applied sciences

Special Issue Reprint

Emerging Technologies of Human-Computer Interaction

Edited by
Tibor Guzsvinecz

mdpi.com/journal/applsci

MDPI

Emerging Technologies of Human-Computer Interaction

Emerging Technologies of Human-Computer Interaction

Guest Editor

Tibor Guzsvinecz

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editor
Tibor Guzsvinecz
Department of Information
Technology and
Its Applications
University of Pannonia
Veszprém
Hungary

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Applied Sciences* (ISSN 2076-3417), freely accessible at: https://www.mdpi.com/journal/applsci/special_issues/627T226B9S.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-9267-9 (Hbk)
ISBN 978-3-7258-9268-6 (PDF)
https://doi.org/10.3390/books978-3-7258-9268-6

© 2026 by the authors. Articles in this reprint are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The reprint as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editor . . . . . **vii**

Preface . . . . . **ix**

Tibor Guzsvinecz
AI, Avatars, and Affective Touch: Emerging Frontiers in Human–Computer Interaction
Reprinted from: *Appl. Sci.* **2026**, *16*, 4990, https://doi.org/10.3390/app16104990 . . . . . **1**

Nils Haneklaus, László Simon Horváth, Hendrik Brink, Kim Brink-Flores, Hilda Dinah Kyomuhimbo, Tzong-Ru Lee, et al.
Negotiating Virtually and Face-to-Face: Experience from a Serious Game Conducted in Person and via Smartphone Application
Reprinted from: *Appl. Sci.* **2026**, *16*, 3300, https://doi.org/10.3390/app16073300 . . . . . **8**

Mayra Carrión-Toro, David Morales-Martínez, Marco Santórum, Luis Bernal, Julián Galindo, Jose Aguilar and Patricia Acosta-Vargas
Rethinking Usability in Serious Games: Designing an Instrument That Evaluates What Really Matters in Learning Contexts
Reprinted from: *Appl. Sci.* **2025**, *15*, 10084, https://doi.org/10.3390/app151810084 . . . . . **27**

Jasem Alostad, Fayez Eid Alazmi, Ali Alfayly and Abdullah Jasim Alshehab
Detection of Fault Events in Software Tools Integrated with Human–Computer Interface Using Machine Learning
Reprinted from: *Appl. Sci.* **2025**, *15*, 10030, https://doi.org/10.3390/app151810030 . . . . . **58**

Cristina Gasch, Alireza Javanmardi, Ameer Khan ,Azucena Garcia-Palacios and Alain Pagani
Exploring Avatar Utilization in Workplace and Educational Environments: A Study on User Acceptance, Preferences, and Technostress
Reprinted from: *Appl. Sci.* **2025**, *15*, 3290, https://doi.org/10.3390/app15063290 . . . . . **75**

Xiaoxiao Wang, Jingjing Zhang, Huize Wan and Yuan Yao
Wormhole Whispers: Reflecting User Privacy Data Boundaries Through Algorithm Visualization
Reprinted from: *Appl. Sci.* **2025**, *15*, 2034, https://doi.org/10.3390/app15042034 . . . . . **93**

Wang Guo, Kun Su, Bin Jiang, Kang Xie and Jie Liu
CMDAF: Cross-Modality Dual-Attention Fusion Network for Multimodal Sentiment Analysis
Reprinted from: *Appl. Sci.* **2024**, *14*, 12025, https://doi.org/10.3390/app142412025 . . . . . **111**

Sheng He, Hao Zeng, Mengru Xue, Guanghui Huang, Cheng Yao and Fangtian Ying
Affective Stroking: Design Thermal Mid-Air Tactile for Assisting People in Stress Regulation
Reprinted from: *Appl. Sci.* **2024**, *14*, 9494, https://doi.org/10.3390/app14209494 . . . . . **125**

Juan Carlos Muria-Tarazón, Juan Vicente Oltra-Gutiérrez, Raúl Oltra-Badenes and Santiago Escobar-Román
Uncovering Research Trends on Artificial Intelligence Risk Assessment in Businesses: A State-of-the-Art Perspective Using Bibliometric Analysis
Reprinted from: *Appl. Sci.* **2025**, *15*, 1412, https://doi.org/10.3390/app15031412 . . . . . **149**

About the Editor

Tibor Guzsvinecz

Tibor Guzsvinecz is an associate professor at the Department of Information Technology and its Applications, Faculty of Information Technology, University of Pannonia, Hungary. He received his PhD in Information Science in 2021. His research connects human–computer interaction, video games and data science. His recent work covers large-scale analysis of player reviews on Steam, including sentiment, toxicity and player engagement, as well as cybersickness and spatial ability in virtual reality, and gamification. He has authored 153 scientific publications, including 60 journal articles and 69 conference papers, and has an h-index of 16. He has edited three books on video games, gamification and game theory, and has guest edited five Special Issues, among them both editions of "Emerging Technologies of Human-Computer Interaction" in *Applied Sciences*. He serves on the editorial boards of *Humanities* and *Social Sciences Communications* and *Discover Computing*. His awards include the Outstanding Young Researcher Award of the Veszprém Regional Committee of the Hungarian Academy of Sciences, a Faculty Award for Outstanding Publication Activity, a Dean's Commendation, an IEEE Section Chair Award and a best section paper award. He is vice-chair of the Informatics and Computer Science Working Group of the same Regional Committee, and a member of the international organizing committees of the IEEE International Conference on Cognitive Infocommunications and the IEEE International Conference on Cognitive Aspects of Virtual Reality.

Preface

This Reprint brings together the articles published in the Special Issue "Emerging Technologies of Human-Computer Interaction" in *Applied Sciences*. The Special Issue was launched with a simple aim: to provide a common venue for research that studies new interaction technologies from the side of the people who use them, rather than from the side of the technology alone.

The contributions come from different disciplines and rely on different methods, including controlled experiments, questionnaire development, prototype design, computational modeling and bibliometric analysis. What connects them is the question of how an emerging technology changes the experience of learning, working, relaxing or making decisions, and what new risks it brings with it.

The Reprint is addressed to researchers and graduate students in human–computer interaction and neighboring fields, and to practitioners who design interactive systems for education, the workplace and health. I hope readers will find useful results in this Reprint, and also open questions worth following up.

The response to this Special Issue led to a second edition, which continues the same line of work. I thank all authors for their contributions, the reviewers for their careful work, and the Editorial Office of *Applied Sciences* for their support throughout the process.

Tibor Guzsvinecz
Guest Editor

applied sciences

MDPI

Editorial

AI, Avatars, and Affective Touch: Emerging Frontiers in Human–Computer Interaction

Tibor Guzsvinecz

Department of Information Technology and Its Applications, University of Pannonia, Gasparich M. utca 18/A, 8900 Zalaegerszeg, Hungary; guzsvinecz.tibor@zek.uni-pannon.hu

1. Introduction

The relationship between humans and computers has undergone a transformation over the past two decades. What was once a narrow exchange mediated by keyboards, mice, and two-dimensional screens has become a multimodal dialog spanning immersive virtual environments, gesture and gaze recognition, voice and language interfaces, haptic feedback, embodied avatars, and increasingly autonomous AI agents. Human–computer interaction (HCI), as a research field, no longer studies a single medium of communication. It studies an entire ecology of interfaces through which people now think, work, learn, play, and relate to one another.

This expansion is driven by several converging forces. First, the maturation of artificial intelligence (AI), particularly large language models, multimodal foundation models, and deep learning architectures for perception, has shifted computers from passive instruments toward active conversational and analytical partners [1,2]. Second, the widespread availability of consumer-grade virtual and augmented reality hardware has made immersive interaction practical outside the laboratory [3]. Third, the growing recognition that interaction is not purely cognitive but also affective, embodied, and socially situated has pushed HCI researchers to engage with topics traditionally associated with psychology, design, and ethics [4,5]. Fourth, the increasing deployment of intelligent systems in high-stakes contexts (workplaces, classrooms, healthcare, and public services) has forced the field to grapple with questions of usability, accessibility, privacy, trust, and risk that go far beyond classical interface design [6].

These shifts pose a methodological challenge as much as a technological one. Studying contemporary HCI requires combining engineering and machine learning with controlled user studies, qualitative inquiry, bibliometric and theoretical analysis, and human-centered design research [7]. No single approach is sufficient on its own, and the most informative contributions tend to arise where these methods are deliberately integrated.

The Special Issue "Emerging Technologies of Human–Computer Interaction" was conceived to capture this breadth. Rather than restricting submissions to a single subdomain, the call invited research on novel interaction techniques, immersive and smart environments, adaptive interfaces, wearable technologies, gesture recognition, user experience and user-centered design, video games, assistive technologies, and AI-driven interaction models. The eight articles accepted for publication reflect this diversity: they illustrate where HCI research stands today and which directions are gaining momentum.

This editorial summarizes the contributions and situates them within four broad themes that emerged organically from the accepted papers:

- serious games and simulation-based learning;
- immersive interaction and embodiment;
- affective, haptic, and emotion-aware interfaces;
- and AI integration and the human factors of intelligent systems.

2. Overview of Published Articles

2.1. Serious Games and Simulation-Based Learning

Two contributions address the design and evaluation of serious games, a domain where HCI intersects with education, behavioral change, and policy-relevant communication [8,9].

Haneklaus et al. [10] reported on the development and evaluation of a smartphone-based version of the Phosphorus Negotiation Game (P-Game), a serious game originally designed for face-to-face delivery in sustainability education [11,12]. Using a mixed-methods design that combined pre- and post-game surveys with qualitative analysis of open-ended responses, the authors compared learning outcomes across face-to-face and virtual modalities. Self-reported knowledge of phosphorus science increased across all groups (between 70.7% and 92.3%, depending on condition), with comparable gains in negotiation knowledge. Importantly, the virtual P-Game achieved a System Usability Scale (SUS) score above 80, indicating high user satisfaction. Beyond the encouraging quantitative results, the qualitative results revealed that the two modalities differed in experiential character: face-to-face delivery emphasized social and performance-oriented dimensions, while the online format favored structured participation and reflective processing. The study makes a practical case for virtual serious games as a complement, rather than a replacement, to in-person formats [13,14] and offers a transferable methodology for comparative evaluation using Reflexive Thematic Analysis [15] and the SUS [16].

Carrión-Toro et al. [17] tackled a long-standing gap in the field: the lack of usability evaluation instruments specifically designed for serious games [18,19]. While general-purpose usability questionnaires such as the SUS exist in abundance, they rarely capture the pedagogical dimensions and game-specific design elements that determine whether a serious game actually supports learning [20,21]. Following a systematic literature review based on the Kitchenham and Charters methodology [22], the authors developed a dedicated questionnaire (SGU) covering gameplay mechanics, interface design, narrative engagement, and gamification elements. The instrument was applied across multiple educational case studies, demonstrating its diagnostic potential and revealing recurring usability issues that traditional instruments had failed to surface [23]. The contribution is both conceptual, offering a richer model of what usability means in learning contexts, and practical, providing a tool that designers and researchers can adopt directly.

2.2. Immersive Interaction, Embodiment, and Privacy

Two articles address how users experience and interpret themselves and others within increasingly mediated digital environments [24,25].

Gasch et al. [26] investigated the use of virtual avatars across professional, educational, and recreational settings [27,28]. Through two studies involving 82 participants in total, the authors examined how variations in avatar realism, gender, and identity affect perceptions of embodiment, technostress, privacy, and credibility. The results are complex: hyper-realistic avatars resembling the user enhanced perceived embodiment and credibility in professional and educational contexts [29], whereas non-realistic avatars were preferred in recreational settings, particularly when interacting with strangers [30]. Privacy concerns intensified when avatars were controlled by others [31], and clear gender effects emerged.

https://doi.org/10.3390/app16104990

Women expressed greater concern about appearance, while men preferred same-gender avatars in professional environments. These results have implications for the design of virtual reality (VR) platforms intended for workplace collaboration, online education, and remote interviewing, where the choice of avatar style is rarely neutral.

Wang et al. [32] approached privacy from a different angle, treating it as a problem of legibility rather than policy. Their "Wormhole Whispers" project uses a metaphorical visual language, drawing on imagery of cosmic wormholes and planetary systems, to visualize how user data flows through online platforms and how algorithms process that data. The work is rooted in the observation that ordinary users, despite signing endless privacy agreements [33], possess little intuitive understanding of what actually happens to their data [34]. By translating abstract algorithmic processes (clustering, collaborative filtering, association rule learning) into interactive, participatory visual experiences [35], the authors offer an alternative path to data literacy that does not depend on technical fluency. The contribution sits at the intersection of HCI, information design, and digital rights [36] and exemplifies how interface design can become a tool for civic empowerment.

2.3. Affective and Haptic Interaction

One particularly notable contribution examines the role of touch in human–computer relationships [37,38].

