., Ardaneswari, G., Handari, B.D.: Tree-based ensemble methods and their applications for predicting students' academic performance. Int. J. Adv. Sci. Eng. Inf. Technol. **13**(3), 919–927 (2023). https://doi.org/10.18517/ijaseit.13.3.16880 https://doi.org/10.18517/ijaseit.13.3.16880 https://doi.org/10.18517/ijaseit.13.3.16880
4. Buenaño-Fernández, D., Gil, D., Luján-Mora, S.: Application of machine learning in predicting performance for computer engineering students: a case study. Sustainability **11**(10), 2833 (2019). https://doi.org/10.3390/su11102833
5. Christou, V., et al.: Performance and early drop prediction for higher education students using machine learning. Expert Syst. Appl. **225**, 120079 (2023). https://doi.org/10.1016/j.eswa.2023.120079
6. Deeva, G., De Smedt, J., Saint-Pierre, C., Weber, R., De Weerdt, J.: Predicting student performance using sequence classification with time-based windows. Expert Syst. Appl. **209**, 118182 (2022). https://doi.org/10.1016/j.eswa.2022.118182
7. Elor, Y., Averbuch-Elor, H.: To smote, or not to smote? CoRR abs/2201.08528 (2022). arXiv:2201.08528

8. Hashim, A.S., Awadh, W.A., Hamoud, A.K.: Student performance prediction model based on supervised machine learning algorithms. IOP Conf. Ser. Mater. Sci. Eng. **928**(3), 032019 (2020). https://doi.org/10.1088/1757-899X/928/3/032019
9. Hu, Q., Rangwala, H.: Course-specific Markovian models for grade prediction. In: Phung, D., Tseng, V.S., Webb, G.I., Ho, B., Ganji, M., Rashidi, L. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 29–41. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-93037-4_3
10. Hu, Q., Rangwala, H.: Academic performance estimation with attention-based graph convolutional networks. arXiv:2001.00632 (2019)
11. Iqbal, Z., Qadir, J., Mian, A.N., Kamiran, F.: Machine learning based student grade prediction: a case study. CoRR abs/1708.08744 (2017). arXiv:1708.08744
12. Kim, B.H., Vizitei, E., Ganapathi, V.: GritNet: student performance prediction with deep learning. arXiv:1804.07405 (2018)
13. Manrique, R., Nunes, B.P., Marino, O., Casanova, M.A., Nurmikko-Fuller, T.: An analysis of student representation, representative features and classification algorithms to predict degree dropout. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp. 401–410. LAK19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3303772.3303800
14. Rastrollo-Guerrero, J.L., Gómez-Pulido, J.A., Durán-Domínguez, A.: Analyzing and predicting students' performance by means of machine learning: a review. Appl. Sci. **10**(3), 1042 (2020). https://doi.org/10.3390/app10031042
15. Sultana, S., Khan, S., Abbas, M.A.: Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. Int. J. Electr. Eng. Educ **54**(2), 105–118 (2017). https://doi.org/10.1177/0020720916688484
16. Yağcı, M.: Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. **9**(1), 11 (2022). https://doi.org/10.1186/s40561-022-00192-z
17. Zeineddine, H., Braendle, U., Farah, A.: Enhancing prediction of student success: automated machine learning approach. Comput. Electr. Eng. **89**, 106903 (2021). https://doi.org/10.1016/j.compeleceng.2020.106903

Simplifying Decision Tree Classification Through the AutoDTrees Web Application and Service

Matthaios Zografos and Stefanos Ougiaroglou(✉)

Department of Information and Electronic Engineering, School of Engineering, International Hellenic University, 57400 Sindos, Thessaloniki, Greece
stoug@ihu.gr

Abstract. Various algorithms are utilized for the purpose of classification, with Decision Trees being one of the most popular. This is due to their easily understandable structure and simple way of operation, which led to their adoption in a variety of applications. However, the usage of Decision Trees often becomes challenging due to the fact that users need to be familiar with Machine Learning, have programming skills or knowledge of specialized scientific software. The present paper aims to address these issues by presenting AutoDTrees, a web-based application that offers the ability to utilize the Decision Trees in a simple and fast way. This can be done through a user-friendly interface, while an open-source Web API is provided for developers. AutoDTrees allows any user to select the preferred dataset and define parameters, in order to build Decision Tree models. Then, the model effectiveness can be evaluated by using the k-fold cross-validation and presenting detailed metrics. Users are then able to save the pre-trained model and reuse it for predicting unclassified instances or visualizing the Decision Tree. AutoDTrees was evaluated in terms of user experience using the System Usability Scale (SUS), with the results indicating that it can be a useful tool for a wide range of users, regardless of their experience level.

Keywords: Decision trees · Classification · AutoML · Web application · Web service

1 Introduction

The concept of classification in Data Mining and Machine Learning refers to the process of classifying each element of a dataset into predefined groups, based on their characteristics [1]. Data classification is a supervised learning method. This type of methods requires a model to be constructed using training datasets that contain labeled data. The model can then be used in order to classify new instances [5].

Classification is a field with significant scientific interest that applies in various domains. Examples include medicine, for more effective disease diagnosis,

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 162–173, 2024.
https://doi.org/10.1007/978-3-031-63031-6_14

commerce, for optimizing the promotion of products and services [1], economics, for loan approval, bankruptcy prediction and fraud detection, as well as in business, facilitating decision-making processes [9].

Decision Trees are primarily used for data classification and can be considered as one of the most popular algorithms in that field. This is due to their tree-like structure which is easily understandable, even for individuals with less experience in Machine Learning and Data Mining. Additionally, the algorithm is characterized by its simple way of operation as well as high performance in model building [9].

The use of Decision Trees often becomes challenging. The main obstacle is that users need to be familiar with Machine Learning and specialized software. Also, programming skills are required. Additionally, it is important to note that the available software and programming tools often have certain limitations, such as the requirement to purchase a subscription and the need to download and install software packages. Therefore, the utilization of the algorithm can become more complex for a user and resource-intensive for their personal computer.

The aforementioned issues constituted the motivation for the present work. The contribution of the present paper is the development of AutoDTrees, an Automated Machine Learning (AutoML) [8] driven application that enables a wide range of users such as students, developers, researchers and data scientists, to perform Decision Tree classification in a simple and fast way. AutoDTrees is a free and open-source web-based application that offers a variety of features. More specifically, it allows users to select the preferred training datasets and determine the various parameters in order to build Decision Tree models. Additionally, the k-fold cross-validation method is utilized to evaluate the model's effectiveness, providing a comprehensive report consisting of various metrics. Based on the results, the user subsequently has the option to save the pretrained model for future handling, such as predicting new instances, as well as visualizing the decision tree. The set of application features is provided through an open-source Web API, while a user-friendly interface has been developed, eliminating the need for programming tools utilization.

The rest of the paper is organized as follows: Sect. 2 briefly reviews the well known Decision Tree algorithms. In Sect. 3, we present the AutoDTrees application, describing its architecture and features and presenting the web interface and the web service. Section 4 details the outcomes obtained from the System Usability Scale (SUS) questionnaire. Finally, Sect. 5 concludes the paper and presents ideas for future work.

2 Decision Tree Classification

Decision trees operate by recursively partitioning the data into subsets based on the most significant attributes, ultimately forming a tree-like data structure where each internal node represents an attribute, each branch denotes a decision rule, and each leaf node corresponds to a class label. Therefore, decision trees partition the attribute space into regions and assign a class label to each region. It is worth mentioning that decision trees are intuitive to interpret and implement.

Building an optimal decision tree involves selecting the best attribute at each node to split the data. Several algorithms have been developed for constructing decision trees. The most well known are:

– ID3 [12] is one of the earliest algorithms for decision tree construction developed by Ross Quinlan. It uses information gain as the criterion for selecting attributes and operates in a top-down manner. ID3 is effective for categorical data but struggles with continuous attributes.
– C4.5 [13] is an extension of ID3 which can handle both continuous and categorical attributes. Instead of information gain, C4.5 uses gain ratio, which normalizes the information gain by the split information to avoid bias towards attributes with many values.
– CART [3] is also a popular decision tree algorithm. It can handle both classification and regression tasks. CART constructs binary trees by recursively partitioning the data into two subsets based on the value of a chosen attribute. The splitting criterion in CART is typically based on either Gini impurity for classification or mean squared error reduction for regression.
– Random Forest [2] is an ensemble learning method based on decision trees. It builds multiple decision trees during training and combines their predictions through voting.
– Gradient Boosting Machines (GBM) [6] is also an ensemble approcah. It builds decision trees sequentially, with each tree correcting the errors of the previous one. It minimizes a loss function using gradient descent during tree construction.

It is wothh mentioning that AutoDTrees utilizes CART algorithm provided by the Python Scikit-learn library.

3 The AutoDTrees Application

3.1 Description

The AutoDTrees application provides an integrated environment that enables users to perform Decision Tree classification. More specifically, as presented in Fig. 1, AutoDTrees allows registered users to upload the preferred training datasets in csv format and select the fields to be used as features, as well as the field to be used as the class attribute for classification. Additionally, users will be able to define the parameters "max_depth" (the maximum depth of the tree) and "min_samples_leaf" (the minimum number of samples per leaf node) of the classifier, as well as the parameter "k" of the k-fold cross-validation method. Subsequently, the Decision Tree algorithm and the k-fold cross-validation method will be executed in order to successfully build and evaluate the model, providing the results of the metrics. AutoDTrees uses the Decision Tree implementation provided by the Python Scikit-learn library [11]. It is worth noting that the aforementioned implementation is based on the CART algorithm; however, it does not support categorical features.

The application also supports model saving for future use, either for predicting new instances or for visualizing the Decision Tree. In the first case, the user must select the preferred model and upload a dataset containing new instances for classification. After performing classification, the results are displayed and can also be exported to a csv file. Furthermore, users are provided with the option to view metrics for evaluating result quality, which is available only if the selected dataset already includes the class field. In the second case, an option to view the Decision Tree visualization is provided, as well as exporting it to a png file.

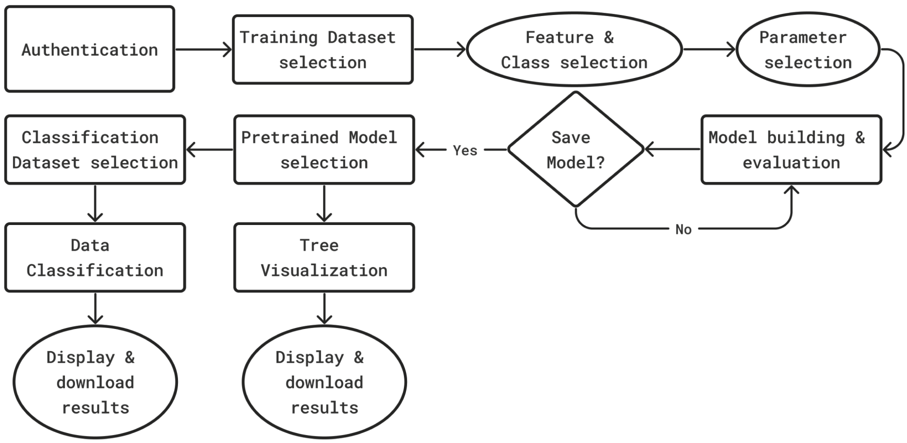


Fig. 1. The application flowchart.

The application's aforementioned features are conveniently accessible via a user-friendly interface as well as an open-source Web API. Specifically regarding the Web API, it facilitates the application's use by the developers, enabling either its expansion or the development of new applications, by leveraging part of the capabilities of the current implementation. For instance, an external application may use a pre-trained model hosted in AutoDTrees to carry out a classification task. It can be done through a request to the appropriate end-point of the API.

Following the architecture of WebApriori [10] for association rule mining and kClusterHub [7] for partition-based clustering, AutoDTrees consists of five main components, as shown in Fig. 2. Concerning the REST API, it is developed using PHP and handles the execution of most functions. It collaborates with other application components as needed. The MySQL Database primarily supports functions such as registration, authentication, and user management, with occasional use for tasks involving the storage and retrieval of pretrained models. Additionally, the Web API interacts with the server's file system to store user files, including datasets and pretrained models. Moreover, the Decision Tree algorithm and various Machine Learning tasks are executed by invoking Python modules and utilizing the capabilities of the Scikit-learn library. Lastly, the interface provides users access to application features through a user-friendly environment. The interface was designed and developed using the Bootstrap framework and JavaScript with the jQuery library. Additionally, AJAX is used for executing the API calls.

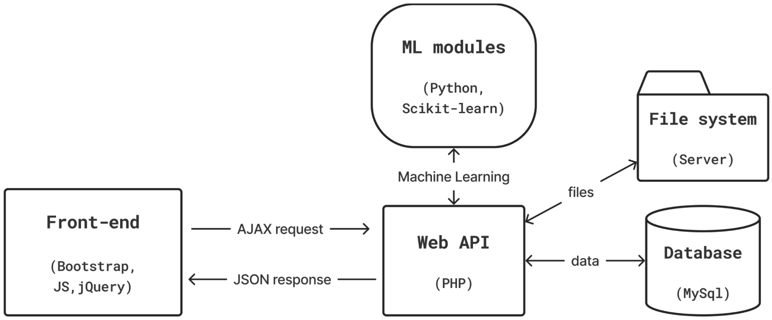


Fig. 2. The architecture of AutoDTrees.

It is noteworthy that AutoDTrees has been deployed on a web server at the Department of Information and Electronic Engineering of the International Hellenic University[1]. Additionally, since the application was developed using Git, its source code is easily accessible via GitHub[2], enabling users to deploy it on their own servers.

Utilizing the application, users will notice that there are two types of datasets, private and public. Private datasets are only available to the user who uploaded them. Conversely, public datasets become accessible to all registered users. This feature aims to assist users in constructing effective Decision Tree models and facilitates scenarios such as dataset sharing between professors and students for laboratory exercises. For security reasons, the permission to upload and delete public datasets is restricted to users who have been authorized by the system administrator. Therefore, depending on their permission level, users fall into four categories:

- **Administrator:** Grants users the permission to upload and delete public datasets.
- **Public dataset creators:** Registered users authorized by the administrator to upload and delete public datasets.
- **Typical users:** Registered users that are able to utilize the existing public or private datasets for model building. However, they cannot upload new public datasets or delete the existing ones.
- **Non-registered users:** They are not able to fully utilize the application's features.

[1] https://kclusterhub.iee.ihu.gr/autodtrees.
[2] https://github.com/manthoszog/AutoDTrees.

3.2 The Web Interface

Firstly, users are required to register and confirm their email address before accessing AutoDTrees via the web interface. After successfully completing this step, they will be able to log in and fully utilize the application. The main features are divided into two pages, one for model building and another for utilizing pretrained models. Concerning the first page, users are initially able to upload their preferred training datasets in csv format or select an existing one from the list (Fig. 3a). After selecting a file, a preview is shown in a tabular format (Fig. 3b), enabling the user to either download or delete the file. In the case of uploading or deleting a public dataset, the system first checks if the user has the necessary permissions.

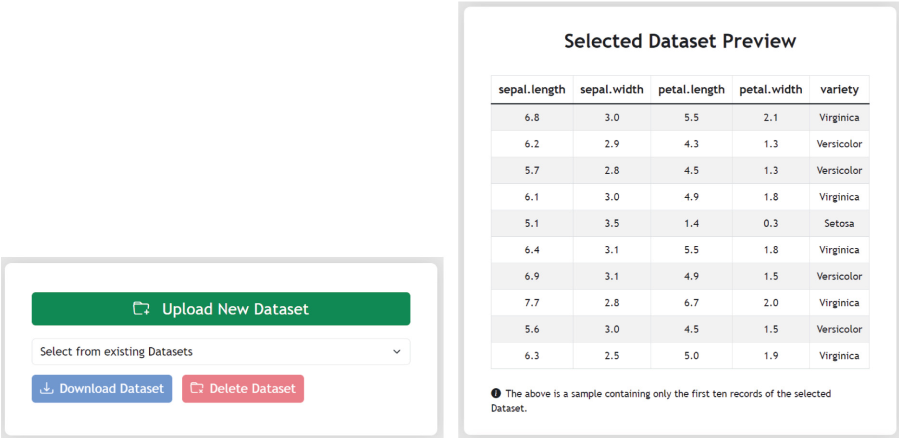
As depicted in Fig. 4a, a supplementary section appears at the bottom of the page. This section is dedicated to defining the parameters for building a Decision Tree model. Initially, users can select the dataset fields they intend to utilize as features[3], along with specifying the class field. Following this, users are prompted to set the parameters for the classifier and the k-fold cross-validation method. They have the option to input preferred values or utilize the default settings. To enhance comprehension and user guidance, tooltips are provided, offering relevant information throughout the process.

Upon clicking the "Build Model" button, AutoDTrees executes the model building and evaluation process, presenting the results of the metrics, as depicted in Fig. 4b. Subsequently, users have the option to save the model to their account for potential future use. To do so, they need to insert a name and click the "Save Model" button.

Upon entering the second page, a list containing all the pretrained models created by the user is displayed. After selecting a model, its content is shown, including the features and the class field which were used for constructing it. Additionally three buttons are displayed. The first button allows downloading the model in pkl format, while the second enables deletion of the model. The last one is utilized for visualizing the Decision Tree, allowing the user to either display the graph or export it in png format (Fig. 6).

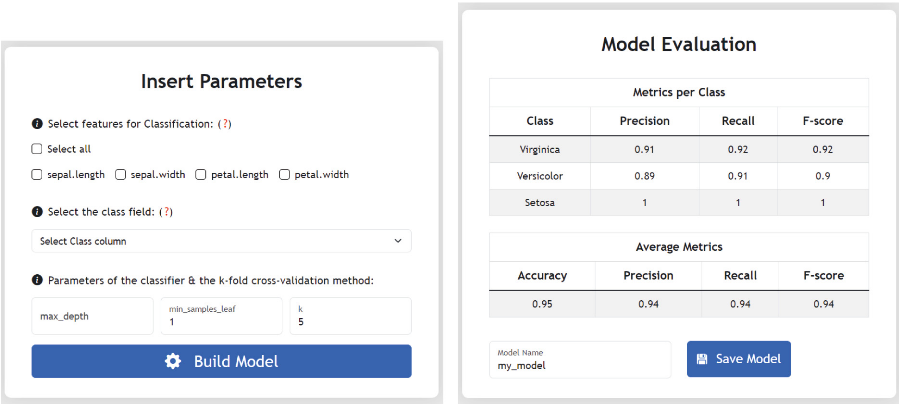
Consequently, a supplementary section appears at the bottom of this page. In this section users are prompted to upload a dataset containing new instances for classification. After selecting a file, a preview is shown in a tabular format. After clicking the "Classify Data" button, the application performs data classification. It then presents a preview of the results and provides the option to export the entire classified dataset in csv format (Fig. 5a). Moreover, users have access to metrics for assessing result quality, which are only available if the selected dataset already contains the class attribute (Fig. 5b). If the previously mentioned condition is not met, then this specific button will remain disabled.

[3] As already mentioned, categorical features are not available due to the Scikit-learn implementation.



(a) Dataset upload and selection (b) Dataset preview

Fig. 3. Using training datasets

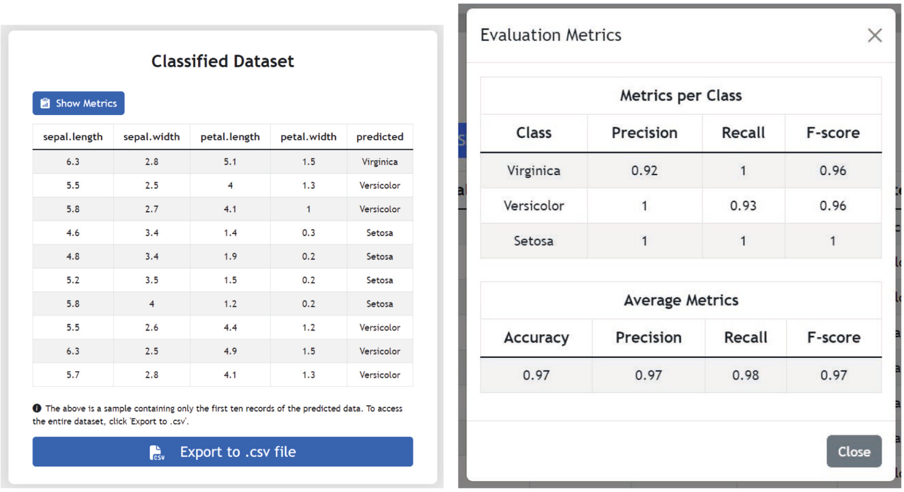


(a) Parameter selection (b) Evaluation metrics

Fig. 4. Model building

3.3 The Web Service

Firstly, users are required to register and confirm their email address before accessing the AutoDTrees web service. After successfully completing this step, they will be able to get an API token and utilize the various features. The web service has been structured as a REST API. In order to perform a specific function, a user or application must initiate a call to the corresponding endpoint accompanied by the personal API token, while the results are returned in JSON format.



(a) Classification results preview (b) Result metrics

Fig. 5. Classification results

As presented in Table 1, the AutoDTrees Web API consists of 19 endpoints, each serving various features. These features include user account operations such as registration, logging in, editing user settings and account deletion, as well as training/classification dataset manipulation functionalities like upload, deletion and retrieval. Additionally, there are specific endpoints dedicated to pretrained model operations, such as deletion and content retrieval. Furthermore, four endpoints enable programmers to perform key Decision Tree tasks, including model building and evaluation, model saving, data classification and tree visualization.

It is worth noting that the web interface features the "API Docs" webpage providing instructions for utilizing the AutoDTrees API endpoints. More specifically, this page showcases examples of requests along with their corresponding responses. An example of building and evaluating a model by calling the corresponding endpoint is presented below:

```
Request Body Example:
{"token": "cf9ed2a453796dcdd42ea95b24e55985",
"folder": "public", "file": "iris.csv",
"checkVal": ["sepal.length","sepal.width","petal.length","petal.width"],
"selected": "variety", "max_depth": 10, "min_samples_leafInt": 1,
"kFoldsInt": 5}
Response Example:
{"labels": ["Setosa", "Versicolor", "Virginica"],
"pre_per_label": [1.0, 0.93, 0.93], "rec_per_label": [1.0, 0.9, 0.93],
"fsc_per_label": [1.0, 0.9, 0.93], "avg_pre": 0.95, "avg_rec": 0.94,
"avg_fsc": 0.94, "avg_acc": 0.95}
```

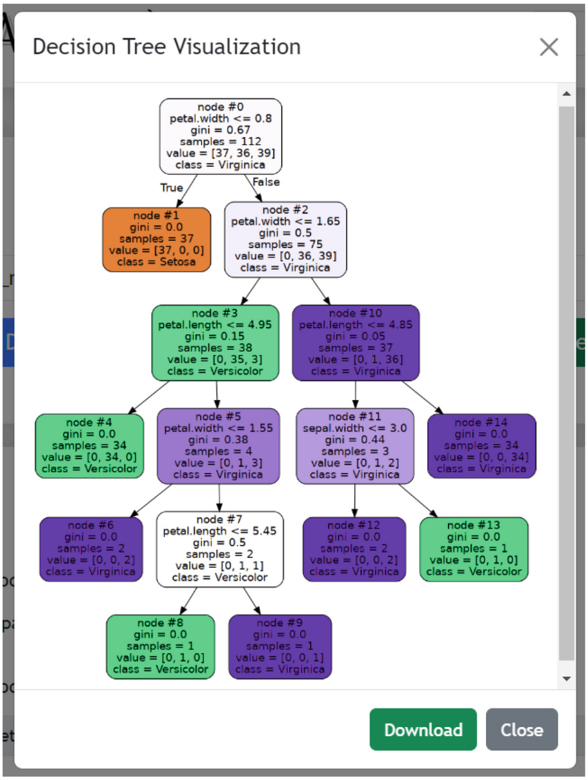


Fig. 6. Tree visualization.

Also, an example of saving a model by calling the corresponding endpoint is presented below:

```
Request Body Example:
{"token": "cf9ed2a453796dcdd42ea95b24e55985",
"folder": "public", "file": "iris.csv",
"checkVal":["sepal.length","sepal.width","petal.length","petal.width"],
"selected": "variety", "max_depth": 10, "min_samples_leafInt": 1,
"model_name": "my_model"}
Response Example:
{"message": "Model successfully saved."}
```

Table 1. AutoDTrees Web API endpoints

No.	HTTP Method	Endpoint
1	POST	register.php
2	POST	login.php
3	POST	edit-account.php
4	DELETE	delete-account.php
5	GET	get_datasets.php
6	POST	upload_dataset.php
7	GET	get_dataset_content.php
8	DELETE	delete_dataset.php
9	GET	get_unclassified_datasets.php
10	POST	upload_unclassified_dataset.php
11	GET	get_unclassified_dataset_content.php
12	DELETE	delete_unclassified_dataset.php
13	GET	get_models.php
14	GET	get_model_content.php
15	DELETE	delete_model.php
16	GET	visualize_tree.php
17	POST	cross_validation.php
18	POST	save_model.php
19	POST	classifyData.php

4 Usability Testing

Within this Section, we will present and analyze the results stemming from the assessment of the application, particularly focusing on user experience. To facilitate this process, a questionnaire was distributed among users of the AutoDTrees application. The participants predominantly consist of undergraduate students attending a Data Mining course at the Department of Information and Electronic Engineering.

The questionnaire is based on the System Usability Scale (SUS), a method designed for easy and quick evaluation of system usability. It comprises ten questions, each offering five response options ranging from “Strongly Disagree” to “Strongly Agree”, or correspondingly, the numerical scale of 1 to 5. Specifically, the ten questions encompassed in the SUS questionnaire are as follows [4]:

1. I think that I would like to use this website frequently.
2. I found the website unnecessarily complex.
3. I thought the website was easy to use.
4. I think that I would need the support of a technical person to be able to use this website.

5. I found the various functions in this website were well integrated.
6. I thought there was too much inconsistency in this website.
7. I would imagine that most people would learn to use this website very quickly.
8. I found the website very cumbersome to use.
9. I felt very confident using the website.
10. I needed to learn a lot of things before I could get going with this website.

To calculate the final score derived from the System Usability Scale questionnaire, the following steps are followed: Initially, responses are mapped onto a $1-5$ scale (1 indicating "Strongly Disagre", 5 indicating "Strongly Agree"). Next, for odd-numbered questions (1, 3, 5, 7, 9), 1 is subtracted from the user's response, while for even-numbered questions (2, 4, 6, 8, 10), the user's response is subtracted from 5. The resulting values are then summed and multiplied by 2.5 to determine the user's score on a $0-100$ scale [4]. The final score is calculated as the average of individual scores. A SUS final score above 80/100 is commonly viewed as excellent.

In the context of the AutoDTrees usability testing, a total of 30 participants provided responses. The SUS score of 81.5 indicates that users are highly satisfied with the experience of utilizing the AutoDTrees application. The detailed findings are presented in Table 2. Only three users (users 10, 11, and 25) appear to be less satisfied.

Table 2. The SUS score of AutoDTrees.

User	Score	User	Score
1	100	16	85
2	97.5	17	85
3	85	18	92.5
4	77.5	19	87.5
5	75	20	90
6	100	21	62.5
7	100	22	77.5
8	72.5	23	90
9	100	24	80
10	52.5	25	47.5
11	50	26	80
12	60	27	75
13	67.5	28	87.5
14	100	29	87.5
15	85	30	95
Final Score:		**81.5**	

5 Conclusions and Future Work

Decision Trees are widely recognized for their effectiveness in data classification, yet their usage is often challenging due to specialized knowledge requirements. To overcome these obstacles, we presented AutoDTrees, a user-friendly web application that simplifies the utilization of Decision Trees by offering a variety of features. These include model building and evaluation, as well as saving the model and utilizing it in tasks such as data classification and decision tree visualization. The application features are accessible either via an open-source Web API or a user-friendly interface, eliminating the necessity of using programming tools. In conclusion, the application's positive user feedback, as measured by the System Usability Scale (SUS) method, underscores its potential as a valuable tool for users of all levels of expertise.

As part of future work, we aim to enhance the AutoDTrees application by incorporating additional Decision Tree implementations, such as the ID3 and C4.5 algorithms, thereby enabling the utilization of categorical features.

References

1. Akinola, S., Oyabugbe, O.: Accuracies and training times of data mining classification algorithms: an empirical comparative study. J. Softw. Eng. Appl. **8**(9), 470–477 (2015). https://doi.org/10.4236/jsea.2015.89045
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
3. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.: Classification and Regression Trees. Chapman and Hall/CRC (1984)
4. Brooke, J.: SUS: a quick and dirty usability scale. Usability Eval. Ind. **189**, 4–7 (1995)
5. Edelstein, H.A.: Introduction to Data Mining and Knowledge Discovery Third Edition. Two Crows Corporation (1999)
6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. stat. **29**, 1189–1232 (2001)
7. Gratsos, K., Ougiaroglou, S., Margaris, D.: kClusterHub: an automl-driven tool for effortless partition-based clustering over varied data types. Future Internet **15**(10), 341 (2023). https://doi.org/10.3390/fi15100341
8. He, X., Zhao, K., Chu, X.: AutoML: a survey of the state-of-the-art. Knowl. Based Syst. **212**, 106622 (2021). https://doi.org/10.1016/j.knosys.2020.106622
9. Kyrkos, E.: Business Intelligence and Data Mining. Kallipos, Open Academic Editions (2015)
10. Malliaridis, K., Ougiaroglou, S., Dervos, D.A.: WebApriori: a web application for association rules mining. In: Kumar, V., Troussas, C. (eds.) Intelligent Tutoring Systems, pp. 371–377. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-49663-0_44
11. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
12. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986). https://doi.org/10.1023/A:1022643204877
13. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)

LBKT: A LSTM BERT-Based Knowledge Tracing Model for Long-Sequence Data

Zhaoxing Li[1(✉)], Jujie Yang[2], Jindi Wang[2], Lei Shi[3], Jiayi Feng[4], and Sebastian Stein[1]

[1] School of Electronics and Computer Science, University of Southampton, Southampton, UK
zhaoxing.li@soton.ac.uk, ss2@ecs.soton.ac.uk
[2] Department of Computer Science, Durham University, Durham, UK
{jujie.yang,jindi.wang}@durham.ac.uk
[3] Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, UK
lei.shi@ncl.ac.uk
[4] Beijing Jiaotong University, Beijing, China
jyfeng@bjtu.edu.cn

Abstract. The field of Knowledge Tracing (KT) aims to understand how students learn and master knowledge over time by analyzing their historical behaviour data. To achieve this goal, many researchers have proposed KT models that use data from Intelligent Tutoring Systems (ITS) to predict students' subsequent actions. However, with the development of ITS, large-scale datasets containing long-sequence data began to emerge. Recent deep learning based KT models face obstacles such as low efficiency, low accuracy, and low interpretability when dealing with large-scale datasets containing long-sequence data. To address these issues and promote the sustainable development of ITS, we propose a **L**STM **B**ERT-based **K**nowledge **T**racing model for long sequence data processing, namely **LBKT**, which uses a BERT-based architecture with a Rasch model-based embeddings block to deal with different difficulty levels information and an LSTM block to process the sequential characteristic in students' actions. LBKT achieves the best performance on most benchmark datasets on the metrics of ACC and AUC.

Keywords: Knowledge Tracing · BERT · Student Modelling · Long-Sequence Data Processing · Intelligent Tutoring Systems

1 Introduction

As one of the widely applied intelligent educational technologies, Knowledge Tracing (KT) has drawn a lot of attention. KT is the field of modelling students' learning trajectories and predicting their sequential actions based on historical interaction data between students and ITS [2]. With the development of ITS, large-scale datasets such as *EdNet* [5] and *Junyi Academy* [4] began to emerge. In these datasets, long-sequence student interaction data were gathered as an increasing number of students used the ITS for an extended period. The long- and

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 174–184, 2024.
https://doi.org/10.1007/978-3-031-63031-6_15

short-sequence data in these datasets are unbalanced, which satisfies the long-tail distribution [18]. For instance, within the EdNet dataset, a substantial amount of student action sequences are included, ranging from the shortest sequence that may comprise just a single action to the longest sequence that encompasses 40,157 actions. Notably, the average action sequence length of the EdNet dataset is 121.5, indicating a moderate length of data sequences overall. However, it is important to note that the distribution of sequence lengths is highly skewed, and this unbalanced distribution has an impact on the overall performance of the KT models. Although the quantity of short-sequence data is larger than the long-sequence data, the latter is of more weight than the former in prediction tasks [15].

In general, KT models could be divided into three categories: probabilistic KT models, logistic KT models, and deep learning based KT methods (DKT) [2]. Traditional probabilistic KT models and logistic KT models are forced to confront difficulties such as decreased processing efficiency and increased memory usage as growing amounts of longer sequence data are released. Deep learning based KT models are known to suffer from inefficiencies when processing long-sequence action data problems, including issues related to accuracy, speed, and memory usage [18]. Therefore, allowing the processing of very long sequence data is key to achieving high performance for next-generation KT models. Moreover, due to the black-box nature of traditional deep learning methods, the current deep learning based KT models also struggle with the lack of interpretability [8].

To address the above issues, in this paper, we propose LBKT, a novel **L**STM **B**ERT **K**nowledge **T**racing model, for processing long sequence data. The model combines the strength of the Bidirectional Encoder Representations from Transformers (BERT) model in capturing the relations of complex data [7] with the strength of the LSTM model in handling long sequential data to improve its performance on large-scale datasets containing long-sequence data (here, the long-sequence data indicates a length longer than 400 interactions). Moreover, we utilise a Rasch model-based embedding method to process the difficulty level information in the historical behaviour data of students. The Rasch model is a classic yet powerful model in psychometrics [21], which could be utilised to construct raw questions and knowledge embeddings for KT tasks [8]. Rasch model based embedding could improve the model's performance and interpretability. The experimental results show that our proposed LBKT outperforms the baseline models in five datasets on metrics ACC and AUC. Moreover, it is faster at processing long-sequence data at two long-sequence datasets we extract from the two large-scale datasets. Furthermore, we use t-SNE as the visualisation tool to demonstrate the interpretability of the embedding strategy.

The main contributions of our paper lie in the following two aspects:

1. We propose LBKT, a novel **L**STM **B**ERT **K**nowledge **T**racing model for long sequence data processing. The LBKT leverages the power of BERT, Rasch-based embedding strategies, and LSTM.
2. The experimental results show that LBKT outperforms the baseline models on five ITS datasets on the metric of AUC(assist12, assist17, algebra06, EdNet, and Junyi Academy).

2 Related Work

2.1 Knowledge Tracing

Knowledge Tracing (KT) models and predicts students' mastery levels over time in Intelligent Tutoring Systems, using observable behaviors to infer hidden knowledge states [1]. It aims to personalize feedback and instruction, enhancing learning outcomes. KT methods are categorized into probabilistic, logistic, and deep learning-based models [6,29,31].

Probabilistic models, like Bayesian Knowledge Tracing (BKT), utilize Hidden Markov Models or Bayesian Belief Networks to track learning states, but struggle with complexity and multi-skill scenarios [6,10,27,30]. Logistic models apply logistic regression to predict mastery levels, incorporating factors like prior performance and response time [3,10,20,28].

Deep learning-based KT, leveraging advancements like self-attention mechanisms and Transformer architectures, has introduced models such as SAKT and SAINT+ for higher performance through sequence prediction and attention to temporal learning dynamics [8,19,22]. BERT-based KT models, though innovative, have not surpassed state-of-the-art KT methods in handling long-sequence, large-scale datasets [11,25].

2.2 Transformer-Based Model and Application

Transformers, with self-attention mechanisms, have revolutionized NLP and image generation, exemplified by BERT and GPT [7,26]. BERT's bidirectional training and large pre-training corpus have set new benchmarks in understanding natural language, with applications extending into image processing, recommendation systems, and music generation [7,9,23]. Despite their success, BERT variants in KT have not achieved superior performance on complex, long-sequence datasets [12–14,16,17,25].

3 Methodology

3.1 Proposed Model Architecture

We propose a novel model, LBKT, for the task of knowledge tracing on large-scale datasets containing long-sequence data. While previous BERT-based KT models have shown remarkable success in capturing the relations of complex data, they also have inefficiencies when dealing with long sequence student action data [25]. On the other hand, LSTM models have been proven to excel in handling long sequential data. In response to these challenges, we propose a novel KT model that combines the strengths of both the BERT and LSTM models to improve performance on large-scale datasets containing long-sequence data (where long-sequence data indicates a length longer than 400 interactions). The Rasch embedding (also known as the 1PL IRT model) is a method to represent questions and concepts in a mathematical space [21]. The embeddings are created using a vector that summarizes the variation in questions covering a concept and a scalar difficulty parameter that controls how far a question deviates from the concept it covers. The embeddings are used as raw embeddings for questions and responses, which is a way to track a learner's knowledge state. By leveraging the strengths of a BERT-based model, Rasch model-based embeddings, and long short-term memory (LSTM) unit, our proposed model architecture has the potential to effectively process and understand relationships among different features in long-sequence data, as illustrated in Fig. 1.

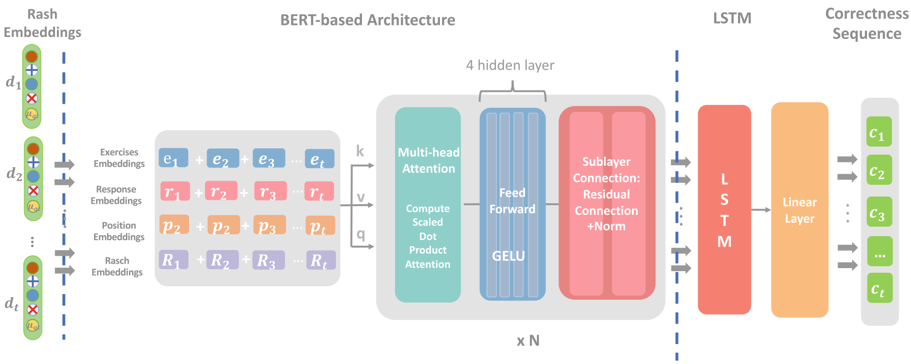


Fig. 1. The architecture of LBKT.

The first component of LBKT is the Rasch model-based embeddings proposed by Ghosh [8]. The Rasch model-based embeddings consist of difficulty level embeddings E_{d} and question embeddings E_{q}. These embeddings are multiplied and added to the BERT token embeddings and the *sin* and *cos* positional embeddings to build the final embeddings, as shown in the following equation:

$$E = E_{\mathrm{Rasch}} + E_{\mathrm{Bert\ Token}} + E_{\mathrm{Position}} \tag{1}$$

where the Rasch model-based embeddings E_{Rasch} are defined as:

$$E_{\mathrm{Rasch}} = E_{\mathrm{d}} + E_{\mathrm{d}} \times E_{\mathrm{q}} \tag{2}$$

The segment embeddings, which are typically used to represent information about the segment in the BERT model, are replaced by the Rasch embeddings mentioned above in our model's architecture. Rasch model-based embeddings are able to more accurately estimate students' knowledge states, as explained earlier, making them a key contributor to the effectiveness of LBKT for knowledge tracing tasks.

The second component of LBKT is a BERT-based block, which consists of 12 Transformer blocks. Each includes a multi-head attention mechanism, a feed-forward network (FFN), and sublayer connections. The multi-head attention mechanism uses the "Scaled Dot Product Attention" method as implemented in BERT, along with queries Q, keys K, values V, and an attention mask for padded tokens. The FFN has a feedforward hidden layer with a size of four times that of the model's hidden layer and uses the GELU activation function rather than RELU.

The sublayer connections in the Transformer block include a residual connection followed by layer normalization. The formulas for the attention mechanism and the FFN are as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

$$\text{FFN}(x) = \text{GELU}(W_1 x + b_1)W_2 + b_2 \tag{4}$$

In the third component of LBKT, we use a neural network (NN) linear transformation instead of the attention projection typically used in conjunction with the LSTM unit. This is based on our observed improved performance with the NN linear transformation in our experiments. It should be noted that this choice is not necessarily related to the length or complexity of the sequence but rather to the specific characteristics of the data and the task at hand.

Overall, LBKT is a model that is tailored specifically for use in the field of knowledge tracing. It combines the natural language processing capabilities of the BERT model with the ability to accurately estimate knowledge states using Rasch model-based embeddings and the ability to effectively handle long sequences of data using the LSTM unit and the NN linear transformation. This makes it an ideal choice for the task of knowledge tracing in large-scale datasets containing long-sequence data with unbalanced data distribution.