He et al. [39] designed and evaluated a haptic device that uses mid-air air pressure combined with Peltier-element thermal feedback to produce affective stroking sensations on the user's forearm. Drawing on the literature on social touch and stress regulation [40,41], the authors first identified intensity and velocity parameters that produced subjectively comfortable and pleasant sensations. They then conducted an experiment with 49 participants randomly assigned to a Stroking Group or a Control Group, examining both perceived stress and physiological indicators (RMSSD and LF/HF heart rate variability metrics [42]). Participants in the stroking condition reported significantly greater relaxation than controls, and physiological indicators trended in the expected direction (though not all changes reached statistical significance). The work is notable not only for its technical novelty (non-invasive, noiseless, and contactless haptic stimulation) but also for its broader argument that vibrotactile biofeedback is not the only viable haptic approach to well-being applications [43]. Affective, computer-mediated touch deserves a place in the HCI toolbox.

2.4. AI Integration and Human Factors of Intelligent Systems

Three contributions address the integration of artificial intelligence and machine learning into human-facing systems, each from a different angle. The growing embeddedness of AI in everyday interfaces raises fundamental questions about transparency, interpretability, and the distribution of decision-making authority between humans and systems [44–46].

Guo et al. [47] proposed a Cross-Modality Dual-Attention Fusion Network (CMDAF) for multimodal sentiment analysis, the task of inferring affective states from combinations of text, audio, and visual signals [48,49]. The architecture integrates representations across modalities at multiple levels and uses a Transformer-based projection module to separate modality-consistent and modality-inconsistent features [50], further refined through multi-scale contrast learning. Evaluated on three benchmark datasets (MOSI, MOSEI, and CH-SIMS), the model achieved state-of-the-art performance on the majority of metrics. While the contribution is primarily technical, its relevance to HCI is direct: as conversational and emotion-aware interfaces become more common [4], the underlying perception models must reliably integrate signals that arrive simultaneously through multiple channels.

https://doi.org/10.3390/app16104990

Alostad et al. [51] addressed software defect prediction (SDP) for tools integrated with human–computer interfaces [52,53]. Their hybrid model combines Restricted Boltzmann Machines for nonlinear feature extraction with Logistic Regression for classification, validated across 21 benchmark datasets from the PROMISE and OpenML repositories. The RBM-LR model consistently outperformed baseline classifiers, including Random Forest, XGBoost, and SVM, across accuracy, precision, recall, F1-score, and AUC, with statistical significance confirmed via paired *t*-tests. The work emphasizes a balance between interpretability and predictive performance [54], a trade-off that becomes increasingly important as AI components are embedded in user-facing software, where failures directly degrade the user experience.

Finally, Muria-Tarazón et al. [55] provided a bibliometric analysis of artificial intelligence risk assessment in business contexts [56,57]. Drawing on Web of Science Core Collection data, the authors documented a compound annual growth rate of 31.20% in publications between 2018 and 2024, with particularly sharp acceleration following the public release of large language models such as ChatGPT and Gemini. Beyond the quantitative evidence of an emerging research front, the review identifies the most influential authors, journals, and institutions in this space and surfaces the critical themes (data privacy, ethics, bias, hallucinations) that organizations must address as they integrate AI into operations [58,59]. The contribution serves as a useful entry point for newcomers to the field and a reminder that the human factors of AI extend well beyond individual interaction into organizational and societal scales [60].

3. Cross-Cutting Observations

Several patterns are worth highlighting across the eight contributions. First, methodological pluralism is now the norm rather than the exception. The papers in this Special Issue draw on controlled experiments, mixed-methods designs combining quantitative and qualitative analysis, systematic literature reviews, design-research approaches, deep learning model development, and bibliometric analysis [61,62]. None of these methods alone would have been adequate for the questions addressed; the combinations are what enable meaningful contributions.

Second, the human is back at the center. After a decade in which much of the AI literature optimized models in isolation from the people who would eventually use them, the work collected here reflects a clear return to human-centered concerns: usability, embodiment, stress, privacy, trust, and learning [46,63]. This is not a rejection of technical sophistication, since several papers contain methodological innovations, but rather a recoupling of technical and human dimensions.

Third, emerging technologies bring emerging risks. Privacy, technostress, algorithmic opacity, and AI-related organizational risk all appear in this collection [5,36]. As interfaces become more capable and more pervasive, identifying and mitigating their unintended consequences becomes part of the HCI research agenda itself.

Fourth, simulation and play remain powerful methodological vehicles. Two of the eight articles engage with serious games not as marginal artifacts but as serious research instruments [64,65], a sign of the maturation of game-based research in HCI and adjacent fields.

Fifth, cross-domain transfer is emerging as a recurring design strategy. Several contributions in this issue borrow methods, frameworks, or theoretical constructs from adjacent fields such as neuroscience, organizational psychology, software engineering, and bibliometrics and apply them to HCI problems. This cross-pollination reflects the maturing

https://doi.org/10.3390/app16104990

interdisciplinarity of the field and suggests that the most productive future contributions will continue to draw on knowledge produced outside HCI's traditional boundaries [66,67].

4. Conclusions

The articles assembled in this Special Issue capture a representative slice of human–computer interaction research at a moment of considerable change. They span serious games, immersive avatars, affective haptics, multimodal AI, software reliability, privacy visualization, and the organizational study of AI risk. What unites them is not a single technique or theme but a shared commitment to understanding how emerging technologies meet, shape, and are shaped by human experience.

A second edition of this Special Issue is already underway, reflecting the continued momentum of the field. The editor hopes that the contributions collected here will serve as a useful reference for researchers, practitioners, and designers working at the intersection of technology and human behavior and that they will stimulate further inquiry into the many open questions that emerging HCI technologies continue to raise.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Xu, W. Toward human-centered AI: A perspective from human-computer interaction. *Interactions* **2019**, *26*, 42–46. [CrossRef]
2. Šumak, B.; Brdnik, S.; Pušnik, M. Sensors and Artificial Intelligence Methods and Algorithms for Human–Computer Intelligent Interaction: A Systematic Mapping Study. *Sensors* **2022**, *22*, 20. [CrossRef]
3. Rakkolainen, I.; Farooq, A.; Kangas, J.; Hakulinen, J.; Rantala, J.; Turunen, M.; Raisamo, R. Technologies for Multimodal Interaction in Extended Reality: A Scoping Review. *Multimodal Technol. Interact.* **2021**, *5*, 81. [CrossRef]
4. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
5. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
6. Lazar, J.; Feng, J.H.; Hochheiser, H. *Research Methods in Human-Computer Interaction*, 2nd ed.; Morgan Kaufmann: Cambridge, MA, USA, 2017.
7. Connolly, T.M.; Boyle, E.A.; MacArthur, E.; Hainey, T.; Boyle, J.M. A systematic literature review of empirical evidence on computer games and serious games. *Comput. Educ.* **2012**, *59*, 661–686. [CrossRef]
8. Deterding, S.; Dixon, D.; Khaled, R.; Nacke, L. From game design elements to gamefulness: Defining gamification. In Proceedings of the 15th International Academic MindTrek Conference, Tampere, Finland, 28–30 September 2011; pp. 9–15. [CrossRef]
9. Jung, M.; Sim, S.; Kim, J.; Kim, K. Impact of personalized avatars and motion synchrony on embodiment and users' subjective experience: Empirical study. *JMIR Serious Games* **2022**, *10*, e40119. [CrossRef]
10. Haneklaus, N.; Horváth, L.S.; Brink, H.; Brink-Flores, K.; Kyomuhimbo, H.D.; Lee, T.-R.; Mišík, M.; Roubík, H.; Kiselicki, M.; Szabó, P.; et al. Negotiating Virtually and Face-to-Face: Experience from a Serious Game Conducted in Person and via Smartphone Application. *Appl. Sci.* **2026**, *16*, 3300. [CrossRef]
11. Hallinger, P.; Wang, R.; Chatpinyakoop, C.; Nguyen, V.T.; Nguyen, U.P. A Bibliometric Review of Research on Simulations and Serious Games Used in Educating for Sustainability, 1997–2019. *J. Clean. Prod.* **2020**, *256*, 120358. [CrossRef]
12. Stanitsas, M.; Kirytopoulos, K.; Vareilles, E. Facilitating Sustainability Transition through Serious Games: A Systematic Literature Review. *J. Clean. Prod.* **2019**, *208*, 924–936. [CrossRef]
13. Oehlschläger, P.; Merz, M.A. Effectiveness of In-Person Versus Online Negotiation Teaching for Practitioners. *Negot. J.* **2023**. [CrossRef]
14. Dwivedi, Y.; Hughes, D.L.; Coombs, C.; Constantiou, I.; Duan, Y.; Edwards, J.; Gupta, B.; Lal, B.; Misra, S.; Prashant, P.; et al. Impact of COVID-19 Pandemic on Information Management Research and Practice: Transforming Education, Work and Life. *Int. J. Inf. Manag.* **2020**, *55*, 102211. [CrossRef]
15. Braun, V.; Clarke, V. Reflecting on Reflexive Thematic Analysis. *Qual. Res. Sport Exerc. Health* **2019**, *11*, 589–597. [CrossRef]
16. Lewis, J.R. The System Usability Scale: Past, Present, and Future. *Int. J. Hum. Comput. Interact.* **2018**, *34*, 577–590. [CrossRef]

https://doi.org/10.3390/app16104990

17. Carrión-Toro, M.; Morales-Martínez, D.; Santórum, M.; Bernal, L.; Galindo, J.; Aguilar, J.; Acosta-Vargas, P. Rethinking Usability in Serious Games: Designing an Instrument That Evaluates What Really Matters in Learning Contexts. *Appl. Sci.* **2025**, *15*, 10084. [CrossRef]
18. Moreno-Ger, P.; Torrente, J.; Hsieh, Y.G.; Lester, W. Usability testing for serious games: Making informed design decisions with user data. *Adv. Hum.-Comput. Interact.* **2012**, *2012*, 369637. [CrossRef]
19. Da Silveira, A.C.; Martins, R.X.; Vieira, E.A.O. E-Guess: Evaluación de usabilidad para juegos educativos. *RIED. Rev. Iberoam. Educ. Distancia* **2021**, *24*, 245–263.
20. Brooke, J. SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
21. Djaouti, D.; Alvarez, J.; Jessel, J.P. Classifying serious games: The G/P/S model. In *Handbook of Research on Improving Learning and Motivation Through Educational Games; Advances in Game-Based Learning*; IGI Global: Hershey, PA, USA, 2011; pp. 118–136. [CrossRef]
22. Kitchenham, B.; Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Technical Report EBSE-2007-01; Keele University: Keele, UK, 2007.
23. Yanez-Gomez, R.; Cascado-Caballero, D.; Sevillano, J.L. Academic methods for usability evaluation of serious games: A systematic review. *Multimed. Tools Appl.* **2017**, *76*, 5755–5784. [CrossRef]
24. Bujić, M.; Macey, A.L.; Järvelä, S.; Hamari, J. Playing with Embodied Social Interaction: A Thematic Review of Experiments on Social Aspects in Gameful Virtual Reality. *Interact. Comput.* **2021**, *33*, 583–595. [CrossRef]
25. Andreasson, R.; Alenljung, B.; Billing, E.; Lowe, R. Affective Touch in Human–Robot Interaction: Conveying Emotion to the Nao Robot. *Int. J. Soc. Robot.* **2018**, *10*, 473–491. [CrossRef]
26. Gasch, C.; Javanmardi, A.; Khan, A.; Garcia-Palacios, A.; Pagani, A. Exploring Avatar Utilization in Workplace and Educational Environments: A Study on User Acceptance, Preferences, and Technostress. *Appl. Sci.* **2025**, *15*, 3290. [CrossRef]
27. Cummings, J.J.; Bailenson, J.N. How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychol.* **2016**, *19*, 272–309. [CrossRef]
28. Roth, D.; Latoschik, M.E. Construction of the Virtual Embodiment Questionnaire (VEQ). *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 3546–3556. [CrossRef] [PubMed]
29. Yee, N.; Bailenson, J. The Proteus Effect: The effect of transformed self-representation on behavior. *Hum. Commun. Res.* **2007**, *33*, 271–290. [CrossRef]
30. Nimrod, G. Technostress: Measuring a new threat to well-being in later life. *Aging Ment. Health* **2018**, *22*, 1080–1087. [CrossRef] [PubMed]
31. Bente, G.; Rüggenberg, S.; Krämer, N.C.; Eschenburg, F. Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations. *Hum. Commun. Res.* **2008**, *34*, 287–318. [CrossRef]
32. Wang, X.; Zhang, J.; Wan, H.; Yao, Y. Wormhole Whispers: Reflecting User Privacy Data Boundaries Through Algorithm Visualization. *Appl. Sci.* **2025**, *15*, 2034. [CrossRef]
33. Luger, E.; Moran, S.; Rodden, T. Consent for all: Revealing the hidden complexity of terms and conditions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 2687–2696. [CrossRef]
34. Barth, S.; De Jong, M. The privacy paradox–Investigating discrepancies between expressed privacy concerns and actual online behavior–A systematic literature review. *Telemat. Inform.* **2017**, *34*, 1038–1058. [CrossRef]
35. Naps, T.; Eagan, J.; Norton, L. JHAVÉ—An environment to actively engage students in Web-based algorithm visualizations. In Proceedings of the Thirty-first SIGCSE Technical Symposium On Computer Science Education, Austin, TX, USA, 7–12 March 2000; pp. 109–113.
36. Royakkers, L.; Timmer, J.; Kool, L.; Van Est, R. Societal and ethical issues of digitization. *Ethics Inf. Technol.* **2018**, *20*, 127–142. [CrossRef]
37. Radhakrishnan, U.; Kuang, L.; Koumaditis, K.; Chinello, F.; Pacchierotti, C. Haptic Feedback, Performance and Arousal: A Comparison Study in an Immersive VR Motor Skill Training Task. *IEEE Trans. Haptics.* **2023**, *17*, 249–262. [CrossRef]
38. Gibbs, J.K.; Gillies, M.; Pan, X. A Comparison of the Effects of Haptic and Visual Feedback on Presence in Virtual Reality. *Int. J. Hum.-Comput. Stud.* **2022**, *157*, 102717. [CrossRef]
39. He, S.; Zeng, H.; Xue, M.; Huang, G.; Yao, C.; Ying, F. Affective Stroking: Design Thermal Mid-Air Tactile for Assisting People in Stress Regulation. *Appl. Sci.* **2024**, *14*, 9494. [CrossRef]
40. Morrison, I. Keep Calm and Cuddle On: Social Touch as a Stress Buffer. *Adapt. Hum. Behav. Physiol.* **2016**, *2*, 344–362. [CrossRef]
41. Gallace, A.; Spence, C. The Science of Interpersonal Touch: An Overview. *Neurosci. Biobehav. Rev.* **2010**, *34*, 246–259. [CrossRef]
42. Shaffer, F.; Ginsberg, J.P. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health* **2017**, *5*, 258. [CrossRef]
43. Eid, M.A.; Al Osman, H. Affective Haptics: Current Research and Future Directions. *IEEE Access* **2015**, *4*, 26–40. [CrossRef]

https://doi.org/10.3390/app16104990

44. Amershi, S.; Begel, A.; Bird, C.; DeLine, R.; Gall, H.; Kamar, E.; Nagappan, N.; Nushi, B.; Zimmermann, T. Software Engineering for Machine Learning: A Case Study. In Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), Montreal, QC, Canada, 25–31 May 2019; pp. 291–300. [CrossRef]
45. Cai, C.J.; Reif, E.; Hegde, N.; Hipp, J.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viegas, F.; Corrado, G.S.; Stumpe, M.C.; et al. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 4–9 May 2019; pp. 1–14. [CrossRef]
46. Shneiderman, B. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *Int. J. Hum.-Comput. Interact.* **2020**, *36*, 495–504. [CrossRef]
47. Guo, W.; Su, K.; Jiang, B.; Xie, K.; Liu, J. CMDAF: Cross-Modality Dual-Attention Fusion Network for Multimodal Sentiment Analysis. *Appl. Sci.* **2024**, *14*, 12025. [CrossRef]
48. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
49. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Volume 2019, p. 6558.
50. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Toronto, ON, Canada, 9–14 July 2023; pp. 7617–7630.
51. Alostad, J.; Alazmi, F.E.; Alfayly, A.; Alshehab, A.J. Detection of Fault Events in Software Tools Integrated with Human–Computer Interface Using Machine Learning. *Appl. Sci.* **2025**, *15*, 10030. [CrossRef]
52. Malhotra, R. A systematic review of machine learning techniques for software fault prediction. *Appl. Soft Comput.* **2015**, *27*, 504–518. [CrossRef]
53. Hall, T.; Beecham, S.; Bowes, D.; Gray, D.; Counsell, S. A systematic literature review on fault prediction performance in software engineering. *IEEE Trans. Softw. Eng.* **2011**, *38*, 1276–1304. [CrossRef]
54. Lessmann, S.; Baesens, B.; Mues, C.; Pietsch, S. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Trans. Softw. Eng.* **2008**, *34*, 485–496. [CrossRef]
55. Muria-Tarazón, J.C.; Oltra-Gutiérrez, J.V.; Oltra-Badenes, R.; Escobar-Román, S. Uncovering Research Trends on Artificial Intelligence Risk Assessment in Businesses: A State-of-the-Art Perspective Using Bibliometric Analysis. *Appl. Sci.* **2025**, *15*, 1412. [CrossRef]
56. Noyons, E.C.M.; Moed, H.F.; Van Raan, A.F.J. Integrating Research Performance Analysis and Science Mapping. *Scientometrics* **1999**, *46*, 591–604. [CrossRef]
57. Passas, I. Bibliometric Analysis: The Main Steps. *Encyclopedia* **2024**, *4*, 65. [CrossRef]
58. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* **2021**, *54*, 1–36. [CrossRef]
59. Siau, K.; Wang, W. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *J. Database Manag.* **2020**, *31*, 74–87. [CrossRef]
60. Fosso Wamba, S.; Bawack, R.E.; Guthrie, C.; Queiroz, M.M.; Carillo, K.D.A. Are We Preparing for a Good AI Society? A Bibliometric Review and Research Agenda. *Technol. Forecast. Soc. Change* **2021**, *164*, 120482. [CrossRef]
61. Creswell, J.W.; Creswell, J.D. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed.; SAGE: Thousand Oaks, CA, USA, 2018.
62. Hornbæk, K. Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Hum.-Comput. Stud.* **2006**, *64*, 79–102. [CrossRef]
63. Norman, D.A. *The Design of Everyday Things: Revised and Expanded Edition*; Basic Books: New York, NY, USA, 2013.
64. Freitas, S.D. Are Games Effective Learning Tools? A Review of Educational Games. *J. Educ. Technol. Soc.* **2018**, *21*, 74–84.
65. Zyda, M. From visual simulation to virtual reality to games. *Computer* **2005**, *38*, 25–32. [CrossRef]
66. Rogers, Y. *HCI Theory: Classical, Modern, and Contemporary*; Synthesis Lectures on Human-Centered Informatics; Springer: Berlin/Heidelberg, Germany, 2012.
67. Grudin, J. A moving target: The evolution of human-computer interaction. In *Human-Computer Interaction Handbook*, 3rd ed.; Jacko, J.A., Ed.; CRC Press: Boca Raton, FL, USA, 2012.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Article

Negotiating Virtually and Face-to-Face: Experience from a Serious Game Conducted in Person and via Smartphone Application

Nils Haneklaus [1,2,*], László Simon Horváth [3], Hendrik Brink [4], Kim Brink-Flores [5], Hilda Dinah Kyomuhimbo [4,6], Tzong-Ru Lee [7], Matúš Mišík [8], Hynek Roubík [9], Martin Kiselicki [10], Patrícia Szabó [3], Tibor Guzsvinecz [11] and Cecilia Sik-Lanyi [3]

1 Research Lab for Global Futures and Complex Systems, University for Continuing Education Krems, Dr.-Karl-Dorrek-Straße 30, 3500 Krems, Austria
2 Faculty of Earth Sciences, Geography and Astronomy, University of Vienna, 1090 Vienna, Austria
3 Department of Electrical Engineering and Information Systems, Faculty of Information Technology, University of Pannonia, Egyetem u. 10, 8200 Veszprem, Hungary
4 Department of Chemical Engineering, University of Pretoria, Pretoria 0002, South Africa
5 Department of Music, University of Pretoria, Pretoria 0002, South Africa
6 Department of Chemistry, Kyambogo University, Kampala P.O. Box 1, Uganda
7 Department of Marketing, National Chung Hsing University, 145 Xingda Rd., Taichung City 402, Taiwan
8 Institute of Political Science, Slovak Academy of Sciences, Dúbravska Cesta 9, 841 04 Bratislava, Slovakia
9 Department of Food and BioResource Technology, Faculty of Tropical AgriSciences, Czech University of Life Sciences Prague, 165 00 Prague, Czech Republic
10 Faculty of Economics—Skopje, Ss. Cyril and Methodius University in Skopje, 1000 Skopje, North Macedonia
11 Department of Information Technology and Its Applications, Faculty of Information Technology, University of Pannonia, 8900 Zalaegerszeg, Hungary
* Correspondence: nils.haneklaus@donau-uni.ac.at

Abstract

Serious games and negotiation simulations such as the Phosphorus Negotiation Game (P-Game) are increasingly used to support sustainability-oriented education. To broaden accessibility, a smartphone-based version of the face-to-face P-Game was developed and is presented here. A comparative design integrating quantitative pre–post survey measures with analysis of open-ended responses was employed to examine self-reported knowledge gains and learning experiences among participants who completed the P-Game in face-to-face workshops and those who played the virtual version. Both formats were associated with significant increases in participants' perceived understanding of phosphorus science and negotiation science/practice. Self-reported knowledge of phosphorus science increased by 92.3% (global face-to-face), 70.7% (Hungarian face-to-face), and 88.4% (online), with comparable gains observed in negotiation science and practice across groups. Qualitative findings complemented these results, indicating that while learning gains were broadly similar, the modes differed in experiential emphasis: face-to-face delivery elicited performance-oriented and socially embedded reflections, whereas the online format was more frequently described in terms of structured participation and reflective processing. User satisfaction with the virtual P-Game was high, reflected by a System Usability Scale (SUS) score above 80. Overall, the findings suggest that the virtual P-Game represents a viable and accessible complement to traditional face-to-face implementation, maintaining educational impact while extending reach. Further research with larger and more diverse participant samples is recommended to strengthen generalizability and explore long-term learning outcomes in sustainability contexts.

Keywords: Phosphorus Negotiation Game (P-Game); serious game; negotiation simulation; smartphone application; sustainability education; digital learning tools

1. Introduction

Serious games [1–4] and negotiation simulations [5–8] are becoming increasingly popular teaching tools to foster sustainable environmental education by providing an interactive and engaging learning environment. These tools offer participants a practical way to explore complex socio-environmental issues, encouraging them to analyze different perspectives, develop critical thinking skills, and practice decision-making in simulated real-world scenarios. Such approaches align with the growing need for innovative educational methods that promote experiential learning and address global sustainability challenges.

Most recently, researchers from the University of Continuing Education Krems in Austria [9] introduced the Phosphorus Negotiation Game (P-Game), a face-to-face serious negotiation game designed to support science-policy decision-making regarding sustainable phosphorus fertilizer production. Phosphate ore, one of the five most-mined materials globally, is an essential raw material for agricultural production due to its phosphorus (P) content. However, phosphate ores often contain relevant concentrations of radiotoxic uranium, which, if not recovered during processing, can largely transfer to the final fertilizer product [10–12]. The presence of uranium poses potential environmental and health risks, making the decision to recover or leaving it in the fertilizer a complex policy and economic challenge. In the P-Game, participants take on stakeholder roles to debate the advantages and disadvantages of recovering uranium during mineral fertilizer production. The goal of the game is to persuade others using well-researched arguments while navigating scientific, economic, and political considerations.

Since its inception, the P-Game has been successfully conducted with 788 participants across more than 20 countries, providing valuable insights into the decision-making processes of diverse stakeholder groups. To further expand its accessibility and impact, a virtual reality platform [13] and a smartphone application were developed, allowing the game to reach a broader audience beyond the constraints of in-person participation. Such digital adaptations of negotiation simulations could potentially enhance learning by offering a more flexible and inclusive approach, accommodating diverse geographical locations and enabling participants to engage asynchronously at their own pace [6,14]. In addition, a virtual setting may enable the inclusion of artificial participants [15].

Most negotiation classes or training courses have traditionally been conducted in person; however, the rapid advancement of technology has led to a significant increase in online or virtual negotiation courses. These formats have gained popularity due to their inherent flexibility [16,17], cost and time efficiency [18,19], and continuous technological improvements that increasingly simulate in-person interactions with higher fidelity [20]. While the trend towards virtual learning already existed before the COVID-19 pandemic, the global crisis accelerated the shift to online education, prompting a critical examination of its effectiveness in comparison to face-to-face methods [21].

A key debate in educational research concerns the effectiveness of virtual teaching and negotiation training, which was recently summarized by Oehlschläger and Merz [22]. One of the theoretical frameworks often referenced in this debate is the Media Richness Theory [23,24], which suggests that learning outcomes may be influenced by the richness of the communication medium used. According to this theory, richer media, such as in-person communication, can provide more effective learning experiences because they convey mul-

 https://doi.org/10.3390/app16073300

tiple cues (e.g., facial expressions, gestures, tone of voice) that enhance understanding and reduce ambiguity [25,26]. In contrast, virtual environments may require participants to invest more cognitive effort in navigating the interface and managing digital communication, potentially diverting cognitive resources from the actual learning process.

The P-Game is grounded in British Parliamentary Style (BPS) debating, an interactive debating format that fosters engagement by enabling direct interaction between students and instructors [27]. This format encourages active participation, critical analysis, and structured argumentation while allowing for constructive feedback and reinforcement of contributions [28]. The question remains whether such an interactive approach can be effectively transferred to a virtual setting without losing the essential elements that drive engagement and learning outcomes. The present study seeks to explore this issue by comparing self-reported learning gains and participant experiences across face-to-face and virtual versions of the P-Game.