3.2 Experiment Setting

Datasets. We used five benchmark datasets to validate the effectiveness of the LBKT model, including assist12[1], assist17[2], algebra06[3], EdNet [5][4], and Junyi Academy [4][5]. In general datasets, such as assist 12 and assist 17, it could be challenging to identify and extract large amounts of long-sequence data. Therefore, we validated the speed performance of every model on two datasets with long-sequence student action data extracted from EdNet and Junyi Academy. The mean action sequence length of EdNet is 121.5. The mean interaction length of Junyi Academic is 104.7. Here, we define the longer action sequence as longer than 100 records. We extract 200 students' action sequences that include interactions longer than 100 actions from each dataset as the long-sequence dataset to validate the performance of different KT models. Lastly, we selected different

1 https://sites.google.com/site/assistmentsdata/home.
2 https://sites.google.com/site/assistmentsdata/home.
3 https://pslcdatashop.web.cmu.edu/KDDCup.
4 https://github.com/riiid/ednet.
5 https://pslcdatashop.web.cmu.edu/Files?datasetId=1275.

lengths of action sequences from Ednet to test the speed performance of each model. We selected four groups with average records lengths of 100, 200, 300, and 400, respectively. Each of these groups included 50 students.

Baseline Models. We compared our LBKT to three state-of-the-art models, BEKT [25], AKT [8], DKVMN [24], as well as the two top baseline models in the Riiid Answer Correctness Prediction Competition provided by Kaggle[6], including SSAKT [32], and LTMTI [5].

Evaluation Metrics and Validation. We used the accuracy (ACC) and the area under the curve (AUC) as performance metrics to compare the models' performance in five datasets.

Hyperparameters for Experiments. To compare with each model, the same parameters were used for model training. The batch size was set to 64, and the train/test split was 0.8/0.2. The model used an embedding size of 128 and the Adam optimizer with a learning rate of 0.001. The loss function used was the Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss). The scheduler was set to OneCycleLR with a maximum learning rate of 0.002. Dropout was also being used at a rate of 0.2. The training ran for a total of 100 epochs, with early stopping set to 10 epochs. If the validation loss does not decrease for the first three epochs, the training stops, in order to prevent overfitting and save resources. The maximum sequence length was 200, with an eight-attention head. Hidden sizes were 128 for BERT, 512 for FFN, and 128 for LSTM. The Transformer block/encoder layer was set to 12.

4 Results and Discussion

4.1 Overall Performance

LBKT outperforms four baseline models on most metrics in the experiments on five benchmark datasets. Table 1 shows the overall performance of each model. We used five-fold cross-validation to estimate their performances. LBKT performed the best on EdNet and Junyi Academy datasets on both ACC and AUC metrics. It also achieved the best performance on the ACC metric on assist12 and AUC on assist17. On algebra06, AKT achieved the best performance on the ACC metric, BEKT achieved the best performance on the AUC metric, and LBKT achieved the second-best performance on both metrics. This result indicates that LBKT is an efficient KT model on most datasets, especially large-scale datasets containing long-sequence interaction data. This was affected by the unique architecture of our LBKT model. The LSTM block enables the model to learn the sequential features of the long sequence and gives more importance to the recent actions of the students, which prevents the model from giving too much weight to the long-ago and low-relevance actions and thus improving the training efficiency.

Table 2 shows the performance comparison on the two large-scale datasets. On both datasets, LBKT achieved the best training efficiency. It was 4.29x faster

[6] https://www.kaggle.com/code/datakite/riiid-answer-correctness.

Table 1. Comparison of different KT models on five benchmark datasets. The best performance is denoted in bold.

Dataset	Metrics	LBKT	BEKT	SSAKT	LTMTI	AKT	DKVMN
assist12	ACC	**0.799**	0.786	0.675	0.813	0.769	0.756
	AUC	0.768	**0.813**	0.741	0.785	0.753	0.701
assist17	ACC	0.792	0.795	0.771	0.796	0.733	**0.797**
	AUC	**0.814**	0.801	0.735	0.683	0.803	0.709
algebra06	ACC	0.801	0.797	0.795	0.811	**0.831**	0.800
	AUC	0.799	**0.815**	0.774	0.791	0.814	0.793
EdNet	ACC	**0.803**	0.781	0.761	0.799	0.756	0.800
	AUC	**0.815**	0.795	0.798	0.802	0.798	0.796
Junyi Academy	ACC	**0.832**	0.807	0.777	0.797	0.791	0.790
	AUC	**0.851**	0.831	0.845	0.812	0.799	0.769

than BEKT on EdNet and 4.77x faster than BEKT on Junyi Academy. Compared with the second-best model, AKT, LBKT was 1.32x faster on EdNet and 1.42x faster on Junyi Academy. For the memory cost, LBKT was about one-third of BEKT and lower than LTMTL on both datasets. Although the memory cost of LBKT was not the smallest, LBKT has achieved the best results in both ACC and AUC metrics running on the same GPU. This allows LBKT to run on middle-range GPUs. To improve the training efficiency, we used a last input as the query method in the Transformer block instead of the whole sequence, which decreased the complexity of the encoder to improve training speed and reduce memory cost.

Table 2. Performance comparison on the two large-scale datasets, EdNet and Junyi Academy. The best performance is denoted in bold.

Model	EdNet			Junyi Academy		
	speed ↑	speed ratio ↑	memory ↓	speed↑	speed ratio ↑	memory ↓
BEKT	4.93	1.00x	16.7 GB	4.85	1.00x	16.6 GB
SSAKT	7.13	1.44x	**3.4 GB**	6.22	1.28x	**3.2 GB**
LTMTI	13.8	1.32x	7.69 GB	12.1	1.19x	8.82 GB
AKT	17.1	3.25x	4.32 GB	16.4	3.35x	4.37 GB
DKNMN	5.97	2.34x	7.68 GB	4.67	3.75x	8.53 GB
LBKT	**21.3**	**4.29x**	6.09 GB	**22.2**	**4.77x**	6.08 GB

4.2 Analysis of Embedding Strategy

In this section, We used t-SNE as the visualisation tool to show the interpretability of LBKT's embedding strategy. Figure 2-*left* shows the results of No-Rasch-embedding, and Fig. 2-*right* shows the Rasch embedding strategy. We can see that, in the No-Rasch-embedding scenario, the difficult questions' embeddings (dark blue vectors) mixed with the easy questions' embeddings (yellow to light blue vectors). In Fig. 2-*right*, the difficult level embeddings were separated to avoid mixing with easy level embeddings.

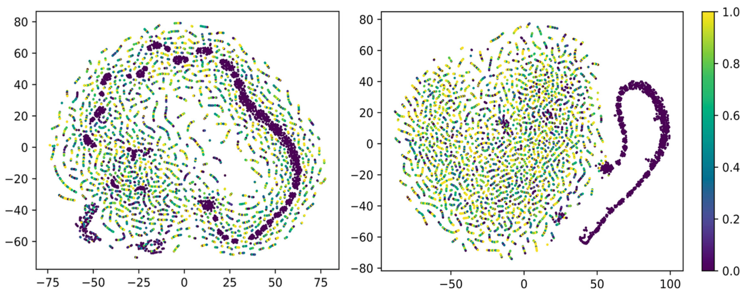


Fig. 2. Visualisation of the embedding vector using t-SNE: *without* Rasch embeddings (on the left) and *with* Rasch embeddings (on the right). The colour bar is the predicted probability of the outputs.

Questions at a higher difficulty level are typically associated with longer sequence data, as students spend more time and steps on difficult exercises, which results in longer interaction sequences. Rasch model-based embeddings could divide different difficulty-level parts before the start of the model training and not mix them with other difficulty-level embeddings. As a result, it might increase training efficiency to converge faster.

5 Conclusion

In this study, we have developed LBKT, which employs a BERT-based architecture with an LSTM block for processing long-sequence data, and Rasch model-based embeddings for different difficulty levels of questions. Experiments show that LBKT outperforms baseline models on most benchmark datasets. We also conducted the speed performance experiment on the two large-scale datasets containing long-sequence data. The results suggest that LBKT could process long-sequence data faster and is more resource-efficient. Furthermore, we conducted an analysis of the embedding strategy using t-SNE. The result shows that Rasch embedding could process the difficulty-level features effectively.

Acknowledgments. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems (https://ccais.ac.uk/) and through the AutoTrust Platform Grant (EP/R029563/1).

References

1. Abdelrahman, G., Wang, Q., Nunes, B.: Knowledge tracing: a survey. ACM Comput. Surv. **55**(11), 1–37 (2023)
2. Abdelrahman, G., Wang, Q., Nunes, B.P.: Knowledge tracing: a survey. ACM Comput. Surv. **55**, 1–37 (2022)
3. Cen, H., Koedinger, K., Junker, B.: Learning factors analysis – a general method for cognitive model evaluation and improvement. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 164–175. Springer, Heidelberg (2006). https://doi.org/10.1007/11774303_17
4. Chang, H.S., Hsu, H.J., Chen, K.T.: Modeling exercise relationships in e-learning: a unified approach. In: EDM, pp. 532–535 (2015)
5. Choi, Y., et al.: EdNet: a large-scale hierarchical dataset in education. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 69–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7_13
6. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adap. Inter. **4**(4), 253–278 (1994). https://doi.org/10.1007/BF01099821
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
8. Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2330–2339 (2020)
9. Jiang, Z.H., Yu, W., Zhou, D., Chen, Y., Feng, J., Yan, S.: ConvBERT: improving BERT with span-based dynamic convolution. Adv. Neural. Inf. Process. Syst. **33**, 12837–12848 (2020)
10. Käser, T., Klingler, S., Schwing, A.G., Gross, M.: Dynamic Bayesian networks for student modeling. IEEE Trans. Learn. Technol. **10**, 450–462 (2017)
11. Lee, U., Park, Y., Kim, Y., Choi, S., Kim, H.: MonacoBERT: monotonic attention based convBERT for knowledge tracing. arXiv preprint arXiv:2208.12615 (2022)
12. Li, Z.: Deep reinforcement learning approaches for technology enhanced learning. Ph. D. thesis, Durham University (2023)
13. Li, Z., Jacobsen, M., Shi, L., Zhou, Y., Wang, J.: Broader and deeper: a multi-features with latent relations BERT knowledge tracing model. In: Viberg, O., Jivet, I., Muñoz-Merino, P., Perifanou, M., Papathoma, T. (eds.) European Conference on Technology Enhanced Learning, pp. 183–197. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42682-7_13
14. Li, Z., Shi, L., Cristea, A., Zhou, Y., Xiao, C., Pan, Z.: SimStu-transformer: a transformer-based approach to simulating student behaviour. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) International Conference on Artificial Intelligence in Education, pp. 348–351. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11647-6_67

15. Li, Z., Shi, L., Cristea, A.I., Zhou, Y.: A survey of collaborative reinforcement learning: interactive methods and design patterns. In: Proceedings of the 2021 ACM Designing Interactive Systems Conference, pp. 1579–1590 (2021)
16. Li, Z., Shi, L., Wang, J., Cristea, A.I., Zhou, Y.: Sim-GAIL: a generative adversarial imitation learning approach of student modelling for intelligent tutoring systems. Neural Comput. Appl. **35**(34), 24369–24388 (2023)
17. Li, Z., Shi, L., Zhou, Y., Wang, J.: Towards student behaviour simulation: a decision transformer based approach. In: International Conference on Intelligent Tutoring Systems, pp. 553–562. Springer (2023). https://doi.org/10.1007/978-3-031-32883-1_49
18. Liu, Y., Zhou, J., Lin, W.: Efficient attentive knowledge tracing for long-tail distributed records. In: 2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD), pp. 104–109. IEEE (2021)
19. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. arXiv preprint arXiv:1907.06837 (2019)
20. Pavlik Jr, P.I., Cen, H., Koedinger, K.R.: Performance factors analysis–a new alternative to knowledge tracing. In: Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009, July 6-10, 2009, Brighton, UK, vol. 200, pp. 531–538 (2009)
21. Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Tests. (1993)
22. Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., Choi, Y.: Saint+: integrating temporal features for EdNet correctness prediction. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 490–496 (2021)
23. Sun, F., et al.: BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1441–1450 (2019)
24. Sun, X., Zhao, X., Ma, Y., Yuan, X., He, F., Feng, J.: Muti-behavior features based knowledge tracking using decision tree improved DKVMN. In: Proceedings of the ACM Turing Celebration Conference-China, pp. 1–6 (2019)
25. Tiana, Z., Zhengc, G., Flanaganb, B., Mic, J., Ogatab, H.: BEKT: deep knowledge tracing with bidirectional encoder representations from transformers. In: Proceedings of the 29th International Conference on Computers in Education (2021)
26. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
27. Villano, M.: Probabilistic student models: Bayesian belief networks and knowledge space theory. In: International Conference on Intelligent Tutoring Systems, pp. 491–498. Springer (1992). https://doi.org/10.1007/3-540-55606-0_58
28. Wang, J., Ivrissimtzis, I., Li, Z., Shi, L.: Comparative efficacy of 2D and 3D virtual reality games in American sign language learning. In: The 31st IEEE Conference on Virtual Reality and 3D User Interfaces. Newcastle University (2024)
29. Wang, J., Ivrissimtzis, I., Li, Z., Shi, L.: Impact of personalised AI chat assistant on mediated human-human textual conversations: exploring female-male differences. In: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, pp. 78–83 (2024)
30. Wang, J., Ivrissimtzis, I., Li, Z., Zhou, Y., Shi, L.: Exploring the potential of immersive virtual environments for learning American sign language. In: European Conference on Technology Enhanced Learning, pp. 459–474. Springer (2023). https://doi.org/10.1007/978-3-031-42682-7_31

31. Wang, J., Ivrissimtzis, I., Li, Z., Zhou, Y., Shi, L.: User-defined hand gesture interface to improve user experience of learning American sign language. In: International Conference on Intelligent Tutoring Systems, pp. 479–490. Springer (2023). https://doi.org/10.1007/978-3-031-32883-1_43
32. Zhang, X., Zhang, J., Lin, N., Yang, X.: Sequential self-attentive model for knowledge tracing. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) ICANN 2021. LNCS, vol. 12891, pp. 318–330. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86362-3_26

Educational Support for Automated Classification of UML Diagrams Using Machine Learning

Irina-Gabriela Nedelcu, Veronica Opranescu, Beatrice-Nicoleta Chiriac, and Anca Daniela Ionita(✉)

National University of Science and Technology POLITEHNICA Bucharest, Splaiul Independentei 313, 060042 Bucharest, Romania
{irina.nedelcu,veronica.opranescu}@stud.acs.upb.ro, {beatrice.chiriac,anca.ionita}@upb.ro

Abstract. As engineering is very much based on modeling, this is also important for education in this field, and teachers sometimes have to check a very large number of models and determine if they are valid or not. In software engineering, for modeling in conformity with the standard Unified Modeling Language, attempts have been made to automatically classify diagrams and determine whether they conform to this language. This paper shows an approach based on machine learning and possible improvements made using a feature-based dataset, with the objective of more accurately categorizing designated labels. Employing a specialized neural network tailored for feature-based learning, the study endeavors to enhance classification accuracy and efficiency. Comparative analysis against a pre-existing model trained on a diagram images dataset reveals better results in predictive outcomes.

Keywords: Neural networks · Unified Modeling Language · Teaching tools

1 Introduction

Artificial Intelligence (AI) has definitely changed the education tools, with impact on a more efficient and personalized learning for students, but also on grading and assessment support for teachers, gaining many advantages from pre-trained models [1][2]. Educational tools concepts are frequently identified in the smart campus context [3], where they may be used to optimize learning experiences, automate administrative processes, and improve collaboration between students and teachers. Among their features, one can notice:

– virtual and augmented reality [4] for design visualization and prototyping, enabling students to manipulate 3D models in real-time,
– e-learning platforms, also known as Learning Management Systems [5], for organizing course materials, assignments, and communication channels, and

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 185–192, 2024.
https://doi.org/10.1007/978-3-031-63031-6_16

– chatbot or virtual assistant recommendation services [6] for instant support on course information, accessing resources and professional guidance for students, within decision processes [7], integrating machine learning techniques.

This paper approaches a teachers' perspective in the domain of software engineering education, where the use of a standard object-oriented modeling language, like Unified Modeling Language (UML) [8], is part of any university curriculum. UML has 14 types of diagrams, of which some commonly used types are class diagrams, state machine diagrams, use case diagrams, and activity diagrams. This paper continues with the description of the problem in Sect. 2, then it compares related work with results obtained prior to the current experiment, in Sect. 3. Section 4 proposes a feature dataset used for the neural network training and it also evaluates the results. Section 5 represents the conclusion and highlights the next steps.

2 Problem Description

Making use of a deep learning model, designed for classifying UML diagrams, serves as a first step with diverse follow-ups in education. The capacity to facilitate automatic classification is particularly beneficial for educational environments. The model could then be integrated to identify similarities and differences among diagrams submitted by different students, providing insights into common misconceptions or areas where additional instruction may be needed. Tavares et al. highlight the existence of UML diagrams as a visual component in many educational books and courses [9].

A deep learning model will promote fairness and objectivity towards students' educational process by relying on an automated model, educators reduce bias and subjectivity in grading. The model treats all diagrams equally, adhering to predefined classification rules. Educators can review flagged similarities or differences and provide targeted feedback.

The model can contribute to educational insights where data from the model can offer insights into common misconceptions or areas where students struggle. Are sequence diagrams consistently misunderstood? Do students tend to omit certain elements in state machine diagrams? Such information informs instructional adjustments and targeted teaching interventions.

The upcoming research endeavors to address this limitation by shifting the focus towards the pivotal stage of feature extraction as experimented in [10]. By simplifying and highlighting crucial aspects of images, feature extraction empowers classification models to distinguish between different objects, enabling them to generalize and make informed decisions, ultimately bolstering the accuracy and reliability of image classification systems.

3 Related Work

Previous experiments concluded that feature-based classification is improving the prediction score [11]. Feature extraction remains fundamental in enhancing image classification across diverse domains, leveraging recent advancements in research. Notable contributions in this space include [12], which introduced deep convolutional neural networks (CNNs) for image classification. This seminal work emphasized the critical role of feature extraction through convolutional layers in achieving remarkable performance.

Another significant advancement is seen in [13]. This paper focused on the Bag of Visual Words (BoVW) model for action recognition, showcasing the importance of robust feature extraction techniques in complex video data analysis. Moreover, recent work like [14] highlighted the efficacy of Convolutional Deep Belief Networks for scene labeling, emphasizing the significance of learned representations in image classification tasks.

By comparing research papers aiming to investigate the UML diagrams classification using images, we identified the following results as similarities and differences in respect to similar research papers (see Table 1).

Table 1. Research Characteristics Comparison

Study	Nedelcu et al. [15]	Shcherban et al. [16]	Gosala et al. [17]	Chen et al. [18]
Image is a UML diagram or not	Yes	Yes	No	No
Diagram type classification	3 categories: class, activity, sequence	10 categories: class, activity, use case, sequence, communication, component, deployment, object, package, and state machine	1 category: class	1 category: class
Evaluated neural networks	2	7	1	1
Number of samples for each category	Balanced, 500	Imbalanced, 204 (the smallest dataset) to 720 (the biggest dataset)		Balanced, 650

Nithyashree also proposes a method to identify features from class diagram [19]. This approach could be used to expand the dataset of features we shall use for the following research. However, the approach needs enhancements to make use of it and it could not be used in our experiments.

4 Feature Extraction for UML Diagrams Classification

This section describes key parts of using a deep learning model with a feature-based dataset. It explains the need and what benefits have been observed in existing research papers.

4.1 Steps to Follow

Integrating feature extraction into one's classification process can significantly enhance the accuracy and robustness of the model in classifying images containing UML diagrams (see Fig. 1). A similar process is described in [20]. To optimize the classification process, it is vital to extract key features from UML diagrams. These features encompass essential components like classes, arrows indicating inheritance or associations, varied shapes, textual elements, and any other distinct attributes that delineate diverse UML diagram types. Gathering these specific features is critical in enabling the model to discern and classify UML diagrams accurately, capturing the nuanced characteristics unique to each diagram type.

Fig. 1. Summary of feature-based classification

By integrating these extracted features into one's deep learning model, one can potentially capture nuanced information present in UML diagrams, enabling the model to make more informed and accurate classifications. It is essential to maintain a balance between feature richness and complexity, ensuring that the added features contribute meaningfully to the classification task, without overwhelming the model.

4.2 Dataset

In addition to our previous research [15], each UML diagram was extended to 500 samples for the training dataset. To create a data set of features, we used the UML class diagrams data set to extract features such as classes, associations, or inheritance. The other group of data set features contained parts of UML and non-UML component extracted from the previous dataset.

Unified Modeling Language Specification Version 2.5.1 [21] was used as an official source, to identify the notations that should be present in the feature dataset. The class diagram is widely described across the entire specification document and is covered in many sections. We thus extracted images for the features from Table 2.

In addition to the features from the UML specification document, we used a platform built based on the official documentation, which and explains the concepts in a summarized manner [22].

Each feature image contains a detail that is part of a UML diagram and conforms to the concrete syntax of the language. Each image was resized to 30 by 30 pixels, but we extended the data set from 500 samples to 1000 per feature.

In previous research [15] we used documentation to extend the existing data set. The chosen data is that augmentation techniques can be applicable to the extracted features such as translation, colorization, and resizing. These techniques were also applied to the existing data set to expand the images from 1000 to 2000 samples.

Table 2. UML class diagram features

UML Class Diagram Features		
Class	Enumeration	Composite aggregation / Composition
Abstract class	Association	Association Navigability
Interface	Operation	Generalization
Nested classifier	Constraint	Dependency
Object	Multiplicity	Required interface

Downsizing in terms of image size is expected to speed up the training process, whereas increasing the number of samples is expected to increase accuracy when the model sees certain features. This approach may solve the challenges of small data sets, improve the accuracy score and manage to better classify what the image contains.

4.3 Neural Network Architecture

For the given dataset, we applied a customized deep neural network architecture designed to handle a feature-rich dataset. The architecture, called multi-input Multi-Layer Perceptron (MLP), processes multiple sets of features separately, before combining them for classification, as explained in [23]. Customized architectures like this allow the network to learn and extract complex relationships from different sets of engineered features, potentially enhancing the classification performance on feature-rich datasets. Multi-input MLP is a type of neural network architecture that handles multiple inputs, processes them through independent paths (sub-networks), and then merges or concatenates the information before making final predictions. It is particularly useful for scenarios where different types of features need distinct processing before combining for a classification or regression task. In the architecture of neural networks with independent processing paths, for each stream of input data undergoes individual processing within its dedicated sub-network or layer set.

This independent processing enables the network to learn nuanced and specialized representations, unique to each input data stream. The key advantage of multi-input MLPs is their ability to handle heterogeneous data types or feature sets by allowing each set of features to be processed separately before integration. This architecture is particularly beneficial when dealing with feature-rich datasets or when different types of information need distinct processing pathways before being combined for the classification task. Shcherban et al. also talk about improving image classification by focusing on features [16].

4.4 Examples of Results

Throughout the research phases, multiple experiments were conducted, with one designated as the baseline test at each stage. In each evaluation, classification was performed on a fixed set of 25 samples to assess the improvement of each approach. The original model, obtained prior to this research, correctly classified 22 samples, a performance matched by the new model. Moreover, the accuracy score increased by approximately 1.8% during the second iteration on average.

Previously, the model achieved a 91% F1 score [15], whereas the resulting model from this research achieved a 92% F1 score, reflecting an expected increase due to higher accuracy observed across testing samples. The scores of each model indicate their suitability for a production-ready environment. Any misclassification, if identified, or any correct classification with low accuracy scores, should be manually reviewed by a human operator.

Feature-based classification, observed across numerous instances, exhibits an elevated accuracy in prediction scores. This trend may be seen in the examples from Fig. 2, amalgamating outcomes from prior experiments. Specifically, in the context of class diagrams, the classification achieved a 92% accurate prediction score.

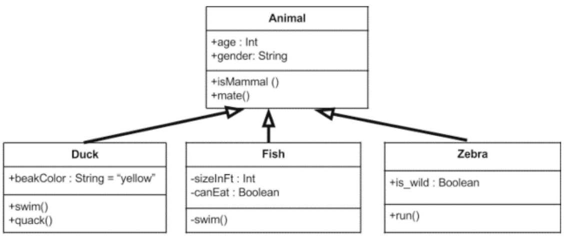
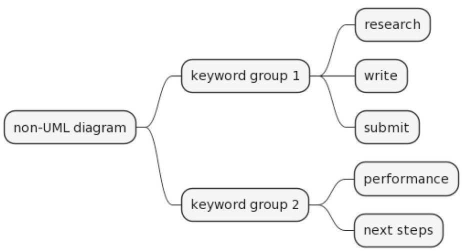
Image	Is Class diagram?
	Yes (92%)
	No (88%)

Fig. 2. Example of feature-based classification for non-UML diagram

5 Conclusion

The paper presented our experiments of UML diagrams classification based on images and utilizing a feature-based dataset, exploring its efficacy during training, and evaluating its influence on the prediction score. Notably, enhancements were observed in class diagram categorization; however, the model fails for some samples that were successfully classified with previous networks. For the samples used for model validation we obtained the following results: precision 0.95, recall 0.91, F1 score 0.93. This performance is better than in our previous research [15]. The accuracy score suggests that the model is suitable for integration into educational systems. In addressing specific problems, students may be expected to realize certain types of UML diagrams, and the model can determine whether their work is valid. This enables teachers to pinpoint areas where students may have misconceptions, allowing for targeted classroom instruction.

Any misclassification may be the result of the overlapping features shared between these diagram types, which suggests a limitation in the model ability to recognize connections crucial for improved classification. Addressing this issue in future work, by minimizing or eliminating overlapping components, expected to elevate the classification score and mitigate potential confusion within the model. As the next steps we aim to expand the model to include other previously investigated categories, such as sequence and activity diagrams.

Acknowledgments. This work was supported by a grant from the National Program for Research of the National Association of Technical Universities – GNAC ARUT 2023".

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Grassini, S.: Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. Educ. Sci. **13**(7), 692 (2023). https://doi.org/10.3390/educsci13070692
2. Nguyen, N.D.: Exploring the role of AI in education. Lond. J. Soc. Sci. **17**, 84–95 (2023). https://doi.org/10.31039/ljss.2023.6.108
3. Maresca, P., Coccoli, M., Guercio, A., Stanganelli, L.: Smarter universities: a vision for the fast changing digital era. J. Vis. Lang. Comput. **25**, 1003–1011 (2014). https://doi.org/10.1016/j.jvlc.2014.09.007
4. Elmqaddem, N.: Augmented reality and virtual reality in education. Myth or reality? Int. J. Emerg. Technol. Learn. **14**(03), 234 (2019). https://doi.org/10.3991/ijet.v14i03.9289
5. Coşkunçay, D., Alkış, N., Özkan-Yıldırım, S.: A structural model for students' adoption of learning management systems: an empirical investigation in the higher education context. Educ. Technol. Soc. **21**, 13–27 (2018)
6. Hwang, G.-J., Chang, C.-Y.: A review of opportunities and challenges of chatbots in education. Interact. Learn. Environ. **31**, 4099–4112 (2021). https://doi.org/10.1080/10494820.2021.1952615
7. Opranescu, V., Nedelcu, I., Ionita, A. D.:Automating students' decision processes in a smart campus. In: 2023 13th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, pp. 1–6 (2023).https://doi.org/10.1109/ATEE58038.2023.10108094
8. Ho-Quang, T., Chaudron, M.R., Samuelsson, I., Hjaltason, J., Karasneh, B., Osman, H.: Automatic classification of UML class diagrams from images. In: 2014 21st Asia-Pacific Software Engineering Conference, vol. 1, pp. 399–406 (2014)

9. Tavares, J.F., Costa, Y.M. G., Colanzi, T.E.: Classification of UML diagrams to support software engineering education. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW), Melbourne, Australia, 2021, pp. 102–107 (2021). https://doi.org/10.1109/ASEW52652.2021.00030
10. Adegun, A.A., Viriri, S., Tapamo, J.R.: Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis. J Big Data **10**, 93 (2023). https://doi.org/10.1186/s40537-023-00772-x
11. Yim, J., Jeongwoo, J., Jung, H., Kim, J.: Image classification using convolutional neural networks with multi-stage feature. In: Kim, J.H., Yang, W., Jo, J., Sincak, P., Myung, H. (eds.) Robot Intelligence Technology and Applications 3: Results from the 3rd International Conference on Robot Intelligence Technology and Applications, pp. 587–594. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-16841-8_52
12. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems, vol. 25 (2012). https://doi.org/10.1145/3065386
13. Wang, L., Peng, X., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput. Vis. Image Underst. **150** (2014). https://doi.org/10.1016/j.cviu.2016.03.013
14. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013). https://doi.org/10.1109/TPAMI.2012.231
15. Nedelcu, I.G., Caramihai, S.I., Mocanu, S., Ionita, A.D.: UML diagrams classificatio with MIMO and SISO Neural Networks. U.P.B. Sci. Bull. Ser. C **85**(4) (2023), ISSN 2286–3540
16. Shcherban, S., Liang, P., Li, Z., Yang, C.: Multiclass classification of UML diagrams from images using deep learning. Int. J. Softw. Eng. Knowl. Eng. **31**, 1683–1698 (2021)
17. Gosala, B., Chowdhuri, S.R., Singh, J., Gupta, M., Mishra, A.: Automatic classification of UML class diagrams using deep learning technique: convolutional neural network. Appl. Sci. **11**, 4267 (2021)
18. Chen, F., Zhang, L., Lian, X., Niu, N.: Automatically recognizing the semantic elements from UML class diagram images. J. Syst. Softw. **193**, 111431 (2022)
19. Nithyashree, V.: Step-by-step guide for image classification on custom datasets (2021). https://www.analyticsvidhya.com/blog/2021/07/step-by-step-guide-for-image-classification-on-custom-datasets/. Accessed 2023
20. Unified Modeling Language 2.5.1., December 2017, Object Management Group, OMG Document Number: formal/2017–12–05
21. Fakhroutdinov, K.: UML class diagrams reference (2017). https://www.uml-diagrams.org/class-reference.html. Accessed 21 Jan 2024
22. Castro, W., Oblitas, J., Santa-Cruz, R., Avila-George, H.: Multilayer perceptron architecture optimization using parallel computing techniques. PLoS ONE **12**(12), e0189369 (2017). https://doi.org/10.1371/journal.pone.0189369
23. Chollet, F., Allaire, J.J.: Image classification on small datasets with Keras (2017). Accessed 2023

Model Decomposition of Robustness Diagram with Loop and Time Controls to Petri Net with Considerations on Resets

John Albert R. Castillo(✉) and Jasmine A. Malinao

Division of Natural Sciences and Mathematics, University of the Philippines Tacloban College, Tacloban City, Philippines
{jrcastillo6,jamalinao1}@up.edu.ph

Abstract. Robustness Diagram with Loop and Time Controls is a multidimensional workflow model that can capture all workflow dimensions, i.e. process, resource, and case. In contrast to other workflow models, such as Class Diagram and Petri Net, tools that support its automated implementation and verification are not yet well-established. As such, previous literature introduced the mapping of the Robustness Diagram with Loop and Time Controls to Petri Net, focusing on the reachability and reuse of substructures therein. However, this mapping is still incomplete since it lacks the support for component reuse of the reset-bound subsystems. Furthermore, this mapping results in unsound and inconsistent Petri Net. This study proposes a mapping of the Robustness Diagram with Loop and Time Controls to Petri Net that addresses the gaps in the literature. Using Petri Net components and extensions, this paper introduces a novel mapping of the reset-bound subsystems that allows component reset and reuse and modifies some structures in PN that lead to its unsoundness and inconsistency. Lastly, the proposed mapping algorithm's overview, validation, and analysis are presented.

Keywords: Workflows · Worflow models · Robustness Diagram with Loop and Time Controls · Petri Net · Resets · Soundness

1 Introduction

The Robustness Diagram with Loop and Time Controls (RDLT) is a multidimensional workflow model that can capture all three workflow dimensions, i.e. process, resource, and case [1], making it a powerful tool for representing simple and complex systems. RDLT introduces the concept of reset-bound subsystems (RBS), allowing the simulation of reset and component reuse in other workflow models. The RDLT can also restrict the number of traversals along an arc using its L-attribute. Real-world applications of RDLTs include profiling the efficiency of the Philippine Integrated Disease Surveillance and Response (PIDSR) system [2], and modeling the Adsorption Chiller system [3], among other things. Unlike other models, such as Class Diagram (CD) and Petri Net (PN), no automated

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 193–201, 2024.
https://doi.org/10.1007/978-3-031-63031-6_17

tool supports the implementation, verification, and simulation of RDLT. This prompted the exploration of [4] in mapping RDLT to CD and PN. However, the mapping from RDLT to PN did not include the L and M-attributes of the RDLT. To solve this, [5] formulated a novel mapping of RDLT to PN that takes into consideration the L and M-attributes of an RDLT. This paper also reported the soundness of the output PN using the mapping. Unfortunately, this mapping could not fully capture the reset functionality of the RBS. This implies that tokens inside the RBS equivalent PN are not replenished once this structure is exited. To contribute to the development of RDLT, this paper aims to formulate a mapping algorithm of RDLT to PN with considerations of the full functionality of RBS. This paper also proposes modifications in mapping some RDLT structures, such as MIX-join and SPLIT structures, that cause the output PN to be inconsistent with the input RDLT in terms of its property and behavior. Lastly, this paper compares the soundness of the input RDLT and the output PN. This paper is structured as follows: Sect. 2 introduces PN, Sect. 3 introduces RDLT and some of its properties used in this research, Sect. 4 focuses on analyzing previous(proposed) RDLT to PN mapping, and Sect. 5 contains the conclusions and recommendations.

2 Petri Net

2.1 Definition [3]

A PN is a graph consisting of two node types, transitions and places. These nodes are linked together using directed arcs. A place p is considered an input(output) place of a transition t if there is a directed arc from $p(t)$ to $t(p)$. t is enabled if all of its input places have at least one token. An enabled transition may fire, producing tokens in its output places. A firing sequence σ is a composition of all transitions fired to reach a marking. A marking M_n is reachable from M_1 in a PN, denoted as $M_1 \xrightarrow{*} M_n$ if and only if there is a firing sequence usable to reach M_n from M_1. A PN is considered a Worfklow Net (WfN) if it has two additional places, i and o, where $i(o)$ has no input(output) nodes.

2.2 Classical Soundness in PN [6]

A Workflow Net PN is classical sound if it satisfies the condition of option to complete, proper completion, and liveness. The first condition requires that every reachable marking will eventually reach the final marking o. The second condition requires that all the other places within the net contain no tokens, whenever a token reaches o. The last condition requires that every transition is part of at least one firing sequence executed in the net.

2.3 Components and Variants of PN

A **reset arc** [5] is an arc with a double-tip arrow. Unlike the regular arc, it does not impact the enabling of a transition. If a place p is connected to a transition t with a reset arc, all the tokens in p will be removed once t is fired. Meanwhile, an **arc weight** [7] signifies the number of tokens required(produced) by an input(output) arc of a transition.

3 Robustness Diagram with Loop and Time Controls

3.1 Definition [3]

A Robustness Diagram with Loop and Time Controls (RDLT) is a graph representation $R = (V, E, T, M)$ where:

- V is a finite set of vertices, where $V_{type} : V \rightarrow \{'b',' e',' c'\}$ where 'b', 'e', and 'c' means the vertex is either a "boundary object", an "entity object", or a "controller", respectively.
- A finite set of arcs $E \subseteq (V \times V) \backslash E'$ where $E' = \{(x, y)|x, y \in V, V_{type}(x) \in \{'b',' e'\}, V_{type}(y) \in \{'b',' e'\}$ with the following attributes with user-defined values,
  - $C : E \rightarrow \Sigma \cup \{\epsilon\}$ where Σ is a finite non-empty set of symbols and ϵ is the empty string. $C(x, y) \in \Sigma$ means that $C(x, y)$ is a condition to be satisfied to proceed from x to y. Meanwhile, $C(x, y) = \epsilon$ means there is no condition imposed by (x, y) or signifies that x is the owner of y.
  - $L : E \rightarrow \mathbb{Z}$ is the maximum number of traversals allowed on the arc.
- Let T be a mapping such that $T((x, y)) = (t1, \ldots, tn)$ for every $(x, y) \in E$ where $n = L((x, y))$ and $t_i \in \mathbb{N}$ is the time a check or traversal is done on (x, y) by some algorithm's walk on R.
- $M : V \rightarrow \{0, 1\}$ indicates whether $u \in V$ is a center of an RBS. An RBS is a substructure G_u of R that is induced by a center $u \in V$, i.e. $M(u) = 1$, and the set of controllers owned by u. (x, y) is an in-bridge of G_u if x is not a vertex in G_u but y is. Conversely, (x, y) is an out-bridge of G_u if x is a vertex in G_u, but y is not. Lastly, a pair of arcs (a, b) and (c, d) are type-alike with respect to y if (a, b) and (c, d) are both in/out-bridges of y, or both are not.

3.2 Vertex-Simplified RDLT [3]

Vertex simplification converts all types of vertices in the RDLT into a controller type. If there is an RBS, two RDLTs will be produced, the level-1 and level-2 vertex-simplification of R.

3.3 Extended RDLT [3]

An extended RDLT R' includes two additional vertices, i and o. These are connected to the source and sink vertices of the original RDLT, respectively.

3.4 Activity Extraction in RDLT [3]

The activity extraction algorithm is used to extract an activity profile. An activity profile $S = \{S(1), S(2), \ldots, S(K)\}, k \in \mathbb{N}$, is a set of reachability configurations $S(j)$ that contains the set of arcs traversed at time step j.

3.5 Expanded Reusability in RDLT [8]

The Expanded Vertex Simplification Algorithm (EVSA) generates at least two RDLTs, level-1 and level-2 vertex-simplified RDLTs, called R_1 and R_2, respectively. R_1 is derived from all the vertices and arcs outside the RBS with the addition of abstract arcs [8]. Meanwhile, R_2 is created from all the components inside an RBS. The L-value of each abstract arc is computed by determining the maximum number of times that each arc is used by an activity in R while taking into account the resets that can affect the reusability of this arc. The overall sum for the reusability of each arc (x, y) in R is called the expanded reusability of (x, y), denoted as $eRU(x, y)$. The algorithm copies the C and L-attributes of each arc in the original RDLT to its corresponding arc in R_1 and R_2 whenever such arc is not an abstract arc. Conversely, for each abstract arc (x', y') in R_1 that represent (x, y) in R, the C-value is set to ϵ and the L-value is set to the minimum expanded reusability of all the arcs along the path inside the RBS of R that (x', y') represents in R_1.

3.6 Processes in RDLT [9]

A simple path from vertex x to y with no cycles or duplicate vertices is called a process P. If two processes have common start and end vertices with no shared intermediate vertices, they are considered siblings.

3.7 Classical Soundness in RDLT [3]

An RDLT R is considered classical sound if it holds the condition of proper termination, and liveness. The first condition requires that every vertex reached in an activity will eventually lead to the final target vertex. Meanwhile, the second condition requires that each arc in R is included in at least one activity profile.

3.8 Existing Mapping of RDLT to PN

In the paper [4], two algorithms were proposed to decompose RDLT to CD and PN. Focusing on the RDLT to PN mapping, the input RDLT will undergoes a pre-processing step through vertex simplification and adding two extra vertices with no incoming and outgoing arcs, respectively. Then, the algorithm maps the vertices and arcs of the pre-processed RDLT into its equivalent PN component. This study was able to represent all possible structures composing an RDLT into 9 structures. These structures were used to verify the correctness of the mapping algorithm by extracting their activity profiles and comparing them with the firing sequences of its output PN. As a limitation, however, this mapping algorithm did not take into consideration the L and M-attribute of the input RDLT. Recent literature [5] focuses on developing a novel mapping for the L- and M-attributes of an RDLT to PN. Drawing inspiration from the approach recommended in [4], each PN of the RDLT structures were modified by adding an auxiliary place with tokens equal to the L-value of an arc in the RDLT. If a looping arc leads back to a transition, a reset arc is not connected from the auxiliary place to that transition. Otherwise, a reset arc is connected from the auxiliary place to that transition. If this transition fires, all the tokens inside the auxiliary place are removed. All the auxiliary places are connected to the final transition of the PN with reset arcs. Once this final transition fires, all the tokens in the auxiliary places will be removed. This is to ensure that the tokens inside the auxiliary places do not affect the soundness of the output PN. For the M-attribute, vertex simplification is performed on the input RDLT resulting in two vertex-simplified RDLTs, level-1 and level-2. The former is converted into an extended RDLT, adding two extra vertices. The vertices and arcs in the two RDLTs are then mapped to their equivalent PN components. A check is then performed to see if an arc is a bridge of an RBS. If it is an in-bridge, the input place that enables the center of the RBS in the level-1 PN is connected to the same transition in the level-2 PN. If it is an out-bridge, the transition representing the vertex with an out-bridge in the level-2 PN is connected to the output places in the level-1 PN. This creates an XOR split connecting the level-1 and level-2 PN through the in- and out-bridges. If analysis is to be performed on the level-1 RDLT, the path through the level-1 PN is followed. Otherwise, if analysis is to be performed on the level-2 RDLT (RBS), the path through the level-2 PN is followed. The combined activity profiles(firing sequences) generated in the level-1 RDLT(PN) and level-2 RDLT(PN) represent a behavior of the whole RDLT(PN).

4 Improving RDLT to PN Mapping

4.1 Mapping of RDLT Components into PN

Mapping of RBS. To provide a mapping of RDLT to PN that allows the reuse of an RBS, the mapping in [5] is extended. The proposed mapping in this study visits each RBS that has at least one out-bridge and creates a place P_{cons} and a transition T_{rr} for that subsystem. P_{cons} is connected to T_{rr} with an arc and a reset arc. Then, each transition that was mapped from a vertex that has an out-bridge is connected to P_{cons}. Afterward, transition T_{rr} is connected to each auxiliary place inside the level-2 PN with arcs that have weights equal to the original number of tokens inside the auxiliary place it is connected to. Conversely, each auxiliary place inside the level-2 PN is connected to T_{rr} with a reset arc.

Mapping of MIX-JOIN. The proposed mapping in this paper will modify the corresponding PN structure of MIX-JOIN in [5] by connecting a reset arc from place P_{zm} to transition T_z. This will handle the instances where two tokens are produced in P_{zm}, whenever both transitions $T_{\epsilon z}$ and T_{Jz} fire, which leads to a PN that is not classical sound.

Mapping of SPLIT Structures. This paper proposes an additional mapping of the RDLT SPLIT structures in [5], by looking at the relationship between the processes at a SPLIT point and the type of JOIN they are merged on. A check will be performed to see if the processes at a SPLIT point are siblings joined by an OR-JOIN, or if the processes are non-siblings at all. If so, the PN mapping of the SPLIT structures will be modified in a way that the SPLIT point T_x will only have a single output place P_{xsplit}. Otherwise, the PN mapping for the RDLT SPLIT structures used in [5] will be followed.

4.2 Proposed Mapping Algorithm of RDLT to PN

Algorithm 1 is the proposed algorithm for mapping an input RDLT to its corresponding Petri Net that takes into consideration all the added/modified components mentioned in Sect. 4.1. For simulation, the RDLT in Fig. 1 is used as an input to the mapping algorithm. The output PN is shown in Fig. 2.

4.3 Validation and Analysis of Proposed Mapping Algorithm

To prove the correctness of the proposed mapping, the activity profiles in an input RDLT will be extracted and then compared to the firing sequences of the output PN. To aid this validation, Theorem 4 in [4] is used to prove that if vertex y is reachable from vertex x in an input RDLT, then there exists a firing sequence $\sigma = t_1, \ldots, t_n$ in the output PN mapped from R, where $t_1 = T_x$ and $t_n = T_y$.

Theorem 1. *The space complexity of Algorithm 1 is $O(v^2)$, where v is the number of vertices in the RDLT R.*

Proof. In an RDLT, the maximum number of arcs is v^2 where v is the number of vertices. Since the algorithm visits each arc of the RDLT and creates its corresponding PN component, the maximum space complexity is $O(v^2)$.

Theorem 2. *The time complexity of Algorithm 1 is $O(v^2)$, where v is the number of vertices in the RDLT R.*

Proof. Similarly, the quadratic growth of the number of arcs in an RDLT determines the runtime of the algorithm. Thus, the time complexity of the algorithm is $O(v^2)$.

Theorem 3. *Given an input RDLT R that is classical sound, a corresponding output Petri Net that is also classical sound can be mapped using Algorithm 1.*

Proof. By definition, an RDLT is classical sound if it satisfies the condition of proper termination and liveness. Meanwhile, a PN is classical sound if it satisfies the condition of option to complete, proper termination, and liveness. The mapped PN from the RDLT (Fig. 2) using Algorithm 1 is also classical sound in the context of PNs since it satisfies the condition of option to complete, proper termination, and liveness. Based on the analyses, it can be concluded that given an input RDLT that is classical sound, an output Petri Net generated using Algorithm 1 is also classical sound.

Algorithm 1 Proposed RDLT to PN Mapping Algorithm

Given RDLT R, perform the following preprocessing steps
Step 1. Apply Expanded Vertex Simplification on R.
Step 2. Convert into Extended RDLT R'.
Input: Extended RDLT R'
Output: PN equivalent R'

for each $x \in V'$ **do**
  create T_x
for each $x \in SplitPts$ **do**
  if processes at x are siblings with an OR-JOIN merge point or processes at x are nonsiblings **then**
    create P_{xsplit} and connect T_x to P_{xsplit}
for each $y \in V'$ where $\exists(x, y) \in E'$ where $x \in V'$ **do**
  create P_{ym} and connect to T_y
  if $y = o$ **then**
    create P_o and connect T_o to P_o
  if $\exists(x, y) \in E'$ where $C'((x, y)) \in \Sigma$ **then**
    create T_{Jy} and P_{Jy}
    connect T_{Jy} to P_{ym} and P_{Jy} to T_{Jy}
    if T_x is connected to P_{xsplit} **then**
      connect P_{xsplit} to T_{Jy}
for each $(x, y) \in E'$ where $C'((x, y)) = \epsilon$ **do**
  create $T_{\epsilon y}$ and $P_{\epsilon ny}$
  connect $T_{\epsilon y}$ to P_{ym}, $P_{\epsilon ny}$ to $T_{\epsilon y}$
  if T_x is not connected to P_{xsplit} **then**
    create $P_{\epsilon y}$
    connect $P_{\epsilon y}$ to $T_{\epsilon y}$ and T_x to $P_{\epsilon y}$
  else
    connect P_{xsplit} to $T_{\epsilon y}$
for each $(x, y) \in E'$ **do**
  if $C'((x, y)) \in \Sigma$ **then**
    if T_x is not connected to P_{xsplit} **then**
      if $\nexists P_{C'((x,y))y}$ **then**
        create $P_{C'((x,y))y}$
        connect T_x to $P_{C'((x,y))y}$ and $P_{C'((x,y))y}$ to T_{Jy}
        if $\exists T_{\epsilon y}$ **then**
          create $P_{C'((x,y))\epsilon}$
          connect $P_{C'((x,y))\epsilon}$ to $T_{\epsilon y}$ and $T_{\epsilon y}$, T_x to $P_{C'((x,y))\epsilon}$
          connect $P_{C'((x,y))\epsilon}$ to T_y and T_o with reset arc
          if $\exists T'_y$ **then**
            connect P_{ym} to T'_y with a reset arc
      else
        connect T_x to $P_{C'((x,y))y}$
  else
    connect T_x to $P_{\epsilon y}$
for each RBS $G_u \in R$ **do**
  if $\exists(x, y) \in E$ where $x \in V_{Gu}$ and $y \notin V_{Gu}$ (an out-bridge) **then**
    create P_{cons} and T_{rr}
    connect P_{cons} to T_{rr} with a normal arc and a reset arc
    for each arc $(x, y) \in E$ **do**
      if arc is an out-bridge **then**
        if T'_x is not connected to P_{cons} **then**
          connect T'_x to P_{cons}
for each arc $(x, y) \in E$ of R **do**
  if arc is an in-bridge **then**
    connect P_{ym} to T'_y
  if arc is an out-bridge **then**
    connect T'_x to each output place of T_x
for each auxiliary place $(P_{\epsilon ny}, P_{Jy}, P_{\epsilon ny,i})$ of an arc (x, y) **do**
  connect auxiliary place to T_o with reset arc
  place tokens in auxiliary place according to initial token placement rules
  if auxiliary place is inside level-2 PN **then**
    connect auxiliary place to T_{rr} with a reset arc
    connect an arc from T_{rr} to the auxiliary place with weight equal to the L-attribute of arc (x, y)
  if a looping arc does not exist for y **then**
    connect auxiliary place to T_y with a reset arc
create a place P_{im} and connect it to each source transition in the level-1 PN

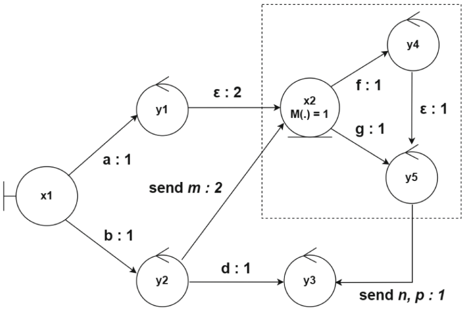


Fig. 1. Classical sound RDLT.

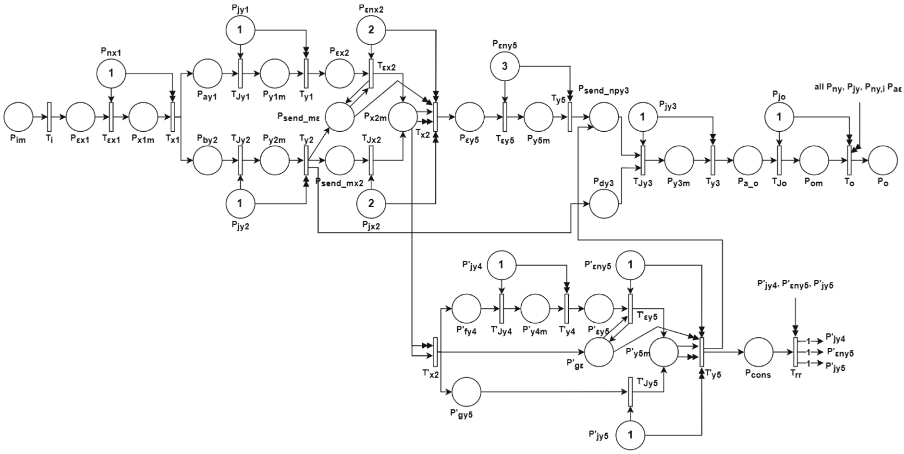


Fig. 2. Generated Petri Net using the proposed mapping algorithm.

5 Conclusions and Future Work

This study addressed some of the existing gaps in RDLT to PN mapping. One of the most significant contributions of this paper is mapping the reset behavior of the RBS. The algorithm creates a place and a transition for each RBS with an out-bridge. These components are responsible for the resets and replenishment of tokens in the auxiliary places inside the level-2 PN. This paper also modified the PN mapping of MIX-JOIN by adding a reset arc connected from P_{zm} to T_z. As a result, the MIX-JOIN no longer produces a firing sequence that results in a non-classical sound PN. Lastly, a new mapping was proposed for the RDLT SPLIT structures by checking the relationship between the processes at a SPLIT point and the type of JOIN they are merged on. As a recommendation, future studies can focus on mapping multiple abstract arcs in the level-1 PN. One approach is to check if the processes at a SPLIT point are abstract arcs and create a transition and an auxiliary place. These transitions are connected to a single input and output place, creating an OR-SPLIT-JOIN structure where only one path can be traversed.

References

1. Hollingsworth, D.: Workflow Management Coalition: The Workflow Reference Model. TC00-1003(19), 224 (1995)
2. Lopez, J.C.L., Bayuga, M.J., Juayong, R.A., et al.: Workflow models for integrated disease surveillance and response systems. In: Theory and Practice of Computation, pp. 141–152. CRC Press (2020)
3. Malinao, J.: On building multidimensional workflow models for complex systems modelling. Dissertation, Technische Universit$\ddot{a}$t Wien. reposiTUm (2017)
4. Yiu, A., Garcia, J., Malinao, J., et al.: On model decomposition of multidimensional workflow diagrams. In: Proceedings of the Workshop on Computation: Theory and Practice (2018)
5. Sulla, C.N.N., Malinao, J.: Mapping of robustness diagram with loop and time controls to petri net with considerations on soundness. In: Kabassi, K., Mylonas, P., Caro, J. (eds.) NiDS 2023. LNNS, vol. 784, pp. 338–353. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-44146-2_36
6. Van Der Aalst, W.M.P., Van Hee, K.M., et al.: Soundness of workflow nets: classification, decidability, and analysis. Formal Aspects Comput. **23**, 333–363 (2011)
7. Pertsukhov, P.A., Mitsyuk, A.A.: Simulating petri nets with inhibitor and reset arcs. In: Trudy ISP RAN/Proceedings, pp. 151–162. ISP RAS (2019). https://doi.org/10.15514/ISPRAS-2019-31(4)-10
8. Malinao, J., Juayong, R.A.: Reset profiles and classical soundness in robustness diagrams with loop and time controls. In: Proceedings of the Workshop on Computation: Theory and Practice (WCTP), Hokkaido, Japan (2023)
9. Roca, T.N, Malinao, J.: Well-handledness in robustness diagram with loop and time controls. Accepted at 20th International Conference on Intelligent Tutoring Systems (ITS 2024). Lecture Notes in Computer Science (LNCS) (to appear)
10. Calvo, G.B., Malinao, J.A.: Mapping hierarchies and dependencies from robustness diagram with loop and time controls to class diagram. In: Kabassi, K., Mylonas, P., Caro, J. (eds.) NiDS 2023. LNNS, vol. 783, pp. 23–42. Springer, Cham (2023) . https://doi.org/10.1007/978-3-031-44097-7_3

Well-Handledness in Robustness Diagram with Loop and Time Controls

Therese Nuelle Roca(✉) and Jasmine A. Malinao

Division of Natural Sciences and Mathematics, University of the Philippines Tacloban College, Tacloban, Philippines
{troca,jamalinao1}@up.edu.ph

Abstract. This study unveils and formalizes well-handledness in Robustness Diagram with Loop and Time Controls(RDLT). This model property has not yet been established for this particular workflow model. The fundamental concept of well-handledness, which is to ensure a balance between the AND/OR-split and AND/OR-join pairs, is adopted from Petri Net(PN). However, due to the multidimensional nature of RDLT, we extended its definition to consider the L-attributes. Meanwhile, the MIX-join structure that is unique to RLDT has been discounted in this study's analysis and formalization of well-handledness. We have also established the structural and behavioural profiles of well-handled RDLTs, unravelling their relationship with relaxed and classical soundness. Lastly, algorithms for verifying balanced and well-handled RDLTs are also provided.

Keywords: Well-handledness · RDLT · Model Verification

1 Introduction

The Robustness Diagram with Loop and Time Controls (RDLT) is a workflow model designed to represent complex systems effectively. RDLT can capture all three workflow dimensions, i.e., resource, process, and case, discussed in [4], within a single model. Furthermore, it introduced the reset-bound subsystems, which function akin to cancellation regions in other workflows. Real-world systems, such as the Absorption Chiller System [1] and Philippine Integrated Disease Surveillance and Response (PIDSR) system [5], have already been modelled and analyzed using RDLT. The Business Process Modeling and Notation (BPMN) and activity diagrams are also multidimensional models. However, BPMN face concerns like concept excess and lack of support for explicit differentiation and rules [1]. In light of the complexity of the model, functional decomposition cannot be performed on activity diagrams. Furthermore, RDLT adopts model properties such as soundness, free-choiceness, conflict-freeness, and etc. from Petri Nets and Workflow Nets [4]. This study adopts another model property called well-handledness from Petri Net, which ensures a balance between the AND/OR-splits and AND/OR-joins. In Petri Net, liveness becomes

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 202–210, 2024.
https://doi.org/10.1007/978-3-031-63031-6_18

an implicit effect of being well-handled. However, considering RDLT's multi-dimensional nature, potential deadlocks can still occur even if there is a balance between its AND/OR-splits and AND/OR-joins. Hence, we integrate structural requirements previously employed to avoid deadlocks and unfinished processes to achieve classical soundness into the requirements of well-handledness in RDLT.

1.1 Robustness Diagram with Loop and Time Controls

Definition 1. *RDLT [1,2] An RDLT is a graph representation R of a system that is defined as $R = (V, E, T, M)$ where:*

- *V is a finite set of vertices, where each vertex has a type $V_{type} : V \rightarrow \{$'b', 'e', 'c'$\}$ where 'b', 'e', and 'c' means the vertex is either a "boundary object", an "entity object", or a "controller", respectively.*
- *A finite set of arcs $E \subseteq (V \times V) \backslash E'$ where $E' = \{(x, y) | x, y \in V, V_{type}(x) \in \{$'b', 'e'$\}, V_{type}(y) \in \{$'b', 'e'$\}\}$ with the following attributes with user-defined values,*
    - *$C : E \rightarrow \Sigma \cup \{\epsilon\}$ where Σ is a finite non-empty set of symbols and ϵ is the empty string. Note that for real-world systems, a task $v \in V$, i.e. $V_{type}(v) =$'c', is executed by a component $u \in V, V_{type}(u) \in \{$'b','e'$\}$. This component-task association is represented by the arc $(u, v) \in E$ where $C((u, v)) = \epsilon$. Furthermore, $C((x, y)) \in \Sigma$ represents a constraint to be satisfied to reach y from x. This constraint can represent either an input requirement or a parameter $C((x, y))$ which needs to be satisfied to proceed from using the component/task x to y. $C((x, y)) = \epsilon$ represents a constraint-free process flow to reach y from x or a self-loop when $x = y$.*
    - *We call (x, y) as an ownership arc if x is an object and y is a controller i.e. we say that x is the owner of y.[1].*
    - *$L : E \rightarrow \mathbb{N}$ is the maximum number of traversals allowed on the arc.*
- *Let T be a mapping such that $T((x, y)) = (t1, ..., tn)$ for every $(x, y) \in E$ where $n = L((x, y))$ and $t_i \in \mathbb{N}$ is the time a check or traversal is done on (x, y) by some algorithm's walk on R.*
- *$M : V \rightarrow \{0, 1\}$ indicates whether $u \in V$ is the center of a reset-bound subsystem (RBS). Given a center $u \in V$, where $M(u) = 1$, an RBS is a subgraph G_u of R that is induced by u and its set of owned controllers.*
  *Finally, $(a, b) \in E$ is called an **in-bridge** of b if $a \notin V_{G_u}, b \in V_{G_u}$. Meanwhile, $(b, a) \in E$ is called an **out-bridge** of b if $b \in V_{G_u}$ and $a \notin V_{G_u}$. Arcs $(a, b), (c, d) \in E$ are **type-alike** if $\exists y \in V$ where $(a, b), (c, d) \in Bridges(y)$ with $Bridges(y) = \{(r, s) \in E | (r, s)$ is either an in-bridge or out-bridge of $y\}$ or if $\forall y \in V, (a, b), (c, d) \notin Bridges(y)$.*

To generate activity profiles of a given RDLT R an activity extraction algorithm is introduced in [1]. Notice that the arcs (x_8, x_{10}) and (x_9, x_{10}) are **unconstrained** since they satisfy requirement (1) of the definition of an unconstrained arc in [1].

---

[1] For the purposes of this study we have this as a slight modification to the concept of ownership between two vertices in R. This modification does not affect existing results in RDLT.

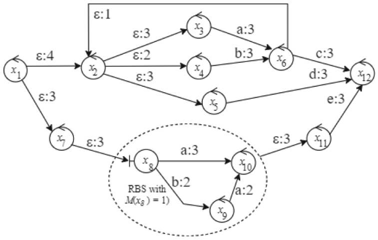


Fig. 1. An example of an RDLT with RBS

1.2 Control Flow and Parallelism in RDLT

The different types of split and join structures in RDLT defined in [1,2] include conditional split(OR-split), parallel split(AND-split), and conditional join(OR-join), AND-join, and MIX-join. These structures serve as the foundation for establishing well-handledness in the context of RDLT. It is also important to note that the definitions of the distinct join structures underscore the type-alikeness of the arcs that form these joins.

1.3 Soundness in RDLT

Classical soundness in RLDT requires the satisfaction of proper termination and liveness [2]. The L-values of the arcs in R significantly influence the realization of these behaviors, making it crucial to consider this arc-attribute when aiming for classical soundness. Relaxed soundness requires R is live and there is at least one terminating activity in R where every AND/MIX-join is resolved.

1.4 *L*-Safeness in RDLT

Another strategy that is used to verify classical soundness for RDLTs with no RBS is L-safeness in [2]. The discussion in [2] highlights the concept of reusability of RDLT components. Additionally, it provided a means of configuring L-values in a way that avoids deadlocks caused by looping back. [2] introduced the concept of a critical arc(CA) which dictactes the reuse of another component called non-critical arc (NCA). The arc with the least L-value among the arcs in cycle c is referred to as a critical arc. Otherwise, it is called an NCA of c. Reusability (RU) of an NCA (x, y) is dictated by sum of the distinct CA's of the cycles that it participates in. An NCA (x, y) is loop-safe if $L(x, y)$ is greater than the sum of all the distinct CAs. Meanwhile, a CA (q, p) is referred to as a safe CA if it has an escape arc (q, r) that is a loop-safe NCA.

Join-safeness is an additional requirement for L-safeness in R, focusing on the importance of the join structures of R, which loop-safe NCAs and safe CAs alone don't address in ensuring classical soundness. Note that we do not list all the requirements of join-safeness from [2] in Definition 2, but only those which shall be reused in this study. The reader is referred to [2,3] for the details of this property.

Definition 2. ***L-safe RDLTs*** *(targeted listing) [2, 3]*
*Let $R = (V, E, T, M)$ be a RDLT with no RBS. R is **JOIN-safe** if for every join merging at $y \in V$ of R where $C(u, y) \neq C(v, y)$, the following hold:*

1. ***Equal L-values of arcs at the AND-JOIN.*** *Every pair of arcs of an AND-JOIN merging at y must have the same L-value.*
2. ***Loop-safe components of every related process.*** *The processes that are involved in an AND- or MIX-JOIN merging at y must have each of their component arcs to be loop-safe (NCAs).*

For an OR-JOIN merging at y, each of the processes that are involved in this JOIN must be composed of loop-safe NCAs and safe CAs, if any.
R is L-safe if every NCA is loop-safe, every CA is safe, and R is JOIN-safe.

1.5 Expanded Vertex Simplifications Algorithm

The Expanded Vertex Simplification Algorithm (EVSA) in [3] builds on top of the results of performing vertex simplification [2] on R. EVSA generates Level 1 and 2 expanded vertex simplifications R'_1 and R'_2 of R by using R_1 and R_2 to preserve the arc attributes, e.g., L-attributes in R. The reusability of arcs in R is now referred to as expanded reusability (eRU). In [2], the internal structures of an RBS in R are represented by abstract arcs in R_1 whose L-values are referred to as derived L-values upon applying EVSA on R.

1.6 Antecedent and Consequent Sets

Definition 3. ***(Antecedent and Consequent Sets of x)*** *[1]*
*Given a vertex $x \in V$ of RDLT R, $\overline{\alpha}_x$ is called the **maximal antecedent set** of x that contains the maximal set of vertices in R where for every $v \in \overline{\alpha}_x$, there exists an elementary path from the source of R to x that passes through v first. Meanwhile, $\overline{\Omega}_x$ is called the **maximal consequent set** of x that contains the maximal set of vertices in R where for every $y\overline{\Omega}_x$, there is a path from x to $u \in \overline{\alpha}_x$, that passes through $y \in V \backslash \overline{\alpha}_x$, where $(u, x) \in E$.*

1.7 Well-Handledness in Petri Net

Well-handledness in Petri Nets, and in particular Workflow Nets(WF-nets), is used to formalize the concept of the good constructions of WF-nets. It is a model property that ensures balance between the AND/OR-splits and AND/OR-joins. Verification methods like the one discussed in [6] help ensure well-handledness in WF-nets. Its formal definition is given by Definition 13 in [4].

2 Methodology

2.1 Well-Handledness in RDLT

RDLT's well-handledness adopts specific structural requirements from L-safeness discussed in [2,3] that are crucial for ensuring classical soundness. But before delving into the specifics of well-handledness, we first introduce novel concepts essential in understanding its definition and related requirements. To facilitate the discussion in this section and the subsequent sections, we assume that R has one source and one sink. In verifying whether an RDLT R with at least one RBS is well-handled or not, the preprocessing step known as EVSA is done to conduct separate and level-based analysis of well-handledness, which is then generalized for the original RDLT.

Definition 4. ***(Process P from x to y)***
A process P from x to y is an elementary path in R where $P = x_1, x_2, \ldots, x_n$, $n \in \mathbb{N}$, $(x_i, x_{i+1}) \in E, i = 1, 2, \ldots, n-1$, such that $x = x_1$ and $y = x_n$.

Definition 5. ***(Process Completion)***
Let $S = \{S(1), S(2), \ldots S(k)\}$ $k \in N$, be an activity in R.
Let $P = x_1, \ldots x_n$ be a process in R, and let the $Arcs(P)$ be the set of arcs traversed through P.
*A **process builder** $\tau_{t_s,t_e} \subseteq S$ is a set of reachability configurations in S whose set of components compose P, where P is initiated at time t_s and ends at time t_e in S, that is, $\biguplus_{(x,y)\in S(j), S(j)\in\tau_{t_s,t_e}, 1\le j\le k}\{(x,y)\} \supseteq \biguplus_{(u,v)\in Arcs(P)}\{(u,v)\}$, where $\uplus$ is the multiset union operator.*

*Let $\tau_{t_{s1},t_{e1}}$ and $\tau'_{t_{s2},t_{e2}}$ be the process builders of processes P and P' where P and P' both begin at vertex x and end at y in R. That is, P and P' forms a **split-join pair** from x to y.*

If for every pair of processes starting at x forms an AND(OR)-split and ends at y forming an AND(OR)-join, we call this pair as an AND(OR)-split-join pair.

*Without loss of generality, we say process P **completes** in its split-join pair x and y if either of the following holds for every activity S in R,*

- *Case 1: For an AND-split-join pair for x and y, if P is initiated in S, then $\tau_{t_{s1},t_{e1}} \subseteq S$ and $\tau'_{t_{s2},t_{e2}} \subseteq S$ fully build P and P' in S, respectively, where $(t_{s1} = t_{s2}) \le (t_{e1} = t_{e2})$.*
- *Case 2: For an OR-split-join pair for x and y, if P is initiated in S, then $\tau_{t_{s1},t_{e1}} \subseteq S$ fully builds P in S and for every P' in this pair, where $P' \ne P$, $\tau'_{t_{s2},t_{e2}} = \emptyset$ in S, where $\tau'_{t_{s2},t_{e2}}$ builds P'.*

Definition 5 means that process completion of P in a split-join pair of x and y is achieved for every activity S in R, either (1) for an AND-split-join pair: if P is initiated for each time t_{e1} in S, then the join at y is always resolved such that all of the processes in this pair are initiated as well each time, and thereafter, would simultaneously reach y at time $t_{e1} = t_{e2}$; (2) for an OR-split-join pair, if P is initiated for each time t_{e1} in S, then no other processes P' in this join can be initiated in S until y is reached at time t_{e1}.

Definition 6. ***(Sibling processes)*** *Given a split at $x \in V$ with its processes P and P', where $P = x_1, x_2 \ldots, x_n$, $P' = y_1, y_2 \ldots, y_m$, P and P' are called **sibling processes** if $x_1 = y_1$, $x_n = y_m$, and P and P' have no arcs in common.*

Definition 7. ***(Closed split-join structure)*** *A split-join pair of x and y is a **closed structure** of x and y if the following hold:*

1. *for every pair of processes P and P' that end at y, there exists a split point x where P and P' start at, and P and P' are sibling processes, and*
2. *for every process P'' that start at x but does not end at y, where y forms either an AND- or MIX-join, P'' ends at a descendant y' of y, where $y \in \overline{\alpha}_{y'}$.*

Definition 8. ***(Complementary split and join)*** *Given a set of sibling processes P_1, $P_2 \ldots, P_n$, we say that the split-join pair starting and ending at x and y, respectively, is **complementary** if this split-join pair is a closed structure of x and y and P_1, $P_2 \ldots, P_n$ forms an OR-split(AND-split) at x and ends with an OR-join(AND-join) at y. Furthermore, the adjacent vertices of every split and join point of sibling processes that are evaluated for complementarity are fully inside or outside the RBS.*

Remark 1. We require adjacent vertices of the split and join points in sibling processes to be fully inside or outside the RBS, as internal structures of an RBS are ignored in R'_1 of R and are instead represented by abstract arcs.

Definition 8 follows the good constructions of WF-nets as described in [4]. In RDLT, an AND-split generates two disjoint flows, each invoking a separate process that operates in parallel. Complementing it with an OR-join is not semantically correct, and will result to an inconsistent split and join behavior.

Definition 9. ***(Balanced RDLT)*** *An RDLT R with or without RBS is balanced if every split-join pair is complementary.*

Theorem 1. *Let R be a connected RDLT with at least one RBS. Furthermore, Let $R'_1 = (V'_1, E'_1, T'_1)$ and $R'_2 = (V'_2, E'_2, T'_2)$ be the expanded vertex simplifications of R where $R'_2(R'_1)$ has no split-join pairs that can be evaluated for complementarity. If $R'_1(R'_2)$ is balanced, then R is balanced.*

Proof. To prove this theorem, we analyze the structures of R'_1 and R'_2 resulting from its premise.

– **Structure 1 and 2:** An AND-join (OR-join) merging at y. This structure can be depicted by $(u, y) \in E'$ and $(v, y) \in E/E'$. This also means that (u, y) and (v, y) are not type-alike, where (u, y) is a part of an RBS R and (v, y) is not. We skip complementarity evaluation for sibling processes involving (u, y) in R'_1 since there is a loss of information e.g., original C-attributes. We then check complementarity for other sibling processes in $R'_1(R'_2)$, then conclude if $R'_1(R'_2)$ is balanced.

Thus, if $R_1'(R_2')$ is balanced, then R is balanced.■

Theorem 2. *Let R be a connected RDLT with at least one RBS. Furthermore, Let $R_1' = (V_1', E_1', T_1')$ and $R_2' = (V_2', E_2', T_2')$ be the expanded vertex simplifications of R where both R_1' and R_2' have split-join pairs that can be evaluated for complementarity. If R_1' and R_2' are balanced, then R is balanced.*

Proof. To prove this theorem, we examine the structures of R_1' and R_2' resulting from its premise.

- **Structure 1 and 2:** An AND-split-join(OR-split-join) pair in R_2' with a split at x' merging at y' that are fully inside the RBS, and $(x', y') \in E/E'$ is an arc of P_1, where $P_1 = x_1x_2, ...x_n, n \in \mathbb{N}$. Furthermore, $P_2 = y_1y_2, ...y_m, m \in \mathbb{N}$ is a sibling process of P_1 where x' and y' are not adjacent vertices of that split and join point of P_1 and P_2. In this case, it necessary to verify if both R_1' and R_2' are balanced before concluding that R is balanced, since the split-join pairs in R_2' are not captured in R_1' and vice versa.

Thus, if R_1' and R_2' are balanced, then R is balanced.■

Definition 10. ***(Well-handled RDLT)*** *An RDLT R with no RBS is well-handled iff R has loop-safe NCAs, safe CAs, equal L-values at AND-joins, loop-safe components of every related process and is balanced.*

In contrast to Petri Nets, achieving well-handledness in RDLT is not solely focused on ensuring that an RDLT R has complementary pairs of splits and joins. RDLT's definition of well-handledness extends to considering the L-attributes of arcs within R to guarantee that the configuration of the L-values of these arcs will not impede the completion of processes initiated by AND/OR-splits in R. It Petri Nets (PN), structural liveness becomes an implicit effect of being well-handled as shown in Lemma 2 of [4]. RDLT's well-handledness requires additional considerations due to the influence of L-values on unfinished processes as discussed in [2]. The requirements of well-handledness largely coincide with those for ensuring L-safeness in RDLT, since we consider the impact of loops and join-structures in the activities of R.

Theorem 3. *Let R be a connected RDLT with at least one RBS. Furthermore, Let $R_1' = (V_1', E_1', T_1')$ and $R_2' = (V_2', E_2', T_2')$ be the expanded vertex simplications of R. If both R_1' and R_2' are well-handed, then R is well-handled.*

Proof. Let R be a well-handled RLDT.
We prove this theorem by contradiction. Assume that R is well-handled, but either R_1' or R_2' is not well-handled. Since R is well-handled it would mean that it satisfies all of the requirements of well-handledness in Definition 10. If either R_1' or R_2' fails to satisfy any of this requirements, it would imply a violation on R. Hence, we arrive at contradiction. ■

By performing the required preprocessing step on the RDLT in Fig. 1 and evaluating if it satisfies the requirements of well-handledness, it can be concluded that it is a well-handled RDLT.

2.2 Proposed Algorithm for Balanced RDLT Verification

In verifying if an RDLT R with at least one RBS is balanced, the algorithm applies EVSA on R. Then, it evaluates R'_1 and R'_2 separately if they are balanced. It first locates all the split and join points in R, and from that, it identifies the sibling processes. Lastly, it checks whether all the split-join pairs of sibling processes are complementary, then it concludes that R is balanced if R'_1 and R'_2 are balanced.

2.3 Proposed Algorithm for Well-Handledness Verification in RDLT

In verifying if an RDLT R with at least one RBS is well-handled, the algorithm applies EVSA on R. The algorithm then checks separately if R'_1 and R'_2 satisfy all the requirements of well-handledness. The algorithms in [2] are used to verify the requirements adopted from L-safeness, along with the algorithm for balanced RDLT verification. It concludes that R is well-handled if both R'_1 and R'_2 are well-handled.

2.4 Profiles of Well-Handled RDLTs

This section provides theorems derived from the structural and behavioral profiles of R with respect to well-handledness. From this, the relationship between well-handledness and the existing notions of soundness in RDLT can also be established[2]

Lemma 1. *A well-handled RDLT R with no RBS observes proper termination for each of its activities.*

Lemma 2. *A well-handled RDLT R with no RBS is live.*

Theorem 4. *A well-handled RDLT R with no RBS is classical sound.*

Corollary 1. *A well-handled RDLT R with no RBS is relaxed sound.*

3 Conclusions and Future Work

This study addressed the lack of a formalized concept of well-handledness in RDLT. Well-handedness is a property that can help verify the correctness and detect errors in workflow models like RDLT based solely on its static structure, eliminating the need to simulate or extract activities on the model, which is helpful for workflow designers. Through this research, the structural and behavioral profiles of well-handled RLDTs have also been established. Given the profiles of well-handled RDLTs, the relationship between well-handledness and classical/relaxed soundness in RDLT has also been established. Hence, RDLT models can now be verified for well-handledness. We recommend as future work to extend the analysis of well-handledness to include the MIX-join structure in RDLT and extend the algorithm for verifying balanced RDLTs to cover abstract arcs in checking for complementary pair of splits and joins.

[2] The proofs of these theorems are available upon request.

References

1. Malinao, J.: On building multidimensional workflow models for complex system modelling. Dissertation, Technische Universität Wien, Vienna, Austria (2017)
2. Malinao, J., Juayong, R.A.: Classical soundness in robustness diagram with loop and time controls. Philippine J. Sci. **152**(6B), 2327–2342 (2023)
3. Malinao, J., Juayong, R.A.: Reset profiles and classical soundness in robustness diagrams with loop and time controls. In: Proceedings of the Workshop on Computation: Theory and Practice (WCTP), Hokkaido, Japan (2023)
4. van der Aalst, W.M.P.: Structural Characterizations of Sound Workflow Nets. Computing Science Reports Computing Science Reports Vol. 9623, Eindhoven University of Technology (1996)
5. Lopez, J.C.L., Bayuga, M.J., Juayong, R.A., Malinao, J., Caro, J., Tee, M.: Workflow models for integrated disease surveillance and response systems. In: Theory and Practice of Computation. CRC Press (2020)
6. Hernandez, N.H.S., Juayong, R.A.B., Adorna, H.N.: A grammar for detecting well-handled 2-split, 2-join workflow nets without cycles. In: Nishizaki, Sy., Numao, M., Caro, J., Suarez, M.T. (eds.) Theory and Practice of Computation. Proceedings in Information and Communications Technology, vol. 7. Springer, Tokyo (2013). https://doi.org/10.1007/978-4-431-54436-4_7

Generative Intelligence and Metaverse

Enhancing Reinforcement Learning Finetuned Text-to-Image Generative Model Using Reward Ensemble

Kyungryul Back, XinYu Piao, and Jong-Kook Kim(✉)

School of Electrical Engineering, Korea University, Seoul, South Korea
{rudfuf0822,xypiao97,jongkook}@korea.ac.kr

Abstract. In recent years, advanced diffusion models have shown good performance in converting text prompts into high-quality images. However, aligning the generated images to human preferences remains challenging due to the biases in training. Previous researches have attempted to address this problem by incorporating reinforcement learning and human feedback into the denoising diffusion models. However, such approaches often encounter over-optimization, commonly referred to as the reward hacking problem. This paper introduces a simple and effective ensemble approach that combines multiple reward models to optimize the overall reward structure. This proposed method successfully overcomes the over-optimization problem in the diffusion model's fine-tuning process. Both quantitative and qualitative results demonstrate the effectiveness of the proposed approach to generate an image that is a more realistic representation.

Keywords: Generative Model · Text-to-Image · Reinforcement Learning

1 Introduction

Text-to-image generative models have advanced significantly in recent years. Many studies have been introduced to create digital images that align with text input and human preferences that includes Generative Adversarial Networks (GANs) [1], Variational Autoencoders (VAEs) [2], and Diffusion Models [3]. In particular, latent diffusion models show good performance in converting more complex text prompts into high-quality images (e.g., [4–6]). Despite these advancements, it is still challenging to generate images that accurately match text inputs and human preferences. For example, the images generated by unusual prompts such as "red colored dog" or reordered prompts such as "horse riding an astronaut" may not match the human preferences [7].

Recent studies, such as Wu, Xu, Kirstain et al. [8–10] and Lee et al. [11], have introduced datasets and reward models based on human-scored rewards to match text inputs and human preferences. These studies demonstrate that the reinforcement learning from human feedback (RLHF) approach has the potential to generate images that are closer to human preferences. Many recent studies

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 213–224, 2024.
https://doi.org/10.1007/978-3-031-63031-6_19

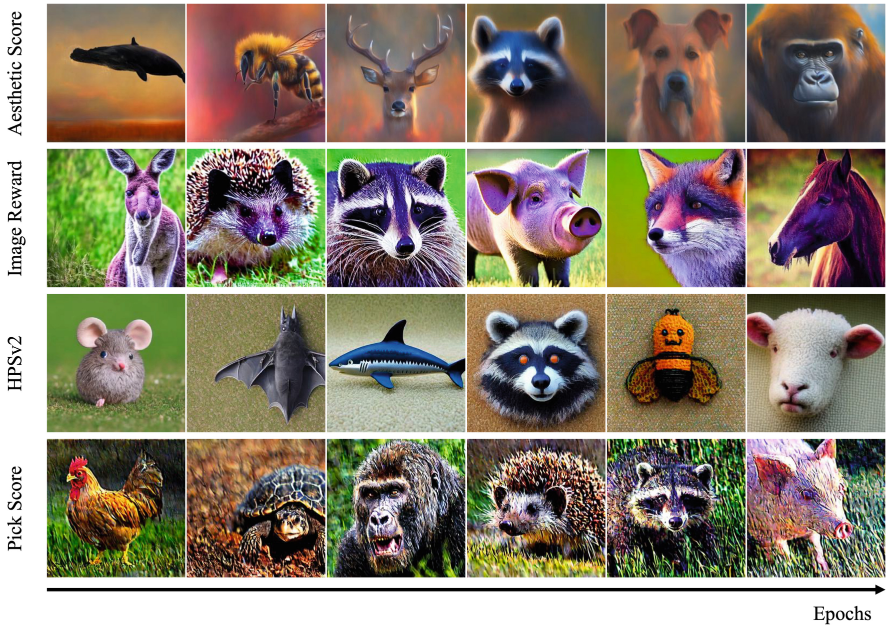


Fig. 1. Results of text-to-image diffusion model trained with different rewards (Aesthetic Score, ImageReward, HPSv2, PickScore) by policy gradient method. Each model shows unique forms of over-optimization over increasing epochs.

have adopted the RLHF-based approach that uses the policy gradient method to optimize expected rewards [12,13]. This approach has shown significant advancements in the text-to-image domain and improved results for the aforementioned challenges. However, despite the success of prior works image fidelity degradation and convergence to a specific style are frequently observed when training the models based on the RLHF approach (as shown in Fig. 1). For example, the images generated by using Aesthetic Score-based reward model shows a trend toward oil painting as the epoch increases. Previous study by Gao et al. [14] identified these problems as a common limitation of RLHF approaches and showed that models trained on small-scale proxy reward models tend to over-optimization and lose efficiency in larger and more complex environments.

This paper proposes a simple and effective ensemble-based reward optimization approach to overcome the over-optimization problem in RLHF text-to-image model. The main idea of the proposed approach is to optimize the reward used in training by integrating multiple reward models and complementing each other by acting as a regularizer of over-optimizing to avoid bias. This integration aims to reduce the over-optimization of RL trained text-to-image stable diffusion model. To identify the most effective combination, this paper also analyzes four different reward models that are widely used in the text-to-image generation domain; LAION-Aesthetic V1 [15], ImageReward [8], PickScore [9], and HPS v2 [10].