Early observations suggest that while in-person debates benefit from real-time interactions and immediate feedback, virtual formats offer advantages in terms of accessibility, flexibility, and opportunities for reflection. Comparisons between the effectiveness of these formats may provide insights similar to those observed in previous studies on telephone negotiations [29], where the absence of non-verbal cues influenced negotiation dynamics but did not necessarily reduce overall outcomes.

Overall, the P-Game serves as a valuable tool for advancing sustainability, offering an interactive learning platform for participants to develop skills and knowledge relevant to complex environmental challenges. The objective of this work is to examine whether the educational benefits observed in face-to-face implementation are similarly reported when the P-Game is delivered via a smartphone application. To address this question, identical pre- and post-game survey instruments were used across both formats to assess self-reported knowledge gains, complemented by qualitative analysis of open-ended responses to explore how participants describe their learning experiences across delivery modes.

2. Materials and Methods

2.1. The Development of the Online Version of the P-Game

For the online version of the P-Game, *Flutter* (Google, Mountain View, CA, USA) [30], *Dart* (Google, Mountain View, CA, USA) [31] and *Firebase* (Google, Mountain View, CA, USA) [32] were used. *Flutter* is an open-source framework that can be used for building natively compiled applications for multiple platforms from a single codebase. *Flutter* was selected for its accessibility features, modern user interface (UI) capabilities, high performance, and extensive ecosystem of development packages. *Dart* offers high performance due to native code compilation, cross-platform support, strong typing for early error detection, and a rich library ecosystem. *Firebase* is a mobile and web application development platform. It provided backend services for application development, user authentication, and real-time data management. The following key *Firebase* services were also used: *Cloud Firestore*, *Cloud Functions*, *Authentication*, and *Cloud Storage*.

For the development, the Spiral Model was used. It was chosen due to its ability to handle risks effectively by combining elements of evolutionary, waterfall, and prototyping models. This iterative approach allowed for continuous refinement based on feedback, making it suitable for complex projects such as the P-Game. The Business Logic Component (BLoC) architecture was used to separate the presentation layer from the business logic and data layers, enhancing maintainability and scalability. The BLoC library's Cubit class provided state management. Additionally, Google's Material Design language ensured a consistent look and feel across platforms.

https://doi.org/10.3390/app16073300

Several P-Game specific features were integrated into the online adaptation. Voice chat was used as the primary communication tool during debates to reduce the influence of visual cues on argument evaluation. Asynchronous participation allowed users to engage in debates at their time of convenience. This promoted inclusivity and enabled participants to prepare more thoroughly before contributing. The developed smartphone application also allows participation from different locations (globally), although this feature was not utilized in the present assessment.

2.2. Details of the Application

Before entering the application, the users accessed a landing page. Here, the users could either log in or sign up before the game. An account could also be created using an already existing Google account. The landing page was accessed via a smartphone application or using a browser on a computer. Both can be seen in Figure 1.

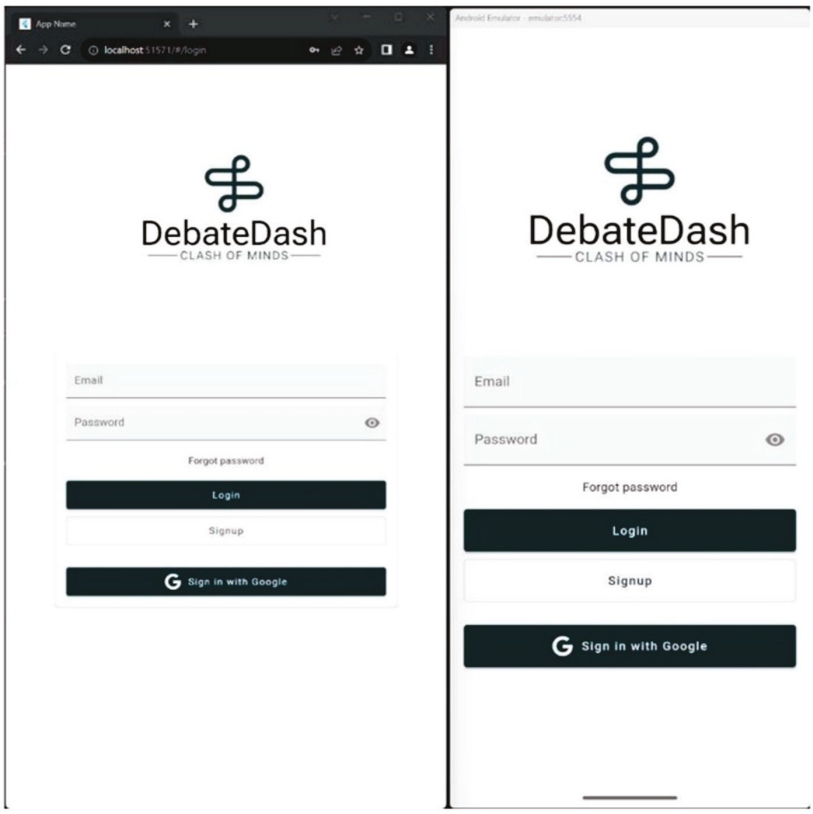


Figure 1. The virtual P-Game landing page using a computer (**left**) or smartphone (**right**).

After (signing up and) logging in, the users were taken to the lobby where they could see different debate rooms. They could also see whether a room was currently used, or whether the debate had already finished in a certain room. Users could add new public and private debate rooms by pressing the "+" button in the bottom right corner of the screen. If the users wanted to join a private debate room, they needed to enter a password. The lobby can be seen in Figure 2.

After entering a selected room, the users were taken to the debating screen (shown in Figure 3), where voice messages could be recorded or replayed. In Figure 3, details about the debate are shown at the top of the screen. According to these details, the debate concerned phosphorus mining. A short introduction about the subject was also provided to help the participants understand the specific topic chosen for the debate. After the game

https://doi.org/10.3390/app16073300

had started, users could send voice recordings to each other, simulating the BPS format that was also used for the in-person P-Game.

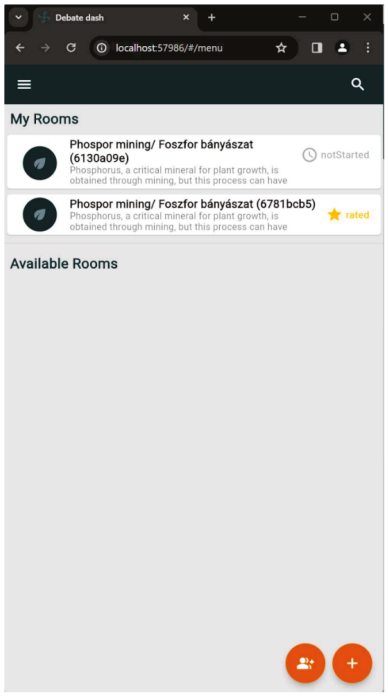


Figure 2. The lobby screen that was used in the virtual P-Game application.

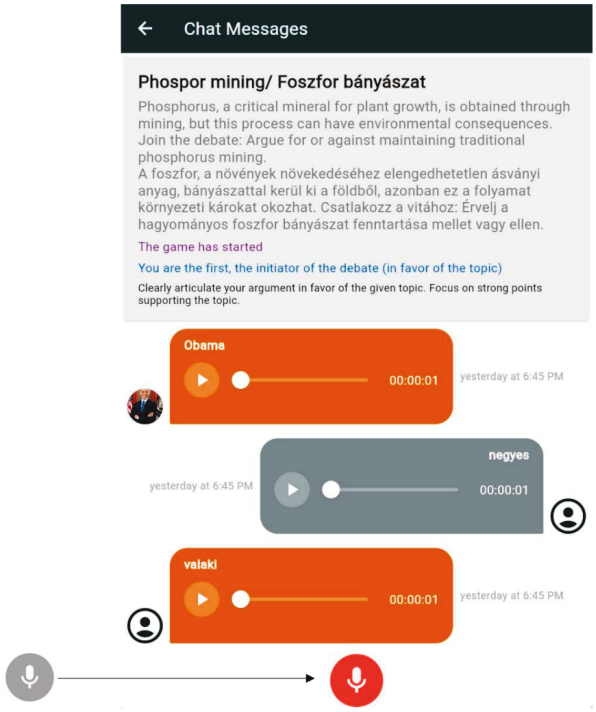


Figure 3. Debating screen with the recorded voice messages used in the virtual P-Game application.

2.3. Data Collection

A mixed-methods design was employed, combining quantitative pre–post measures with qualitative open-ended responses to examine both reported knowledge change and participants' experiential framing of learning.

https://doi.org/10.3390/app16073300

Data were collected on the gaming experience and the system usability. To better understand the gaming experience of participants and to compare the virtual P-Game experience with participant experiences from playing the face-to-face P-Game, the same pre- and post-game surveys were used as previously applied in the in-person P-Game [9]. Minor modifications were made to the survey questionnaire to include additional clarification for the Likert scale and open-ended questions. The content of the pre- and post-game surveys, as illustrated in Figure 4, remained unchanged. Despite the virtual P-Game being conducted on smartphones, participants were requested to complete printed copies of the surveys to maintain consistency with the face-to-face P-Game approach.

(**a**)

Pre P-Game Survey

Do you consider yourself a science or social science student?	Science ☐			Social Science ☐	
What is your major?					
What is your level of education?	BSc-Student ☐	MSc-Student ☐	PhD-Student ☐	PhD Awarded ☐	
Gender	Male ☐	Female ☐	Other ☐		
Age group	<20 ☐	20-25 ☐	25-30 ☐	30-35 ☐	>35 ☐
Nationality					
Country of residence					
How would you rate your knowledge of Phosphorus Science? *(from 1, very poor/little – to 5 very good/much) – write number*	1	2	3	4	5
How would you rate your knowledge of Negotiation Science/Practice? *(from 1, very poor/little – to 5 very good/much) – write number*	1	2	3	4	5
Date					
What would you like to learn about Phosphorus Science and/or Negotiation Science/Practice? (open answer)					

(**b**)

Post P-Game Survey

Do you consider yourself a science or social science student?	Science ☐			Social Science ☐	
What is your major?					
What is your level of education?	BSc-Student ☐	MSc-Student ☐	PhD-Student ☐	PhD Awarded ☐	
Gender	Male ☐	Female ☐	Other ☐		
Age group	<20 ☐	20-25 ☐	25-30 ☐	30-35 ☐	>35 ☐
Nationality					
Country of residence					
How would you rate your knowledge of Phosphorus Science? *(from 1, very poor/little – to 5 very good/much) – write number*	1	2	3	4	5
How would you rate your knowledge of Negotiation Science/Practice? *(from 1, very poor/little – to 5 very good/much) – write number*	1	2	3	4	5
How much and what do you think you learned about Phosphorus Science (open answer)					
How much and what do you think you learned about Negotiation Science/Practice? (open answer)					

Figure 4. Pre-game (**a**) and post-game (**b**) surveys used for the face-to-face and also the virtual P-Game.

Because the study examines perceived knowledge acquisition, negotiation confidence, and experiential engagement, constructs that involve internal cognitive and reflective processes, participants' self-reports represent an appropriate and widely used method in serious game and negotiation simulation research. While self-assessment measures may be influenced by social desirability bias or individual response tendencies, the pre/post design of the surveys strengthens interpretability by focusing on within-participant change rather than absolute scores. Furthermore, the integration of qualitative open-ended responses allows examination of how the participants articulated learning, providing complementary evidence beyond numerical self-ratings.

The surveys therefore included both quantitative and qualitative elements to provide a comprehensive overview of learning outcomes and game experience. Pre- and post-game measures assessed changes in participants' self-reported understanding of phosphorus science and negotiation science/practice. Open-ended questions enabled participants to express their reflections in their own words and were analyzed to identify patterns of meaning that extended beyond scaled responses. Future studies may further strengthen evaluative

https://doi.org/10.3390/app16073300

robustness by incorporating objective knowledge tests or structured performance-based indicators alongside self-reported data.

In addition to the pre- and post-game surveys used for the face-to-face P-Game, a standard System Usability Scale (SUS), shown in Table 1, was used to evaluate the usability of the online P-Game [33].

Table 1. Standard System Usability Scale (SUS) that was used in this work.

No.		Strongly Disagree				Strongly Agree
1	I think that I would like to use this system frequently.	○	○	○	○	○
2	I found the system unnecessarily complex.	○	○	○	○	○
3	I thought the system was easy to use.	○	○	○	○	○
4	I think that I would need the support of a technical person to use this system.	○	○	○	○	○
5	I found the various functions in this system were well integrated.	○	○	○	○	○
6	I thought there was too much inconsistency in this system.	○	○	○	○	○
7	I would imagine that most people would learn to use this system very quickly.	○	○	○	○	○
8	I found the system very cumbersome to use.	○	○	○	○	○
9	I felt very confident using the system.	○	○	○	○	○
10	I needed to learn a lot of things before I could get going with this system.	○	○	○	○	○

Testing of the virtual P-Game was conducted in two phases. First, an initial pilot test was conducted with a small group of participants from a course related to this work to identify and subsequently address major technical issues. Results from this pre-pilot testing study are not reported here. After these preliminary tests, full-scale testing was conducted with a group of 29 voluntary participants in Hungary at the University of Pannonia.