The proposed approach outperforms quantitatively and qualitatively compared to other single reward approaches and effectively mitigates the problem of over-optimization. The contributions in this paper are summarized as follows:

- This paper targets the over-optimization problem often observed in RLHF text-to-image models, and proposes a simple and effective ensemble approach that integrates multiple reward models to address this problem without any complex techniques.
- This paper also analyzes all candidate combinations of four different reward models to determine the most effective combination of reward models that is unbiased and robust to the over-optimization problem.
- Results showed that the proposed approach sufficiently mitigates the over-optimization problem and generates high-quality unbiased images. Thus, it shows that the proposed approach has the potential to generate a variety of digital artworks that are closer to human preferences.

This paper is organized as follows. The background information of diffusion models, reinforcement learning fine-tuning, and reward models are detailed in Sect. 2. Section 3 details our proposed methodology. Experimental results are depicted in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Background

2.1 Diffusion Models

Diffusion probabilistic model [3] can be described as two processes; (i) forward and (ii) reverse gradual noising processes. The forward process starts from the original image x_0 and gradually adds Gaussian noise $\epsilon_t \sim \mathcal{N}(0, I)$, culminating in a completely noisy image x_t. This forward process is formulated as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I) \tag{1}$$

where β_t is a predefined variance schedule. The reverse process, which is the main point of the denoising diffusion probabilistic model, is formulated as:

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \tag{2}$$

where σ_t^2 is the predefined reverse variance schedule and $\mu_\theta(x_t, t)$ is the denoising neural network. This denoising neural network can be reparameterized to predict the Gaussian noise ϵ_t. The reparameterized neural network $\epsilon_\theta(x_t, t)$ is trained using mean squared error (MSE) loss (L_ϵ) as following equation.

$$L_\epsilon = \mathbb{E}_{t,x_t,\epsilon_t} \left\|\epsilon_t - \epsilon_\theta(x_t, t)\right\|^2 \tag{3}$$

Additional text prompt constraints are added to make this denoising neural network into a text-to-image diffusion model. Given a text prompt $z \sim p(z)$, during training, the same noising process q is predicted and both the reparameterized unconditional neural network $\epsilon_\theta(x_t, t)$ and conditional neural network $\epsilon_\theta(x_t, t, z)$ denoising models are trained [4].

2.2 Reinforcement Learning Training of Diffusion Models

In Denoising Diffusion Probabilistic Model (DDPM) framework, the denoising process can be formulated as a Markov Decision Process (MDP). This formulation allows to use reinforcement learning to optimize the diffusion model through policy gradient methods. To maintain the consistency and fair evaluation of previous approaches, this paper defines the fine-tuning of the diffusion model as a multi-step decision making problem, referring to DDPO [12], and uses the policy gradient method to maximize the expected reward. Thus, the expected reward can be defined as:

$$\theta^* = \arg\max_\theta \mathbb{E}_{x_0 \sim p_\theta(\cdot|c)}[r(x_0|c)] \tag{4}$$

where θ represents the parameters of the diffusion model, c represents the conditioning variable, and $r(\cdot)$ represents the reward function. The MDP perspective allows to establish policy gradient methodologies for optimizing diffusion models. The state, action, and reward formulations within the MDP framework are defined as:

$$\begin{aligned}
\mathbf{s}_t &\triangleq (\mathbf{c}, t, \mathbf{x}_t) \\
\pi(\mathbf{a}_t|\mathbf{s}_t) &\triangleq p_\theta^{\text{diffusion}}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}) \qquad (5)\\
p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) &\triangleq (\delta_\mathbf{c}, \delta_{t-1}, \delta_{\mathbf{x}_{t-1}})
\end{aligned}$$

$$\mathbf{a}_t \triangleq \mathbf{x}_{t-1} \tag{6}$$

$$\rho_0(\mathbf{s}_0) \triangleq (p(\mathbf{c}), \delta_T, \mathcal{N}(\mathbf{0}, \mathbf{I})) \tag{7}$$

$$R(\mathbf{s}_t, \mathbf{a}_t) \triangleq \begin{cases} r(\mathbf{x}_0, \mathbf{c}) & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

In this setup, s_t represents the state at time step t, meaning the image x_t and conditioning variable c. The action a_t is the image one timestep before x_{t-1}, and the policy $\pi(a_t|s_t)$ is the parameterized denoising process. The reward function $R(s_t, a_t)$ assigns a reward after each action and state, based on the final image and conditioning. The reward is zero until the last step when the denoising is complete. The cumulative reward of all trajectories is $r(x_0, c)$, which means that the maximization of θ in reinforcement learning is equivalent to maximizing the policy π.

Two policy gradient methods are evaluated by Black et al. [12]: REINFORCE and Proximal Policy Optimization (PPO). Our research focuses on PPO methods (DDPO$_{IS}$) due to its superior performance demonstrate from prior works [12,13]. Multiple reward models r_k examined by various experiments are used for policy gradient methods. The gradient for the reinforcement learning objective in our method is defined as:

$$\nabla_\theta J_{\text{EN}} = \mathbb{E}\left[\frac{p_\theta(x_{t-1}|x_t, c)}{p_{\theta_{\text{old}}}(x_{t-1}|x_t, c)} \nabla_\theta \sum_{k=0}^{N} \log p_\theta(x_{t-1}|x_t, c) r_k(x_0, c)\right] \tag{9}$$

For policy gradient estimation, Monte Carlo estimates of $\nabla_\theta J_{\text{EN}}$ is used, alternating between sampling denoising trajectories $\{x_T, x_{T-1}, \ldots, x_0\}$ and updating parameters via gradient descent. A clipping technique [16] is applied to mitigate excessive deviation of the current policy from the previous iteration.

2.3 Reward Models

To align text-to-image synthesis more closely to human preferences, employing a well designed reward model is crucial. Xu, Wu, Kirstain et al. [8–10] collected a large human-labeled datasets and proposed a scoring function to measure perceived aesthetic quality and preference.

LAION-Aesthetics Score. LAION-Aesthetics predictor [15] derived from 176K human-rated images, uses contrastive language-image pretraining (CLIP) embeddings [17] for aesthetic quality assessment. This model, integral to our methodology, has been validated in the DDPO framework [12] for its effectiveness as a reward model.

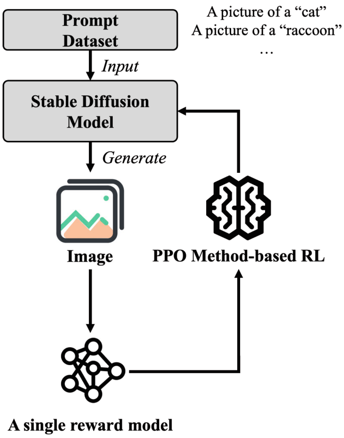
ImageReward. The ImageReward model [8], employed in the DPOK [13], is trained on 137K pairs of expertly curated comparisons. Sourced from the diverse prompts of DiffusionDB [18], this model utilizes a graph-based algorithm to enhance prompt selection diversity, based on language model-driven similarity metrics.

PickScore. PickScore [9] is a CLIP-based model trained on the extensive Pick-a-Pic dataset, featuring over 500,000 examples and 35,000 unique prompts. Developed by Stability AI, this model is designed to align closely with human preferences, contributing significantly to the nuanced understanding of text-to-image synthesis quality.

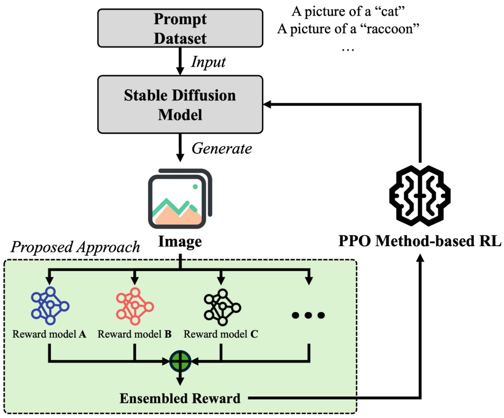
HPS v2. Human Preference Score v2 (HPS v2) [10] is a score model for evaluating user preference in text-to-image generation. Originating from the expansive Human Preference Dataset v2 (HPD v2), which contains 798,090 selections across 433,760 image pairs, HPS v2 is meticulously curated to reduce bias, ensuring a broad representation of human aesthetic judgment (Fig. 2).

3 Proposed Approach

This paper aims to enhance the process of fine-tuning diffusion models using reinforcement learning techniques, in particular focusing on the application of policy gradient methods combined with human preference-based reward models. This paper proposes an ensemble-based reward optimization approach based on the insight that a trained reward model is only an approximation and may not lead to the optimal results [14]. This proposed approach is designed to provide a more robust framework for reward optimization and enhance the effectiveness of policy gradient reinforcement learning methods for fine-tuning diffusion models.



(a) Prior works.



(b) Our proposed approach.

Fig. 2. Difference in the overall process between our proposed approach and the default reinforcement learning of fine-tuning diffusion model.

3.1 Reward Model Ensemble

To mitigate the over-optimization problem commonly known as reward hacking in reinforcement learning policy methods, many approaches such as regularization through augmentation [19] and the adoption of multiple objectives [20] have been proposed. In particular, the ensemble approach that is one of the latter approaches has gained popularity for its simplicity and effectiveness. This paper adopts the ensemble approach and uses a simple logarithmic summation (Eqs. 9 and 10) to combine rewards calculated by each reward model.

$$R_{EN}(x_0, c) = \sum_{k=0}^{N} \log r_k(x_0, c) \tag{10}$$

In this equation, c represents the prompt provided to the diffusion model, and r_k denotes the reward from the k-th model within the ensemble, each of which has been introduced in Sect. 2.3. This integration works through a regulatory mechanism and prevents over-optimization by any single reward. Because each reward models has a different range of scores, the scores are normalized to have similar ranges when combining rewards.

3.2 Combination of Rewards

As described in Sect. 2.3, each reward model has a different bias in training datasets and human preferences. Thus, it is important to examine these reward models in more detail.

LAION-Aesthetics Score. Multi-layer perceptron (MLP) model is trained with Aesthetic Visual Analysis (AVA) dataset [21], containing 250,000 image-text pairs ratings ranging from 1 to 10. The trained aesthetic predictor (Aesthetic Score) favors images of art paintings and painting style images tend to receive high scores (Score above 5.5).

Image Reward. This is aimed at assessing alignment, fidelity, and overall human preference, this score model utilizes human ratings for images sourced from DiffusionDB [18]. Image Reward model cooperated with annotation company in order to collect real human evaluation dataset. However, the model tends to have a preference on specific color, evidenced by our empirical study.

Pick Score. Pick Score has similar approach to the Image Reward, it extends its dataset to include generated images from various diffusion models, soliciting annotations from random users. However, the random method may have potential of reducing the consistency of the reward model.

HPSv2. HPSv2 is bulit upon the methodologies of both the Pick Score and Image Reward. It generates additional images beyond those found in DiffusionDB. It employs annotators to ensure more consistent outcomes, aiming for a higher reliability in model assessments. However, the model tends to have a preference on images with cartoon style since it contained the images and user preferences from Stable diffusion online community (Stable Foundation Discord server [10]).

Preliminary predictions suggested HPSv2 as the superior reward model, given its enhancements in consistency and its complementary function to include the Pick Score and Image Reward. Despite these improvements, an inherent bias within the reward model remains a concern. It was hypothesized that the integration of Pick Score, with its broader variance in user preference backgrounds, and HPSv2 could address the potential problem of overoptimization. Empirical examination showed that the ensembled rewards incorporating HPSv2 showed favorable outcomes, with the combination of HPSv2 and Pick Score emerging as the most effective. This finding showed the importance of using diverse reward signals to capture a wide array of human preferences, thus enhancing the robustness and appeal of generated images.

4 Experiments

4.1 Experimental Setup

To evaluate the proposed approach, the experiments are run on a system consisting of intel i7-8700K 3.7GHz 6-Core processor, 64GB main memory, and NVIDIA GeForce RTX 3090 GPU with 24GB GDDR6 memory. In addition, the experiments that uses more than three reward models are run on the system consisting of NVIDIA Tesla A100 GPU with 40GB memory, two AMD EPYC 7443 2.85GHz 24-Core processors and 128GB main memory. This is because these experiments require more GPU memory.

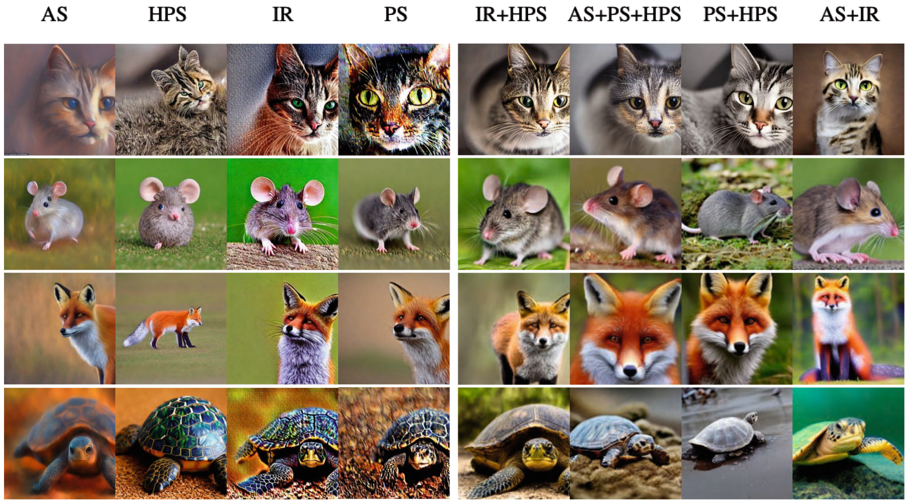


Fig. 3. Comparative visualization of output images after 150 epochs of training. Images generated using ensemble-based reward optimization methods are on the right and images generated from single reward model are on the left.

The performance of the proposed approach is evaluated using the stable diffusion model version 1.4 [4] that is pre-trained on large image-text datasets [15]. A uniform sample of 45 common animals from ImageNet-1000 datasets [22] are used as prompts in the experiments. In the experiments, this approach uses the same setup as the experimental setup in DDPO [12]. The prompts that serve as inputs to the text-to-image model are formulated in a simple format, such as "animal".

4.2 Evaluation Metrics

To evaluate the performance of the RL trained stable diffusion models, five score metrics including Aesthetic Score, ImageReward, Pickscore, HPSv2 and CLIP score are used (see Table 1). For fair comparison, CLIP score [17], a widely used score metrics that measures the similarity between image and text embedding is also used for evaluation. All 15 possible combinations of single reward and reward ensemble are evaluated, which are listed in Table 1.

To compare the overall performance across all metrics, the min-max normalization is applied to all five metrics which adjusts the range of different score values to match each other and is shown in Table 2. The models using specific reward model or reward ensemble that leads to over-optimized results will have an imbalanced score distribution and low overall score. Thus, the sum of min-max scaled scores and their standard deviation is a good indicator of the overall performance and over-optimization.

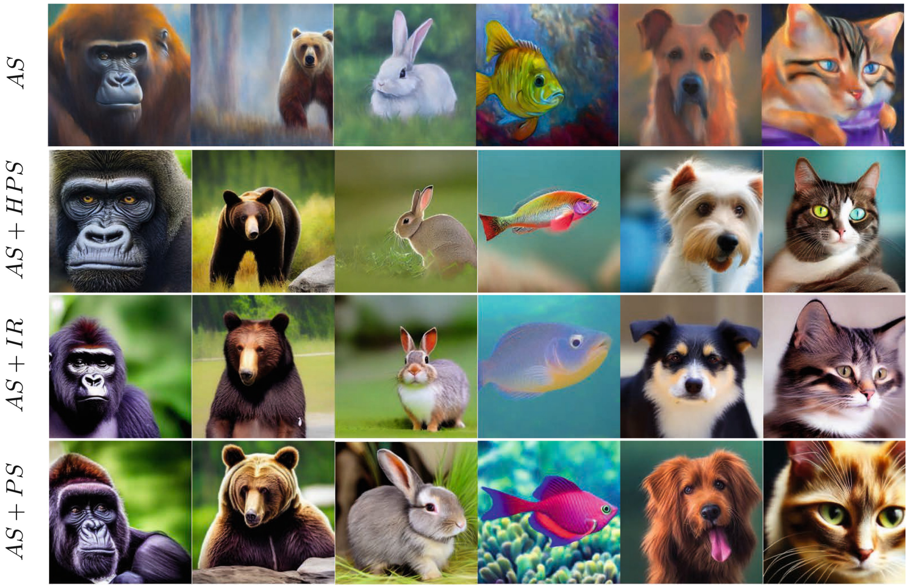


Fig. 4. Comparative visualization of output images after 150 epochs for a single Aesthetic score (AS) reward model and Aesthetic score-based reward ensemble.

4.3 Qualitative Results

Figures 3 and 4 shows that the proposed approach significantly mitigates the over-optimization problem in qualitative terms. In particular, problematic artifacts such as undesired object colors, backgrounds and artistic styles that resemble paintings or cartoons were reduced. The collapsing of the image also reduced and the animals with diverse backgrounds were generated, creating more realistic images.

Specifically, as illustrate in Fig. 4, the integration of Aesthetic reward with other rewards generated realistic images and stability of images enhanced compared with previous single reward trained result. This qualitative improvement highlights the potential of our approach to generate images that align more with intended outcomes, thereby reducing the unintended biases during the training process.

4.4 Quantitative Results

For our quantitative analysis, the stable diffusion models were fine-tuned with different reward models using randomly selected animal prompts. After 150 epochs of training, 32 randomly picked images were generated and the top 10 were selected based on their trained reward scores for further analysis. This selection aimed to identify whether generated images perform well across other metrics. Given that all four reward scores originate from diverse human-annotated

Table 1. Performances of the stable diffusion model for each reward model and reward ensemble. For each metric, top-5 performances are highlighted in bold and the best performances are indicated with an asterisk (*).

Reward Models	Metrics				
	Aesthetic	Image Reward	Pick Score	HPSv2	CLIP
Aesthetic (AS)	**6.108***	0.376	21.14	0.272	0.290
Image Reward (IR)	5.239	**0.918**	20.91	**0.281**	0.281
Pick Score (PS)	5.184	0.510	**22.01**	0.281	0.284
HPSv2 (HPS)	5.158	0.360	21.16	**0.287**	0.283
AS + IR	5.363	0.828	21.33	0.279	**0.299**
AS + PS	**5.789**	0.366	21.18	0.276	**0.294**
AS + HPS	**5.709**	0.546	21.27	**0.282**	0.291
IR + PS	5.324	0.796	21.23	0.280	0.292
IR + HPS	5.362	**1.073***	**21.57**	0.281	0.294
PS + HPS	5.334	0.422	**22.02***	**0.287***	0.291
AS + IR + PS	5.340	0.802	**21.58**	0.279	0.293
AS + IR + HPS	5.325	**0.970**	21.08	0.279	**0.295**
AS + PS + HPS	**5.578**	0.494	**21.75**	**0.283**	**0.297**
IR + PS + HPS	5.110	**0.950**	21.27	0.277	0.293
AS + IR + PS + HPS	**5.387**	**0.912**	21.08	0.276	**0.302***

datasets, high and balanced overall scores suggest that an image closely aligns with human aesthetic preferences. Conversely, imbalanced scores in five metrics indicate over-optimization to the specific reward model, failing to capture the optimal point of preference for an image.

Tables 1 and 2 shows that models fine-tuned with a single reward model are over-optimized, which result in biased performance on certain metrics. On the other hand, our proposed methods achieved high and balanced overall scores with highest min-max scaled scores with low standard deviation. Our proposed methods also outperformed a single reward trained model on the metrics of ImageReward, PickScore and HPSv2. In the case of CLIP metrics, which was used as the neutral evaluation in this experiment, the top-5 ranked models are trained using our ensemble methods. These results show two contributions of our proposed method in (i) generating realistic images that better align with text input, and (ii) effectively mitigating the over-optimization and generate images that better align with human preference. This is consistent with our qualitative results that the proposed ensemble method is helpful in generating images that better align with text input, reducing the unintended biases, and improving the quality of generated images in terms of various human preferences.

Table 2. Min-max scaled scores and standard deviations for each reward model and reward ensemble. It is sorted in descending order by score, then sorted in descending order by standard deviation if the scores are equal.

Rank	Reward models	Score	Std.
1	IR + HPS	3.07	0.237
2	AS + PS + HPS	2.91	0.225
3	PS + HPS	2.79	0.382
4	AS + IR	2.61	0.213
5	AS + IR + PS	2.49	0.144
6	AS + IR + PS + HPS	2.47	0.331
7	AS + IR + HPS	2.36	0.266
8	AS + HPS	2.33	0.155
9	IR + PS	2.17	0.154
10	IR + PS + HPS	2.06	0.276
11	PickScore (PS)	2.02	0.346
12	AS + PS	1.82	0.251
13	Aesthetic (AS)	1.66	0.368
14	ImageReward (IR)	1.51	0.326
15	HPSv2 (HPS)	1.37	0.371

5 Conclusion

This paper targets the over-optimization problem which is often observed in prior reinforcement learning fine-tuning methodologies for text-to-image generation domain. To address this problem, this paper proposes a simple and effective ensemble-based reward optimization approach. By integrating multiple reward models that complement each other, this approach prevents over-optimization during training. The proposed approach uses four different reward models widely used in reinforcement learning methodologies and this paper determined the most effective combination. The results show that the combination of Image Reward and HPSv2 scored best on Image Reward metrics and also achieved highest score in min-max scaled overall score. The combination of PickScore and HPSv2 scored best on both PickScore and HPSv2 metrics and third best in the min-max scaled overall score. Ensemble of all four models scored best on the CLIP metrics. Our proposed approach effectively mitigates the over-optimization problem and performs overall higher performance compared to other single reward approaches. Furthermore, this work contributes to the broader understanding of reinforcement learning applications to the generative models and highlights the importance of diverse reward mechanisms to achieve high-quality, unbiased results.

References

1. Goodfellow, I.J., et al.: Generative adversarial networks (2014)
2. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022)
3. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
5. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation (2022)
6. Ramesh, A., et al.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. PMLR (2021)
7. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
8. Xu, J., et al.: Imagereward: learning and evaluating human preferences for text-to-image generation (2023)
9. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: an open dataset of user preferences for text-to-image generation (2023)
10. Wu, X., et al.: Human preference score V2: a solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
11. Lee, K., et al.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
12. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
13. Fan, Y., et al.: Reinforcement learning for fine-tuning text-to-image diffusion models (2023)
14. Gao, L., Schulman, J., Hilton, J.: Scaling laws for reward model overoptimization (2022)
15. Schuhmann, C., et al.: LAION-5B: an open large-scale dataset for training next generation image-text models (2022)
16. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017)
17. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
18. Wang, Z.J., et al.: Diffusiondb: a large-scale prompt gallery dataset for text-to-image generative models (2023)
19. Yarats, D., Kostrikov, I., Fergus, R.: Image augmentation is all you need: regularizing deep reinforcement learning from pixels (2021)
20. Yang, R., Sun, X., Narasimhan, K.: A generalized algorithm for multi-objective reinforcement learning and policy adaptation (2019)
21. Murray, N., Marchesotti, L., Perronnin, F.: Ava: a large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2408–2415. IEEE (2012)
22. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR 2009 (2009)

Multi-scale Intervention Planning Based on Generative Design

Ioannis Kavouras[1(✉)], Ioannis Rallis[1], Emmanuel Sardis[1], Eftychios Protopapadakis[1,2], Anastasios Doulamis[1], and Nikolaos Doulamis[1]

[1] National Technical University of Athens, Zografou Campus 9, Iroon Polytechniou str, 15772 Zografou, Greece
{ikavouras,sardism}@mail.ntua.gr, irallis@central.ntua.gr, eftprot@uom.edu.gr, {adoulam,ndoulam}@cs.ntua.gr

[2] University of Macedonia, 156 Egnatia Street, 546 36 Thessaloniki, Greece

Abstract. The scarcity of green spaces, in urban environments, consists a critical challenge. There are multiple adverse effects, impacting the health and well-being of the citizens. Small scale interventions, e.g. pocket parks, is a viable solution, but comes with multiple constraints, involving the design and implementation over a specific area. In this study, we harness the capabilities of generative AI for multi-scale intervention planning, focusing on nature based solutions. By leveraging image-to-image and image inpainting algorithms, we propose a methodology to address the green space deficit in urban areas. Focusing on two alleys in Thessaloniki, where greenery is lacking, we demonstrate the efficacy of our approach in visualizing NBS interventions. Our findings underscore the transformative potential of emerging technologies in shaping the future of urban intervention planning processes.

Keywords: Generative Design · Artificial Intelligence · Multi-scale Intervention Planning

1 Introduction

The Urban green space (UGS) availability has long been investigated, because of the importance of green spaces for the health and well-being of urban residents. Generally, there are beneficial associations between green space exposure and reduced stress, positive mood, less depressive symptoms, better emotional well-being, improved mental health and behaviour, and decreased psychological distress in adolescents [14]. Yet, there is significant differentiation, regarding the UGS accessibility, between Northern (above-average availability) and Southern (below-average availability) European cities [7].

Generative artificial intelligence (genAI) has garnered significant attention for its transformative potential across diverse domains, including computer science, creative arts, and language processing. While its efficacy in fields like medicine and healthcare has been demonstrated, its application in engineering domains such as urban planning and architectural design remains unexplored.

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 225–233, 2024.
https://doi.org/10.1007/978-3-031-63031-6_20

In response to this gap, this paper explores the utilization of genAI, specifically generative design methodologies, in addressing critical challenges in intervention planning, particularly within urban environments. Generative design, characterized by advanced algorithms and computational techniques, offers a systematic approach to automating the generation of design scenarios based on predefined parameters and constraints. By extending its application to multiscale intervention planning, including architectural design and urban revitalization, we aim to harness the potential of genAI in transforming urban landscapes.

The primary objective of this study is to showcase the potential of generative AI models in intervention planning applications. To this end, we introduce a simple Graphical User Interface (GUI) Desktop application developed for generating images and implementing generative design in real-world scenarios. Through experimentation and case studies, we demonstrate the feasibility and effectiveness of utilizing generative AI technology in intervention planning, thereby offering insights into its practical implications for shaping future urban environments.

The rest of this paper is organized as follow: (a) Sect. 2 provides a short description of the current literature review; (b) Sect. 3 describes the experimental setup; (c) Sect. 4 presents the experimental results; and (d) Sect. 5 concludes this work.

2 Related Work

The generative design [8] for engineering applications, like urban planning, architectural planning, renovations and other is a relatively newly introduced field. Currently, this technology lacks the necessary applicability of real case experimentation [5], thus further development and testing in practical and complex design scenarios are necessary.

In addition, it is worth mentioning that the generative design applications are not intended to replace the human factor (i.e., intelligence, opinion and design ideas). In essence, genAI models may support the architects, urban planners, etc. by proposing solutions, ideas, and scenarios. Thus, these tools can help the necessary engineers by inspiring then by the early stages of the designing process [15].

An indicative example can be the work of Han et al. [3], where they investigated the performance-based automatic urban design approach based on Deep Reinforcement Learning Generative Design algorithms and computer vision. By comparing their proposed approach with similar conventional approaches [1,11] they observed that their propose methodology is not limited by the number of the design variables, thus it can generate scenarios with different number of features, such as building of any shame.

Another example can be the work of Zhang et al. [13]. They developed a parametric generative algorithm for automatically generating green design scenarios of typical Chinese urban residences based on performance-oriented design flow. Similarly, Gan et al. [2] investigated the automation of the novel BIM-based graph data model based on the generative design of modular buildings. Finally, Wei et al. [12] investigate the application of the generative design approach for module construction.

The aforementioned works indicate that generative design is an emerging trend for various engineering applications and especially in architectural design, urban planning, renovation, and building construction. However, these works approach the generative design problem by developing generative algorithms similar to conventional and state-of-the-art approaches. In this work, we approach the generative design problem from a different angle. For inspiring the architects, urban planners, and designers generally, we investigate the usage of pre-trained image-based generative models for proposing intervention ideas by the generation of photorealistic images over a predetermined area of interest. Thus, the contribution of this work is that we investigate the usage of image-to-image and image inpainting algorithms for multi-scale intervention planning. Moreover, the proposed methodology of this manuscript has not been investigated yet, thus it is important to further enrich the current literature with new methodologies based on emerging and innovative technologies.

3 Experimental Setup

The main scope of this manuscript is to examine the potential of generative AI models, such as image-to-image and image inpainting techniques, for architectural design, urban planning and other forms of multi-scale intervention planning. To achieve this, we developed an application for generating images based on models from HuggingFace repository [4]. The AI Image Generator Application [6] supports both image-to-image [10] and image inpainting [9] technologies.

Figure 1 illustrates the image-to-image workflow, while Fig. 2 illustrates the image inpainting workflow. Both of these techniques need as input a base image and a description text prompt. The image inpainting needs to a mask bitmap (i.e., black and white) image for specifying the generated area of the image. The output in both techniques is a number N generated images.

The proposed methodology has been evaluated over two randomly selected alleys in Thessaloniki city. These areas share some common traits as the lack of green space and the existence of high buildings in the vicinity. Thus, these two case studies can be used effectively for testing image-to-image and image inpainting techniques potential intervention planning. Figure 3 illustrates the input base images. The secluded by the yellow dashed line areas indicate the inpainting masked selection, thus the area which will be changed after the potentially intervention.

The experimental results are evaluated by an architect based on the architectural composition and the realism of the generated images. Moreover, for evaluating the productivity scale of the proposed methodology and the equivalent architects work, the average time per generated image is calculated. Thus, for the experimental setup will be produced 15 images per technique per case study. In total, the final result is compromised by 60 (30 images per case study)

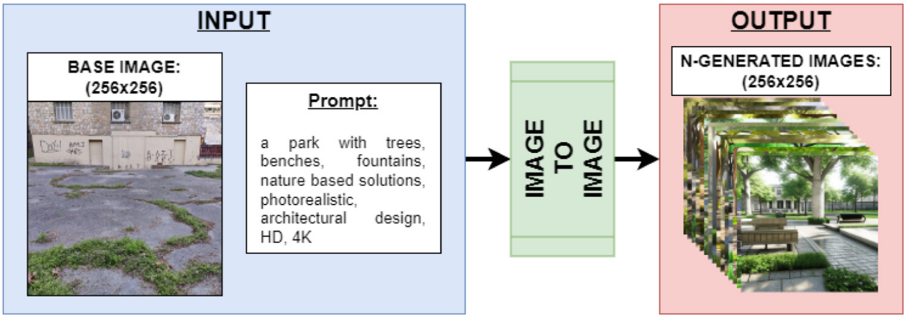


Fig. 1. A brief overview of the Image to Image workflow. **Input:** (a) Base Image; and (b) Description Text Prompt. **Output:** N-Generated Images indicating different scenarios of intervention planning.

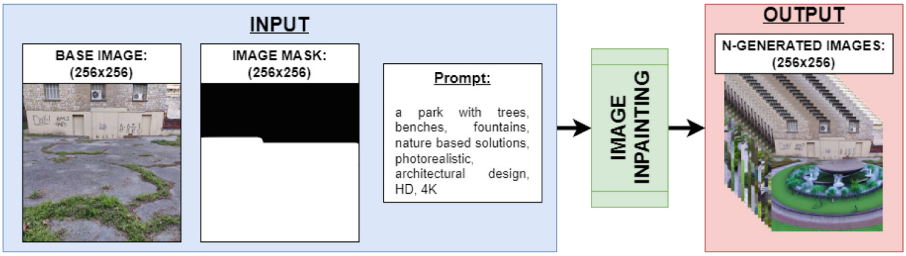


Fig. 2. A brief overview of the Inpainting Image workflow. **Input:** (a) Base Image; (b) Image Mask; and (c) Description Text Prompt. **Output:** N-Generated Images indicating different scenarios of intervention planning.

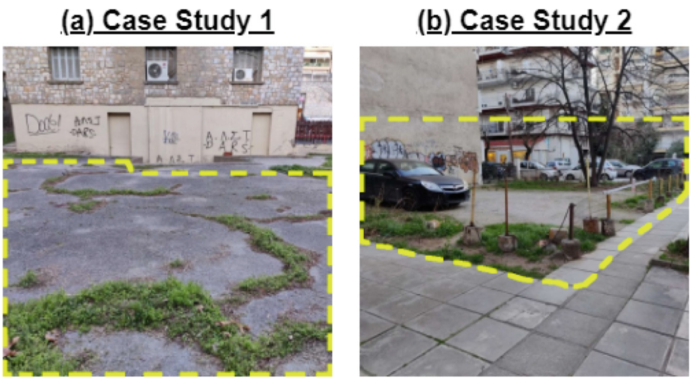


Fig. 3. The images used for the experiments. The area secluded by the yellow dashed line corresponds to the masked area for the image inpainting technique. (Color figure online)

generated images, illustrating different scenarios or possible ideas for intervene to the area of interest.

4 Experimental Results

In the Sect. 3 were presented the experimental setup, which is used for generating several images of photo-realistic alternative interventions over the two case studies (Fig. 3). Figures 4 and 5 present the generated images. For each case study were generated 15 images from image-to-image method and 15 images from image inpainting method.

For the evaluation of the results we asked the opinion of architect about the architectural composition and the implement-ability of each solution. For the image-to-image method the generative images illustrate good looking results, with correct photo-realism and similar to the description prompt in both case studies. However, image-to-image approach alters the environment of the original image and the scale of the intervention, thus the proposed solution are not implementable as is. However, the results can inspire the architect and help him during his architectural composition by providing him different visual ideas.

For the image inpainting approach the generative images alters only the area of intervention, thus the generative solutions in this method are referred to the original area of interest. In many cases, the generative solution respected the scale of the intervention as well, providing realistic and implementable results. However, the majority of the solutions lacked the architectural composition, which means that some of the generative solutions are implementable, but for achieving the best possible result, the architect needs to further process the solution.

In addition, based on architects' design methodology, the related needed time to produce a photo-realistic similar to the generative images varies from 1 to 4 h, depended on the architectural composition, the scale of the area and the depth of details, as well as the software that the photo-realistic will be produced. For the generative methods, the image-to-image approach needed approximately 3 min to produce a 256×256 generative image, using CUDA acceleration and an NVidia 3070 Ti 8GB graphic card. Using the same setting, the image inpainting approach needed approximately 4 min to produce an image. Table 1 summarizes the time differences for generating a photo-realistic images for intervention planning.

In general, the generative AI technology can really help the experts (i.e., architects, urban planners and other engineers in the design field). Even if the generated solutions are not the best possible, compared with and architect's composition, they still can be used for inspiring the experts. A strong advantage of this technology is that they can produce several solutions in a short time. Thus, this technology can be sufficiently used for intervention planning, however further research is necessary.

Fig. 4. The generated results for the Case Study 1. (a) Image-to-Image results; and (b) Image Inpainting results

Fig. 5. The generated results for the Case Study 2. (a) Image-to-Image results; and (b) Image Inpainting results

Table 1. Time duration for generating a photorealistic image of pixel size 256×256.

Generating Method	Time	Images per Hour
Architect	1–4 h	max 1
Image-to-Image	$\sim$3 min	$\sim$20
Image Inpainting	$\sim$4 min	$\sim$15

5 Conclusions

In conclusion, this study delves into the potential of generative AI for multi-scale intervention planning, with a focus on addressing urban green space scarcity. The experimentation in Thessaloniki's alleys, Greece, employing image-to-image and image inpainting methods, highlights genAI's promising role in architectural and urban planning. These models swiftly propose diverse solutions, providing valuable assistance even in preliminary planning phases. Despite challenges posed by current models trained on generalized data, specialized datasets and ongoing research promise to enhance their implementation. By overcoming these obstacles, generative AI can revolutionize intervention planning, fostering sustainable and vibrant urban landscapes.

Acknowledgments. This work is supported by the European Union Funded project euPOLIS "Integrated NBS-based Urban Planning Methodology for Enhancing the Health and Well-being of Citizens: the euPOLIS Approach", under the Horizon 2020 program H2020-EU.3.5.2., grant agreement No. 869448.

References

1. Eilouti, B.: Shape grammars as a reverse engineering method for the morphogenesis of architectural façade design. Front. Archit. Res. **8**(2), 191–200 (2019)
2. Gan, V.J.: BIM-based graph data model for automatic generative design of modular buildings. Autom. Constr. **134**, 104062 (2022)
3. Han, Z., Yan, W., Liu, G.: A performance-based urban block generative design using deep reinforcement learning and computer vision. In: Yuan, P.F., Yao, J., Yan, C., Wang, X., Leach, N. (eds.) CDRF 2020, pp. 134–143. Springer, Singapore (2021). https://doi.org/10.1007/978-981-33-4400-6_13
4. HuggingFace: Hugging face model repository (2023). https://huggingface.co/models
5. Jiang, F., et al.: Generative urban design: a systematic review on problem formulation, design generation, and decision-making. Prog. Plann. 100795 (2023)
6. JohnCrabs: AI image generator app (2024). https://gitlab.com/JohnCrabs/ai-image-generator-app
7. Kabisch, N., Strohbach, M., Haase, D., Kronenberg, J.: Urban green space availability in European cities. Ecol. Ind. **70**, 586–596 (2016)
8. Koenig, R., Miao, Y., Aichinger, A., Knecht, K., Konieva, K.: Integrating urban analysis, generative design, and evolutionary optimization for solving urban design problems. Environ. Plann. B Urban Anal. City Sci. **47**(6), 997–1013 (2020)

9. runwaym: Stable diffusion inpainting model (2023). https://huggingface.co/runwayml/stable-diffusion-inpainting
10. stabilityai: Stable diffusion image-to-image model (2023). https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0
11. Sun, C., Rao, J.: Study on performance-oriented generation of urban block models. In: Yuan, P., Xie, Y., Yao, J., Yan, C. (eds.) CDRF 2019, pp. 179–188. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-8153-9_16
12. Wei, Y., Choi, H., Lei, Z.: A generative design approach for modular construction in congested urban areas. Smart Sustain. Built Environ. **11**(4), 1163–1181 (2022)
13. Zhang, J., Liu, N., Wang, S.: Generative design and performance optimization of residential buildings based on parametric algorithm. Energy Build. **244**, 111033 (2021)
14. Zhang, Y., Mavoa, S., Zhao, J., Raphael, D., Smith, M.: The association between green space and adolescents' mental well-being: a systematic review. Int. J. Environ. Res. Public Health **17**(18), 6640 (2020)
15. Zheng, H., Yuan, P.F.: A generative architectural and urban design method through artificial neural networks. Build. Environ. **205**, 108178 (2021)

Security, Privacy and Ethics in Generative Intelligence

Exploring Inclusivity in AI Education: Perceptions and Pathways for Diverse Learners

Michael Pin-Chuan Lin[1](✉) and Daniel Chang[2]

[1] Mount Saint Vincent University, Halifax, NS, Canada
michael.lin@msvu.ca
[2] Simon Fraser University, Burnaby, BC, Canada
dth7@sfu.ca

Abstract. Artificial Intelligence (AI)'s rise in education brings benefits and challenges. Tools like Grammarly and ChatGPT have been part of our lives, particularly in the education sector. Consequently, it is important to explore how inclusive and adaptable these tools are for students from diverse cultures and languages. This study analyzes survey data from 87 undergraduate students in Canada to identify key themes such as the need for better language translation capabilities, cultural and linguistic inclusivity, and reducing bias and Eurocentrism in AI technologies. These findings form the basis of the Adaptive and Inclusive AI Learning (AIAL) theory, which emphasizes the importance of adaptability, inclusivity, and responsiveness in AI educational technologies to meet the diverse needs for students. The research highlights the need for AI educational tools to advance technologically while remaining culturally sensitive and accessible, in order to promote educational equity and for diverse learners. It suggests a collaborative approach involving cross-cultural developers, educators, and policymakers in creating AI technologies that not only enhance educational outcomes but also respect the diversity of the global student community.

Keywords: AI in Education · Inclusivity · Diversity · Educational Technology · Technology-enhanced Learning

1 Introduction and Literature

The creation of Artificial Intelligence (AI) technologies, such as chatbots and adaptive learning platforms, has significantly changed the way we learn and teach nowadays. These innovations bring the potential to improve student engagement, enhance understanding of course material, and facilitate skill development [20]. However, the implementation of AI in educational settings requires a careful examination of its impact on equity and inclusivity [5, 15, 19]. Given the diverse nature of education, it is crucial to assess whether AI technologies can meet the needs of our learners, ensuring our practice does not overlook and exclude our student needs with respect to their language, culture, or learning preferences. Concerns about potential biases, ethical issues, and equity of technological benefits underscore the importance of thoroughly examining the effectiveness and inclusiveness of these technologies [7, 10, 12].