2.4. Data Analysis

Data from playing the virtual P-Game with 29 participants at the University of Pannonia in Veszprem in 2024 were compared to data from face-to-face P-Game sessions conducted with a different cohort of 30 participants in Hungary (19 students from the University of Pannonia in Veszprem in 2023 and 11 students from the University of Debrecen in 2024). Additionally, the obtained results were analyzed alongside data from all reported face-to-face P-Game rounds, which were conducted with a total of 788 participants across 22 countries between 2022 and 2024 [9].

The comparative analysis included descriptive statistics, percentage change calculations with uncertainty propagation [34], and effect size assessments using Hedges' G parameter [35]. These analyses were computed using Microsoft Excel 365 (Microsoft Corporation, Redmond, WA, USA). Statistical significance was assessed using two-way ANOVA in GraphPad Prism version 10.4.1 (GraphPad Software, Inc., La Jolla, CA, USA) to determine differences in self-reported knowledge scores across participant groups, delivery modes, gender, and academic backgrounds.

For comparisons between post-game differences relative to the pre-game baseline within each subgroup, Dunnett's multiple comparisons test was used, as it is well-suited for comparing multiple conditions against a common control. For within-group pre- vs. post-game comparisons, Šídák's multiple comparisons test was applied, allowing for robust

https://doi.org/10.3390/app16073300

pairwise analysis while controlling for the family-wise error rate. This combined approach ensured statistical validity while providing detailed insights into learning outcomes.

Data obtained from the SUS are presented separately, since such data were not obtained during the previous face-to-face P-Game rounds.

2.5. Qualitative Analysis of Open-Ended Responses

To complement the quantitative pre/post measures, open-ended survey responses from both delivery modes were analyzed using Reflexive Thematic Analysis (RTA) [36,37]. RTA was selected because the data comprised short form written reflections, where the aim was to identify patterned meanings in participants' accounts of learning and experience, rather than to quantify response categories or frequency counts.

An inductive data-driven coding approach was employed, allowing themes to emerge through iterative engagement with the dataset rather than being imposed a priori. No pre-defined codebook was used. Coding focused on semantic content, including explicit statements regarding perceived knowledge gains, negotiation skills, engagement, and platform evaluation. Analysis followed Braun and Clarke's six phases: familiarization with the dataset, initial coding, theme development, theme review, theme definition, and interpretive synthesis. An independent researcher completed the coding and analysis of the qualitative data.

Two qualitative corpora were examined:

- In-person Hungary pre- and post-game responses (n = 22);
- Online software pre- and post-game responses and usability comments (n = 19–21).

A comparative matrix was constructed to examine:

- Pre–post shifts within each delivery mode;
- Differences in thematic emphasis between face-to-face and online formats.

This comparative approach enabled interpretation of how the delivery context shaped participants' descriptions of learning, negotiation practice, and engagement. Qualitative findings were interpreted alongside quantitative results to enable methodological triangulation and a more nuanced understanding of reported learning outcomes across delivery modes.

3. Results and Discussion

3.1. Comparison of the Face-to-Face and Virtual P-Game Experience

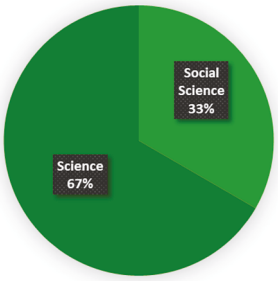
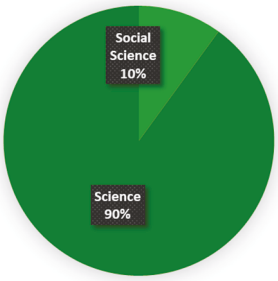
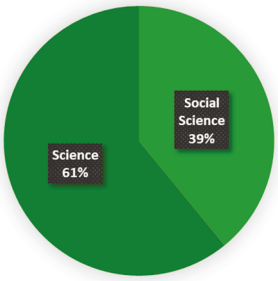
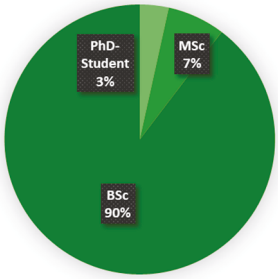
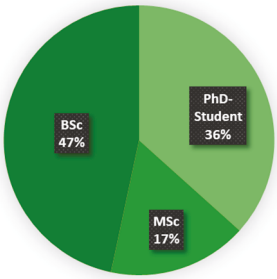
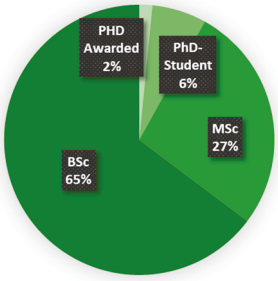
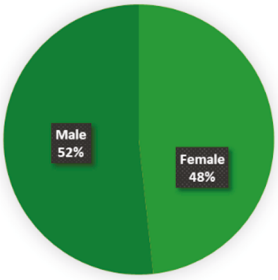
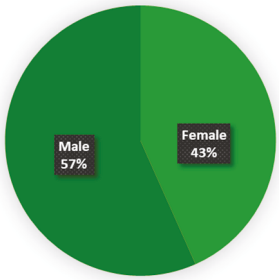
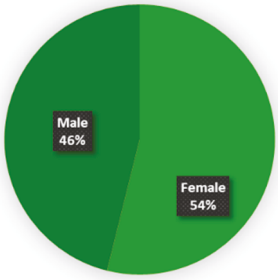
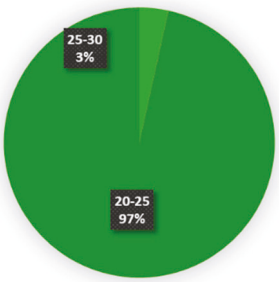
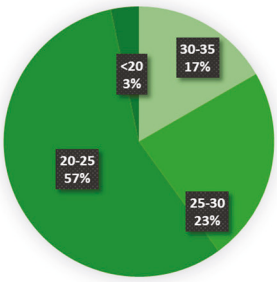
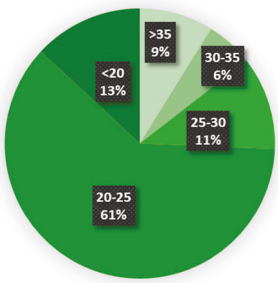
Table 2 provides an overview of the participants' backgrounds. The self-reported educational background of the virtual P-Game group in Hungary differed somewhat from that of the face-to-face P-Game group in Hungary, while it was more similar to that of the global face-to-face P-Game group. With regard to level of education, the virtual P-Game group from Hungary showed the largest share of BSc students, which was also reflected in the younger age distribution of the participants. The gender distribution was in fairly good agreement across all three groups.

It is worth mentioning that the participants of the in-person games self-reported a considerable number of nationalities, resulting in potentially diverse cultural backgrounds. The P-Game was inspired by the Mercury Game [38], and the researchers developing and conducting the Mercury Game recommended including participants with diverse cultural backgrounds to understand whether the developed game methodology can successfully be used with participants from different contexts. This was achieved in the global face-to-face P-Game, which was played in 22 countries with participants self-reporting more than 60 nationalities (Austria, Bangladesh, Belize, Bosnia and Herzegovina, Bulgaria, Cape Verde, China, Congo, Croatia, Czech, Egypt, Ethiopia, France, Germany, Ghana, Greece, Haiti, Hungary, Indian, Indonesia, Iranian, Iraq, Ireland, Italy, Japan, Kenya, Kingdom of

https://doi.org/10.3390/app16073300

Eswatini, Lebanon, Lesotho, Lithuania, Malaysia, Mexico, Mongolia, Moroccan, Myanmar, Nepal, Nigeria, Nigerian, Pakistan, Palau, Palestine, Philippines, Poland, Romania, South Africa, Russia, Rwanda, Serbia, Slovakia, Slovenia, South Sudan, Syria, Taiwan, Tanzania, Thailand, Tunisia, Uganda, Ukraine, United States of America, Vietnam, Zambia, Zimbabwe). The face-to-face P-Game in Hungary was also played with a relatively diverse group of students that self-reported 14 nationalities (Bangladesh, Cape Verde, China, Egypt, Hungary, Indonesia, Iraq, Kenya, Lebanon, Malaysia, Morocco, Palestine, Syria, Thailand). In contrast, the focus of the virtual P-Game was not on recruiting participants with diverse cultural backgrounds, and the virtual game was ultimately conducted with a more homogeneous group, largely (>96%) consisting of Hungarian students.

Table 2. Self-reported data of the participants from the pre-and post-game surveys.

	Virtual *P-Game* Hungary (*n* = 29, This Study)	**Face-to-Face *P-Game* Hungary (*n* = 30)**	**Face-to-Face *P-Game* Globally (*n* = 788)**
Background of the participants			
Educational background			
Level of education			
Gender			
Age			

https://doi.org/10.3390/app16073300

Table 2. *Cont.*

	Virtual *P-Game* Hungary (*n* = 29, This Study)	Face-to-Face *P-Game* Hungary (*n* = 30)	Face-to-Face *P-Game* Globally (*n* = 788)
Self-reported knowledge on phosphorus science			
Pre-game	1.72 ± 1.36	1.50 ± 0.86	1.56 ± 0.98
Post-game	3.24 ± 1.61	2.56 ± 0.94	3.00 ± 1.04
%Change [1]	+88.4 ± 82.5%	+70.7 ± 48.1%	+92.3 ± 66.2%
Hedges' G	1.02	1.16	1.43
Self-reported knowledge on negotiation science/practice			
Pre-game	2.34 ± 1.39	1.97 ± 1.07	1.83 ± 1.03
Post-game	3.64 ± 1.38	3.03 ± 0.96	3.12 ± 1.02
%Change [1]	+55.6 ± 39.1%	+53.81 ± 33.83%	+70.5 ± 45.9%
Hedges' G	0.94	1.04	1.26

[1] The SD of the % Change was calculated applying the principles of uncertainty propagation [34].

Despite these differences, comparable improvements in self-reported knowledge of phosphorus science and negotiation science/practice were observed across all three groups. Specifically, the self-reported knowledge of phosphorus science increased by 88.4 ± 82.5% in the virtual P-Game group from Hungary, 70.7 ± 48.1% in the Hungarian face-to-face P-Game group, and 92.3 ± 66.2% in the global face-to-face P-Game group, with corresponding Hedges' G effect sizes of 1.03, 1.18, and 1.43, respectively. The relatively large standard deviations across groups suggest substantial variability in reported learning gains, which may reflect differences in prior knowledge, engagement levels, and individual learning styles. Nevertheless, the consistently positive mean improvements indicate that both the virtual and face-to-face P-Game formats support learning about complex challenges such as the dilemma of unconventional uranium recovery from phosphate ores in mineral fertilizer production.

The Hedges' G values suggest a strong impact of the intervention across all groups. Values greater than 0.8 are generally considered large, indicating that the P-Game was associated with substantial improvements in participants' self-reported knowledge [39].

The self-reported knowledge improvements in negotiation science/practice showed a slightly different trend. The virtual P-Game group reported an increase of 55.6 ± 39.1%, whereas the face-to-face groups reported increases of 53.81 ± 33.83% (Hungary) and 70.5 ± 45.9% (global), with corresponding Hedges' G values of 0.94, 1.04, and 1.26. The relatively high pre-game survey values in the virtual P-Game group suggest that participants may have had prior exposure to the content, potentially explaining the observed differences in reported knowledge increases. This may also be attributed to information sharing from previous face-to-face P-Game rounds. This highlights the need for further studies with larger and more diverse samples to better understand the impact of the virtual format on negotiation knowledge gains. Conducting the P-Game at different institutions could also help clarify whether the observed differences stem from prior exposure or other contextual factors.

Overall, the Hedges' G values for both phosphorus science and negotiation science/practice are closely aligned, indicating a consistently strong effect of the P-Game intervention across both domains. This suggests that the P-Game supports learning in

https://doi.org/10.3390/app16073300

both areas, with variations likely attributable to contextual differences between participant groups rather than a fundamental disparity in the game's educational impact.

Both the online and face-to-face P-Game formats demonstrated their capacity to enhance participants' understanding of phosphorus science and negotiation skills. From a facilitation perspective, the face-to-face format was characterized by higher levels of direct engagement and motivation among participants. The qualitative evidence indicated that face-to-face debates fostered a competitive and dynamic environment, encouraging active participation and immediate feedback loops. This heightened interaction may contribute to stronger motivation and engagement, reinforcing the importance of interpersonal interaction in educational settings.

Conversely, the virtual P-Game format offered distinct advantages, which were highlighted in the feedback received from the participants. The flexibility of asynchronous participation allowed users to engage with the material at their own pace, facilitating reflection and, in some cases, a more thorough engagement with the topic. This flexibility may be particularly beneficial for participants with scheduling constraints or those in different time zones. Additionally, the virtual format's location independence provides opportunities for broader participation, potentially fostering a wider exchange of ideas and perspectives that could enhance the learning experience through cross-cultural insights.