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 237–249, 2024.
https://doi.org/10.1007/978-3-031-63031-6_21

As the influence of AI in education continues to grow, its ethical implications have become an important topic of discussion in education [10]. The increasing developmental progress in AI capabilities has received a lot of attention and concerns, particularly in creating new Large Language Models (LLMs) for educational purposes. Specifically, one of the concerns is that education may be vulnerable to flawed human logic, data errors, biases, and the trial-and-error nature of AI [12, 21, 26]. The key question is whether the rapid advancement in educational AI can align with ethical considerations that cover the needs and limitations of students, teachers, policymakers and the academic environment [3, 29]. This is why there have been some emergent recent objectives that intend to establish more specific AI conducts, such as "Fairness," "Accountability," "Transparency," and "Ethics," collectively referred to as the "FATE" framework [14, 21, 31].

In a recent examination of AI's role in higher education, Zawacki-Richter and others have reviewed 146 papers on AI applications [33]. They found that these papers explored various applications, such as student profiling, intelligent tutoring, personalized learning systems, and automated assessment and feedback. This review illuminated a major trend: the bulk of AI research is published in specialized AI, educational technology, or computing journals, rather than in journals dedicated to higher education discourse. Additionally, Zawacki-Richter and others highlighted concerns that AI might contribute to "digital colonialism," [33] a phenomenon that Kwet describes as the extension and reinforcement of existing societal inequalities through technology [17]. Similarly, Memarian and Doleck's review is based on the "FATE" framework in the context of AI in education [21], and their findings indicate that while "fairness" has been comprehensively defined both descriptively and quantitatively. The crucial aspects of "accountability" and "transparency" remain relatively absent within the existing literature on AI in education [21]. Therefore, the juxtaposition of these perspectives shows an urgent need for a comprehensive exploration of the intersection between AI and higher education from student perspectives, encompassing both technical and social dimensions.

When exploring learner perspectives with AI, the principle of transparency emerges as a cornerstone for integrating AI into educational settings, as delineated by Memarian and Doleck's FATE framework [21]. This principle not only involves the ability of AI to clarify its internal mechanics and reveal its guiding policies, but it also has a significant impact on student learning. By understanding learners' experiences, we can also identify possible biases in their learning processes. These biases can then be applied to the creators of AI systems, potentially resulting in educational inequalities [13, 22]. Further, the issue of transparency is particularly important in avoiding perpetuating stereotypes or marginalizing individuals from diverse cultural and linguistic backgrounds [27, 32]. Transparency encompasses technical aspects as well as ethical considerations regarding student data, consent, and the right to seek redress [30]. By focusing on transparency and how students use AI in their learning processes, educators can promote more equitable and accountable AI systems in education. Additionally, it is essential to provide comprehensive information to data subjects and transparently communicate the usage, benefits, and drawbacks of AI applications in line with the goal of promoting transparency in AI-enabled learning experiences [16, 25].

Our study aims to contribute to the ongoing conversation about AI in education by focusing specifically on inclusivity, diversity, and the crucial aspects of accountability and transparency in AI educational technologies. Building upon the insights shared by Memarian and Doleck [21], our goal is to develop an exploratory yet evidence-based conceptual framework that emphasizes the shared responsibility of both AI systems and educators in shaping an inclusive and adaptable learning environment.

While Big Data and AI technologies offer the potential for personalized adaptive learning and efficient learning management [2], achieving this mission depends on the responsible and fair use of these technologies. Research (i.e., [21, 25]) has shown that addressing the accountability of AI in education requires attention at both micro and macro levels, taking into account the diverse legal, regulatory, and moral frameworks across different geographical regions.

The incorporation of AI into teaching practices introduces a shift in the dynamics of accountability. Educators who engage with AI tools enter a symbiotic relationship where understanding and responsibility are shared among the students, the teacher, and the AI system [1]. This shift demands a thorough grasp of the capabilities and limitations of both the educator and the AI system, ensuring that accountability is appropriately assigned and managed. Therefore, our study aims to deepen the comprehension of accountability in AI-enhanced education by exploring how AI can be developed and utilized in ways that align with diverse educational policies and address the unique needs of various student demographics from the student's perspective. Through this exploration, we hope to contribute to an educational digital transformation that is not only inclusive and effective but also accountable, ensuring fair advantages for learners from all walks of life. Together, we try to answer the research questions below:

1. To what extent are current AI educational tools like Grammarly or ChatGPT perceived as being inclusive and responsive to the cultural and linguistic diversity among students?
2. What specific modifications or enhancements can be implemented in AI educational technologies to better accommodate the needs of a diverse student population?

2 Methods

2.1 Participants

The study collected 87 undergraduate students' survey responses. These students primarily enrolled in two distinct education courses at a comprehensive university located in Coastal Western Canada. There were 58 students enrolled in a year 3 elective education course. 29 students were year 2 general education students intending to pursue teacher education. Demographic information was not gathered in this exploratory study as it was not deemed a primary factor influencing the research outcomes. This year 2 course focused specifically on developing students' understanding of research methodology in the field of education. Both courses spanned a duration of 13 weeks, and the participants were engaged in an array of educational activities and curriculum relevant to their respective courses.

2.2 Procedure and Instrument

Participants were invited to complete an end-of-course survey. This survey was part of a broader initiative by the second author's department, aimed at examining and enhancing the learning experience in online courses. Prior to the commencement of the study, the purpose of this initiative was reviewed by the institutional research ethics board and obtained the status of exemption due to its intention to improve future course offerings and practitioners' teaching strategies. The survey was anonymous and aimed at collecting student perceptions of online learning experiences and their independent learning. The survey administration occurred in the 12th week of the courses, aligning with the culmination of the core curriculum, and allowing students to reflect on their entire course experience and their AI use. Upon completion of the survey, students were awarded bonus points as part of their course grades. The survey instrument consisted of 5 open-ended items. The survey questions were exploratory in nature and were designed to understand how students use AI and explore the intersection of AI technologies and constructivist learning theories. Constructivism posits that learning is an active, contextual process where learners construct knowledge based on their experiences and interactions [9]. This theoretical framework is crucial in guiding the development of these questions, so it helps us understand how AI tools like ChatGPT or other generative AI tools can be integrated into learning environments. It was administered through the learning management system (i.e., Canvas), a platform familiar to the participants due to its extensive use in their courses. The survey was designed to elicit detailed, qualitative responses regarding the students' experiences with AI technologies in their learning, their perceptions of the inclusivity and effectiveness of these technologies, and their suggestions for improvements.

The open-ended nature of the survey items was chosen to encourage in-depth and personal responses, providing rich qualitative data for analysis [23]. The items were specifically crafted to align with the research questions, focusing on students' perceptions of AI technologies in education, their experiences regarding the inclusivity and responsiveness of these technologies, and their recommendations for enhancements.

The responses gathered from the survey were intended to be analyzed using appropriate qualitative data analysis methods, such as thematic analysis for the two research questions. This methodology was aimed at gaining a comprehensive understanding of the participants' views and experiences, thereby contributing valuable exploratory insights into the role and impact of AI technologies in educational settings.

2.3 Data Analysis

In this study, data analysis was conducted qualitatively using thematic analysis. The thematic analysis began with a detailed examination of the open-ended survey responses, identifying initial codes that encapsulated key aspects of the students' experiences and perceptions of AI technologies in their learning [4, 11, 24]. These codes were then carefully grouped into broader themes, reflecting the overarching patterns in the data. This method allowed rooms for an in-depth exploration of how AI technologies impacted students' learning experiences, focusing on aspects such as engagement, understanding, skill development, and perceptions of inclusivity and responsiveness to diverse needs.

Two researchers held weekly meetings to discuss the codes and agreed on the patterns and codes assigned to each entry in the dataset.

3 Results

3.1 To What Extent Are Current AI Educational Tools like Grammarly or ChatGPT Perceived as Being Inclusive and Responsive to the Cultural and Linguistic Diversity Among Students?

The exploration of learners' perceptions of AI educational tools like Grammarly or ChatGPT in relation to cultural and linguistic diversity among students revealed a range of themes, as noted in Table 1.

When discussing **Language and Translation Capabilities**, participants recognized the potential of AI in facilitating multilingual communication. One pointed out, "ChatGPT claims to be a multilingual chatbot that supports over 50 different languages," highlighting AI's language support. Nevertheless, another participant offered a more nuanced perspective based on personal experience, stating, "I personally have never used chatgpt in my mother tongue, Swahili. However, I do know that it works cause I just tested it out," implying the necessity for further enhancements in AI's linguistic adaptability.

With respect to **Cultural and Linguistic Inclusivity**, learners identified limitations in AI's ability to accommodate various cultural nuances in language. An observation made was, "Different types of English and the different nuances in its are normally corrected into and formal English instead of the AI picking up that it is another form of English." This statement reflects the challenge for AI in recognizing and valuing linguistic diversity.

The issue of **Bias and Eurocentrism** was highlighted by learners who felt that AI predominantly caters to English-speaking Western societies. This sentiment was captured in the statement: "In other words, these technologies are mainly served by English-speaking Western societies, which can be seen from their language Settings and login software restrictions," underlining the need for more inclusive AI tools.

The theme of **Limitations in Understanding Cultural Contexts** was mentioned by students. Moreover, the theme of **Need for Inclusivity** emphasized AI's role in supporting a wide range of learners. Quotes such as "As an international student myself, English is not my mother tongue… I could seek AI for help by asking it to interpret the requirements in simple words," and "In an elementary school I volunteer in, there is a child who is nonverbal, and he uses an AI program to communicate with everyone," illustrate the use of AI tools in various educational settings to accommodate to different linguistic and educational needs.

These themes and the accompanying quotes provide valuable insights into the current capabilities and limitations of AI educational tools in addressing cultural and linguistic diversity. They underscore the importance of improving the inclusivity, language capabilities, and accessibility of AI to effectively serve a diverse student population.

Table 1. Themes for student perceptions of inclusive and responsive to cultural and linguistic diversity

Identified Themes	N	Sample Quotes
Language and Translation Capabilities	13	"ChatGPT can be used to simplify information from large materials. This can be especially helpful for students with disabilities or whose first language is not English. Further-more, ChatGPT is good at translating from various languages, such as French. I have also noticed that ChatGPT is language-inclusive. I can read and write Punjabi and have used the Punjabi keyboard on ChatGPT to see if it can generate responses, and it works!" (Student 40)
Cultural and Linguistic Inclusivity	20	"I can assume that most of its information is pulled from the West and, therefore, has an English-based system and pulls from North American websites or articles. However, I think it also has the capability to translate languages and search different diverse cultures if prompted." (Student 18)
Bias and Eurocentrism	11	"I don't think these artificial technologies can meet the diverse needs of all students. First of all, most of the artificial technology is developed by Western developed countries. In other words, these technologies are mainly served by English-speaking Western societies, which can be seen from their language Settings and login software restrictions" (Student 9)
Limitations in Understanding Cultural Contexts	13	"both systems have difficulties when taking into account the different needs of students from different backgrounds, whether they be linguistically or culturally. Even while ChatGPT is flexible in its language usage, it might not fully understand linguistic or cultural quirks, which makes it less useful for those who are not native English speakers or come from diverse cultural backgrounds" (Student 55)
Need for Inclusivity	16	"Sometimes I have difficulty understanding assignment requirements or certain terminologies, I could seek AI for help by asking it to interpret the requirements in simple words. Also, ChatGPT supports conversation in Chinese. It is a great study tool even outside of the academic setting" (Student 44)

N = counts of the instances from dataset

3.2 What Specific Modifications or Enhancements Can Be Implemented in AI Educational Technologies to Better Accommodate the Needs of a Diverse Student Population?

To effectively address the third research question on specific modifications or enhancements in AI educational technologies for a diverse student population, several key themes emerge, underscored by student perspectives. Table 2 lists all the identified themes.

The theme, **Accessibility and Inclusivity,** stands out as a significant, most occurring instance from this data. Students are calling for making AI technologies available and usable for a diverse range of students, including those with disabilities, non-native English speakers, and individuals from various cultural and linguistic backgrounds. Student 14 highlighted that "An improvement that can be made for Al technologies would be having it be multilingual to support all students," emphasizing the need for improving inclusivity.

Secondly, the theme of **Enhanced Educational Support** occurs 18 times in the dataset. Some students mentioned the potential of AI technologies to support educational processes, whether through personalized learning, providing detailed explanations, or offering tutoring in multiple languages. One student's statement, "I think that making AI technologies more responsive to the needs of a diverse student population, it could help generate smaller aspects of a paper," highlights the importance of responsiveness to diversity.

Ethical and Cultural Sensitivity emphasized the need for AI to be developed and used in ways that are ethically sound and culturally sensitive, recognizing and respecting the diversity of users' backgrounds and perspectives. This involves accommodating various learning styles and abilities, with one student suggesting, "Invest in creating solutions that work for everyone, every student and every need."

Technological Affordability and Accessibility is a crucial theme that focuses on the economic aspects of accessing AI technologies, emphasizing the need for affordable and flexible pricing models to make these tools accessible to a wider range of students. Student 1 mentioned, "Currently, there is only a free 30 day trial for ChatGPT and is currently $20 USD per month…ChatGPT should have a similar pricing structure to make it more accessible to more students in terms of cost." This quote highlights the importance of making it economically accessible to users and students.

Addressing the challenges of **Transparency and Trust in AI**, students express the need for clarity about how AI technologies generate their responses, including the sources of their information, to build trust and reliability among users. A student's insight (Student 74), "Right now, I don't know how AI generates its answers, what data sets it uses to pool its ideas from… It would help if their answers came with greater transparency about where they got their insights and information from," underscores the importance of transparency and responsible disclosure of training data.

In summary, these themes collectively call for AI educational technologies that are financially accessible, transparent, culturally sensitive, linguistically inclusive, and personalized to individual user contexts. These enhancements are essential for establishing an equitable and effective learning environment for all students.

Table 2. Themes for students' opinions on AI improvements for diverse student needs.

Identified Themes	N	Sample Quotes
Accessibility and Inclusivity	24	"Improvements I would suggest would be to allow people to upload pictures of a math problem or question so that ChatGPT can see the picture and help answer the question in detail." (Student 15)
Enhanced Educational Support	18	"I think AI has the capability to be a very, very powerful educative tool…it flourishes is in its ability to teach in any language, which would be very helpful for ESL learners or exchange students." (Student 5)
Ethical and Cultural Sensitivity	16	"I think AI needs to focus entirely on reducing the tediousness of schoolwork… The point is to support learning, not harm learning." (Student 7)
Technological Affordability and Accessibility	10	"I feel like they do have the means to already make it so inclusive but seeing the rising popularity has made it so they've locked these changes and the only way to get these improvements is by paying for access." (Student 42)
Transparency and Trust in AI	14	"Right now, I don't know how AI generates its answers, what data sets it uses to pool its ideas from… It would help if their answers came with greater transparency about where they got their insights and information from." (Student 74)

N = counts of the instances from dataset

4 Deriving the Adaptive and Inclusive AI Learning (AIAL) Theory

The theory of Adaptive and Inclusive AI Learning (AIAL) is conceptualized to address the explicit needs and challenges identified through our survey data. It underscores the necessity for AI educational technologies to progress in adaptability, inclusivity, and responsiveness. This need is directly reflected in the survey responses, where students articulated the desire for AI tools that understand diverse cultural contexts and linguistic nuances.

The theory proposes expanding the scope of **Adaptive Learning Enhancement** to highlight the indispensable role of AI technologies in providing personalized educational support that is linguistically adaptable and culturally sensitive. This expansion is rooted in direct observations from our survey, particularly students' feedback on the current limitations of AI in recognizing linguistic diversity and cultural nuances. For instance, one student's critique of AI's challenges in accommodating various English dialects illuminates the theory's focus on linguistic versatility (Sect. 3.1, "Language and Translation Capabilities"). By integrating multilingual capabilities and recognizing cultural

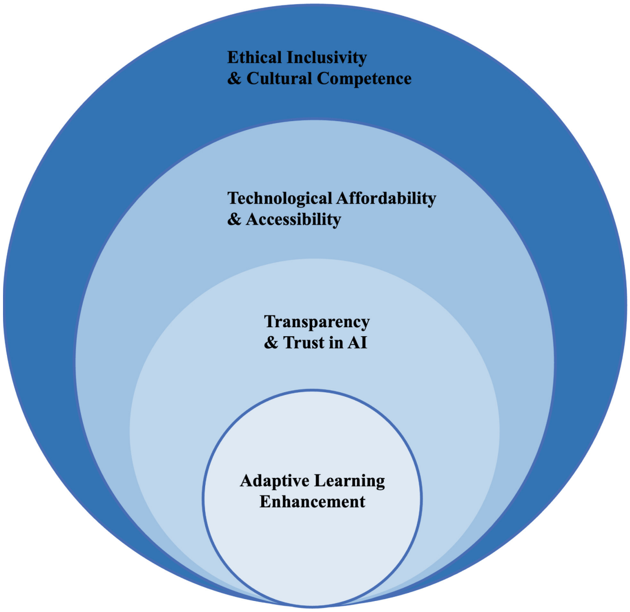


Fig. 1. Representation of AIAL Theory

diversity within AI tools, as evidenced by student feedback, we can bridge linguistic gaps and create more inclusive learning experiences for ESL learners and students with unique educational needs [5, 6].

Another major aspect is **Ethical Inclusivity and Cultural Competence**, which echoes the survey participants' call for AI tools developed with a deep appreciation for cultural and linguistic diversity. This component of the theory emerges from the learners' highlighted experiences of bias and a Eurocentric focus on existing AI technologies, advocating for the support of multilingual capabilities and the integration of cultural sensitivities in AI education tools. By minimizing biases and ensuring equitable access for students worldwide, AI tools can be more inclusive and accessible [27, 28, 32]. It also calls for the integration of ethical principles in AI development, such as careful selection and diversification of datasets, robust evaluation protocols to identify and rectify biases, and transparency in AI methodologies and data sources. These measures align AI educational tools with educational goals, promote critical thinking, and enhance learning outcomes while respecting diverse cultures, languages, and perspectives [8, 13, 27].

The theory also emphasizes **Technological Affordability and Accessibility**, directly correlating with students' concerns over economic barriers to AI tool access as reported in the survey findings. It proposes diverse pricing strategies and institutional licenses to democratize access and ensure that students from varied socioeconomic backgrounds can benefit from these technologies, addressing direct feedback regarding the current prohibitive costs of AI educational tools [8, 13].

Moreover, the AIAL theory highlights the importance of **Transparency and Trust in AI**, advocating for clear explanations of AI methodologies and data sources in response to students' requests for greater transparency about how AI generates its responses. His principle aims to establish trust and reliability, fostering an environment where AI technologies are recognized as credible and valuable educational aids [5, 19].

In short, the AIAL theory offers a strategic blueprint for advancing and embedding AI within educational settings, as shown in Fig. 1. It is a response to the articulated needs of a global learner demographic, emphasizing the importance of a synergistic effort among technologists, educational professionals, and policymakers to create AI applications that are accessible, ethically principled, culturally attuned, and committed to fostering an inclusive and equitable learning environment for all students.

5 Discussion and Implications

Based on the AIAL theory and the findings, the discussion surrounding the use of AI in education highlights a more nuanced understanding of inclusivity, cultural competence, and the need for continuous adaptation to diverse learner needs [19]. These insights not only reinforce the significance of the AIAL theory but also necessitate its expansion to address emerging challenges and opportunities in AI-driven education.

Recent research into students' perceptions of AI tools like Grammarly and ChatGPT reveals a complex and dynamic landscape where technological potential intersects with the practical realities of cultural and linguistic diversity. While AI's ability to bridge language barriers and facilitate learning in multiple languages is recognized, the feedback emphasizes the need for AI systems to go beyond mere translation. It calls for a deeper integration of cultural nuances and a more authentic representation of linguistic diversity, in line with the AIAL theory's major focus on Ethical Inclusivity and Cultural Competence. This expanded focus brings together previously distinct elements into a unified approach that acknowledges the interconnected nature of ethics, culture, and inclusivity in AI development [27, 28, 32].

Students' experiences highlight a significant gap in AI's responsiveness to the various forms of English and other languages, indicating that current AI technologies often default to standard forms and overlook the rich diversity of dialects and cultural expressions. This observation echoes the AIAL theory's call for AI to respect and value linguistic diversity, ensuring that AI educational tools are genuinely inclusive and capable of serving a global student population [13, 27].

Moreover, the themes of Technological Affordability and Accessibility, as well as Transparency and Trust in AI, as evidenced by the new findings, emphasize the importance of making AI tools economically accessible and transparent in their operations [18]. This reinforces the AIAL theory's assertion that equitable access to AI technologies is crucial for educational equity. The theory's advocacy for clear methodologies and data sourcing in AI development aligns with students' desire for transparency, enhancing trust and reliability in AI as an educational resource [8, 13, 18].

The implications of these findings extend to developers, educators, and policymakers, urging a collaborative effort to refine AI technologies in ways that resonate with the diverse needs of learners. It underscores the need to incorporate user feedback and diverse datasets in the AI development process to mitigate biases and enhance cultural and linguistic inclusiveness. Additionally, the AIAL theory establishes a critical foundation for subsequent scholarly inquiry, steering explorations at the confluence of AI, educational practice, and diversity. It delineates pathways for rigorous examination into the strategic development and deployment of AI to extend beyond mere academic support to significantly enrich the educational experiences of a diverse student body. This approach aspires to cultivate a learning landscape that is inclusively designed, equitable in access, and deeply sensitive to the broad spectrum of cultural and linguistic identities, contributing towards a more holistically inclusive, equitable, and culturally responsive educational ecosystem.

6 Future Directions

Advancing AI in education, as proposed by the AIAL theory, requires a focus on developing AI systems that are not only technically proficient but also culturally nuanced. These systems should be designed to understand and integrate a wide range of cultural norms, languages, and expressions. This necessitates a collaborative endeavor across AI developers, educators, and stakeholders to ensure these technologies are accessible and affordable, with a special focus on reaching learners from marginalized backgrounds.

Furthermore, the ethical deployment of AI within educational frameworks is a critical area of research. This involves creating AI solutions that uphold academic integrity and mitigate over-reliance, while also encouraging ethical engagement from both students and teachers. It is important to highlight the adaptability of AI to meet the changing demands of educational contexts, particularly in supporting individuals with disabilities and diverse learning preferences. By pursuing these objectives, AI can become a valuable resource for equitable, impactful, and conscientious education, aligning with the inclusive and ethically mindful aspirations of the AIAL theory.

Acknowledgments. This research was funded by ML's institutional conference funding from Mount Saint Vincent University. Part of this research draws on DC's funding supported by the Institutional Explore Grant from Social Sciences and Humanities Research Council (SSHRC) [Project R832319].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bearman, M., Ryan, J., Ajjawi, R.: Discourses of artificial intelligence in higher education: a critical literature review. High. Educ. **86**(2), 369–385 (2023). https://doi.org/10.1007/s10734-022-00937-2
2. Beerkens, M.: An evolution of performance data in higher education governance: a path towards a 'big data'era? Qual. High. Educ. **28**(1), 29–49 (2022). https://doi.org/10.1080/13538322.2021.1951451
3. Borenstein, J., Howard, A.: Emerging challenges in AI and the need for AI ethics education. AI and Ethics **1**, 61–65 (2021). https://doi.org/10.1007/s43681-020-00002-7
4. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa

5. Chang, D.H., Lin, M.P.-C., Hajian, S., Wang, Q.Q.: Educational design principles of using AI chatbot that supports self-regulated learning in education: goal setting, feedback, and personalization. Sustainability **15**(17), 12921 (2023). https://doi.org/10.3390/su151712921
6. Chen, L., Chen, P., Lin, Z.: Artificial intelligence in education: a review. IEEE Access **8**, 75264–75278 (2020). https://doi.org/10.1109/ACCESS.2020.2988510
7. Chi, N., Lurie, E., Mulligan, D.K.: Reconfiguring diversity and inclusion for AI ethics. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 447–457 (2021). https://doi.org/10.1145/3461702.3462622
8. Garg, S., Sharma, S.: Impact of artificial intelligence in special need education to promote inclusive pedagogy. Int. J. Inform. Educ. Technol. **10**(7), 523–527 (2020). https://doi.org/10.18178/ijiet.2020.10.7.1418
9. Gibson, D., Kovanovic, V., Ifenthaler, D., Dexter, S., Feng, S.: Learning theories for artificial intelligence promoting learning processes. Br. J. Edu. Technol. **54**, 1125–1146 (2023). https://doi.org/10.1111/bjet.13341
10. Holmes, W., et al.: Ethics of AI in education: towards a community-wide framework. Int. J. Artif. Intell. Educ. **32**(3), 504–526 (2022). https://doi.org/10.1007/s40593-021-00239-1
11. Holloway, I., Todres, L.: The status of method: flexibility, consistency and coherence. Qual. Res. QR **3**(3), 345–357 (2003). https://doi.org/10.1177/1468794103033004
12. Jobin, A., Ienca, M.: The global landscape of AI ethics guidelines. Nature Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
13. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023). https://doi.org/10.1016/j.lindif.2023.102274
14. Khosravi, H., et al.: Explainable artificial intelligence in education. Comput. Educ. Artif. Intell. **3**, 100074 (2022). https://doi.org/10.1016/j.caeai.2022.100074
15. Kılınç, S.: Embracing the future of distance science education: opportunities and challenges of ChatGPT integration. Asian J. Distance Educ. **18**(1), 205–237 (2023). https://doi.org/10.5281/zenodo.7857396
16. Kong, S.C., Cheung, W.M.Y., Zhang, G.: Evaluating an artificial intelligence literacy programme for developing university students' conceptual understanding, literacy, empowerment and ethical awareness. Educ. Technol. Soc. **26**(1), 16–30 (2023). https://www.jstor.org/stable/48707964
17. Kwet, M.: Digital colonialism: US empire and the new imperialism in the Global South. Race & Class **60**(4), 3–26 (2019). https://doi.org/10.1177/0306396818823172
18. Lin, M.P.C.: A Proposed Methodology For Investigating Chatbot Effects in Peer Review. (Unpublished doctoral dissertation). Faculty of Education, Simon Fraser University, Canada (2020)
19. Lin, M.P.C., Chang, D.: CHAT-ACTS: a pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning. Comput. Educ. Artif. Intell. **5**, 100167 (2023). https://doi.org/10.1016/j.caeai.2023.100167
20. Lin, M.P.C., Chang, D.: Enhancing post-secondary writers' writing skills with a chatbot: a mixed-method classroom study. J. Educ. Technol. Soc. **23**(1), 78–92 (2020). https://doi.org/10.30191/ets.202001_23(1).0006
21. Memarian, B., Doleck, T.: Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: a systematic review. Comput. Educ. Artif. Intell. **5**, 100152 (2023). https://doi.org/10.1016/j.caeai.2023.100152
22. Moran, T.C.: Racial technological bias and the white, feminine voice of AI VAs. Commun. Critical/Cult. Stud. **18**(1), 19–36 (2021). https://doi.org/10.1080/14791420.2020.1820059
23. Moser, A., Korstjens, I.: Series: practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. Eur. J. Gener. Pract. **24**(1), 9–18 (2018). https://doi.org/10.1080/13814788.2017.1375091

24. Nemorin, S., Vlachidis, A., Ayerakwa, H.M., Andriotis, P.: AI hyped? A horizon scan of discourse on artificial intelligence in education (AIED) and development. Learn. Media Technol. **48**(1), 38–51 (2023). https://doi.org/10.1080/17439884.2022.2095568
25. Pagallo, U.: From automation to autonomous systems: a legal phenomenology with problems of Accountability. In: 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, pp. 17–23 (2017)
26. Raji, I.D., Scheuerman, M.K., Amironesei, R.: You can't sit with us: exclusionary pedagogy in AI ethics education. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 515–525 (2021). https://doi.org/10.1145/3442188.3445914
27. Salas-Pilco, S., Xiao, K., Oshima, J.: Artificial intelligence and new technologies in inclusive education for minority students: a systematic review. Sustainability **14**(20), 13572 (2022). https://doi.org/10.3390/su142013572
28. Sætra, H.S.: Generative AI: here to stay, but for good? Technol. Soc. **75**, 102372 (2023). https://doi.org/10.1016/j.techsoc.2023.102372
29. Sikdar, S., Lemmerich, F., Strohmaier, M.: Getfair: generalized fairness tuning of classification models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 289–299. https://doi.org/10.1145/3531146.3533094
30. Ungerer, L., Slade, S.: Ethical considerations of artificial intelligence in learning analytics in distance education contexts. In: Learning Analytics in Open and Distributed Learning: Potential and Challenges, pp. 105–120. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-0786-9_8
31. Woolf, B.: Introduction to IJAIED special issue, FATE in AIED. Int. J. Artif. Intell. Educ. **32**(3), 501–503 (2022). https://doi.org/10.1007/s40593-022-00299-x
32. Yang, H., Kyun, S.: The current research trend of artificial intelligence in language learning: a systematic empirical literature review from an activity theory perspective. Austral. J. Educ. Technol. **38**(5), 180–210 (2022). https://doi.org/10.14742/ajet.7492
33. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int. J. Educ. Technol. High. Educ. **16**(1), 39 (2019). https://doi.org/10.1186/s41239-019-0171-0

Generative Intelligence for Applied Natural Language Processing

A Rule-Based Chatbot Offering Personalized Guidance in Computer Programming Education

Christos Papakostas(✉), Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou

Department of Informatics and Computer Engineering, University of West Attica, Egaleo, Greece
{cpapakostas,ctrouss,akrouska,csgouro}@uniwa.gr

Abstract. In the field of education, the integration of Artificial Intelligence technologies, particularly chatbots, has transformed education by offering personalized guidance and support to learners. This paper introduces a well-designed rule-based chatbot system tailored for computer programming education, specifically Java. The architecture comprises essential components such as the User Input Handler, Rule Engine, Response Generator, and Knowledge Base, each carefully crafted to fulfill distinct functions. The chatbot is programmed with a wide range of rule categories, covering basic greetings and farewells to detailed explanations of syntax errors, logical guidance, variable declarations, loop structures, common pitfalls, and valuable resource recommendations. The study evaluated the chatbot's efficacy in personalized guidance for computer programming education, engaging 50 undergraduate students. Feedback was collected through a structured questionnaire covering user satisfaction, query accuracy, response time, and personalized guidance. Results indicate high user satisfaction and proficiency in addressing programming queries promptly. Strengths include personalized guidance aligned with individual learning needs, while limitations in conversational depth and adaptability, alongside challenges like technical issues and user engagement, were noted. Suggestions for refinement are provided to enhance effectiveness and applicability in educational contexts.

Keywords: Chatbot · Computer Programming Education · Rule-based Chatbot · Java · Personalized Guidance

1 Introduction

The advent of online education has ushered in a transformative era, breaking down geographical barriers and democratizing access to knowledge. As a result, learners worldwide are embracing the flexibility and convenience offered by virtual learning environments [1–5]. In this dynamic educational landscape, the integration of artificial intelligence (AI) stands out as a key enabler, poised to revolutionize the online learning experience.

Recognizing the importance of tailoring the learning environment to users, AI emerges as a pivotal technology in shaping the future of education. The ability to

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 253–264, 2024.
https://doi.org/10.1007/978-3-031-63031-6_22

understand and adapt to the specific needs, preferences, and learning styles of each learner has the potential to significantly enhance the efficacy of educational interventions [6–10]. Through intelligent algorithms and machine learning techniques, AI can not only personalize content delivery but also offer real-time feedback and guidance, fostering a more engaging and effective learning experience.

Within the realm of AI tools in education, chatbots have emerged as versatile and interactive companions, capable of providing valuable support to learners [11]. Chatbots, as AI-driven conversational agents, hold the promise of enhancing the educational journey by offering personalized guidance, answering queries, and facilitating a more natural and intuitive interaction. Their possibilities extend beyond routine tasks, encompassing adaptive learning scenarios, immediate feedback mechanisms, and the ability to scaffold complex concepts.

Various techniques have been employed to implement chatbots in education, showcasing a diverse landscape of approaches tailored to meet the unique demands of different educational settings, including machine learning (ML), natural language processing (NLP) and rule-based approaches [12]. While ML and NLP approaches offer advanced capabilities, their complex underlying mechanisms might introduce challenges in terms of transparency and interpretability, a fact that is not met in rule-based approaches. Rule-based chatbots operate on predefined rules and decision-making pathways, allowing for explicit control over their behavior. The clarity of these rules facilitates alignment with pedagogical goals, ensuring that the chatbot's responses align with educational objectives. This explicit control makes rule-based approaches particularly relevant in educational settings where a clear and understandable interaction is crucial for effective learning support.

This paper contributes to the growing discourse by presenting a rule-based chatbot specifically designed to offer personalized guidance in the domain of computer programming education, and specifically the programming language Java. By incorporating rule categories such as greetings (initiating a conversation with personalized messages), farewells (concluding interactions courteously), default responses (providing fallback messages), syntax error explanations (assisting users in rectifying syntax discrepancies), logic error guidance (helping users identify and rectify logic errors), variable declaration explanations (clarifying variable declaration syntax), loop structure help (providing guidance on loop structures), common mistakes warnings (alerting users to potential coding pitfalls), and resource recommendations (suggesting relevant learning materials), this chatbot aims to address the unique challenges posed by programming education. Ultimately, it seeks to enrich the educational journey for learners in this critical field by delivering tailored and effective support.

2 Related Work

As the literature on chatbots in education burgeons, researchers have undertaken comprehensive explorations into the diverse applications and impacts of these AI tools, uncovering their potential across a spectrum of educational domains. For instance, in the domain of language learning [13–17], chatbots have been leveraged to facilitate immersive language practice, offering learners a dynamic conversational experience to enhance their linguistic proficiency. The application of chatbots in language education often involves sophisticated natural language processing (NLP) techniques to comprehend and generate contextually relevant responses, fostering interactive language acquisition.

In the domain of programming education [18–22], chatbots have emerged as invaluable aids, guiding learners through the intricacies of coding syntax, logic, and problem-solving. Techniques such as machine learning-based approaches have been employed to create chatbots that adapt to the diverse coding challenges presented by learners. These systems analyze patterns in coding errors and misconceptions, tailoring their feedback to address specific programming issues effectively. Rule-based approaches, on the other hand, provide a structured and explicit framework for guiding learners through the step-by-step processes of coding, ensuring clarity and alignment with pedagogical goals.

Moreover, the choice of implementation techniques is influenced by the desired level of personalization within educational interactions [11, 12, 23]. Machine learning-based chatbots excel in adapting to individual learning styles, providing tailored feedback based on user behavior and performance data. Rule-based chatbots, while transparent and interpretable, offer explicit control over the learning pathway, ensuring that educational content is delivered in a structured manner aligned with curriculum objectives.

3 Logical Architecture

The logical architecture of the rule-based chatbot involves the following components: User Input Handler, Rule Engine, Response Generator and Knowledge Base (Fig. 1).

The User Input Handler plays a crucial role in the chatbot's operation. Its primary responsibility lies in receiving and processing user inputs. Functionally, it excels at identifying trigger phrases or patterns within the input, aiming to discern the user's intent effectively. The culmination of its processing efforts results in the extraction of relevant information, which is then meticulously passed on to the Rule Engine for further evaluation.

At the heart of the chatbot's decision-making process, the Rule Engine assumes the responsibility of evaluating user input and contextual cues. Its functionality revolves around the application of a meticulously crafted set of predefined rules. These rules, designed to detect triggers and patterns in user queries, serve as the basis for generating appropriate responses. The Rule Engine orchestrates the seamless flow of information by effectively communicating with the Response Generator, paving the way for the creation of personalized and context-aware responses.

The Response Generator is the creative force behind crafting appropriate and engaging responses. Its responsibility encompasses the creation of responses based on the outcomes of rule evaluations performed by the Rule Engine. In terms of functionality, it employs a range of techniques, including response templates, dynamic content generation, and predefined messages. Its interactions are characterized by a symbiotic relationship with the Rule Engine, where it receives clear instructions on how to tailor responses, ensuring a coherent and personalized dialogue with the user.

The Knowledge Base serves as the reservoir of static information, syntax rules, and predefined responses. Functionally, it acts as a wellspring of knowledge that the chatbot can tap into to enhance its responses and support user queries. Interactions with the Knowledge Base are orchestrated by the Rule Engine, ensuring that the chatbot has access to a comprehensive reference to provide accurate and informative responses to users.

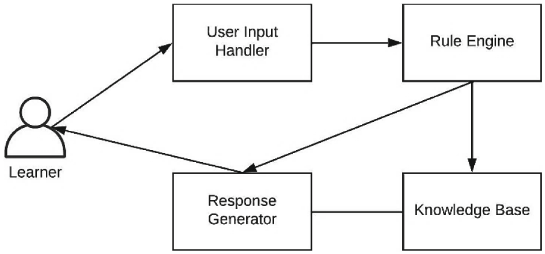


Fig. 1 Logical Architecture.

4 Rule Engine in the Rule-Based Chatbot

The Rule Engine in the Rule-based Chatbot section governs the decision-making process based on predefined rules and logic. The rules are analyzed in the following section.

Greeting Rule: The "Greeting" rule within the Rule Engine of the chatbot serves as a fundamental component of user engagement and experience. When a user initiates a conversation, this rule triggers the chatbot to respond with a personalized greeting. For example, if a user says "Hello!" the chatbot might respond with "Welcome, [User's Name]! How can I assist you today?" This simple yet crucial interaction sets the tone for the entire conversation, laying the foundation for a friendly and welcoming atmosphere.

Farewell Rule: The "Farewell" rule in the Rule Engine serves a vital role in ensuring a courteous and respectful conclusion to user interactions with the chatbot. When the chatbot detects cues indicating that the user is ending the conversation or expressing a desire to say goodbye, it promptly responds with a farewell message. For instance, if a user says, "That's all for now, thanks!" the chatbot might reply with "Goodbye! Feel free to reach out if you have any more questions." This proactive acknowledgment of the user's departure demonstrates the chatbot's attentiveness and consideration for the user's intentions.

Default Response Rule: The "Default Response" rule within the Rule Engine of a chatbot serves as a crucial fallback mechanism to handle scenarios where the user's input does not match any specific triggers or patterns. In such cases, the chatbot responds with a predefined default message, prompting the user to provide more information about their query or request. For instance, if a user says, "I need help," the chatbot might reply with "Sure! Could you please provide more details about what you need assistance with?" This default message serves several important functions, contributing to a seamless and effective user experience.

Syntax Error Explanation Rule: The "Syntax Error Explanation" rule in the Rule Engine of a chatbot plays a pivotal role in assisting users when they encounter syntax errors in their code. For example, if a user reports a syntax error like 'missing semicolon at the end of the line,' the chatbot might analyze the code snippet surrounding the reported error location and provide specific guidance on how to rectify it. This detailed explanation helps users improve their understanding of programming concepts and language syntax, ultimately leading to better coding practices and more robust code solutions.

Logic Error Guidance Rule: The "Logic Error Guidance" rule within the Rule Engine of a chatbot is a crucial component in assisting users when they encounter logic errors in their code. For instance, if a user mentions that their program is running but not giving the correct output, the chatbot might analyze the code section surrounding the reported error and offer detailed feedback on how to rectify the logic error. By providing clear and actionable advice, the chatbot empowers users to identify and rectify logic errors in their code more effectively, ultimately leading to better coding practices and more robust code solutions.

Variable Declaration Explanation Rule: The "Variable Declaration Explanation" rule within the Rule Engine of a chatbot serves as an essential tool for assisting users in understanding the concept of variable declaration in Java programming. For example, if a user inquiries about variable declaration in Java, the chatbot might initiate a detailed explanation of the syntax for declaring variables, including the data type, variable name, and optional initialization value. By providing this information, the chatbot helps users understand the structure and format of variable declarations in Java.

Loop Structure Help Rule: The "Loop Structure Help" rule within the Rule Engine of a chatbot serves as a valuable resource for users seeking assistance with Java loop structures. For instance, if a user queries the chatbot about Java loop structures, the chatbot might initiate a detailed explanation of the different types of loops available in Java, such as the "for," "while," and "do-while" loops. By offering comprehensive information and guidance on loop structures, the chatbot equips users with the knowledge and skills they need to write efficient and functional code.

Common Mistakes Warning Rule: The "Common Mistakes Warning" rule within the Rule Engine of a chatbot serves as a proactive mechanism to assist users in identifying and avoiding common coding mistakes that may lead to syntax errors in their code. For example, if the chatbot detects a missing semicolon at the end of a line in the user's code, it might issue a warning alerting the user to the potential mistake and provide instructions on how to correct it. By alerting users to potential pitfalls such as missing semicolons, the chatbot enhances the user's awareness and helps them develop a more disciplined and error-conscious approach to coding.

Resource Recommendation Rule: The "Resource Recommendation" rule within the Rule Engine of a chatbot serves as a valuable resource for users seeking additional learning materials to deepen their understanding of computer programming concepts. For instance, if a user expresses a need for additional learning resources, the chatbot might recommend relevant materials based on the user's query and context, such as books, tutorials, and coding exercise platforms. By offering personalized recommendations for curated educational materials, the chatbot helps users find resources that meet their specific needs and preferences, ultimately enabling them to deepen their understanding of programming concepts and advance their skills.