The motivational aspects of both formats differed. Face-to-face participants often cited the energy and immediacy of live discussions as key motivational factors. In contrast, online participants appreciated the ability to prepare more thoroughly for their speeches, as well as the reduced pressure associated with not having to perform in person in front of peers, which may lessen performance-related anxiety for some participants. Communication patterns also differed: face-to-face interactions benefitted from non-verbal cues and immediate feedback, while online interactions relied primarily on voice recordings. The latter potentially reduced immediacy but allowed for more considered and structured responses.

3.2. Comparative Analysis of the Pre- and Post-Game Surveys

The bar charts (Figure 5) illustrate self-reported pre- and post-game knowledge (P and Neg) across three participant groups: in-person global, in-person Hungarian, and online participants. The analyses presented in Figure 5 are broken down into overall averages, gender, and academic fields, with statistical significance assessed using a two-way ANOVA followed by Dunnett's multiple comparisons test. Significance levels are indicated by (* $p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), and **** ($p < 0.0001$).

It is important to note that the in-person global study included 788 participants from 22 countries, while the Hungarian in-person and online groups each consisted of approximately 30 participants. Pre- vs. post-game comparisons were conducted exclusively within the online group, and Hedges' G effect size values provided additional context regarding the magnitude of observed differences.

For pre-game phosphorus science knowledge (Figure 5a), the two-way ANOVA revealed no significant differences across any subcategories, indicating that participants across all groups, regardless of delivery mode, gender, or academic field, began with a comparable baseline self-reported knowledge of phosphorus science. In contrast, pre-game negotiation science/practice knowledge (Figure 5c) showed significant differences, with the online group reporting higher scores overall and particularly within the social sciences subgroup (****). These findings suggest that online participants, especially those in social sciences, may have had higher prior exposure to or confidence in negotiation-related knowledge. However, given the small sample size of the online group, these results should be interpreted cautiously.

https://doi.org/10.3390/app16073300

The post-game results for phosphorus science knowledge (Figure 5b) revealed significant differences in the science and social sciences subgroups, where online participants reported higher post-game scores than either of the in-person groups. Similarly, post-game negotiation science/practice knowledge (Figure 5d) showed that the online group reported higher post-game scores overall (*) and within the social sciences subgroup (**). These results suggest that the online delivery mode may offer specific advantages for social sciences participants in improving phosphorus- and negotiation-related knowledge. However, the generalizability of these findings is limited by the smaller size of the online cohort compared to the global in-person dataset.

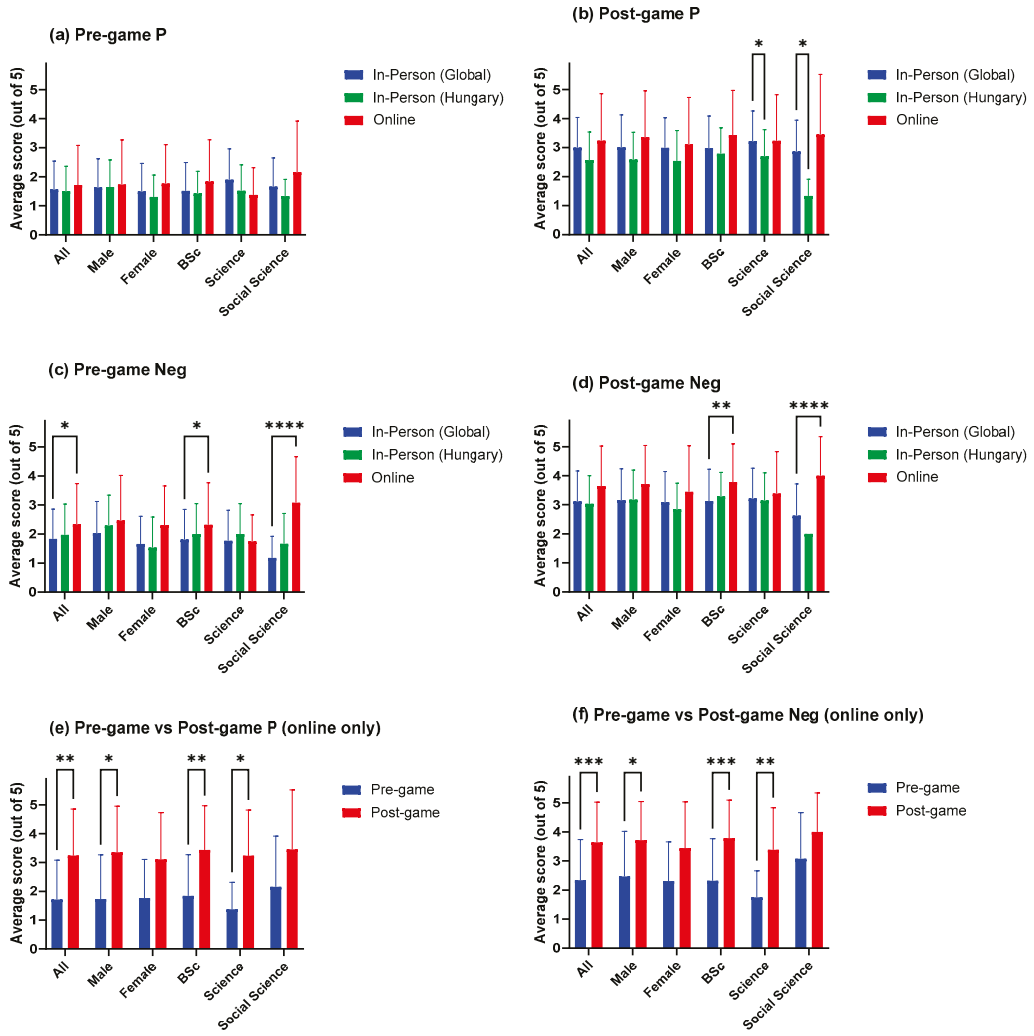


Figure 5. Comparison of pre-game and post-game self-reported knowledge scores for phosphorus science (P) and negotiation science (Neg) across different participant groups and demographics. Panels (**a**,**b**) show the pre-game and post-game scores for phosphorus science, respectively, while panels (**c**,**d**) present the corresponding scores for negotiation science. Panels (**e**,**f**) compare pre- and post-game scores within the online P-Game group for phosphorus science and negotiation science, respectively. The data are categorized by gender (male, female), educational level (BSc), and educational background (Science, Social Sciences). Statistical significance was determined using Dunnett's multiple comparisons test following a two-way ANOVA. Significance levels are indicated by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$.

The pre-game vs. post-game comparisons (Figure 5e,f), conducted exclusively within the online group, revealed significant gains in both phosphorus science and negotiation

https://doi.org/10.3390/app16073300

science/practice knowledge within most subcategories. The observed improvements, supported by statistical analysis, reinforce the effectiveness of the online P-Game format in supporting learning across diverse demographic groups.

For overall knowledge on phosphorous science, a large effect size (g = 1.025) suggests strong learning gains, with the largest gains observed in science participants (g = 1.44) and the smallest in social sciences participants (g = 0.68) (Table 3). Similarly, for knowledge on negotiation science/practice, a large effect size (g = 0.94) was observed, with science participants showing the largest improvement (g = 1.36) and social sciences participants the lowest (g = 0.63). These effect sizes suggest that while the activity was effective across all groups, the magnitude of improvement varied, with science participants benefiting the most.

Table 3. S Hedges' G Effect Sizes for self-reported knowledge gains across different participant demographics in the online P-Game format, based on the pre- and post-game surveys.

Group	Knowledge on Phosphorous Science	Knowledge on Negotiation Science/Practice
All	1.03	0.94
Male	1.04	0.86
Female	0.91	0.77
BSc	1.08	1.06
Science	1.44	1.36
Social Science	0.68	0.63

These findings highlight that the online format may be particularly effective for participants in science fields, as evidenced by the large effect sizes for both phosphorus science and negotiation science/practice knowledge improvements in this subgroup. Social sciences participants, while reporting smaller effect sizes, still demonstrated meaningful gains. The small sample size of the online group (approximately 30 participants) remains a limitation for generalizability, while the global in-person study's larger cohort of 788 participants from 22 countries provides a more robust dataset for analysis. Overall, the results emphasize the activity's adaptability and effectiveness across different delivery modes, with notable benefits observed in the online format for specific subgroups, particularly those in science fields.

3.3. Qualitative Thematic Findings from Open-Ended Responses

To complement the quantitative pre–post measures, open-ended responses were analyzed using Reflexive Thematic Analysis (RTA) [36,37]. The analysis explored how participants described learning, engagement and negotiation experience across face-to-face and online delivery modes. While both delivery formats showed positive self-reported learning gains, thematic patterns indicated that participants framed their learning experiences differently depending on mode of delivery. The qualitative findings largely supported the quantitative trends, providing interpretive depth regarding how learning was experienced across modes.

3.3.1. In-Person Delivery Themes

Five themes characterized responses from the in-person Hungary workshops.

Theme 1: Rapid conceptual orientation

Participants frequently described entering with minimal baseline knowledge and leaving with a foundational understanding of the topic (e.g., "now I feel I have the basics").

https://doi.org/10.3390/app16073300

This suggests that the face-to-face format supported accelerated entry into an unfamiliar socio-technical domain.

Theme 2: Memorable consequence-linked scientific learning

Learning claims were often anchored to vivid high-salient scientific details, particularly those involving uranium recovery, heavy metals, and environmental or health risks. These "sticky" recall anchors suggest that scientific learning was framed through consequence, trade-off, and uncertainty.

Theme 3: Negotiation learning as embodied performance

Negotiation learning was described in terms of confidence, persuasion, rebuttal, and argument craft. Participants emphasized speaking under pressure, structuring arguments, and responding to opposing claims (e.g., "if you are confident, you can sell even weaker statements"). This indicates that negotiation learning was experienced as enacted performance rather than abstract theory.

Theme 4: Critical realism about power and politics

Several participants reflected on how debate and negotiation function in real institutions, including strategic positioning, persuasion norms, and the influence of power. These reflections suggest that the simulation encouraged systems-level awareness beyond procedural debating skills.

Theme 5: Affective and social engagement as learning catalysts

Enjoyment, challenge, and social participation were frequently described as meaningful aspects of the learning experience (e.g., "really rewarding activity," "I had fun," "develop my social skills"). These responses highlight that affective engagement and group participation may contribute to learning outcomes in face-to-face delivery.

3.3.2. Online Delivery Themes

Five themes characterized responses from the online smartphone-based implementation.

Theme A: Baseline uncertainty and conceptual ambiguity

Pre-test responses frequently expressed unfamiliarity with the topic (e.g., "What is this?", "Never heard of it"). Compared to the in-person dataset, online baseline responses were often shorter and more minimal, suggesting variability in engagement or interpretation of the task.

Theme B: Generalized learning gains with fewer specific anchors

Post-test responses frequently reflected broad evaluative statements (e.g., "learned a lot," "new perspective"), with fewer specific scientific details compared to the in-person responses. However, occasional concrete anchors were present (e.g., uranium as a by-product of phosphate mining).

Theme C: Negotiation framed as listening and structure

Negotiation learning in the online mode was described more metacognitively, including attentive listening, understanding the structure of formal debate, and incremental improvement in confidence. Compared to face-to-face delivery, fewer responses emphasized real-time persuasion and rebuttal craft.

Theme D: The platform as a pedagogical actor

Online participants explicitly evaluated the interface and software features (e.g., comments on graphical user interface, usability, and suggestions for enhancements). This indicates that the technology itself became part of the learning experience, shaping engagement and reflection.

Theme E: Engagement polarization

https://doi.org/10.3390/app16073300

Some responses reflected strong engagement, while others indicated confusion or disengagement. This polarization suggests that online delivery may require clearer onboarding and scaffolding to ensure consistent participation and interpretation of the game context.

3.3.3. Cross-Mode Comparison

Overall, the thematic comparison suggests that face-to-face delivery elicited more vivid consequence-linked scientific recall and more performative descriptions of negotiation learning, while online delivery was more frequently associated with reflective structure-oriented negotiation learning and explicit evaluation of the software platform. These findings support the quantitative results by providing insights into how learning experiences were interpreted and articulated across delivery formats.

3.4. Results from the System Usability Scale (SUS)

The SUS score of this first group of participants was slightly higher than 81. According to standard SUS evaluation guidelines [33], a score above 68 is considered acceptable, whereas scores exceeding 80.3 are deemed excellent. The relatively high usability score suggests that the online platform is user-friendly and effective in delivering a positive game experience. However, it is important to acknowledge that this initial evaluation was conducted with technologically proficient BSc students, which may have influenced the overall score.

Participants from diverse educational and technical backgrounds may perceive and evaluate the usability of the platform differently. Therefore, continuous improvements and refinements will be essential to maintain and enhance usability, especially as new features and updates are introduced to the platform. Future assessments should include broader participant demographics to ensure a more comprehensive understanding of the platform's accessibility and effectiveness across different user groups.

4. Conclusions, Limitations and Outlook

The development of the virtual P-Game, as evidenced by self-reported knowledge gains, effect sizes, usability scores, and qualitative thematic findings, represents a meaningful and successful extension of its face-to-face counterpart. Both delivery formats were associated with substantial increases in participants' perceived understanding of phosphorus science and negotiation science/practice. The qualitative analysis of the open-ended responses suggests that these gains were not merely numerical shifts but were reflected in how participants articulated their learning experiences. In-person participants frequently described vivid consequence-linked scientific insights and embodied negotiation performance, whereas online participants more often emphasized reflective understanding, listening skills, and structured argumentation. Together, these findings indicate that while learning outcomes were broadly comparable across modes, the experiential texture of learning differed in meaningful ways.

The virtual format also expanded accessibility and flexibility, allowing participation independent of location and enabling asynchronous engagement. This increased accessibility broadens the reach of the P-Game and enhances inclusivity for participants who may be unable to attend in-person sessions. At a broader level, the virtual P-Game demonstrates how digital transformation can extend sustainability education while maintaining core pedagogical integrity. Rather than replacing the face-to-face format, the online version appears to offer a complementary pathway, supporting similar learning gains through different interactional affordances.

https://doi.org/10.3390/app16073300

However, significant differences were observed in pre- and post-game self-reported knowledge across groups. The online participants, particularly in the social sciences subgroup, reported higher pre-game scores in negotiation science/practice knowledge, suggesting potential prior exposure to the topic, which may have influenced post-game knowledge increases. This underscores the need for future studies to account for pre-existing knowledge differences when interpreting learning outcomes.

While both formats produced statistically significant improvements, qualitative patterns suggest that the face-to-face format may foster heightened affective engagement and real-time persuasive performance, whereas the online format may encourage reflective processing and structured contribution. These distinctions align with established theories of social presence and mediated communication and indicate that the delivery mode shapes not only outcomes but also how participants experience and describe their learning [28,40,41].

A key limitation of this study is that learning outcomes were assessed primarily through self-reported survey measures rather than objective performance-based indicators. Although the pre–post design reduces the influence of individual response tendencies by focusing on within-participant change, self-report data may still be affected by social desirability bias, differences in interpretation of scale items, or variability in response effort. Future research could strengthen evaluation by incorporating objective knowledge checks (e.g., short quizzes on phosphorus concepts), behavioral indicators (e.g., participation frequency, response length, turn-taking patterns), or structured rubrics for assessing negotiation quality (e.g., evidence use, rebuttal effectiveness, argument coherence). Such measures would provide a valuable additional layer of validation and allow closer comparison between perceived and demonstrated learning outcomes.

The nature of the written qualitative responses collected across the two delivery modes represents a related limitation. Differences in social presence and facilitation between face-to-face gameplay and asynchronous online participation may shape not only learning outcomes but also how participants articulate their experience. The online cohort showed more variability in the depth and detail of open-ended responses, which is a recognized feature of self-administered survey formats [42–45] and may reflect differences in engagement or in the interpretation of the prompts.

A further limitation concerns the sample size and the lack of diversity among participants. The online format was tested with a relatively small sample of fewer than 30 participants, predominantly from Hungary, which does not represent a globally diverse population. Expanding the sample size and including participants from more diverse geographical, educational, and cultural backgrounds in future studies could yield more generalizable results. Another limitation is the relatively short testing period. A longer duration could facilitate deeper interactions, potentially leading to more comprehensive insights into learning dynamics and knowledge retention.

The face-to-face format allowed for immediate observable feedback and dynamic social interaction, which were less directly replicated in the online environment. The qualitative responses suggest that in digital settings, onboarding clarity, interface design, and platform usability become integral components of the pedagogical experience.

Additionally, technological barriers, such as access to reliable internet and device compatibility, could have influenced participants' ability to fully engage with the virtual P-Game. Differences in technology access may introduce variability into the user experience that was not captured in this study. Future studies should aim to enhance the platform's support for various devices and connectivity conditions to ensure a more consistent and inclusive experience for all users.

https://doi.org/10.3390/app16073300

In conclusion, the findings suggest that both virtual and in-person formats can effectively support sustainability-oriented learning in phosphorus science and negotiation practice. While the modes differ in interactional dynamics and experiential emphasis, the virtual P-Game demonstrates strong potential as an accessible and scalable complement to traditional implementation. Further research with larger more diverse samples and multimodal evaluation strategies will be essential to fully understand how delivery context shapes learning processes and long-term educational impact.

Author Contributions: Conceptualization N.H., T.G., C.S.-L., data curation N.H., H.B., K.B.-F., formal analysis N.H., L.S.H., H.B., K.B.-F., H.D.K., T.-R.L., M.M., H.R., M.K., P.S., T.G., C.S.-L., funding acquisition N.H., H.B., T.-R.L., M.M., C.S.-L., investigation N.H., L.S.H., H.B., K.B.-F., H.D.K., T.-R.L., M.M., H.R., M.K., P.S., T.G., C.S.-L., writing—original draft N.H., H.B., K.B.-F., T.G., writing—review and editing N.H., H.B., K.B.-F., H.D.K., T.-R.L., M.M., H.R., M.K., P.S., T.G., C.S.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research, Development and Innovation Office of Hungary [2021-1.2.4-TÉT-2021-00007] and the Federal Ministry of Education, Science and Research (BMBWF) through Austria's Agency for Education and Internationalization (OeAD) [Africa UniNet P119, P169, APPEAR Project341, HU 06/2022, HU 06/2024, MK 13/2024]. APPEAR is a programme of the Austrian Development Organization. The project received further support from the Austrian Science Fund (FWF) [10.55776/ESP1514224], the National Research Foundation of South Africa (NRF) [CSRP220420402 and CPRR240404212107] and the Ministry of Education and Science of the Republic of North Macedonia [20-667/21]. H. R. acknowledges support from Horizon Europe project BIO-CAPITAL [101135150].

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Research Ethics Committee of the University of Pannonia (Ikt.sz.: RK/6/2/2026).

Informed Consent Statement: Freely-given and informed consent was obtained for each participant. They were reassured that their participation is voluntary and that they were free to withdraw at any time. In addition, all information was gathered anonymously and handled confidentially. The study design assured adequate protection of study participants, and neither included clinical data about patients nor configured itself as a clinical trial.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors are grateful to the students who participated in this work.

Conflicts of Interest: The authors have no relevant financial or non-financial interests to disclose.

References

1. Madani, K.; Pierce, T.W.; Mirchi, A. Serious Games on Environmental Management. *Sustain. Cities Soc.* **2017**, *29*, 1–11. [CrossRef]
2. Stanitsas, M.; Kirytopoulos, K.; Vareilles, E. Facilitating Sustainability Transition through Serious Games: A Systematic Literature Review. *J. Clean. Prod.* **2019**, *208*, 924–936. [CrossRef]
3. Ahmadov, T.; Karimov, A.; Durst, S.; Saarela, M.; Gerstlberger, W.; Wahl, M.F.; Karkkainen, T. A Two-Phase Systematic Literature Review on the Use of Serious Games for Sustainable Environmental Education. *Interact. Learn. Environ.* **2025**, *33*, 1945–1966. [CrossRef]
4. Hallinger, P.; Wang, R.; Chatpinyakoop, C.; Nguyen, V.T.; Nguyen, U.P. A Bibliometric Review of Research on Simulations and Serious Games Used in Educating for Sustainability, 1997–2019. *J. Clean. Prod.* **2020**, *256*, 120358. [CrossRef]
5. Reckien, D.; Eisenack, K. Climate Change Gaming on Board and Screen: A Review. *Simul. Gaming* **2013**, *44*, 253–271. [CrossRef]
6. Kim, J.M.; Hill, R.W.; Durlach, P.J.; Lane, H.C.; Forbell, E.; Core, M.; Marsella, S.; Pynadath, D.; Hart, J. *BiLAT: A Game-Based Environment for Practicing Negotiation in a Cultural Context*; IOS Press: Amsterdam, The Netherlands, 2009; Volume 19.

7. Douglas, B.D.; Brauer, M. Gamification to Prevent Climate Change: A Review of Games and Apps for Sustainability. *Curr. Opin. Psychol.* **2021**, *42*, 89–94. [CrossRef]
8. Duchatelet, D.; Jossberger, H.; Rausch, A. Assessment and Evaluation of Simulation-Based Learning in Higher Education and Professional Training: An Introduction. *Stud. Educ. Eval.* **2022**, *75*, 101210. [CrossRef]
9. Haneklaus, N.; Kaggwa, M.; Misihairabgwi, J.; Abu El-Magd, S.; Ahmadi, N.; Ait Brahim, J.; Amasi, A.; Balláné Kovács, A.; Bartela, Ł.; Bellefqih, H.; et al. The Phosphorus Negotiation Game (P-Game): First Evaluation of a Serious Game to Support Science-Policy Decision Making Played in More than 20 Countries Worldwide. *Discov. Sustain.* **2025**, *6*, 1. [CrossRef]
10. Haneklaus, N.H. Unconventional Uranium Resources from Phosphates. *Encycl. Nucl. Energy* **2021**, *2021*, 286–291. [CrossRef]
11. Haneklaus, N.; Sun, Y.; Bol, R.; Lottermoser, B.; Schnug, E. To Extract, or Not to Extract Uranium from Phosphate Rock, That Is the Question. *Environ. Sci. Technol.* **2017**, *51*, 753–754. [CrossRef]
12. Haneklaus, N.H.; Mwalongo, D.A.; Lisuma, J.B.; Amasi, A.I.; Mwimanzi, J.; Bituh, T.; Ćirić, J.; Nowak, J.; Ryszko, U.; Rusek, P.; et al. Rare Earth Elements and Uranium in Minjingu Phosphate Fertilizer Products: Plant Food for Thought. *Resour. Conserv. Recycl.* **2024**, *207*, 107694. [CrossRef]
13. Gyetvai, L.; Lovas, B.V.; Kiss, M.; Talas, M.; Halmosi, B.; Ara, J.; Sik-Lanyi, C.; Haneklaus, N.; Guzsvinecz, T.; Szűcs, J. Development of a Negotiation-Based Serious Game in Virtual Reality to Help Teach Responsible Consumption and Production. In Proceedings of the 2022 1st IEEE International Conference on Cognitive Aspects of Virtual Reality (CVR), Virtual, 11–12 May 2022; pp. 21–26.
14. Gratch, J.; DeVault, D.; Lucas, G. The Benefits of Virtual Humans for Teaching Negotiation. In *Proceedings of the Intelligent Virtual Agents*; Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 283–294.
15. Salvi, F.; Horta Ribeiro, M.; Gallotti, R.; West, R. On the Conversational Persuasiveness of GPT-4. *Nat. Hum. Behav.* **2025**, *9*, 1645–1653. [CrossRef] [PubMed]
16. Abrami, P.C.; Bernard, R.M.; Bures, E.M.; Borokhovski, E.; Tamim, R.M. Interaction in Distance Education and Online Learning: Using Evidence and Theory to Improve Practice. *J. Comput. High. Educ.* **2011**, *23*, 82–103. [CrossRef]
17. Parlamis, J.; Mitchell, L. Teaching Negotiations in the New Millennium: Evidence-Based Recommendations for Online Course Delivery. *Negot. J.* **2014**, *30*, 93–113. [CrossRef]
18. Thi Hue Dung, D. The Advantages and Disadvantages of Virtual Learning. *IOSR J. Res. Method Educ.* **2020**, *10*, 45–48. [CrossRef]
19. Means, B.; Toyama, Y.; Murphy, R.; Bakia, M.; Jones, K. *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies*; Center for Technology in Learning: Amstelveen, The Netherlands, 2010.
20. Geiger, I. From Letter to Twitter: A Systematic Review of Communication Media in Negotiation. *Group Decis. Negot.* **2020**, *29*, 207–250. [CrossRef]
21. Dwivedi, Y.; Hughes, D.L.; Coombs, C.; Constantiou, I.; Duan, Y.; Edwards, J.; Gupta, B.; Lal, B.; Misra, S.; Prashant, P.; et al. Impact of COVID-19 Pandemic on Information Management Research and Practice: Transforming Education, Work and Life. *Int. J. Inf. Manag.* **2020**, *55*, 102211. [CrossRef]
22. Oehlschläger, P.; Merz, M.A. Effectiveness of In-Person Versus Online Negotiation Teaching for Practitioners. *Negot. J.* **2023**. [CrossRef]
23. Daft, R.; Lengel, R. Information Richness: A New Approach to Managerial Behavior and Organization Design. *Res. Organ. Behav.* **1983**, *6*, 73.
24. Daft, R.; Lengel, R. Organizational Information Requirements, Media Richness and Structural Design. *Manag. Sci.* **1986**, *32*, 554–571. [CrossRef]
25. Sanford, D. Course Format and Learning: The Moderating Role of Overall Academic Performance. *Int. J. Manag. Educ.* **2017**, *15*, 490–500. [CrossRef]
26. Hillman, D.C.A.; Willis, D.J.; Gunawardena, C.N. Learner-interface Interaction in Distance Education: An Extension of Contemporary Models and Strategies for Practitioners. *Am. J. Distance Educ.* **1994**, *8*, 30–42. [CrossRef]
27. Eckstein, J.; Bartanen, M. British Parliamentary Debate and the Twenty-First-Century Student. *Commun. Stud.* **2015**, *66*, 458–473. [CrossRef]
28. Parrish, C.; Guffey, S.; Williams, D.; Estis, J.; Lewis, D. Fostering Cognitive Presence, Social Presence and Teaching Presence with Integrated Online—Team-Based Learning. *TechTrends* **2021**, *65*, 473–484. [CrossRef] [PubMed]
29. Galin, A.; Gross, M.; Gosalker, G. E-Negotiation versus Face-to-Face Negotiation What Has Changed—If Anything? *Comput. Hum. Behav.* **2007**, *23*, 787–797. [CrossRef]
30. Flutter Flutter Official Webpage. Available online: https://flutter.dev/ (accessed on 23 December 2024).
31. Dart Dart Official Webpage. Available online: https://dart.dev/ (accessed on 23 December 2024).
32. Firebase Firebase Official Webpage. Available online: https://firebase.google.com/ (accessed on 23 December 2024).