The pseudocode of the chatbot is presented as follows:

```
if user_initiates_conversation(input):
    return personalized_greeting()
if user_ends_conversation(input):
    return farewell_message()
if no_specific_trigger_matched(input):
    return default_response()
if syntax_error_detected(input):
    return syntax_error_explanation(input)
if logic_error_detected(input):
    return logic_error_guidance(input)
if user_inquires_about_variable_declaration(input):
    return variable_declaration_explanation()
if user_queries_about_loop_structures(input):
    return loop_structure_help()
if common_mistakes_detected(input):
    return common_mistakes_warning()
if user_needs_learning_resources(input):
    return resource_recommendation()
return default_response() # Default response for any other input
```

5 Evaluation

This section presents an assessment of the rule-based chatbot's performance and efficacy in providing personalized guidance in computer programming education. To evaluate its effectiveness and potential, a comprehensive evaluation process was conducted with a diverse group of participants. Users interacted with the chatbot through a web-based platform designed for computer programming education. This platform incorporates advanced features, including the integration of rule-based algorithms, to personalize the learning experience for users.

The evaluation involved 50 undergraduate students of informatics and computer engineering within the University. To ensure unbiased findings, the age and gender of the participants were randomly sampled. The demographic analysis, depicted in Table 1, provides insights into the characteristics of the evaluation participants.

Although no strict minimum interaction duration was imposed, users typically engaged with the chatbot for several minutes each. The testing period did not exceed a maximum of 10 min to prevent fatigue or loss of interest. This flexible approach allowed participants to explore the chatbot's functionalities at their own pace, accommodating individual preferences and levels of programming expertise. It ensured that users could interact with the chatbot comfortably without feeling rushed or constrained by time limitations.

Table 1. Sample population.

Measure	Item	Frequency	Percentage (%)
Sample size		50	100.0
Gender	Male	37	74.0
	Female	13	26.0
Age	18–19	24	48.0
	20–21	12	24.0
	Over 22	14	28.0
Motivation	Every student displayed a keen interest in the course, as they wanted to earn a pass degree		

The chatbot's performance and effectiveness was evaluated using metrics about user satisfaction ratings, accuracy in identifying and addressing user queries, and response time, and the ability to provide personalized guidance. The delivered questions are presented in Table 2.

Table 2. Evaluation Questionnaire.

Questions Categories	Questions
User Satisfaction	How satisfied were you with the chatbot's overall performance in assisting you with computer programming queries?
	Did you find the chatbot's responses helpful and informative in addressing your specific programming questions and concerns?
	Would you recommend the chatbot to others as a useful tool for learning computer programming concepts?
Accuracy in Identifying and Addressing	How accurately did the chatbot understand and interpret your programming queries or concerns?
	Were the chatbot's responses relevant and effective in addressing the specific issues you encountered while programming?
	Did you encounter any instances where the chatbot misunderstood your query or provided incorrect information?
Response Time and Personalized Guidance	How would you rate the chatbot's response time in providing answers to your programming questions?
	Did you feel that the chatbot offered personalized guidance tailored to your individual learning needs and preferences?
	Were you able to find relevant educational resources or materials recommended by the chatbot to further deepen your understanding of programming concepts?

5.1 Descriptive Analysis

The feedback from the students provided valuable insights into their interaction with the rule-based chatbot. This input can be instrumental in identifying areas for improvement and tailoring future iterations of the chatbot to better meet the needs of users in computer programming education.

The evaluation process involved 50 undergraduate students from various academic disciplines within the university. Participants interacted with the rule-based chatbot through a web-based platform designed for computer programming education. The evaluation spanned a range of user queries and scenarios to assess the chatbot's performance comprehensively. Feedback from participants was collected through a questionnaire, which included Likert-scale questions. The Likert-scale questions focused on user satisfaction, accuracy in addressing queries, response time, and the provision of personalized guidance.

The authors conducted a descriptive analysis of the questionnaire responses, summarizing the data through frequency calculations for each question (see Fig. 2). This analysis offers a comprehensive understanding of the users' perceptions and experiences with the chatbot, highlighting areas of strength and areas that may require refinement. By systematically analyzing the feedback, the authors can gain valuable insights into the effectiveness of the chatbot's personalized guidance and identify opportunities for enhancing its functionality to further support users in their programming endeavors.

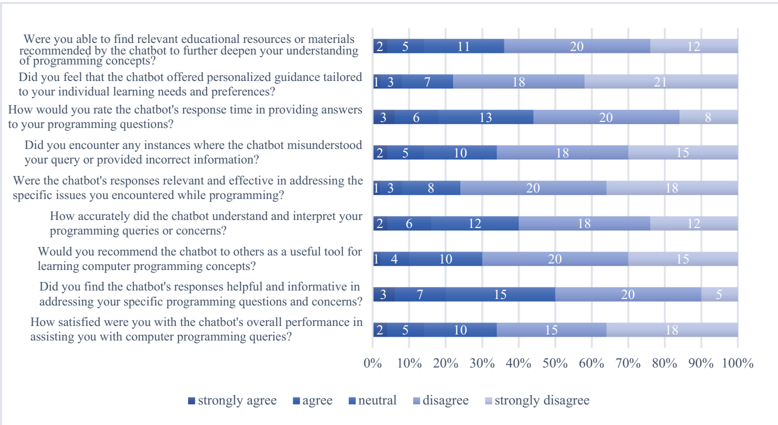


Fig. 2 Frequency of the answers in the nine questions.

The evaluation results indicate positive feedback from users regarding the chatbot's performance and effectiveness in addressing various user queries and scenarios.

In terms of user satisfaction, the majority of participants expressed high levels of satisfaction with the chatbot's overall performance in assisting them with computer programming queries. Over 90% of respondents rated their satisfaction levels as either "Satisfied" or "Very Satisfied." Participants found the chatbot's responses to be helpful and informative, addressing their specific programming questions and concerns effectively.

Nearly all respondents agreed that the chatbot's responses were relevant and effective. A significant percentage of participants indicated that they would recommend the chatbot to others as a useful tool for learning computer programming concepts, highlighting its perceived value and utility in educational settings.

The chatbot demonstrated a high level of accuracy in understanding and interpreting participants' programming queries or concerns. The vast majority of respondents indicated that the chatbot accurately addressed their specific issues without misunderstanding. Users reported that the chatbot's responses were consistently relevant and effective in addressing the specific programming challenges they encountered. Very few instances of misunderstanding or incorrect information provision were reported.

Participants rated the chatbot's response time favorably, with the majority indicating that the chatbot provided answers to their programming questions promptly. This suggests that the chatbot was able to maintain an efficient dialogue with users, minimizing response delays. Users perceived the chatbot to offer personalized guidance tailored to their individual learning needs and preferences. The chatbot's recommendations for educational resources were particularly well-received, with many participants indicating that they found relevant materials to deepen their understanding of programming concepts.

5.2 Statistical Analysis

In addition to the descriptive analysis, a *t*-test was conducted to compare the responses to two representative questions from the Likert-scale questionnaire. The two representative questions were as follows:

- Question 1: How satisfied were you with the chatbot's overall performance in assisting you with computer programming queries?
- Question 2: Did you find the chatbot's responses helpful and informative in addressing your specific programming questions and concerns?

These questions are representative of user satisfaction and perceived helpfulness of the chatbot's responses, which are crucial aspects to evaluate its effectiveness in assisting users with programming queries.

The *t*-test compared the mean scores of participants' responses to these two questions to determine if there was a significant difference between them. The null hypothesis (H0) posited that there would be no difference in mean scores, while the alternative hypothesis (H1) suggested that there would be a significant difference.

In Table 3, "Experimental Group" represents participants who interacted with the chatbot, while "Control Group" represents those who did not. The mean and variance for each group are calculated for both Question 1 and Question 2. The *t*-statistic and *p*-value for each question are also provided, along with the critical *t*-value for a two-tailed test at a specified significance level (.05).

Table 3. *t*-Test results.

	Question 1		**Question 2**	
	Experimental group	Control group	Experimental Group	Control group
Mean	4.76	2.52	4.74	2.10
Variance	0.19	0.25	0.20	0.46
t-Stat	2386		2306	
P two-tail	<0001		<0001	
t Critical two-tail	1.98		1.98	

In Table 3, both questions yielded *p*-values below .05, the null hypothesis is rejected, indicating a significant difference in responses between the experimental and control groups for both questions. These results underscore the substantial impact of interacting with the chatbot on users' perceptions of satisfaction and perceived helpfulness, emphasizing the pivotal role of the chatbot in enhancing the overall learning experience in computer programming education.

6 Conclusions

The evaluation of the rule-based chatbot in computer programming education highlighted several strengths that underscore its effectiveness. Notably, the chatbot achieved high levels of user satisfaction, demonstrated accuracy in addressing programming queries, and provided prompt responses. Users appreciated the personalized guidance tailored to their individual learning needs, enhancing their overall learning experience.

The evaluation of the rule-based chatbot in computer programming education, while positive, also highlighted certain limitations. One significant constraint is the chatbot's limited conversational depth and adaptability, which might hinder its ability to engage users in more complex or nuanced discussions. Additionally, the presence of technical issues and challenges in maintaining user engagement over time suggests areas that require further improvement to enhance the chatbot's effectiveness and reliability in educational contexts. Despite these, the chatbot exhibited significant potential as a valuable tool in computer programming education.

Moving forward, addressing the identified weaknesses and challenges will be crucial for enhancing the chatbot's performance. Future research will focus on refining the chatbot based on user feedback to maximize its effectiveness and relevance in educational settings. Additionally, efforts will be made to improve conversational depth and adaptability, ensuring that the chatbot can cater to a wider range of user queries and learning contexts. By iteratively refining the chatbot and addressing user needs, its potential as an effective educational tool in computer programming education can be fully realized.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gaffas, Z.M.: Students' perceptions of e-learning ESP course in virtual and blended learning modes. Educ. Inf. Technol. **28**, 10329–10358 (2023)
2. Krouska, A., Troussas, C., Virvou, M.: A literature review of Social Networking-based Learning Systems using a novel ISO-based framework. Intelligent Decision Technologies **13**(1), 23–39 (2019). https://doi.org/10.3233/IDT-190362
3. Gao, Y., Zhu, X.: Research on the learning experience of virtual simulation class experimental teaching and learning based on the perspective of nursing students. BMC Nurs. **22**, 367 (2023). https://doi.org/10.1186/s12912-023-01534-z
4. Troussas, C., Krouska, A., Alepis, E., Virvou, M.: Intelligent and adaptive tutoring through a social network for higher education. New Rev. Hypermedia Multimedia **26**(3–4), 138–167 (2020). https://doi.org/10.1080/13614568.2021.1908436
5. Sewell, J.L., Joshi, M., Thanh, C., et al.: Pre-clerkship teaching and learning in the virtual learning environment: lessons learned and future directions. Med. Sci. Educ. **32**, 1313–1317 (2022). https://doi.org/10.1007/s40670-022-01694-8
6. Krouska, A., Troussas, C., Sgouropoulou, C.: Applying genetic algorithms for recommending adequate competitors in mobile game-based learning environments. In: Kumar, V., Troussas, C. (eds.) ITS 2020. LNCS, vol. 12149, pp. 196–204. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49663-0_23
7. Jia, F., Sun, D., Looi, C.: Artificial Intelligence in Science Education (2013–2023): Research Trends in Ten Years. J. Sci. Educ. Technol. **33**, 94–117 (2024). https://doi.org/10.1007/s10956-023-10077-6
8. Virvou, M., Troussas, C., Caro, J., Espinosa, K.J.: User modeling for language learning in Facebook. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 345–352. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32790-2_42
9. Mao, J., Chen, B., Liu, J.C.: Generative artificial intelligence in education and its implications for assessment. TechTrends **68**, 58–66 (2024). https://doi.org/10.1007/s11528-023-00911-4
10. Troussas, C., Chrysafiadi, K., Virvou, M.: Machine learning and fuzzy logic techniques for personalized tutoring of foreign languages. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS, vol. 10948, pp. 358–362. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2_67
11. Alaswad, S., Kalganova, T., Awad, W.: Investigating the value of using emotionally intelligent artificial conversational agents to carry out assessments in higher education: review. In: International Conference on IT Innovation and Knowledge Discovery (ITIKD), Bahrain (2023)
12. Nagarhalli, T.P., Vaze, V., Rana, N.K.: A review of current trends in the development of chatbot systems. In: 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 706–710, India (2020)
13. Troussas, C., Krouska, A., Virvou, M.: MACE: mobile artificial conversational entity for adapting domain knowledge and generating personalized advice. Int. J. Artif. Intell. Tools **28**(04), 1940005 (2019). https://doi.org/10.1142/S0218213019400050
14. Oralbayeva, N. et al.: K-Qbot: language learning chatbot based on reinforcement learning. In: 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 963–967, Sapporo, Japan (2020)
15. Shi, N., Zeng, Q., Lee, R.: Language chatbot–the design and implementation of English language transfer learning agent apps. In: IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, pp. 403–407 (2020)
16. Troussas, C., Krouska, A., Virvou, M.: Integrating an adjusted conversational agent into a mobile-assisted language learning application. In: 2017 IEEE 29th International Conference

on Tools with Artificial Intelligence (ICTAI), pp. 1153–1157 (2017). https://doi.org/10.1109/ICTAI.2017.00176
17. Hou, L.: Mandarin learning system based on chatbot. In: International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), pp. 1–6, India (2023)
18. Lin, Y.-H., Tsai, T.: A conversational assistant on mobile devices for primitive learners of computer programming. In: IEEE International Conference on Engineering, Technology and Education (TALE), Indonesia, pp. 1–4 (2019)
19. Lin, Y.-H.: Chatbot script design for programming language learning. In: IEEE 5th Eurasian Conference on Educational Innovation (ECEI), pp. 123–125, Taiwan (2022)
20. Troussas, C. Krouska, A., Mylonas, P., Sgouropoulou, C.: Personalized learner assistance through dynamic adaptation of chatbot using fuzzy logic knowledge modeling. In: 18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP 2023), pp. 1–5, Cyprus (2023). https://doi.org/10.1109/SMAP59435.2023.10255169
21. Banić, B., Konecki, M., Konecki, M.: Pair programming education aided by ChatGPT. In: 46th MIPRO ICT and Electronics Convention (MIPRO), pp. 911–915, Croatia (2023)
22. Verleger, M. and Pembridge, J.: A pilot study integrating an AI-driven chatbot in an introductory programming course. In: IEEE Frontiers in Education Conference (FIE), USA (2018)
23. Singh, S., Thakur, H.K.: Survey of various AI chatbots based on technology used. In: 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1074–1079, India (2020)

Deploying ChatGPT for Automated Tagging of Greek Dialogue Data of University Students

Andreas Gavros(✉), Stavros Demetriadis, and Stergios Tegos

Aristotle University of Thessaloniki, Thessaloniki, Greece
andreaga@csd.auth.gr

Abstract. In this study, we propose a methodology of automated labelling for dialogues. We analyze dialogue data by deploying ChatGPT 3.5 model. Large language models (LLMs) have recently been in the focus of the scientific world and public since the release of ChatGPT. The approaches we used in this study were zero-shot, one-shot and few-shot learning. The model was asked to assign a label to each turn of dialogue using a tweaked version of the Issue-Based Information System (IBIS), and its labels were compared to those assigned by human raters, to evaluate its performance. Several different versions of prompts were used to investigate their effect on the model's performance, and it was also investigated whether the number of different labels affected the model's predictions. The results provide evidence that LLMs are not still able to fully process highly contextualized human dialogues but can provide worthwhile results if the assigned task is simplified. Also, we discuss on the reliability of labels created by a model, compared to labels created by a human annotator.

Keywords: Automated Tagging · Large Language Models · Prompt Engineering

1 Introduction

Dialogues, especially everyday dialogues where there is no well-defined structure, are a field that is quite difficult to be analyzed by a human, let alone by an artificial intelligence model. So, trying to form a pattern of dialogue analysis, can be a relatively challenging field of research [1]. Dialogues between different social/age groups have different levels of complexity, as different social teams tend to develop a code of communication that is difficult to be fully understood by people that do not have close connections to the group into consideration [2].

In this research we investigate dialogues between undergraduate university students, that take place in an internet chat environment. Analysis of these dialogues can provide valuable information on how to effectively process dialogues of any form and gain valuable information [3]. We used a set of labels to annotate each turn of dialogues. The resulting format will be a structure that can be used to process these dialogues automatically using machine learning models, LLMs, etc. [4]. In this work, we report on our effort to automate the process of annotating these dialogues with preselected labels. This raises the question of what tools can be used for this task, as the dialogue data is

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 265–277, 2024.
https://doi.org/10.1007/978-3-031-63031-6_23

in most cases so labyrinthine that it is quite difficult for a model created from scratch to find a pattern in the data [5]. A possible answer to this challenge, is the automated labeling process to be handled by a model, that was trained on the dialogue data [6].

Automated dialogue analysis and subsequently automated dialogue labelling in non-English languages is a relatively challenging task, as the vast majority of LLMs (if not all), were trained almost exclusively on English data and non-English language data are not widely available [7]. This structural and inherent weakness of the models leads to a reduced understanding of idioms, abbreviations, acronyms and generally complex concepts used in everyday dialogue in any language that only a native or experienced speakers of that language are able to understand [8].

This research, after completing this introduction, continues with an overview of the relevant literature on dialogue analysis and automated tagging. Then, we present the methodology we developed to analyze the dialogue data. In results section, we provide a comparative analysis of the results based on the different approaches employed. Finally, we discuss how feasible it is to apply the methodology presented in the present research and provide our conclusions.

2 Background

The main objective of this research is to develop a methodology that will automate the dialogue tagging process. Data tagging is the process of assigning a value (categorical or numerical) to a set of data (image, video or text as in our case) [9]. Automated tagging is the scientific branch of Natural Language Processing (NLP) that deploys different methodologies to automate this process. This field is becoming an integral aspect of dialog systems research, as annotated text data can be easier processed by models and more accurate/useful information can be extracted [10]. Existing research literature for automatic dialogue tagging is limited in number, therefore, it remains unclear how effective LLMs or other approaches are in evaluating text data [11].

Many case studies try to perform an automated summarization of a dialogue, while other focus on the prediction of emotional responses in a human dialogue, by employing label prediction [12]. Zhu et al. Deploy ChatGPT to annotate five different datasets in a variety of tasks (stance detection, hate speech detection, etc.) with significant success [13]. Most research works make use of a well-defined dataset, that may result to high performance of the models. In this study however we process dialogues of young people in informal context which are more complex and unstructured. This structure creates difficulties in their processing, as these dialogues often contain short expressions, inside jokes, etc., which are easily understood between people who are connected by ties of friendship, but incomprehensible, even to a human observer [14].

Most research attempts focusing on automated labeling of text data, take advantage of recent developments in LLMs. LLMs, demonstrate excellent capabilities in processing the context of a dialogue. Recent studies note that compared to currently available state-of-the-art methods, LLMs can achieve relatively competitive performance in few-shot or zero-shot-learning dialogue problems [15].

This approach theoretically enables us to annotate large volumes of text. However, there is concern in whether the annotation done automatically can be of the same quality or come close to that performed by a human. Ollion, et.al. Suggest that the effectiveness of LLMs remains incomplete in text annotation tasks and in many cases the best approach would be the deployment of models fine-tuned with human annotations [16].

Various approaches are available to improve the performance of the model. Prompt engineering emerges as one of the most direct avenues for this pursuit. A prompt serves as a comprehensive directive supplied to an LLM, tailored to enhance, and expand its capabilities with precision and finesse [17]. Weber and Reichardt suggest that the prompt can help shape or guide the output of an LLM, instilling specific rules and guidelines that establish the conversation's context and the desired output format [18].

In this work we will explore the following research questions. The first is if the labels provided automatically by the model can reach the quality of those provided by a human annotator. The second is to investigate the effect of prompt engineering in the performance of the models and the extend of this effect. Finally, we will explore different approaches to whether our framework can be generalized and made reproducible.

3 Methodology

3.1 Dataset

We use a dialogue dataset that was developed by [disclosed information]. The dataset was created by the dialogues of 72 undergraduate computer science students (18 females and 54 males). The participants' ages ranged from 19 to 23 and all of them were native Greek speakers. The frame of the dialogues was a university course assignment in a second-year course. Students were following a course on Human-Computer Interaction (HCI) and were asked to participate in an online task, where they would collaborate to answer a series of questions, based on their understanding of the HCI concepts taught. This resulted in 2646 students' contributions/turns of dialogue, which were saved in 48 different dialogues, stored as spreadsheets [19].

3.2 Choice of Labels

Labels need to be assigned to each turn of dialogue. A prerequisite is that these labels are well defined, have distinct differences and have sufficient representation in the data sample, for a model to be effectively trained. Obviously, it is not feasible for the labels to fully adapt to our dialogue data but provide the best possible representation of the dialogue's nature. Based on the above, we employed an adapted version of Issue-Based Information System framework (IBIS), which is constructed around three basic elements: "Issue", "Position" and "Argument" [20]. IBIS framework is a widely accepted way to organize information related to text, so we fitted it to the needs of our case study. Two new categories were added to this scheme: explicit position/argument [21], so that there is more discretization in the data. The tagging system chosen is presented in Fig. 1. Example of the student dialogues.

Table 1 An example of the dialogues that were processed is also available in Fig. 1.

good morning [Off Task]
my name is good [Off Task]
nice to met you! [Off Task]
basically I know who you are, let's enter the interface for a few minutes to browse and then we'll talk ok? [Position]
Yes nice [Argument]
let's take the first screenshot first [Position]
fine, do you want to start with a comment or should I mention something? [Issue]
First of all, the existence of many sub-menus makes it quite difficult for the user to browse (and because I have browsed this page many times, it is very difficult since it requires a lot of attentio n because sometimes the menu is lost if you move the mouse even a little outside the area) so I agree with the law of the tunnel this delays a lot and makes it difficult for the user both in terms of time wasted and in terms of satisfaction [Explicit Position]
what do you believe; [Issue]
I agree in parts about the sub-menus that they are quite dense, however the initial grouping and categorization of the products I think is apt [Explicit Argument]
and gives a pretty good initial direction to the user [Argument]
the fact that apart from sound and image the television is also an electrical device does not caus e confusion/ [Explicit Position]

Fig. 1. Example of the student dialogues

Table 1. Explanation of the dialogue-turn labels

Category	Description
Off-task	Statements with solely a social purpose or those unrelated to the purpose of the assigned task (such as “Hello or Hi”)
Repetition	Reiterations of earlier turns of dialogue
Team Management	Management-oriented explanations utilized for errand coordination (e.g., “Let’s yield our answer”)
Common understanding	Brief (ordinarily one- or two-word) expressions building up common understanding on the subject (e.g., “OK,” “I see”)
Issue	What must be done or settled to continue with the in general errand (e.g., “What other laws are relevant?”)
Position	Suppositions often related to the determination of the issue raised (e.g., “Fitts’ law applies here”)
Argument	Conclusions supporting/protesting to a position (“You are right”)
Explicit Position	Positions unequivocally showing thinking on space concepts (e.g., “According to Hyman, the response time increments logarithmically as the number of alternatives increases”)
Explicit Argument	Much as explicit positions, contentions showing explicit thinking on different concepts (e.g., “I oppose this idea, Hick’s law cannot be utilized for haphazardly requested…”)

3.3 Labelling Process

This process is quite important and plays a key role in the success of the model trained on the data. Here we assign the labels at each turn of the dialog. This process can be done by one or more human raters. It requires a good understanding of the dialogues, as well as the context in which they take place. In the case of a single rater, the process is simplified. In the case where there are two or more annotators, it is necessary to create a strategy of resolving disagreements that will arise in the attribution of labels. In this study we opted for a human annotator and these labels were then inspected by another human agent. In case of disagreement on a label then a consensus was reached to reach a common decision on the attribution of the label.

3.4 Data Processing

LLM Choice
As NLP is a rapidly growing field, especially in the last year and a half with large language models becoming widely available (ChatGPT, BARD, etc.) finding a model to use for our data is not a difficult task. It is also possible to use different models to conduct a comparative study. Some objective difficulties that also determine certain criteria in the choice of the model to be used are availability, cost, and computational resources available. Here, we chose ChatGPT, as it fitted the criteria we set and is a robust model, capable of processing dialogue data and providing reliable results [1].

Language Choice
In the first tests we performed, Greek was used, but the model was not even able to process the data, giving at best, incomplete results and at worst, no results (crash). It may seem obvious that the language in which the training of the model will be conducted will be the original language, but the experience we gained challenges this conclusion. As ChatGPT is automatically translating the input language (Greek in this case) to English, we believe that there is a loss of meaning in this process as dialogues originate from very young people, making extensive use of slang, clipped expressions, greeklish (Greek words written in Latin characters), English terms with Greek characters, spelling mistakes, use of language referring to friendly communication codes, etc.

Therefore, we decided to preprocess our data in an attempt to rise the performance of the model. We chose to manually edit the dialogues to remove some of the problematic features mentioned, then translate them into English using DeepL API and finally feed the dialogues to the model. Undoubtedly, this caused a loss of information and meaning that was reflected in model's performance (although there is no objective way to quantify this claim, since we did not even have results using the Greek language).

Model Training Techniques
In this study, we used three different approaches [22]:

- Zero-shot learning: The model is expected to generalize its knowledge and make predictions about labels it has never encountered.
- One-Shot learning: This technique focuses on training models based on a single example or very limited data for each class.
- Few-shot learning: Model is provided with limited examples for each label.

3.5 Training Strategies

Data Grouping
As the results in the primary 9 classes were not particularly encouraging, as will be shown,

we decided to adopt a different approach to the problem. This approach is reducing the number of available classes by grouping certain classes based on their conceptual proximity. In our case this was a non-complicated task, as certain classes have a direct link (e.g. position - explicit position). We performed a clustering into three classes of labels, to investigate if there would be a noticeable improvement in the performance of the model. The clustering was the following:

- Off Task, Repetition = Off Task.
- Team Management, Issue, Common Understanding = Team Management.
- Position, Argument, Explicit Position, Explicit Argument = Position.

As the results were encouraging, we decided to conduct additional test cycles, with an increase of the available classes by one in each cycle and investigate its effect on the results. The grouping into 4 classes was the following:

- Off Task, Repetition = Off Task.
- Team Management, Issue, Common Understanding = Team Management.
- Position, Explicit Position = Position.
- Argument, Explicit Argument = Argument.

Prompt Engineering

To enhance the performance of the model, we decided to experiment with providing various versions of prompts to the model, e.g. provide more tagged data at the prompt, and provide an explanation of each tag used. Data from related research shows that LLMs, perform better when the task they are asked to perform is explained in as much detail and clarity as possible [23]. So, the prompt we provided the model consisted of a description of the context in which the dialogue is conducted, a detailed explanation of each label, and finally, a quotation of a model dialogue which has been labelled and is quite indicative of the task the model must perform.

Since we obtained the results from the model, we evaluated the results and provided the model with a follow-up prompt, that was asking the model to perform the same task, but this time we were indicating the most typical cases where the model has misevaluated (according to our judgement). The resulting updated labels were compared with the original labels of the human reviewer and new metrics are derived. We compared these metrics with the metrics from the first prompt and obtained the results.

4 Results

In this study, the most popular metrics in machine learning were used to evaluate the results produced by the model, which were: Accuracy, Recall, F-1 score and Precision. We obtained the metric values by comparing the labels created by the model with the labels that were assigned to the dialogue by human raters.

4.1 Few-Shot Learning - The 'Full' Approach

The model's performance in this approach was particularly low, which demonstrates the difficulty of the task and the need to develop different strategies of analyzing the data and obtaining results. The results of this test cycle are presented in Table 2. In most categories the evaluation metrics have relatively low values, in some label categories (Position-Argument) the metrics have mediocre values and only in one category (Off-task) we have high metric values.

Table 2. Evaluation metrics for each label – 1st approach

Category	Precision	Recall	F1-score
Off-task	0.83	0.77	0.77
Repetition	0.25	0.25	0.25
Team Management	0.86	0.53	0.59
Common understanding	0.13	0.17	0.14
Issue	0.29	0.58	0.31
Position	0.40	0.46	0.41
Argument	0.42	0.23	0.29
Explicit Position	0.12	0.15	0.13
Explicit Argument	0.07	0.06	0.12

4.2 Few-Shot Learning – 2nd Approach

Few-shot learning with additional prompt engineering approach, yielded no significant difference compared to the previous section and was the reason why label clustering was attempted. The model is again not displaying high performance, as depicted in the evaluation metrics. This is a sign that no matter how well we explain the concept of the framework we are using the model is not able to gain additional information that will be translated to a boost in its performance. So, to investigate possible solutions to this problem, we investigated different ways to pre-process our data (Table 3, 4 and 5 and Figs. 2, 3).

Table 3. Evaluation metrics for all labels – 2nd approach

	Accuracy	F1-score	Precision	Recall
Mean for all dialogues	0.40	0.41	0.47	0.40

Table 4. Evaluation metrics for each label – 2nd approach

Category	Precision	Recall	F1-score
Off-task	0.79	0.70	0.74
Repetition	0.00	0.00	0.00
Team Management	0.68	0.31	0.42
Common understanding	0.25	0.44	0.32
Issue	0.25	0.40	0.31
Position	0.41	0.42	0.42
Argument	0.33	0.21	0.26
Explicit Position	0.23	0.40	0.29
Explicit Argument	0.08	0.11	0.10

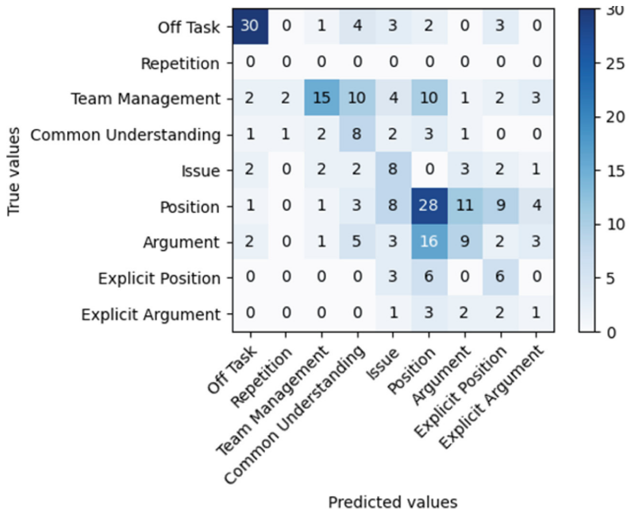


Fig. 2. Overall confusion matrix for each label – 2nd approach

Table 5. Overall results for each label – 3rd approach

	Precision	Recall	F1-score	Accuracy
Off Task	0.63	0.79	0.70	0.731
Position	0.80	0.87	0.83	0.731
Team Management	0.71	0.47	0.57	0.731

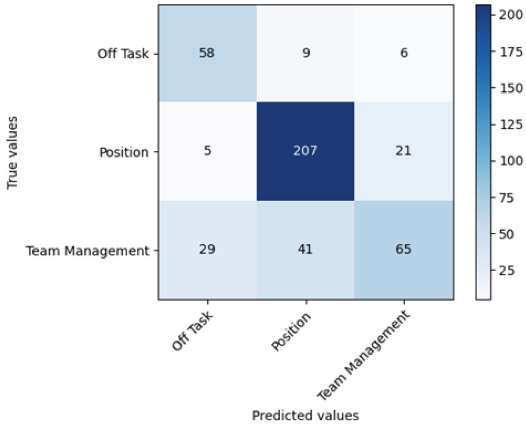


Fig. 3. Confusion Matrix for each label – 3rd approach

4.3 Few-Shot Learning – 3rd Approach

The results in this approach had, a dramatic improvement and this was reflected in all of the dialogues where tests were conducted. Although this may seem as a relatively easy task, the concept of automating the process of dialogue annotating, seems to be a feasible option using the method of data grouping. These three label categories can provide a representative qualitative rendering of the dialogues. We are also observing similar performance of the model in all label categories, so it is safe to assume that the model can successfully process the data effectively, at this 'simplified' level.

4.4 Few-Shot Learning – 4th Approach

Here we investigated if the performance would still be high with the introduction of one more label which now begins to increase the level of complexity for the model, such as that of Argument, where it is necessary to distinguish it from Position. Unfortunately, however, with the use of 4 grouped labels, a significant drop in model performance was again observed, which discouraged further testing with the addition of one more label. In Table 6 and Table 7, we present the overall results, which explain our decision not to pursue further analysis. One immediate conclusion that can be drawn from these results is that the model had low performance metrics in the argument and position labels, which would however would be a difficult decision even for a human reviewer (Fig. 4).

Table 6. Evaluation metrics for all labels – 4th approach

	Accuracy	F1-score	Precision	Recall
Mean for all dialogues	0.54	0.54	0.57	0.54

Table 7. Overall results for each label – 4th approach

	Precision	Recall	F1-score	Accuracy
Off Task	0.53	0.82	0.64	73
Position	0.54	0.51	0.52	161
Team Management	0.74	0.51	0.60	153
Argument	0.36	0.41	0.38	94

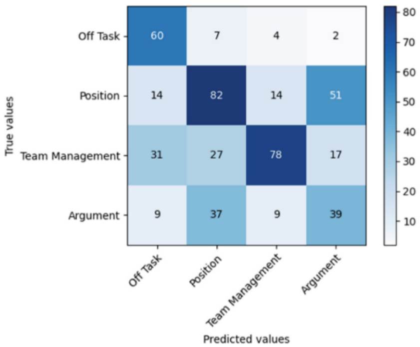


Fig. 4. Confusion Matrix for each label – 4th approach

5 Discussion

In this study we employed an LLM (ChatGPT) to automate the task of assigning labels of the IBIS framework to turns of dialogues conducted by university students in Greek language. A key question raised in the design of the tests conducted was: "Is model-based labeling as reliable as human labeling?". Based on the results so far, the answer that can be given has two levels.

At a first glance the results indicate that LLMs still do not have the same ability to capture the complexity of everyday dialogues, compared to humans. As the complexity of the task that the model was asked to perform increased, its performance dropped. We observed however, that the model demonstrated some level of competence in certain categories. When tasked with automatic tagging using fewer classes, the metrics significantly improved. However, achieving the better result that we were hoping to reach, remains a challenge, at least for now.

Is the answer above entirely reliable? We cannot be sure, by any means. Everyday conversations pose significant challenges for comprehension, even for a human external observer devoid of knowledge regarding the discussed subject. Moreover, possible discrepancies among individuals that will undertake the task of labelling dialogues, is underlying the absence of a singular objective truth in assigning a label to a turn of dialogue. Thus, we consider that at this point we are not yet able to establish a gold standard. In certain instances, the model may be able to provide a more accurate assessment of a given dialogue, that will result in labeling of higher accuracy. At last, nevertheless, the overarching conclusion suggests that humans still have higher ability to process the information that is available in a dialogue, although there is no objective ground truth to determine that statement.

However, this particular question leads to deeper questions about the nature of the research. We quantify the performance of the model based on the comparison made with human labels. But how can we be sure of the human rater's choices if complete convergence between two different raters is not achieved? Therefore, there is the possibility that in some instances the model may yield the correct labels than the human rater, but this cannot be quantified by the metrics we used in this research.

A second research question we ought to answer is the following: "Which is the effect of prompt engineering on model's performance?". In general, attempts to improve the model's performance through interfering with the prompt and changing its form, yielded no discernible boost in metrics. Furthermore, comparisons between few-shot and one-shot learning techniques across nine classes revealed no significant differences in model performance. Similarly, tests incorporating label explanations in the prompt did not yield noticeable improvement in performance metrics. The preliminary tests we conducted, deploying zero-shot learning, showed minimal deviation from both few-shot and one-shot learning approaches. Overall, our findings suggest that future endeavors in automatic tagging may effectively utilize zero-shot learning without significantly impacting the overall results, thus questioning the necessity of having a human intervention reviewing the process at this stage. Furthermore, it is possible that a more detailed explanation of the labels would result to a better model performance. However, this approach meets the limitations of the available characters that can be used in the ChatGPT version we used. In later versions of this LLM such an approach is feasible as it has a higher limit of characters in prompt.

6 Conclusions

This study provided evidence LLMs can excel in automated tagging tasks is the labels have well-defined differences or the labels are clustered. As demonstrated, automated labelling conducted by an LLM such as ChatGPT proves to be viable for streamlined tasks. Thus, provided that the labels remain constrained in quantity and there exists a clear differentiation in their significance, we assert that processing extensive dialogue data using this methodology is indeed feasible.

Our study comes in agreement with similar research attempts, that provided evidence that prompt engineering seems to have low or minimum effect, in enhancing the performance of LLMs for very complicated tasks, like the one this study, where the model was called to find pattern in highly conceptualized dialogues. So, future research may be benefitted by investigating other strategies of boosting the performance of models in similar tasks.

The methodology developed for this case study can have also more practical applications, especially in education, were different labels can be deployed to qualitatively evaluate essays created by the students more effectively and quicker, to assist the work of the professor in providing a better feedback to the students.

Finally, the methodology developed in this study can be used in similar research efforts, since despite the (seemingly) low performance of the model on a task with many labels, we observed that by applying different approaches to the problem, the model was ultimately able to perform particularly well on clustered labels. Variations of this approach can perhaps lead to even better results.

References

1. Ni, J., Young, T., Pandelea, V., Cambria, E.: Chatbots: recent advances in deep learning based dialogue systems: a systematic survey. Artif. Intell. Rev. **56**, 3055–3155 (2022)
2. Redden, S.M., Way, A.: 'Adults don't understand': exploring how teens use dialectical frameworks to navigate webs of tensions in online life. J. Appl. Commun. Res. **45**(1), 21–41 (2015)
3. Tudor Car, L., et al.: Conversational agents in health care: scoping review and conceptual analysis. J. Med. Internet Res. **22**(8) (2020)
4. Duran, N., Battle, S., Smith, J.: Inter-annotator agreement using the conversation analysis modelling schema, for dialogue. Commun. Methods Meas. **16**(3), 182–214 (2022)
5. Yaskorska-Shah, O.: Managing the complexity of dialogues in context: a data-driven discovery method for dialectical reply structures. Argumentation **35**, 551–580 (2021)
6. Nakata, T., Chen, S., Saiki, S., Nakamura, M.: Employing large language models for dialogue-based personalized needs extraction in smart services. In: 12th International Conference on Soft Computing, Artificial Intelligence and Applications (2023)
7. Zhu, W., Lv, Y., Dong, Q., Yuan, F., Xu, J., et al.: Extrapolating large language models to non-English by aligning languages. arXiv e-prints (2023)
8. Bouamor, H., Pino, J., Bali, K.: Don't trust ChatGPT when your Question is not in English: a study of multilingual abilities and types of LLMs. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023)
9. Woodall, P., Giannikas, V., Lu, W., Mcfarlane, D.: Potential problem data tagging: augmenting information systems with the capability to deal with inaccuracies. Dec. Support Syst. **121**(4) (2019)
10. Dandapat, S., Sarkar, S., Anupam, B.: Automatic part-of-speech tagging for Bengali: an approach for morphologically rich languages in a poor resource scenario. In: ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23–30, 2007, Prague, Czech Republic (2007)
11. Zhang, C., D'Haro, L.F., Chen, Y., Zhang, M., Li, H.: A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. arXiv (2024)
12. Xue, H., et al.: E-chat: emotion-sensitive spoken dialogue system with large language models. arXiv (2023)
13. Zhu, Y., Zhang, P., Haq, E., Haq, P., Tyson, G.: Can ChatGPT reproduce human-generated labels? A study of social computing tasks. arXiv (2023)
14. Feng, Y., Lu, Z., Liu, B., Zhan, L., Wu, X.M.: Towards LLM-driven dialogue state tracking. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, pp. 739–755. Association for Computational Linguistics, Singapore (2023)
15. Deng, Y., Liao, L., Chen, L., Wang, H., Lei, W., Chua, T.S.: Prompting and evaluating large language models for proactive dialogues: clarification, target-guided, and non-collaboration. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 10602–10621. Association for Computational Linguistics, Singapore (2023)
16. Ollion, E., Shen, R., Macanovic, A., Chatelain, A.: ChatGPT for text annotation? Mind the hype! SocArXiv (2023)

17. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al.: A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv (2023)
18. Weber, M., Reichardt, M.: Evaluation is all you need. Prompting generative large language models for annotation tasks in the social sciences. A primer using open models. arXiv. (2024)
19. [Disclosed information]
20. Rittel, H. W. J., Kunz, W.: "Issues as elements of information systems: The IRAS model." Working Paper No. 131, Institute of Urban and Regional Development, University of California, Berkeley (1970)
21. Kuraku, S., Samaah, F., Kalla, D., Smith, N.: Study and analysis of chat GPT and its impact on different fields of study. Int. J. Innovative Sci. Res. Technol. **8**(3), 827–833 (2023)
22. Kim, S., Joo, S.J., Kim, D., Jang, J., et al.: The CoT collection: improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. arXiv (2023)
23. Zamfirescu-Pereira, J.D., Wei, H., Xiao, A., Gu, K., Jung, G., et al.: Herding AI cats: lessons from designing a chatbot by prompting GPT-3. In: ACM Designing Interactive Systems (DIS'23). Pittsburgh, PA, USA (2023)

Paraphrase Generation and Identification at Paragraph-Level

Arwa Al Saqaabi[1(✉)], Craig Stewart[2], Eleni Akrida[2], and Alexandra I. Cristea[2]

[1] Qassim University, Buraydah 52571, Saudi Arabia
A.AlSaqabi@qu.edu.sa
[2] University of Durham, Durham DH1 3LE, UK
{craig.d.stewart,eleni.akrida,alexandra.i.cristea}@durham.ac.uk

Abstract. The availability and growth of tools and natural language generation (NLG) models that are used to paraphrase text could be helping to improve students' writing and comprehension skills or a threat to intellectual property and educational integrity specifically when the text has been copied from other authors. These tools can be used by plagiarists to paraphrase individual words, phrases, sentences, and paragraphs. To solve this issue, much work has been done on plagiarism detection (PD) and paraphrase identification (PI) utilising downstream tasks and natural language processing (NLP) methods. These works mainly focus on sentence length and sentence-level paraphrasing. In this paper, we investigate paragraph-length and paragraph-level paraphrasing as the most common method of committing plagiarism is copying and paraphrasing paragraphs from other authors. Here, we construct a novel, large-scale paragraph-level paraphrasing dataset by implementing and examining a state-of-the-art Transformer-based model to reorder and paraphrase sentences without affecting a paragraph's meaning. In a first-of-a-kind study, we consider both intra-sentence and inter-sentence similarity before examining the efficiency of state-of-the-art Transformer-based models in detecting paraphrased paragraphs. We offer a technique that serves as both a tool for honing paraphrasing skills and a means of identifying plagiarism. Our outcomes surpass those presented in the existing literature.

Keywords: Natural Language Generation (NLG) · Natural Language Processing (NLP) · Paragraph-Level Paraphrasing · Transformer-Based Model

1 Introduction

Paraphrase generation is a commonly studied NLG task. On the one hand, such paraphrased text could be used to enhance plagiarism detection, machine translation, and summarisation for NLP downstream tasks. In addition, paraphrasing can be employed to assist comprehension and writing skills development. On the other hand, paraphrase generation could undermine academic integrity if it is misused by students seeking to plagiarise existing work. According to [1] the current state of artificial intelligence (AI) models makes it possible to create highly coherent and contextually suitable paraphrased material that might be used to generate plagiarised content. In addition, [2] concluded

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 278–291, 2024.
https://doi.org/10.1007/978-3-031-63031-6_24

that it is difficult to differentiate artificially paraphrased text from human-written text. Thus, AI models have the potential to be utilized both for improving students' writing skills and simultaneously detecting plagiarism.

Paraphrase generation is the task of generating an output text that preserves the meaning of the input text in other forms of text [3]. Current state-of-the-art research focuses on paraphrasing texts at the sentence-level [[4, 5]] and paragraph-level [6] but utilises sentence-level paraphrasing methods only. These approaches consider the meaning of each sentence independently; they did not determine any semantic relationships between sentences. The novel research presented in this paper paraphrases paragraphs utilising paragraph-level paraphrasing by considering both the intra-sentence and inter-sentence relationships by implementing paraphrase generation Transformer-based models. This is harder and more valuable than sentence-level paraphrasing because it considers the diversity across multiple sentences beyond the lexical and syntactic diversity of a single sentence. This holds practical significance as it is a necessary skill that needs to be cultivated and applied in educational tasks, such as citing the work of others. In addition, according to [7], plagiarists reuse paragraphs not sentences the most frequently.

Paragraph-level paraphrasing includes sentence reordering, sentence splitting, and sentence merging. The initial work in this area was presented in [8], where the authors applied an algorithm to detect paraphrasing (focusing on the paragraph level); however, this work was limited by the fact that very few suitable datasets are available for this type of research. As there are no published paragraph-level paraphrase datasets established using paraphrase generation Transformer-based models, we implement two algorithms based on state-of-the-art Transformer-based models that have become the standard paradigm for most NLG tasks. We perform sentence reordering considering inter-sentence diversity before paraphrasing the paragraphs using state-of-the-art paraphrase generation models. Specifically, we apply the Sentence Order Prediction (SOP) of the ALBERT [9] re-training model and Transformer-based models (BERT [10], RoBERTa [11] and Longformer [12] for paraphrasing. The output paragraphs are generated based on the semantic relations among the source sentences. We generate multiple paraphrased versions for each source, making our approach effective for improving students' writing abilities. In this work, this dataset (ALECS) enables us to investigate the Transformer-based models' ability to distinguish between the source and paraphrased text after reordering and paraphrasing its sentences using a variety of levels within the masked language model (MLM). This research aims to investigate the following research question (RQ):

- RQ: How efficiently can state-of-the-art Transformer-based models discriminate between the original and paraphrased text at the paragraph-level with sentence reordering?

To the best of our knowledge, this study provides the first extensive dataset of paragraphs (ALECS) that have been paraphrased at the paragraph-level using Transformer-based models along with a study of Transformer-based models' performance in detecting paragraph-level paraphrasing.

The remainder of this paper is organised as follows: Sect. 2 outlines the main related endeavors. Section 3 details our methodology, while Sect. 4 contains the dataset evaluation. The experimental results and discussion are reported in Sect. 5. Section 6 summarises the main conclusions and future research directions.

2 Related Work

PI is a main task in NLP that is involved in many downstream tasks such as PD [13] and data augmentation [14]. It is considered a classification task which can be performed at the sub-sentence level, sentence-level, paragraph-level, or document-level. According to [15]:

- Sub-sentence level: the algorithm finds the pertinent sub-expression categories that are contained within a sentence.
- Sentence-level: The appropriate categories of a single sentence are obtained.
- Paragraph-level: A single paragraph's relevant categories are retrieved by the algorithm.
- Document-level: The algorithm uses the entire document to extract the relevant categories.

In this research, we focus on sentence-level and paragraph-level PI.

Most of the existing work has been done at the sub-sentence-level or sentence-level by applying machine learning classification algorithms on hand-crafted features such as syntactic dependency features [14] and lexical features from a Bag of Words [16]. Other works used neural networks that focus on word embedding [17], recurrent neural networks,, or Transformer-based models [10].

These works were implemented on sentence-length datasets such as the Microsoft Research Paraphrase Corpus (MRPC), PAN, and Quora Question Pairs (QQP). A total of 5801 pairs in the MRPC corpus have been manually tagged as paraphrases or non-paraphrases [18]. MRPC was collected from online news collection by using heuristics to identify candidate document pairs and candidate sentences from the documents. The PAN datasets include cases that were obfuscated using elementary automated techniques that did not preserve the text's intended meaning. These heuristics include, for instance, randomly deleting, adding, or changing words or phrases, as well as exchanging words with randomly chosen synonyms, antonyms, hyponyms, or hypernyms [19]. In addition, in QQP[1] question titles from the forum are divided into duplicate-or-not questions. These questions are published on a website where users can post questions and receive answers. The designers of the enormous QQP dataset state that, despite it containing labels made by humans, the labels were not intended to be used for PI tasks.

These datasets have a limitation on their size which makes training neural or Transformer-based models difficult. To solve this limitation, many datasets have been created using a variety of techniques. PARADE [6] created computer science concepts from online user-generated flashcards. They implemented clustering to group each specific term's definitions. They then selected one as the source and the other one as a paraphrased text. A four-label system was used to manually annotate each extracted sample. In addition, [5] created a dataset of sentence-level paraphrasing that was generated by machine translation. They translated the text to another language (Czech)

1 https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs.

and then translated it back to the original text language (English). The quality of the paraphrased text is affected by the efficiency of the translation model used. Despite the differences in style and content quality of the above-mentioned datasets, they all consist of sentence-level paraphrasing text which is not suited for paraphrase identification at the paragraph-level paraphrasing. They mainly apply different algorithms to paraphrase each sentence independently as a result all the works have been done on these datasets focusing on sentence-level. This type of paraphrasing is less common among plagiarists as they tend to paraphrase a paragraph by using sentence reordering, splitting and/or merging with consideration of the paragraph's meaning [20].

Nowadays, paragraph paraphrasing and classification has become possible, especially with Transformer-based models that accept a long input of tokens. However, few studies have considered a paragraph as an input for PI; [20] artificially constructed a dataset using content from Wikipedia, theses, and arXiv articles by paraphrasing text using Transformer-based models. They also applied state-of-the-art Transformer-based models to distinguish between source and paraphrased text. They achieved commendable results. The main difference between this work and ours is that they directly paraphrased text without reordering sentences as we do to achieve document-level paraphrasing in the future.

Two works, [21, 22] considered paragraph-level paraphrasing by applied sentence reordering after paraphrasing the text through back-translation. However, auto-translation for paraphrasing text may still cause errors where a word is translated into a synonym which may not be contextually valid [23]. In [21, 22] graph models are implemented to generate the best order of the sentences based on the paraphrased text ignoring the relation of the source's sentences. Thus, the sentences semantic relations will be affected by the quality of the paraphrasing algorithm used. In addition, they have not yet applied the paraphrase identification method to these datasets.

In our work, we aim to avoid this limitation by using SOP on the original text to generate two different sentence orders for each paragraph based on the source text's inter-sentence similarity and intra-sentence similarity; then, we produce paraphrased paragraphs using a state-of-the-art Transformer-based model to achieve paraphrase generation. To the best of our knowledge, this is the first dataset for training PI classification models that consists of paragraph-level paraphrasing utilising Transformer-based models.

3 Methodology

3.1 Dataset Creation

Our dataset, ALECS, contains text relating to social science domains collected from Wikipedia[2]. We eliminate linguistics articles because of the manner that is used for paraphrasing, and law articles as models appropriate for such articles would need to be trained on legalistic language specifically. The major goal of this step is to create a dataset that can be used to enhance the students' writing skills and distinguish between human-written and machine-paraphrased texts in order to identify plagiarism in academic

[2] https://en.wikipedia.org/wiki/social_science

writing. The total number of paragraphs is 391,205 after filtering the collected text into paragraphs of 50–151 words in length (this is the average paragraph length in English [24]) with at least three sentences. This minimum sentence requirement was put in place as our paraphrased text creation methodology (see below) utilises inter-sentence semantics, and therefore, any fewer than three sentences would have rendered the paraphrased text more akin to texts created using sentence-level methods used by other datasets (see Fig. 1).

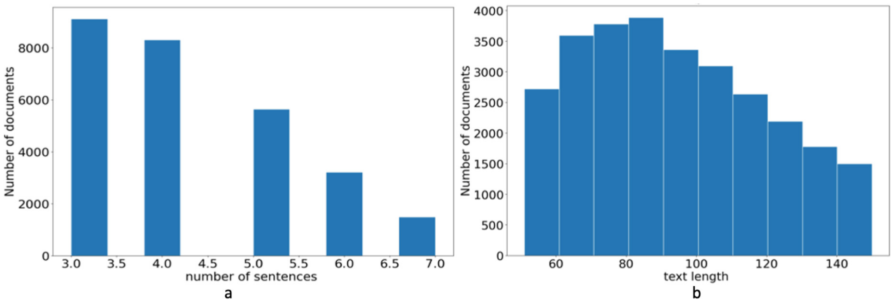


Fig. 1. a) Number of samples containing a specific number of sentences. b) Number of samples in relation to the number of words in the documents.

Inter-sentence Paragraph Coherence Score (Sentences Reordering). Text coherence has been the subject of a lot of research; coherence is described in different terms such as *entity* [25] and *word co-occurrence*[26]. In this work, we implement the SOP of the ALBERT Transformer-based model as it considers inter-sentence coherence and generates a coherence score that represents the validity of the order between two sentences [9]. This model outperforms other Transformer-based models in terms of paragraph coherence [27].

Firstly, we convert each paragraph D into a fully connected directed graph G where the set of sentences S serve as nodes:

$$V(G) = S_1, S_2, \ldots, S_n \tag{1}$$

$$P_{SOP}\left(S_s^i, S_s^j\right) = \begin{Bmatrix} P \geq \varepsilon, i \neq j \\ 0, i = j \end{Bmatrix} \tag{2}$$

Then, we apply two algorithms to reorder the paragraph's sentences depending on the SOP probability. Each algorithm suggests a path that passes over all nodes without repeating and has the highest coherence score based on a coherence measurement approach of the algorithm used. The paragraph's sentences are shuffled based on the suggested path before being evaluated by human evaluators (discussed in Sect. 4) to determine which algorithm is the best and investigate the correlation between the human-written paragraphs and generated paragraphs.

Inter-sentence Shuffling. Assume that G is a fully connected directed graph where the nodes are the document's sentences, and the weight of the edges is measured by the SOP probability. The generated path must pass through each node without repeating.

– *Algorithm SALAC1*

This algorithm gives priority to the nodes that are linked by the strength coherence score that are represented as edges' weights in the graph. SALAC1 determines the order of sentences depending on many conditions that are shown in the flowchart in Fig. 2. Let us assume that we have a paragraph consisting of four sentences and its graph matrix shown in Fig. 3; the strength coherence score is 0.7 which is linked S1 to S2 and S2 to S4. This means S1 and S2 must come before S4 although we could insert other sentences between them without breaking down their relation. Moreover, the weakest coherence score in this example is 0.4 and the rest of the coherence scores are distributed between the strongest and weakest scores. We replace the matrix diameter values with 0 to remove the path from a sentence to itself.

The first step as shown in the flowchart (Fig. 2) considers only sentences that are linked with the highest coherence scores, which are in the example S1-S2-S4. Then, SALAC1 checks if the path is completed or if there are unincluded sentences. Then it removes all the coherence scores that are considered in the previous step leading to a decrease in the highest coherence score from 0.7 to 0.6 in this example. Now SALAC1 selects all sentences that have the highest coherence score, then it inserts them in the path depending on their relations to the sentences already in the path considering their relations to each other (parent or child). In some cases, the sentence has the same coherence score as all the nodes in the graph, which means it could be at any position on the path. In the example (Fig. 3), S3 links to all sentences with a coherence score of 0.5 so we can locate it at the end of the path S1-S2-S4-S3.

Another condition can be seen in the flowchart in Fig. 2: when there is a sentence with a strong link to the second sentence in the path while its relation to the first sentence is weak; in this case, this sentence is a parent to the second sentence (i.e., it should be inserted before the second sentence) but a child to the first one (i.e., it must come after it).

– *Algorithm SALAC2.*

SALAC2 goes over all possible paths in the graph and picks a path with the best coherence score. It calculates the path coherence between a parent node to its child node using Eq. 3.

By implementing these algorithms, we generate for each source paragraph two paragraphs with different sentence orders compared to the source. Then, we calculate the paragraph's coherence score by implementing Eq. 3.

$$COH = \sum\nolimits_{i}^{n-1} \sum\nolimits_{j=i+1}^{n} P_{SOP}\left(S_s^i, S_s^j\right) \tag{3}$$

Intra-sentence Masking (Paraphrasing). In an effort to develop a paragraph-level dataset, we implement three state-of-the-art Transformer-based models to paraphrase paragraphs after applying the sentence shuffling algorithms. For paraphrasing, we implement BERT [10], which is mostly used as a baseline in NLG research, RoBERTa [11],

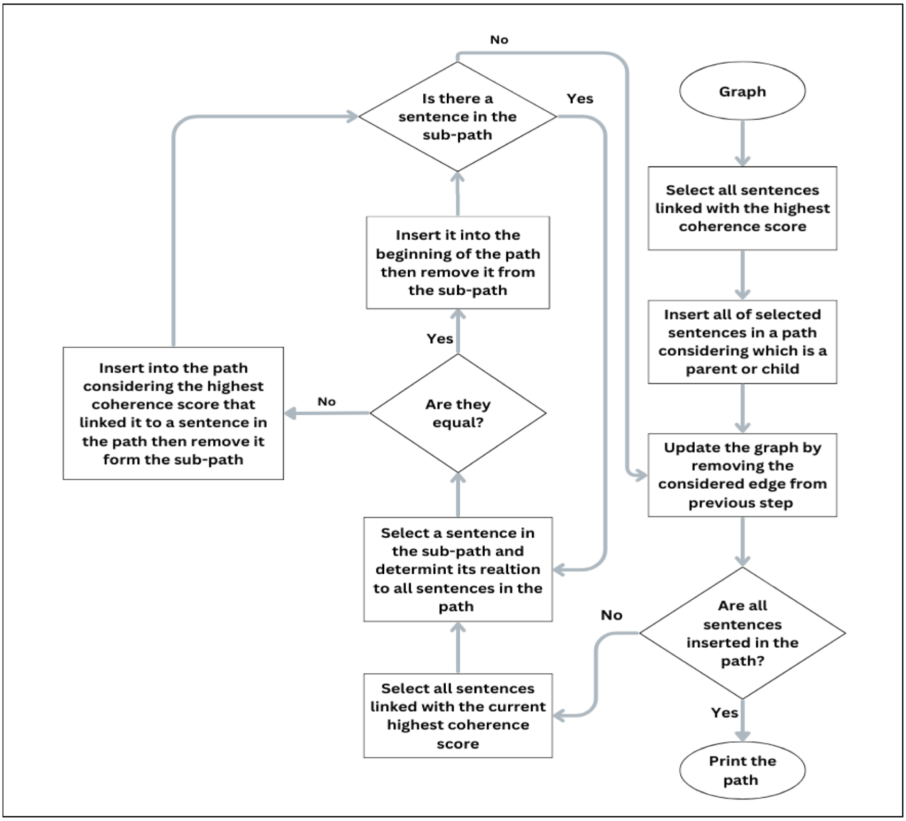


Fig. 2. SALAC1 flowchart algorithm.

	Sentence S1	Sentence S2	Sentence S3	Sentence 4
Sentence S1	0	**0.7**	0.6	0.5
Sentence S2	0.4	0	0.6	**0.7**
Sentence S3	0.6	0.6	0	0.6
Sentence S4	0.4	0.5	0.6	0

Fig. 3. SALAC 1 graph matrix example, scores in bold represent the strength coherence score while underlined scores represent the weakest coherence score.

which is built on BERT to handle longer documents, and Longformer [12], which is mainly developed for long documents. To account for the diversity of our dataset, we apply a variety of levels of the masked language model (MLM) for all three Transformer-based models. It masks a part of the words from a sequence of input or sentences and requires the designed model to predict the most likely word choices to complete the sentence. To avoid producing false information compared to the source, we exclude named entities and punctuation, such as brackets, digits, currency symbols, and quotation marks from paraphrasing as in [20].

4 Evaluation

4.1 Human Evaluation

To demonstrate the efficacy of the task, we perform a manual evaluation study, which is the most common approach in NLG. In the next paragraphs, we explain the human evaluation method applied to this study.

Firstly, based on the task and goal of this study, we must evaluate the quality of the output text of our algorithms by collecting and analysing numerical data. To achieve this, we implement a quantitative study as an intrinsic approach. In more detail, we measure the differences in the text semantics by comparing the generated paragraph to the source. Both documents (source and output) should convey a similar meaning, that is, we aim to maintain the meaning of the source by reordering the paragraph's sentences.

Secondly, we randomly sample 100 paragraphs from the dataset. In NLG, the median number of samples used for human evaluation is 100 [28]. In addition, three evaluators check each sample, and the decision on whether the source and the output were similar in meaning is taken based on majority voting. In terms of the human assessors, six highly educated fluent speakers are selected as the text used in this study is extracted from Wikipedia articles written for a general readership.

The participants are provided with the source texts and the reconstructed paragraphs for each sample, then they are asked to evaluate each of the generated paragraphs in terms of semantic similarity to the source. According to [28], complex concepts cannot be captured in a single arbitrary rating [19], therefore the participants are asked to select a score on a 5-point Likert scale where each score represents a defined value as follows:

5: Almost identical
4: Very similar, with only minor changes to the meaning
3: Similar, with major changes to the meaning
2: Dissimilar, with significant changes to the meaning
1: Extremely different

The experiment was approved by the University's ethics committee and took about three hours to complete.

4.2 Automatic Evaluation

Inter-annotator Agreement (IAA) Correlation. Inter-annotator correlation or agreement (IAA) determines the degree of agreement between the evaluations of different raters. It is commonly used when using multiple annotators. According to [29], the acceptable range of IAA is between 0.3–0.5 in NLG research where the higher the IAA, the more valuable.

In this work, we implement the kappa coefficient as an IAA statistical test; this involves two groups of three evaluators, with each group evaluating 50 samples of different generated texts. The results for both groups of evaluators show a low correlation, namely 0.4. In [28], the authors explained that IAA is more likely to be low when measuring a complex language concept such as semantic similarity as in this study. However, this score reaches an outstanding correlation of 0.8 after categorising the rating scores into two categories depending on their differentiations (5, 4, 3 = A, 2, 1 = B).

Efficiency of the Algorithms. We compare the algorithms' efficiency in relation to the human evaluation results. A total of 300 scores were given by the evaluators for each algorithm. The results in Table 1 show that SALAC1 and SALAC2 are comparable. SALAC1 and SALAC2 generate paragraphs with identical meanings to the source by 39% and 40%, respectively. In contrast, the percentage of samples that have different meanings to the source is very low in all algorithms' results, 1% and 3%, respectively. The difference between SALAC1 and SALAC2 can be noticed in similar and dissimilar samples. To make it clear, we categorise the scores depending on their definition (Table 1 grey columns). SALAC1 is higher by 6% in similar samples and lower by 6% for dissimilar samples compared to SALAC2.

Table 1. Distribution of 300 votes to the scores given by humans.

Score	1	2	3	4	5	1, 2	3, 4, 5
SALAC1	1%	9%	27%	24%	39%	10%	90%
SALAC2	3%	13%	21%	23%	40%	16%	84%

Correlation Between the Paraphrased Paragraph's Similarity Score and the Human-Written Paragraph's Similarity Score. Measuring the correlation between the paraphrased paragraph's similarity score and the human-written paragraph's similarity score is important as we try to generate a paragraph-level paraphrased text based on the human-written paragraph. To achieve this objective, we apply Eq. 3 to measure the coherence score on source paragraphs. We then compare it to the generated paragraph's coherence score for each algorithm.

In Fig. 4, SALAC1 and SALAC2 provide high Pearson's correlation values, namely 0.89 and 0.80, respectively. This high correlation indicates that the implemented algorithms maintain the original text's semantics.

Mask Applied Method. We apply Transformer-based models (BERT, RoBERTa, Longformer) to paraphrase the paragraphs generated by the SALAC algorithms with 0.15, 0.20, and 0.30 MLM. Thus, we have six paraphrased texts for each source paragraph with each Transformer-based model as we apply two algorithms and three MLM levels to consider the variety of abilities to paraphrase a text that usually happens in reality. The highest correlation is obtained by SALAC1 and Longformer as shown in Fig. 5. Thus, Longformer's capacity to handle longer input sequences may be useful in

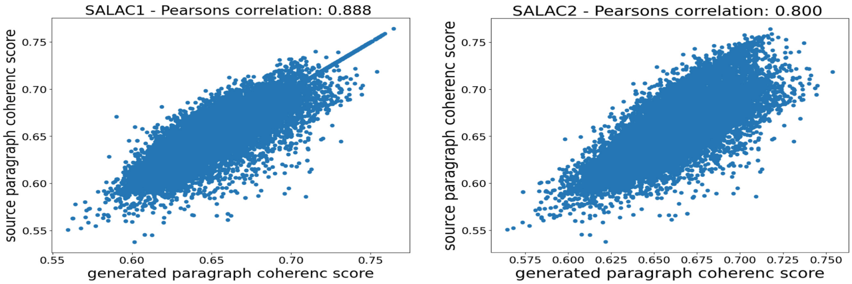


Fig. 4. Correlation of the generated paragraphs to the human-written paragraphs.

producing longer paraphrased texts. For instance, if the input text is a paragraph, Longformer might be better able to capture the overall context of the paragraph and use this context to generate a more accurate and meaningful paraphrased output text.

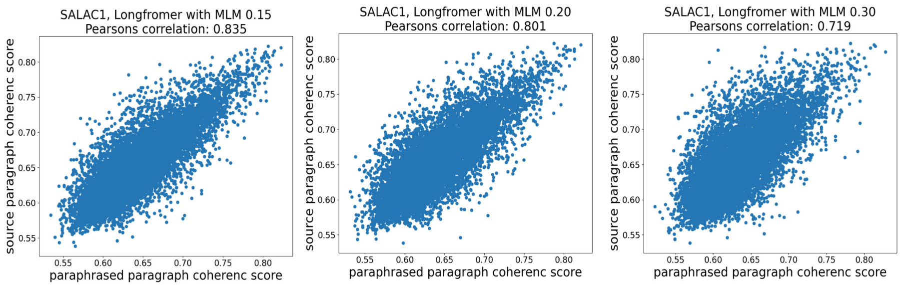


Fig. 5. Correlation of the paraphrased paragraph to the human-written paragraph.

5 Experiment

To address the lack of existing paragraph-level paraphrasing datasets created by paraphrase-generation models, we create the ALECS dataset which is divided into training and testing sets with 938,892 and 234,723 samples, respectively. The main objective is to study the efficiency of the Transformer-based model in detecting paraphrased paragraphs after reordering their sentences. We apply three state-of-the-art Transformer-based models in their default hyperparameters configurations for paragraph paraphrasing and paraphrase identification: RoBERTa [11], an extension of BERT designed to accommodate lengthier documents, Longformer [12], primarily developed for processing extended documents, and BERT [10], often utilised as a baseline in NLP and NLG studies. However, we consider samples paraphrased using Longformer as they show the highest correlation with the human-written paragraphs (Sect. 4.2). In addition, we report only the best results obtained by Longformer, because of restricted number of pages.

5.1 Baseline

We use off-the-shelf BERT as a baseline classifier model, which is commonly implemented in most of the existing work. Moreover, Additionally, we consider the work of [20] as a ground truth on paragraph-level classification for the PI task as this study was performed on a paragraph-sized but sentence-level paraphrasing dataset. Furthermore, we compare the classification results of paragraph-level paraphrasing and sentence-level paraphrasing as in [8].

5.2 Results and Discussion

The results in Table 2 show that the paraphrasing level and MLM levels affect the Transformer-based model's efficiency. For Transformer-based models including BERT and Longformer, MLM is a primary, self-supervised, fine-tuning objective. In paraphrase generation models MLM represents the percentage of paraphrased words. Since 0.15 MLM is the standard percentage of paraphrased words when utilising available paraphrase generation tools [32] and because it's more challenging for the Transformer-based model in detecting paraphrased content, we compare our results to those of others at that level. In general, our result outperforms the existing work result using Longformer with 0.15 MLM by 4%. For BERT, although the fact that we reorder the sentences in the paragraphs then paraphrase them, our output is high as in another work result that directly paraphrases text without changing the sentence order. Additionally, we notice that considering the paragraph-level rather than the sentence-level has a positive impact on the Transformer-based model's output.

From the sentence reordering point of view, we can compare our results to [20]. The main difference is that they carried out paragraph paraphrasing without the sentence reorder step. The results prove that Transformer-based models can distinguish between the source text and reordered-paraphrased paragraphs without providing pair information. Longformer provides the best results at the paragraph and sentence levels. We suppose that the global attention prediction (GAP), a feature used by Longformer, enables the model to learn how to focus on the most important sections of a long text.

To expand on what other researchers have found in terms of how text length affects the machine learning algorithm's capacity [8], we can notice the same effect on the.

Transformer-based model's results: longer text provides more context and semantics thus improving the efficiency of machine learning and Transformer-based models in PI and PD tasks. Specifically, the F1-score of BERT and Longformer results increase by 18% and 23%, respectively, in detecting paragraph-level paraphrasing with 0.15 MLM (see Table 2 for the differences between the results for sentence vs. paragraph length). These percentages decrease as the percentage of paraphrased paragraphs' words increase.

Table 2. Classification results represented as F1 macro scores.

Classifier model	The [20] results		Our results					
	Bert	Longformer	Bert			Longformer		
MLM	0.15	0.15	0.15	0.20	0.30	0.15	0.20	0.30
Paragraph-level length	69	86	83	89	96	90	95	98
Sentence-level length	-	-	65	71	80	67	85	85

6 Conclusion and Limitation

In this work, we investigate the features of Transformer-based models in distinguishing between samples of original paragraphs and their paraphrases at the paragraph-level. Our excellent results with sentence reordering mean that the splitting and merging approach could potentially be used to develop a highly accurate paragraph-level paraphrasing detection approach although this would require a new dataset. To achieve this important objective, we create a large-scale paragraph-level paraphrasing dataset of content from multiple domains mostly related to education. We address the RQ using an experiment that shows high efficiency in detecting even the most difficult sample where the percentage of paraphrased tokens is low (15%) without any information from the source paragraph. Moreover, we report on the impact of text length on the Transformer-based models' efficiency.

In terms of limitation, an examination is conducted on the cutting-edge Transformer-based models, completely omitting ChatGPT due to its inconsistency, which renders it unsuitable for our dataset generation objectives. As for the evaluation methodology, an alternative forum might be explored, but we adhere to the approach advocated by researchers, involving 100 samples and conducting quantitative analysis based on qualitative analysis. Additionally, we implement automatic analysis across the entire dataset.

For future work, based on the findings of our experiment, which show that reordering paragraph sentences does not affect the classifier's capacity to recognise paraphrased paragraphs, we aim to determine how well Large Learning Models (LLMs) can generate and identify paragraph-level paraphrases.

References

1. Ventayen, R.J.M.: OpenAI ChatGPT Generated Results: Similarity Index of Artificial Intelligence (AI) Based Model (2023). Available at SSRN 4332664
2. Becker, J., Wahle, J.P., Ruas, T., Gipp, B.: Paraphrase detection: human vs. machine content. arXiv (2023). http://arxiv.org/abs/2303.13989. Accessed 15 Sept 2023
3. Gupta, A., Agarwal, A., Singh, P., Rai, P.: A deep generative framework for paraphrase generation. AAAI **32**(1), (2018). https://doi.org/10.1609/aaai.v32i1.11956
4. Ganitkevitch, J., Durme, B.V., Callison-Burch, C.: PPDB: The Paraphrase Database (2013)
5. Hu, J.E., Rudinger, R., Post, M., Van Durme, B.: ParaBank: monolingual Bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. arXiv (2019). http://arxiv.org/abs/1901.03644. Accessed 31 May 31, 2022

6. He, Y., Wang, Z., Zhang, Y., Huang, R., Caverlee, J.: PARADE: a new dataset for paraphrase identification requiring computer science domain knowledge. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020). http://arxiv.org/abs/2010.03725. Accessed 21 Sept 2021
7. Foltýnek, T., Meuschke, N., Gipp, B.: Academic plagiarism detection: a systematic literature review. ACM Comput. Surv. **52**(6), 1–42 (2020). https://doi.org/10.1145/3345317
8. Saqaabi, A.A., Akrida, E., Cristea, A., Stewart, C.: A paraphrase identification approach in paragraph length texts. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), Orlando, FL, USA, pp. 358–367. IEEE (2022). (2022). https://doi.org/10.1109/ICDMW58026.2022.00055
9. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. arXiv (2020). http://arxiv.org/abs/1909.11942. Accessed 22 Mar 2023
10. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv (2019). http://arxiv.org/abs/1810.04805. Accessed 21 Mar 2023
11. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv (2019). http://arxiv.org/abs/1907.11692. Accessed 21 Mar 2023
12. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: the long-document transformer. arXiv (2020). http://arxiv.org/abs/2004.05150. Accessed 21 Mar 2023
13. Roe, J., Perkins, M.: What are automated paraphrasing tools and how do we address them? A review of a growing threat to academic integrity. Int. J. Educ. Integr. **18**(1), 15 (2022). https://doi.org/10.1007/s40979-022-00109-w
14. Wan, S., Dras, M., Dale, R., Paris, C.: Using dependency-based features to take the "Para-farce" out of paraphrase. In: Proceedings of the Australasian Language Technology Workshop, p. 8 (2006)
15. Kowsari, J.M., Heidarysafa, M.: Barnes, and brown, 'text classification algorithms: a survey.' Information **10**(4), 150 (2019). https://doi.org/10.3390/info10040150
16. Ferreira, R., Cavalcanti, G.D.C., Freitas, F., Lins, R.D., Simske, S.J., Riss, M.: Combining sentence similarities measures to identify paraphrases. Comput. Speech Lang. **47**, 59–73 (2018). https://doi.org/10.1016/j.csl.2017.07.002
17. Yang, R., Zhang, J., Gao, X., Ji, F., Chen, H.: Simple and effective text matching with richer alignment features. arXiv:1908.00300 [cs] (2019). http://arxiv.org/abs/1908.00300. Accessed 08 Sept 2021
18. Dolan, W.B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Third International Workshop on Paraphrasing (2005)
19. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An evaluation framework for plagiarism detection. In: Coling 2010: Posters, pp. 997–1005 (2010)
20. Wahle, J.P., Ruas, T., Meuschke, N., Gipp, B.: Are neural language models good plagiarists? A benchmark for neural paraphrase detection. In: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, pp. 226–229. IEEE (2021). https://doi.org/10.1109/JCDL52503.2021.00065
21. Lin, Z., Cai, Y., Wan, X.: Towards document-level paraphrase generation with sentence rewriting and reordering. arXiv (2021). http://arxiv.org/abs/2109.07095. Accessed: May 31, 2022
22. Qiu, D.: Document-level paraphrase generation base on attention enhanced graph LSTM. Appl. Intell. **1–13**, 13 (2022)
23. Prentice, F.M., Kinden, C.E.: Paraphrasing tools, language translation tools and plagiarism: an exploratory study. Int. J. Educ. Integr. **14**(1), 11 (2018). https://doi.org/10.1007/s40979-018-0036-7

24. Larock, M.H., Tressler, J.C., Lewis, C.E.: Mastering effective English. Copp Clark Pitman, Mississauga (1980)
25. Elsner, M., Charniak, E.: Extending the entity grid with entity specific features. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 125–129 (2011)
26. Soricut, R., Marcu, D.: Discourse generation using utility-trained coherence models. In: Proceedings of the COLING/ACL on Main conference poster sessions, Sydney, Australia, pp. 803–810. Association for Computational Linguistics (2006). https://doi.org/10.3115/1273073.1273176
27. Shen, A., Mistica, M., Salehi, B., Li, H., Baldwin, T., Qi, J.: Evaluating document coherence modeling. Trans. Assoc. Comput. Linguist. **9**, 621–640 (2021). https://doi.org/10.1162/tacl_a_00388
28. van der Lee, C., Gatt, A., van Miltenburg, E., Krahmer, E.: Human evaluation of automatically generated text: current trends and best practice guidelines. Comput. Speech Lang. **67**, 101151 (2021). https://doi.org/10.1016/j.csl.2020.101151
29. Amidei, J., Piwek, P., Willis, A.: Rethinking the agreement in human evaluation tasks (Position Paper). In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3318–3329 (2018)

Educational Knowledge Graph Creation and Augmentation via LLMs

Gaganpreet Jhajj[1(✉)], Xiaokun Zhang[1], Jerry Ryan Gustafson[1], Fuhua Lin[1], and Michael Pin-Chuan Lin[2]

[1] School of Computing and Information Systems, Athabasca University, Edmonton, AB, Canada
{gjhajj,xiaokunz,oscarl}@athabascau.ca, jgustafson1@learn.athabascau.ca

[2] Faculty of Education, Mount Saint Vincent University, Halifax, NS, Canada
michael.lin@msvu.ca

Abstract. In this study, we explore the efficacy of Generative AI and Large Language Models (LLMs) in the tasks of constructing and completing Educational Knowledge Graphs (EduKGs). Knowledge Graphs (KGs) help represent real-world relationships. This can take the form of modeling course domains and student progression in educational settings. Through this work, we leverage GPT-4 to aid KG construction and align it with predefined learning objectives, course structure, and human interaction in validating and refining the generated KGs. The methodology employed utilized prompting LLMs with course materials and evaluating the generation of KGs through automatic and human assessment. Through a series of experiments, we show the potential of LLMs in enhancing the EduKG construction process, particularly for course modeling. Our findings suggest that LLMs such as GPT-4 can augment EduKGs by suggesting valuable and contextually relevant triplets. This KG creation and augmentation approach shows the potential to reduce the workload on educators and adaptive learning systems, paving the way for future applications in content recommendation and personalized learning experiences.

Keywords: Knowledge Graphs · Large Language Models · Educational Knowledge Graphs

1 Introduction

Knowledge graphs (KGs) are valuable tools for representing real-world relationships between entities based on ontology. Made famous by Google, modern KGs can help represent vast amounts of data and are commonly used in social networking services to represent real-world connections between individuals [1].

KGs have two key areas of interest: completion and construction, which are both often time-consuming and manual tasks. Many different strategies and approaches have been explored in the areas of KG construction and KG completion [2,3].

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 292–304, 2024.
https://doi.org/10.1007/978-3-031-63031-6_25

Educational knowledge graphs (EduKGs) refer to the application of KGs in education. EduKGs can be utilized in various ways and are a key area of interest in modeling both student course progression and course knowledge domains [4–6]. EduKGs can be used to create student-specific models for the educational domain and thus adapt over time to show students' progress. A domain model is a representation of a specific knowledge domain, such as a particular subject matter or field of study [7]. This model typically includes concepts, principles, and relationships that are relevant to the domain, and it is often used to represent the knowledge and expertise of a particular individual or group.

Generative AI (Gen AI) technologies, such as large language models (LLMs), have demonstrated potential in educational-related tasks, particularly question-answer generation. Our objective is to utilize the capabilities of Gen AI and LLMs to assist in constructing EduKGs. LLMs are trained on code from numerous programmers, enabling them to generate useful and effective code for specific contexts, and we aim to leverage this potential in our work [8].

2 Related Work

The process of KG construction can be divided into three sub-stages: knowledge acquisition, knowledge refinement, and knowledge evolution [9]. In our exploration, we focus on these first two steps of knowledge acquisition and knowledge refinement through leveraging Gen AI.

Prior work has shown that advanced LLMs are effective in KG construction [10,11]. Other related work has explored prompting for KG construction [12]. KGs have been used to model entire program curricula, which can be valuable for representing prerequisites, co-requisites, other program dependencies, and shared knowledge between courses [6,13].

KGs are utilized optimally in environments that emphasize the connections and relationships between data. The resulting information in a KG can be drawn and visualized, improving understanding and explainability. KGs are optimized around entity relationships, allowing us to explore the deep, intricate connections between learning objectives, questions, answers, and student interactions. This feature will enable us to provide individualized and adaptive feedback. KGs offer the opportunity to support and produce efficient, effective, and explainable adaptive learning assessments, increasing their transparency and trustworthiness. The applications of KGs in education primarily fall into the categories of assisted instruction, assisted learning, and education assessment [14].

LLM-enhanced KG generation can open up many opportunities as KG creation and triplet authoring are costly manual tasks. This problem can take away from the instructor's capacity, and the adoption of AI technologies hinges on how it impacts the instructor's workload [15]. Such technologies can only be effective in an educational setting if they do not increase the overall workload of the instructor [15]. Recent work on KG creation in a range of Massive Open Online Courses (MOOCs) has shown that using reference data like Wikipedia to determine semantic similarity can aid in KG construction and reduce the manual effort required to create an extensive KG [16]. Systems (e.g., KnowEdu) have shown potential for enhancing teacher and student outcomes on MOOC platforms and in intelligent tutoring systems [17]. LLMs have proven to be effective in KG construction through various tasks, which provide promise for LLM-assisted EduKG generation [18]. Conversely, LLMs have also been used to generate natural language from knowledge graphs, showcasing the versatility of such models in a wide range of tasks applicable to the educational domain [19] and have the potential to be used in adaptive formative assessment [20].

3 Research Problem

KGs can be used in a variety of ways in education, such as student modeling and content recommendation [4–6]. Additionally, KGs are a helpful tool in aiding educators with knowledge management and students with personalized learning [5,14].

What we hope to discover is how these Gen AI technologies can aid educators in KG creation and how we can enable a form of co-creative KG construction. Through this process, educators can provide a starting point for the LLM, which can undertake the task of KG completion. Through co-creative KG construction, this explores the symbiotic relationship between humans and AI systems. Human-AI co-creation has been explored in the context of programming classes with LLMs being used to power the AI engine [21].

Through this work we hope to address the following research questions:

1. How can we use an LLM to aid in EduKG construction?
2. To what extent can an LLM use unstructured or semi-structured text to augment existing KGs?

While much research has been done on automating KG construction to some extent or another, leveraging LLMs inherently requires a human-in-the-loop approach. The output of LLMs can be unpredictable in certain situations. When prompted for Cypher query generation, there is potential for non-query text to be generated, which will cause errors in KG tools such as Neo4j if the process is fully automated [22]. As a result, in this research, we propose a hybrid human AI approach to knowledge management related to generating our KGs. This is essential as curriculum alignment is one key area of interest and concern in AI-driven systems. This revolves around how the created content adheres to the knowledge required for successful course completion. Ensuring that the AI-generated content within our system adheres to our domain's learning outcomes and objectives is essential. Prior studies have shown how Human AI partnerships can be symbiotic in knowledge management tasks [23]. The human-in-the-loop approach uses intuition and reasoning to determine if proposed KG nodes and relationships fit our scope, verify results, and check for inaccuracy.

4 Methodology

4.1 KG Generation Process

To explore LLM-powered KG construction, we carried out a series of experiments. We leveraged Open AI's GPT-4 to create a KG for our Introductory Data Structures and Algorithms course [24].

We first investigated the ability of an LLM with unstructured reference text to generate a KG for our Introductory Data Structures and Algorithms course. For this task, we provided GPT -4 with an initial selection of reference text based on the course overview found on the course webpage and prompted it to create Cypher, the query language used in the graph database Neo4j; we will refer to this as KG1 for our analysis.

We selected Neo4j, a valuable tool for KG modeling [25,26] and a popular open-source graph database management system, considering its capability for processing more intuitive and flexible data modeling and powerful querying with the Cypher query language [25]. Figure 1 shows a high-level overview of our generation of KG1.

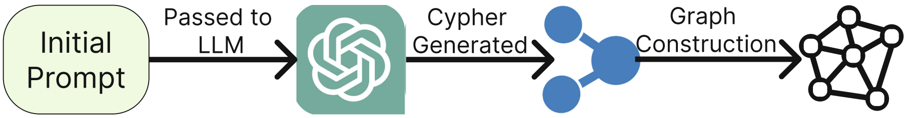


Fig. 1. High level overview of the generation process for KG1.

We also explored how using a base hand-authored KG can impact KG completion and augmentation. Figure 2 has a high-level overview of our augmented KG generation. In this, we can see that we create a KG augmented from text provided directly by the LLM, which we will refer to as KG2. We also create a KG generated from course reference text supplied to the LLM, which will be referred to as KG3, through which the generative KGs grow and improve incrementally with the participation of the instructors or students. Specifically, in the case of the student's participation in the loops, these typically generative KGs could dynamically record and accompany individual students' learning experience in the fine-grained context of the learning activities throughout the lifetime of the course study period.

Figure 3 shows the hand-authored KG mentioned in Fig. 2. As shown in Fig. 2, we took our hand-authored KG and then employed two different prompting strategies.

The first involved prompting GPT-4 to use its existing knowledge of data structures and algorithms to suggest additional nodes to augment each of the seven unit nodes. Unit nodes comprise the major topics of interest in the course, to which augmenting nodes are added via LLMs. After exploring various strategies through prompt engineering, we used GPT-4 to generate textual content representing topics and relationships based on the existing KG.

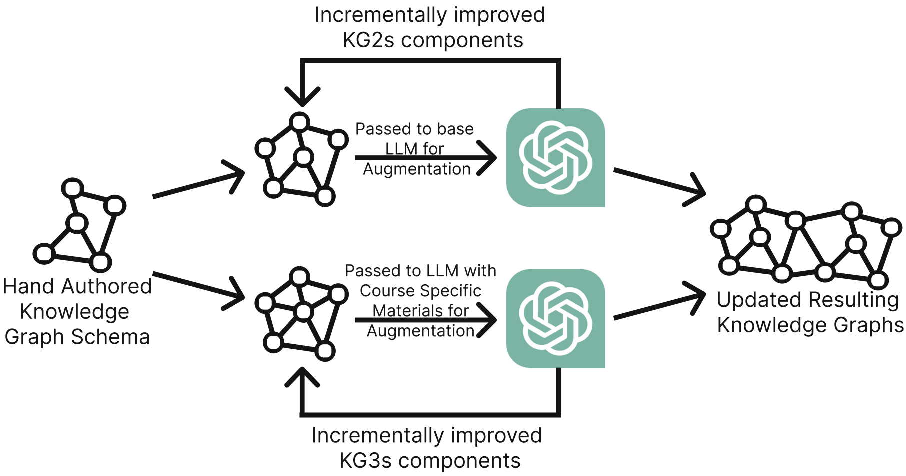


Fig. 2. High level overview of the generation process for our augmented KGs 2 and 3.

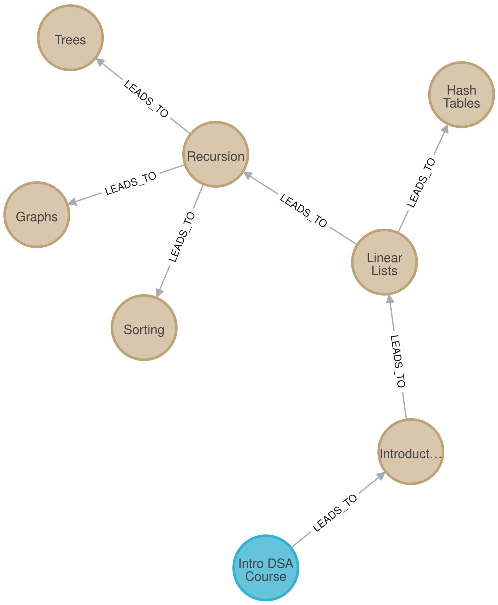


Fig. 3. Hand Authored KG based on Learning Outcomes for a Data Structures and Algorithms course.

GPT-4 provided content for all units in our base KG. After this step we prompted GPT-4 via ChatGPT with Prompt 1 to generate cypher that can be used to augment our initial KG.

Prompt 1: Use the provided text on course content and extract triples for knowledge graph augmentation in neo4j. This triplet and cypher query generation and creation should be consistent with the base knowledge graph provided.

In Prompt 1, for the construction of the fully LLM-augmented KG, the course content refers to the text generated by GPT-4, as shown previously. The base knowledge graph in Prompt 1 is the Cypher that constitutes the KG in Fig. 3.

The second method involved a similar prompting GPT-4 but using course-specific information derived from the study guide and curriculum. The same approach was used as seen in Prompt 1, and the course instructors provided the course content text.

No post-processing was required for the creation of KG1. However, for our augmented KGs 2 and 3, we had to employ post-processing steps to the generated Cypher to ensure that duplicate nodes were not created and that augmented nodes would be mapped onto the hand-authored KG.

4.2 Automatic Evaluation

For the automatic evaluation of our KGs, we determined the semantic similarity of shared nodes between KG2 and KG3 using cosine similarity to compare our embeddings that were generated by BERT using the BERT-base-uncased model [27]. As each augmented KG had a differing number of nodes, the most semantically similar nodes were chosen. Community Detection was performed using the Louvain Community Detection Algorithm and NetworkX in Python.

4.3 Human Evaluation

To evaluate KGs 2 and 3 for their quality, we compared them against a series of evolution questions, as seen in Table 1. We took questions from the quality evaluation section from work by Trajanoska et al. as they provided reasonable indications on the quality of KGs [10]. We also developed some of our own to fit our context of EduKGs; a total of 3 human evaluators rated each EduKG against the evaluation framework.

Table 1. Evaluation Framework for our KGs. Questions evaluated on a scale of 1 to 5 with a score of 5 being the highest.

Questions	Score for KG2	Score for KG3
1. Concise Learning Objectives and Outcomes: Clarity in educational objectives and outcomes is fundamental to ensure learners and educators can easily understand and align with the intended pedagogical goals		
2. Comprehensive Coverage of Subject Matter: Ensuring a broad and deep representation of subject areas guarantees learners access to all necessary knowledge and skills, fostering a well-rounded education		
3. Accuracy of Educational Content: A trustworthy EduKG's foundation is its content's reliability, necessitating rigorous verification to align with current academic standards and knowledge		
4. Consistency in Learning Pathways: Logical and contradiction-free learning paths ensure a coherent educational experience, enabling learners to build upon their knowledge systematically		
5. Contextual information about entities should be captured: This ensures a comprehensive representation of entities, contributing to the completeness of the graph		
6. The triples should not contradict each other: Consistency in the information presented prevents contradictions, which is crucial for the correctness of the knowledge graph		

5 Results

KG1 was created as outlined in Fig. 1, prompting GPT-4 with course materials; however, due to a large tendency for nodes to be interconnected, it was omitted from the results. Figures 4 and 5 showcase the two augmented KGs 2 and 3 created based on the methodology in Fig. 2. Table 2 showcases the semantic similarity between similar nodes. Human evaluation scores can be seen in Table 3. Community detection was done using the Louvain method; the results can be viewed in Table 4.

Table 2. Semantic Similarity between Knowledge Graphs

Unit	KG2 Nodes	KG3 Nodes	Similar Nodes	Semantic Similarity
1	4	12	4	0.965
2	10	11	10	0.875
3	4	6	3	0.867
4	6	3	3	0.966
5	6	5	5	0.860
6	6	7	6	0.778
7	9	8	8	0.876

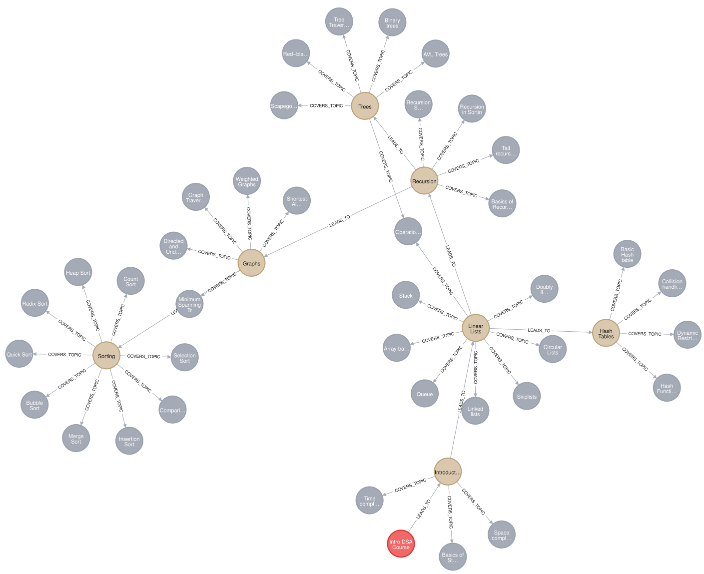


Fig. 4. KG2 created on direct prompting to GPT-4 with LLM generated content.

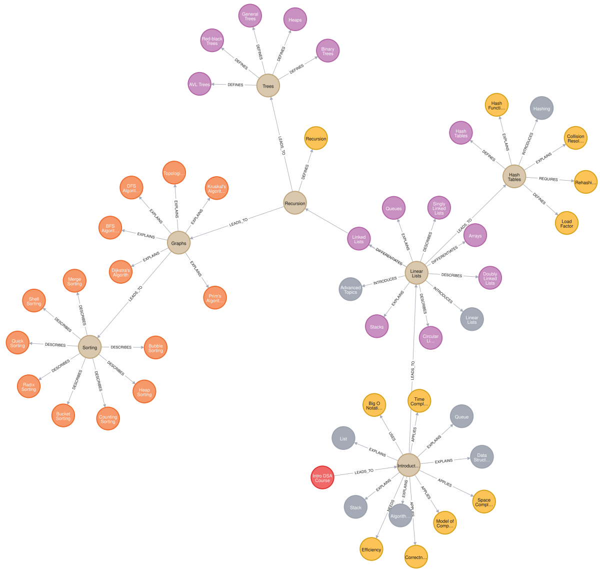


Fig. 5. KG3 created on direct prompting to GPT-4 with course specific learning outcomes.

Table 3. Human Evaluation Scores for Knowledge Graphs 2 and 3

Question	Average Score for KG2	Average Score for KG3
1	3.333	3.667
2	4	4.333
3	4	5
4	4.667	4.667
5	3.667	4.333
6	4.5	4.167

Table 4. Comparison of Community Structures in Knowledge Graphs

Graph	Nodes	Edges	Number of Communities	Community Number: Number of Nodes in the Community
KG2	46	46	7	0: 10, 1: 5, 2: 5, 3: 6, 4: 8, 5: 5, 6: 7
KG3	44	43	5	0: 13, 1: 9, 2: 6, 3: 7, 4: 9

6 Discussion

This work has explored the efficacy of LLM-assisted KG creation and co-creation. The ontology of KG1 did not contain any higher-order nodes indicating any specific learning outcomes. Instead, it showed the relationships that major course segments have with one another. Within the context of the course, we can see a unit as a major knowledge or course segment from which the models suggest augmenting nodes.

Figures 4 and 5 show the resulting KGs from the methodology outlined in Sect. 4.1 and Fig. 2. At a glance, we can see the presence of many more nodes and relationships suggested by GPT-4. KG2 has a total of 45 nodes, while KG3 has 52 nodes. Table 2 showcases the semantic similarity between the KGs. Across both KG2 and KG3, there were 39 similar nodes. This similarity was determined based on the number of nodes and by decreasing semantic similarity until no nodes were left from one KG to compare to the other. The average semantic similarity scores on the shared nodes were relatively high across all seven units. The highest of 0.965 on Unit 1 is due to only four compared nodes, three of which had the same semantic content. Units with higher amounts of similar nodes, such as Units 2 and 7, still had a high semantic similarity across the similar nodes of 0.875 for Unit 2 and 0.876 for Unit 7. This suggests that nodes created by the LLM for KG2 were reasonably comparable to course content provided by the instructor and that LLMs can suggest valuable and valid content to augment KGs.

The human evaluation scores in Table 3 show that across evaluation questions 1 to 5, KG3 was rated higher than KG2. As the creation process for KG3 involved using course materials, the scores related to learning objectives, subject matter, and accuracy are expected to be higher due to the course-specific content that has the reliability that was used in the creation process. In addition, the instructor-created content is the basis for the course-specific content, so it should be expected that KG3 would score higher on more metrics. The most considerable difference in ratings between both KGs is on question 3, which is "Accuracy of Educational content." Again, as KG3 received accurate course contents fed into the LLM, the difference of 1 point between the two averages is understandable. One metric that KG2 scored higher on in the framework was question 6, which relates to the contradiction of triplets. As seen in Fig. 5, KG3 has redundant triplets; an example is the Unit node Recursion pointing to a node Recursion with the relationship "defines." KG3 also has no other nodes connecting to the Unit Recursion.

When exploring the community structures generated by the Louvain method, Table 4 showcases that KG2 had 7 detected communities and KG3 had 5. Both are similar to the number of units of 7 that our base KG had specified, indicating that suggested nodes by GPT-4 fit within the broader community of each unit.

With our task of KGs in domain modeling, it is important to use descriptions for relationships that show the nature of the interaction. In KGs for educational tasks, the relationships connecting nodes play a critical role in describing the connections between course contents, objectives, assessments, learning materials, etc. These allow us to represent the complex educational environment and course contents in a scalable fashion to enable potential personalized learning experiences. Within EduKGs, relationship verbs such as teaches, requires, precedes, and enhances aid in structuring educational content. This allows us to effectively structure education content conducive to a wide array of learning and teaching strategies. Our work shows that LLMs can aid in creating and completing EduKG. However, future post-processing would be required to work towards using such specific domain-related semantics.

7 Conclusion and Future Work

Our findings suggest that Gen AI in the form of LLMs can facilitate synergistic relationships to leverage each one another's strengths [23]. When provided with enough foundational knowledge and context, in this case with the hand-authored KG in Fig. 3, the LLM could suggest functional augmenting nodes and relationships on both KG2 and KG3.

Using LLMs to generate EduKGs opens up future avenues to unify and use both technologies in tandem in real-time scenarios to leverage the general knowledge capabilities of LLMs with the domain-specific knowledge of KGs [28,29].

Leveraging LLMs to aid in EduKG creation can enhance adaptive practicing and learning systems by decreasing the manual time needed to set up such systems and create and maintain student models [30]. This enhancement is a crucial step to decrease the manual time needed by educators for this costly process. In the future, we hope to explore using graph neural networks to aid in various tasks, such as content recommendation, and to explore transfer learning via knowledge graph embeddings [31,32].

Acknowledgments. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Innovates, Alberta Advanced Education, and Athabasca University, Canada.

References

1. Introducing the Knowledge Graph: things, not strings. https://blog.google/products/search/introducing-knowledge-graph-things-not/. Accessed 30 Jan 2024
2. Ain, Q.U., Chatti, M.A., Bakar, K.G.C., Joarder, S., Alatrash, R.: Automatic construction of educational knowledge graphs: a word embedding-based approach. Information **14**, 526 (2023). https://doi.org/10.3390/info14100526
3. Chen, B., Bertozzi, A.L.: AutoKG: Efficient Automated Knowledge Graph Generation for Language Models (2023). https://doi.org/10.48550/arXiv.2311.14740
4. Woda, M., Kubacki-Gorwecki, K.: An effective learning environment. In: Zamojski, W., Kacprzyk, J., Mazurkiewicz, J., Sugier, J., Walkowiak, T. (eds.) Dependable Computer Systems, pp. 285–299. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21393-9_22
5. Troussas, C., Krouska, A.: Path-based recommender system for learning activities using knowledge graphs. Information **14**, 9 (2022). https://doi.org/10.3390/info14010009
6. Yu, X., Stahr, M., Chen, H., Yan, R.: Design and implementation of curriculum system based on knowledge graph. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 767–770 (2021). https://doi.org/10.1109/ICCECE51280.2021.9342370

7. Vélez-Langs, O., Argúello, X.: Adaptation in intelligent tutoring systems: development of tutoring and domain models. In: Jacko, J.A. (ed.) Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments, pp. 527–534. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73110-8_57
8. Mouret, J.-B.: Large language models help computer programs to evolve. Nature **625**, 452–453 (2024). https://doi.org/10.1038/d41586-023-03998-0
9. Zhong, L., Wu, J., Li, Q., Peng, H., Wu, X.: A comprehensive survey on automatic knowledge graph construction. ACM Comput. Surv. **56**, 1–62 (2024). https://doi.org/10.1145/3618295
10. Trajanoska, M., Stojanov, R., Trajanov, D.: Enhancing Knowledge Graph Construction Using Large Language Models (2023). https://doi.org/10.48550/arXiv.2305.04676
11. Liu, J., Zhan, J.: Constructing knowledge graph from cyber threat intelligence using large language model. In: 2023 IEEE International Conference on Big Data (BigData), pp. 516–521 (2023). https://doi.org/10.48550/arXiv.2111.07093
12. Carta, S., Giuliani, A., Piano, L., Podda, A.S., Pompianu, L., Tiddia, S.G.: Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction (2023). https://doi.org/10.48550/arXiv.2307.01128
13. Li, Z., Cheng, L., Zhang, C., Zhu, X., Zhao, H.: Multi-source education knowledge graph construction and fusion for college curricula. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 359–363 (2023). https://doi.org/10.1109/ICALT58122.2023.00111
14. Fettach, Y., Ghogho, M., Benatallah, B.: Knowledge graphs in education and employability: a survey on applications and techniques. IEEE Access **10**, 80174–80183 (2022). https://doi.org/10.1109/ACCESS.2022.3194063
15. Cukurova, M., Miao, X., Brooker, R.: Adoption of artificial intelligence in schools: unveiling factors influencing teacher's engagement. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education, pp. 151–163. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36272-9_13
16. Dang, F.-R., Tang, J.-T., Pang, K.-Y., Wang, T., Li, S.-S., Li, X.: Constructing an educational knowledge graph with concepts linked to wikipedia. J. Comput. Sci. Technol. **36**, 1200–1211 (2021). https://doi.org/10.1007/s11390-020-0328-2
17. Chen, P., Lu, Y., Zheng, V.W., Chen, X., Yang, B.: KnowEdu: a system to construct knowledge graph for education. IEEE Access. **6**, 31553–31563 (2018). https://doi.org/10.1109/ACCESS.2018.2839607
18. Zhu, Y., et al.: LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities (2023). https://doi.org/10.48550/arXiv.2305.13168
19. Axelsson, A., Skantze, G.: Using Large Language Models for Zero-Shot Natural Language Generation from Knowledge Graphs (2023). https://doi.org/10.48550/arXiv.2307.07312
20. Lin, F., De Silva, S.: An approach to generating adaptive feedback for online formative assessment. In: Frasson, C., Mylonas, P., and Troussas, C. (eds.) Augmented Intelligence and Intelligent Tutoring Systems, pp. 88–99. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-32883-1_8
21. Hassany, M., Brusilovsky, P., Ke, J., Akhuseyinoglu, K., Narayanan, A.B.L.: Human-AI Co-Creation of Worked Examples for Programming Classes (2024). https://doi.org/10.48550/arXiv.2402.16235

22. Mittelstadt, B., Wachter, S., Russell, C.: To protect science, we must use LLMs as zero-shot translators. Nat. Hum. Behav. **7**, 1830–1832 (2023). https://doi.org/10.1038/s41562-023-01744-0
23. Jarrahi, M.H., Askay, D., Eshraghi, A., Smith, P.: Artificial intelligence and knowledge management: a partnership between human and AI. Bus. Horiz. **66**, 87–99 (2023). https://doi.org/10.1016/j.bushor.2022.03.002
24. OpenAI: GPT-4 Technical Report (2023). https://doi.org/10.48550/ARXIV.-2303.08774
25. Hao, X., et al.: Construction and application of a knowledge graph. Remote Sens. **13**, 2511 (2021). https://doi.org/10.3390/rs13132511
26. Bai, J., Che, L.: Construction and application of database micro-course knowledge graph based on Neo4j. In: The 2nd International Conference on Computing and Data Science, pp. 1–5. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3448734.3450798
27. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). https://doi.org/10.48550/ARXIV.1810.04805
28. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying large language models and knowledge graphs: a roadmap. IEEE Trans. Knowl. Data Eng. 1–20 (2024). https://doi.org/10.1109/TKDE.2024.3352100
29. Pan, J.Z., et al.: Large Language Models and Knowledge Graphs: Opportunities and Challenges. https://doi.org/10.48550/arXiv.2308.06374. Accessed 12 Feb 2024
30. Kabir, M.R., Lin, F.: An LLM-powered adaptive practicing system. In: AIED 2023 Workshop on Empowering Education with LLMs-the Next-Gen Interface and Content Generation, Tokyo, Japan (2023)
31. Alatrash, R., Chatti, M.A., Ain, Q.U., Fang, Y., Joarder, S., Siepmann, C.: ConceptGCN: knowledge concept recommendation in MOOCs based on knowledge graph convolutional networks and SBERT. Comput. Educ. Artif. Intell. **6**, 100193 (2024). https://doi.org/10.1016/j.caeai.2023.100193
32. Liu, X., Yang, Z., Cheng, J.: Music recommendation algorithms based on knowledge graph and multi-task feature learning. Sci. Rep. **14**, 2055 (2024). https://doi.org/10.1038/s41598-024-52463-z

Semi-automatic Construction of Bidirectional Dialogue Dataset for Dialogue-Based Reading Comprehension Tutoring System Using Generative AI

Sung-Kwon Choi[1(✉)], Jin-Xia Huang[1,2], and Oh-Woog Kwon[1]

[1] Electronics and Telecomuncations Research Institute, 218 Gajeong-Ro, Daejeon, Korea
{choisk,hgh,ohwoog}@etri.re.kr

[2] University of Science and Technology (UST), Daejeon, Korea

Abstract. The goal of this paper is to semi-automatically construct a bidirectional reading comprehension dialogue dataset that enables bidirectional dialogue or debate on reading passages within a dialogue-based reading comprehension tutoring system. To achieve this goal, we developed a process for semi-automatically constructing bidirectional reading comprehension dialogue dataset. Using this process, ten English experts were able to construct 2,951 datasets, with an average difficulty level of 8.24 (high school level) and an average dialogue turn count of 9.75 per passage.

Keywords: Dialogue-based Reading Comprehension Tutoring · Bidirectional Dialogue Dataset · Semi-Automatic Construction Process

1 Introduction

Efforts to integrate AI technology into the field of education are very active. Among the various fields of education, there are emerging chatbots or dialogue systems for foreign language education, aiming to enhance foreign language learning abilities through AI technology. As an example, an English education dialogue system called 'AI PengTalk' was deployed to over 6,000 elementary schools in South Korea in early 2021 and applied to English speaking and listening learning [1]. The advantage of such English education dialogue systems is that learners can receive a learning environment similar to receiving English conversation education from native English teachers without being constrained by time and place.

Reading comprehension is important in foreign language education because it allows individuals to analyze the content of a text and grasp its key points. Especially in non-English-speaking countries, English reading comprehension is considered extremely important as it directly correlates with exam scores. Recently, a dialogue-based reading comprehension tutoring system has been developed [2]. This system was implemented through a reading comprehension tutoring dialogue dataset generated from multi-choice

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 305–313, 2024.
https://doi.org/10.1007/978-3-031-63031-6_26

reading comprehension data. The reading comprehension tutoring dialogue dataset consisted of system-centered dialogue data, composed of system (teacher) questions and student answers. Consequently, students were unable to ask questions when they found the reading passage difficult or did not understand the system's questions. To address this, this paper aims to describe the semi-automatic construction of bidirectional reading comprehension dialogue dataset for a dialogue-based reading comprehension tutoring system, enabling interaction between the system and students to engage in dialogue or debate regarding reading passages.

2 Related Work

Among reading comprehension datasets related to education, there are RACE [3] and ReClor [4], etc. These datasets are in the form of multiple-choice answers to a single question and do not aim to be dialogue datasets for reading comprehension education. Among English machine reading comprehension datasets, CoQA [5], DREAM [6], and ShARC [7] belong to conversational-style reading comprehension data, but they do not include dialogue data similar to the bidirectional dialogue between system and students described in this paper.

DIRECT (Dialogue-based Reading Comprehension Tutoring) [2] is the most similar dataset to the bidirectional dialogue data we aim to construct in this paper. Constructed based on the RACE [3] dataset, DIRECT dataset was composed of the following components: 1) English passages: at a middle school level, 2) Dialogue: system-initiated questions and student responses created based on multiple-choice practice questions, 3) Feedback: system feedback on student response errors, 4) Evidence sentences: sentences extracted from the passages where the responses are based on.

3 Semi-automatic Construction Process of Bidirectional Reading Comprehension Dialogue Dataset

3.1 Semi-automatic Construction Process

DIRECT which was centered around system-initiated questions was manually constructed based on reading practice exercises, primarily focusing on questions posed by the system and responses from the student. In contrast, the bi-directional reading comprehension dialogue dataset (hereinafter referred to as Bi-DIRECT) is semi-automatically constructed based on reading practice exercises. Leveraging generative AI technology, particularly ChatGPT [8], conversations are built bidirectionally, encompassing not only interactions initiated by the system with the student but also those initiated by the student with the system. The semi-automatic construction process unfolded as follows.

In Fig. 1, the left side represents the conventional method of constructing reading comprehension dialogue data, while the right side depicts the semi-automatic construction method using generative AI. Whereas the conventional method consisted of three steps, the semi-automatic construction method consists of five steps: 1) collection of passages and correct answers from open dataset data, 2) measurement of passage readability using Flesch-Kincaid Grade Level (FKGL) [9], 3) design and construction of prompts,

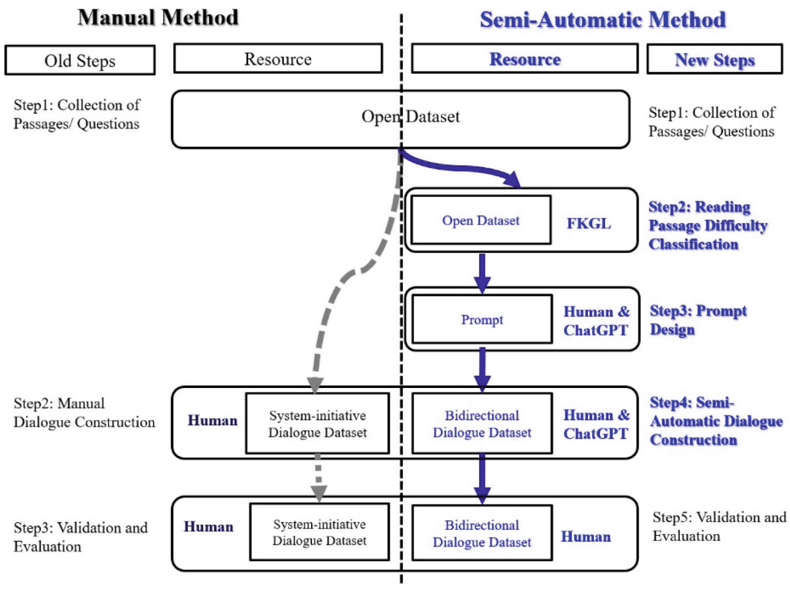


Fig. 1. Semi-Automatic Construction Process of Bi-DIRECT using Generative AI

4) automatic construction of Bi-DIRECT based on prompts using ChatGPT, along with error correction by English experts, and 5) validation and evaluation of automatically constructed dialogue data by English experts.

3.2 Example of Bi-DIRECT

An example of the input to the semi-automatic construction process, open dataset, and the output, Bi-DIRECT, is shown as follows.

Table 1 provides examples of passages and practice questions for high school-level students from RACE [3]. Table 2 presents examples of Bi-DIRECT constructed by the semi-automatic construction process from Table 1. In Table 2, 'Teacher-type' denotes the utterance type of 'Teacher-turn'. 'Source-tag' indicates the sentence number where 'Teacher-turn' appears in the passage. 'Teacher-turn' represents the system's utterance, while 'Student-turn' represents the student's utterance. 'Student-type' indicates the utterance type of 'Student-turn'. Detailed explanations for 'Teacher-type' and 'Student-type' will be provided in Sect. 4.2 on prompt design.

Table 1. An example of the input: RACE

Passage: Animals are natural resources that people have wasted all through our history. Animals have been killed for their fur and feathers, for food, for sport, and simply because they were in the way. Thousands of kinds of animals have disappeared from the world forever. Hundreds more are on the danger list today. About 170 kinds in the United States alone are considered in danger. Why should people care? Because we need animals. And because once they are gone, there will never be any more. Animals are more than just beautiful and interesting. They are more than just a source of food. Every animal has its place in the balance of nature. Destroying one kind of animal can create many problems. For example, when farmers killed large numbers of hawks, the farmers' stores of corn and grain were destroyed by rats and mice. Why? Because hawks eat rats and mice. With no hawks to keep down their numbers, the rats and mice multiplied quickly. Luckily, some people are working to help save the animals. Some groups raise money to let people know about the problem. And they try to get the governments to pass laws protecting animals in danger. Quite a few countries have passed laws. These laws forbid the killing of any animals or plants on the danger list. Slowly, the number of some animals in danger is growing

Exercise			
index	Question	Candicate answers	Answer
1	Animals are important to us mainly because	A."they give us a source of food", B."they are beautiful and lovely", C."they keep the balance of nature", D."they give us a lot of pleasure"	C
2	What has happened to the animals on the earth?	A."Hundreds of kinds of animals have disappeared forever.", B."Many kinds of animals have died out.", C. "About 170 kinds of animals have disappeared forever.", D. "All kinds of animals are in danger."	B
3	Which of the following is NOT true?	A. "People care much about animals because they need them.", B. "Once a certain kind of animals is gone forever, there will never be any more.", C. "Killing rats and mice will cause some problems.", D. "Laws have been passed to forbid the killing of any animals or plants on the danger list in some countries."	C

Table 2. An example of the output: Bi-DIRECT

Teacher-type	Source-tag	Teacher-turn	Student-turn	Student-type
QA1	11	Let's talk about the passage. In the passage, why are animals important to us?	They give us a source of food	Incorrect
Feedback	Other	Not quite. Think about the main point of the passage	Because they keep the balance of nature	Correct
QA2	3	That's right! / Now, what is going on with the animals on Earth?	Hundreds of kinds of animals have disappeared forever	Incorrect
Feedback	Other	Try to be a bit more specific with your answer and try again	Many kinds of animals have died out	Correct
QA3	Full text	Great job! / Which of the following statements is false?	Killing rats and mice will cause some problems	Incorrect
Feedback	15	That's not quite right. According to the passage, killing rats and mice actually caused problems. Do you remember why?	Oh, right! When farmers killed large numbers of hawks, the rats and mice multiplied quickly and destroyed their stores of corn and grain	Correct
Sympathizing	Other	Exactly! Do you have any other questions about the passage or anything else you'd like to discuss?	Why do you think some people don't care about protecting animals?	Critical thinking

4 Bidirectional Reading Comprehension Dialogue Dataset (Bi-DIRECT)

4.1 Reading Passage Difficulty Classification

The conventional reading comprehension tutoring system provided students with random passages regardless of the difficulty level. As a result, when presented with passages of higher difficulty, students with lower English proficiency tended to give up on reading comprehension. To address this issue, it was necessary to classify the difficulty level of passages and provide them to students accordingly. To classify the difficulty level of passages, the readability metric Flesch-Kincaid Grade Level (FKGL) [13] was utilized.

The formula to calculate FKGL was as follows.

$$\text{FKGL} = 0.39\left(\frac{\#total\_words}{\#total\_sentences}\right) + 11.8\left(\frac{\#total\_syllables}{\#total\_words}\right) - 15.59 \quad (1)$$

4.2 Prompt Design

Prior to designing prompts, a feasibility study was conducted to determine the validity of automatically constructing Bi-DIRECT. 50 dialogue datasets were automatically constructed using combinations of zero-shot and few-shot prompting with chatgpt-3.5-turbo and chatgpt-4.0 engines. These automatically generated dialogues were then compared and evaluated against the manually constructed ones. Automatic evaluations based on cosine similarity, TF-IDF, and sentence transformer [10] indicated that the combination of few-shot prompting and the gpt-4.0 engine yielded results most similar to manual construction [11]. Based on this feasibility study, prompts were designed for automatic generation of bidirectional dialogue between the system and students using few-shot prompting and the gpt-4.0 engine. The instructions in the prompts were designed to address types of issues [12] encountered by EFL students in reading comprehension for bidirectional dialogue. These types of issues included empathy, common sense, reasoning, critical thinking, personal experience, vocabulary, usage of words, repeating questions, and random questions.

4.3 Semi-automatically Constructed Bi-DIRECT

Based on reading passages and practice questions, bidirectional reading conversation data was automatically generated through prompts. Then, English experts attached tags to this conversation data. When comparing Bi-DIRECT with existing reading comprehension data, the following can be observed.

Among the reading comprehension data in Table 3, there are five datasets created for educational purposes, and among these, the datasets that have a conversational type are CoQA, DIRECT, and BI-DIRECT. BI-DIRECT features bidirectional dialogue initiative, and it stands out from other reading comprehension dialogue datasets with the highest average number of turns between the system and students.

When comparing BI-DIRECT with DIRECT, in DIRECT, 10 English experts manually constructed 92,233 turns over approximately 3.5 months for passages with a difficulty level of 4.53 (middle school level). In BI-DIRECT, 10 English experts semi-automatically constructed 57,559 turns over approximately 0.5 months for passages with a difficulty level of 8.24 (high school level). The reason for the smaller number of passages semi-automatically constructed, totaling 2,951, is because there was a predetermined budget. Despite the passages being twice as difficult and the average turn length being 3.33 words longer, the same group of English experts were able to construct 57,559 turns in just 1/7 of the time through the semi-automatic construction process, demonstrating the impact of the semi-automatic construction process.

Table 3. Comparison between Bi-DIRECT and existing reading comprehension data (ER: Educational Resources)

Dataset	Data Source	# Passages	# AVG Query	# AVG Aanswer	Dialogue Type	Dialogue Initiative
RACE [3]	ER	27,933	12.0	6.3	Multiple choice	X
CoQA [5]	Books, News, Wikipedia, ER	7,699	6.5	2.9	Conversational	System
DREAM [6]	ER	6,444	8.8	5.3	Multiple choice	System
ShARC [7]	Legal web sites	24,160	8.6	4.0	Conversational	System
DIRECT [2]	ER	5,704	8.1	8.1	Conversational	System
Bi-DIRECT	**ER**	**2,951**	**9.8**	**9.7**	**Conversational**	**System, Student**

5 Experiment

We investigated high-frequency combinations of student turn types (student type) and system turn types (teacher type) in Bi-DIRECT.

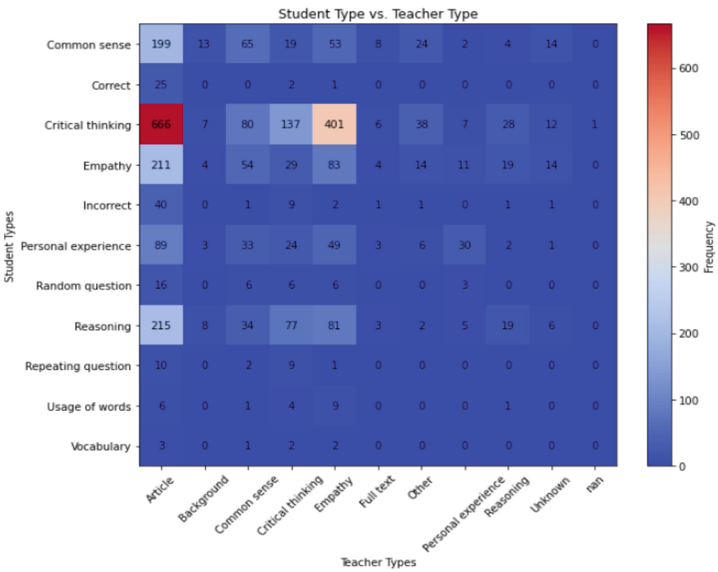


Fig. 2. Heatmap of student question types and system (teacher) response types

In Fig. 2, the heatmap displays the types of questions asked by students and answered by the system (teacher). The most frequent type of response from the system when students ask critical thinking questions is answering from the article, with 666 occurrences. While not shown in the paper due to space constraints, the heatmap of student responses to system questions reveals that the most common type of response from students when the system asks feedback-type questions about their incorrect answers is providing correct responses, with 6,932 occurrences.

6 Conclusion

In this paper, our goal was to describe a method for semi-automatically constructing a large-scale bidirectional reading comprehension dialogue dataset, enabling systems and students to engage in bidirectional conversations or debates regarding reading passages. As a result, we constructed bidirectional reading dialogue data through the semi-automatic construction process. To our knowledge, this process represents the first attempt at semi-automatically constructing dialogue data for reading comprehension education.

However, there are limitations to this study. The first limitation is that there is still a shortage of bidirectional dialogue data for deep-learning purposes. To address this, methods should be devised to automatically generate educational dialogues when only passages are provided without practice questions. The second limitation is that we did not evaluate the performance of bidirectional reading comprehension dialogue data when applied to dialogue-based reading comprehension tutoring system. We need to develop a dialogue-based bidirectional reading comprehension tutoring system within this year.

Acknowledgments. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019–0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

References

1. Jeon, H.B., Chung, H., Kang, B.O., Lee, Y.K.: Survey of recent research in education based on artificial intelligence. Electron. Telecommun. Trendes **36**(1), 71–80 (2021). (in Korean)
2. Huang, J.X., Lee, Y.H., Kwon, O.W.: DIRECT: toward dialogue-based reading comprehension tutoring. IEEE Access **11**, 8978–8987 (2023)
3. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 785–794 (2017)
4. Yu, W., Jiang, Z., Dong, Y., Feng, J.: ReClor: a reading comprehension dataset requiring logical reasoning. In: Proceedings of 8th International Conference Learning Representation, pp. 1–26 (2020)
5. Reddy, S., Chen, D., Manning, C.D.: CoQA: a conversational question answering challenge. Trans. Assoc. Comput. Linguist. **7**, 249–266 (2019)
6. Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., Cardie, C.: DREAM: a challenge data set and models for dialogue-based reading comprehension. Trans. Assoc. Comput. Linguist. **7**, 217–231 (2019)

7. Saeidi, M., et al.: Interpretation of natural language rules in conversational machine reading. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2087–2097 (2018)
8. OpenAI: GPT-3: Generative Pretrained Transformer, vol. 3 (2021). https://beta.openai.com/docs/models/gpt-3
9. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for Navy enlisted personnel. In: Institute for Simulation and Training (1975)
10. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992 (2019)
11. Choi, S.K., Jeon, Y.L., Huang, J.X., Kwon, O.W.: Feasibility study on constructing reading comprehension tutoring dialogue data based on ChatGPT. In: Korea Software Congress 2023, pp. 323–325 (2023). (in Korean)
12. Abeer, A.G., Mahmoud, A.S.: Reading comprehension problems encountered by EFL students at Ajloun National University. Int. J. Lang. Linguist. **6**(1), 6–15 (2021)

Author Index

A. Sifaleras and F. Lin (Eds.): ITS 2024, LNCS 14799, pp. 315–317, 2024.
https://doi.org/10.1007/978-3-031-63031-6