33. Lewis, J.R. The System Usability Scale: Past, Present, and Future. *Int. J. Hum. Comput. Interact.* **2018**, *34*, 577–590. [CrossRef]
34. Kirchner, J. *Data Analysis Toolkit #5: Uncertainty Analysis and Error Propagation*; University of California: Berkeley, CA, USA, 2001.
35. Hedges, L.V. Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *J. Educ. Stat.* **1981**, *6*, 107–128. [CrossRef]
36. Braun, V.; Clarke, V. Reflecting on Reflexive Thematic Analysis. *Qual. Res. Sport Exerc. Health* **2019**, *11*, 589–597. [CrossRef]
37. Braun, V.; Clarke, V. Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
38. Stokes, L.C.; Selin, N.E. The Mercury Game: Evaluating a Negotiation Simulation That Teaches Students about Science-Policy Interactions. *J. Environ. Stud. Sci.* **2016**, *6*, 597–605. [CrossRef]
39. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Routledge: New York, NY, USA, 1988.
40. Short, J.; Williams, E.; Christie, B. *The Social Psychology of Telecommunications*; Wiley: Hoboken, NJ, USA, 1976; ISBN 0-471-01581-4.
41. Garrison, D.R.; Anderson, T.; Archer, W. Critical Thinking, Cognitive Presence, and Computer Conferencing in Distance Education. *Am. J. Distance Educ.* **2001**, *15*, 7–23. [CrossRef]
42. Chan, D. So Why Ask Me? Are Self-Report Data Really That Bad? In *Statistical and Methodological Myths and Urban Legends*; Lance, C.E., Lance, C.E., Vandenberg, R.J., Eds.; Routledge: New York, NY, USA, 2009.
43. Gonyea, R.M. Self-Reported Data in Institutional Research: Review and Recommendations. *New Dir. Institutional Res.* **2005**, *2005*, 73–89. [CrossRef]
44. Dillman, D.A.; Smyth, J.D.; Christian, L.M. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*; Wiley: Hoboken, NJ, USA, 2014.
45. Larsen, M.; Rasinski, K. The Psychology of Survey Response. *J. Am. Stat. Assoc.* **2002**, *97*, 358–359. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Article

Rethinking Usability in Serious Games: Designing an Instrument That Evaluates What Really Matters in Learning Contexts

Mayra Carrión-Toro [1], David Morales-Martínez [1], Marco Santórum [1], Luis Bernal [1], Julian Galindo [1], Jose Aguilar [2] and Patricia Acosta-Vargas [3,*]

[1] Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito 170525, Ecuador; mayra.carrion@epn.edu.ec (M.C.-T.); david.morales03@epn.edu.ec (D.M.-M.); marco.santorum@epn.edu.ec (M.S.); luis.bernal01@epn.edu.ec (L.B.); julian.galindo@epn.edu.ec (J.G.)
[2] Centro de Microcomputación y Sistemas Distribuidos (CEMISID), Universidad de Los Andes, Merida 5101, Venezuela; aguilar@ula.ve
[3] Intelligent and Interactive Systems Laboratory, Universidad de Las Américas, Quito 170503, Ecuador
* Correspondence: patricia.acosta@udla.edu.ec

Abstract: Serious games are increasingly recognized as valuable tools in educational settings due to their capacity to enhance learning through engaging and interactive experiences. However, existing usability evaluation instruments often fail to account for the specific features and pedagogical goals inherent to serious games, thereby limiting their utility in iterative design improvement. This study presents a usability evaluation questionnaire SGU specifically designed for serious games, developed through a systematic literature review (SLR) following the methodology of Kitchenham and Charters. The resulting tool is a standardized questionnaire intended to comprehensively assess key usability dimensions, including gameplay mechanics, interface design, narrative engagement, and gamification elements. The questionnaire was applied in multiple serious game case studies developed within educational contexts, allowing for the identification of recurring usability issues and context-specific design challenges. The findings highlight the instrument's relevance, clarity, and diagnostic potential, demonstrating its practical value for evaluating and refining serious games. This research contributes to the field by addressing a critical gap in usability assessment and offering a structured, context-sensitive approach for improving user experience in educational game design.

Keywords: serious games; usability evaluation; human–computer interaction; user experience; educational games

1. Introduction

The integration of interactive technologies into education has significantly transformed how knowledge is accessed, experienced, and retained [1–3]. Among these technologies, serious games have emerged as powerful tools that combine entertainment and pedagogy, utilizing gameplay mechanics to support instructional objectives [3–5]. Their ability to increase learner motivation, foster active participation, and promote knowledge retention has led to their increasing adoption across diverse educational settings, from primary education to professional training contexts [6–9].

Despite their pedagogical potential, the success of serious games depends heavily on their usability, that is, how effectively, efficiently, and satisfactorily users can interact

with the system to achieve specific learning goals [10–12]. Traditional usability evaluation methods, originally designed for general-purpose software, often fall short when applied to serious games as they often overlook game-specific dimensions such as narrative coherence, gamification elements, player immersion, and learning flow [13,14]. Consequently, evaluations based on these generic instruments may fail to identify critical user experience issues that directly impact learning effectiveness.

Recent research highlights have underscored the lack of consensus regarding appropriate instruments to assess usability in serious games [15]. While some researchers adapt standardized questionnaires like the System Usability Scale (SUS) or Questionnaire for User Interaction Satisfaction (QUIS), others develop custom instruments tailored to specific projects or game genres. However, these approaches are rarely generalizable and often lack validation in a variety of educational contexts. Moreover, most existing tools do not comprehensively account for the pedagogical and interactive features that distinguish serious games from traditional software applications.

To address this gap, this study proposes a usability evaluation instrument SGU—Serious Games Usability—specifically designed to assess serious games. The questionnaire was developed based on a systematic literature review (SLR), following the guidelines of Kitchenham and Charters [16], with the goal of identifying and synthesizing existing methods, tools, and usability dimensions pertinent to serious games.

However, mainstream usability questionnaires such as SUS, CSUQ, or SUMI, while effective in assessing general-purpose software, show important limitations when applied to serious games. These tools primarily capture aspects such as interface efficiency, task completion, and user satisfaction but do not account for pedagogical alignment, narrative immersion, gamification mechanics, or immersive interfaces, dimensions that are central to the effectiveness of serious games in educational contexts. These shortcomings directly motivated the development of the present study, which introduces a domain-specific instrument capable of evaluating both functional and pedagogical aspects of serious games.

The proposed questionnaire was applied in multiple serious game case studies to assess its effectiveness in identifying usability issues and informing design improvements. Through this application, the study demonstrates the clarity, adaptability, and practical relevance of the instrument for both researchers and developers seeking to evaluate serious games within educational frameworks.

The main objective of this research is to provide a validated, structured, and context-sensitive usability questionnaire that fills a critical gap in the evaluation of serious games. The expected contributions include a theoretically grounded instrument, empirical evidence of its application, and a framework for future usability studies in game-based learning environments.

This paper is structured as follows: Section 2 reviews the background on serious games, the main characteristics of serious games, and usability concepts; Section 3 presents the methodology and protocol of the systematic literature review; Section 4 describes the design of the usability questionnaire; Section 5 reports its application in real-world case studies; Section 6 discusses the results, insights, and limitations; and Section 7 concludes with final reflections and suggestions for future research.

2. Background

The evaluation of the usability of serious games presents unique challenges due to the complexity and multifunctional nature of these applications. Unlike conventional software systems, serious games incorporate diverse components, such as narrative structures, gamification mechanics, and immersive interfaces that serve both entertainment and

instructional purposes [17,18]. However, current usability evaluation instruments often fail to fully capture these multidimensional aspects [19,20]. As a result, when tools that are not specifically designed for serious games are applied, important elements of the user experience may be overlooked, leading to incomplete or misleading usability assessments. This section provides a comprehensive overview of the theoretical and empirical foundations related to usability, laying the groundwork for the development of a tailored questionnaire-based evaluation instrument.

2.1. Video Games and Serious Games

A video game is a form of interactive software primarily designed for entertainment purposes, utilizing computing technology to enable real-time human–machine interaction. Gameplay can occur on a wide range of electronic platforms, including televisions, personal computers, consoles, and other similar devices [21]. According to Lewis, a video game can be defined as "a computer-based environment that displays a game on a screen whose rules have been pre-programmed" [22]. Video games typically consist of four fundamental components: aesthetics, mechanics, narrative, and technology [23].

- Aesthetics: This is one of the most crucial layers of a video game as it directly interfaces with the player. It shapes the sensory experience of the game, encompassing what the player sees and hears, and significantly influences immersion and engagement.
- Mechanics: These represent the formal rules of the game. They define what actions the player can perform and establish the objectives to be achieved within the game environment.
- Narrative: The storyline or sequence of events that unfolds throughout the game. Provides context and depth, often giving players the sense of agency through interactive decision-making.
- Technology: Refers to the tools, platforms, and technical infrastructure used to develop and execute the game. Virtually any computing technology may serve as a foundation for video game development.

These fundamental elements, Figure 1, are adapted and expanded in serious games, which apply similar interactive structures to serve educational or training purposes beyond pure entertainment. Serious games are broadly defined as digital games developed for purposes that go beyond entertainment [24]. One of the earliest formal definitions was proposed by American researcher Clark C. in 1970, who described serious games as those designed with explicit educational objectives rather than solely for user enjoyment [25]. Later, in 2008, Julián Álvarez expanded this definition by emphasizing that serious games are digital applications that integrate instructional elements, such as tutoring, teaching, training, communication, and information, with playful and technological components drawn from traditional video games [26]. The purpose of this hybridization is to transform practical or useful content into a fun and engaging experience for the user. When serious games are developed for educational communication, they tend to exhibit a set of distinct features that differentiate them from purely recreational games.

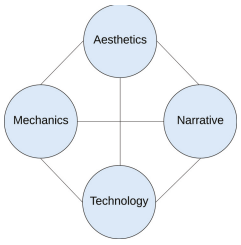


Figure 1. Elements of a video game.

2.2. Characteristics of Serious Games

To develop a robust and context-aware usability evaluation instrument tailored to serious games, it was first essential to identify and define the core criteria that characterize them. This process followed an integrative approach that combined theoretical analysis, empirical evidence, and expert validation [6,27–29].

The identification of criteria began with a comprehensive review of the academic literature on serious game design, usability in educational software, and gamification strategies. Key references included conceptual frameworks, methodological guidelines, and applied case studies from the domains of educational technology and game-based learning [29]. In parallel, these findings were contrasted with principles from established software development methodologies to ensure alignment with usability evaluation practices. Furthermore, the research process included reflective analysis and expert input, supported by the direct examination of various educational serious games.

Through this triangulated approach, the researchers identified a set of recurring and meaningful characteristics considered fundamental for the design and evaluation of serious games, some of which are necessary, while others are sufficient to classify a video game as a serious game [29]. These characteristics were subsequently organized based on their functional relevance and logical necessity within the serious game framework. The following attributes emerged as consistently essential or supportive components in the effective design and assessment of serious games.

1. **Serious Objectives:** This is the defining feature of a serious game. Content must be intentionally designed to support educational or training goals, ensuring that the user engages in meaningful interactions that lead to learning outcomes. Without this component, the game cannot be classified as "serious".
2. **Storytelling:** A well-integrated storyline enhances comprehension, contextualizes tasks, and provides motivation by immersing the player in a progressive virtual world. Required in all serious games, narrative elements often strengthen cognitive and emotional engagement.
3. **Gamification techniques:** Gamification involves applying game-like mechanics to reinforce engagement and motivation [30]. In serious games, these techniques serve to structure progression and reward learning behavior. Common techniques include
    - Points: Numerical values assigned to quantify performance and progress.
    - Levels: Unlockable stages that become accessible as users meet specific objectives.
    - Leaderboards: Ranking systems to introduce competition and benchmarking.
    - Badges: Visual rewards given for completing specific tasks or achieving milestones, serving as indicators of progress and competence.
    - Challenges/Missions: Clearly defined tasks or objectives within the game that guide user behavior and provide structured progression.
    - Onboarding: The process through which a game intuitively teaches players its rules and goals without requiring textual instruction.
    - Customization: Allows players to personalize elements of the game world, such as avatars, names, or environments, enhancing their emotional connection.
4. **Gameplay:** Rules and interactions that define how the player engages with the game. Well-designed mechanics are fundamental to promoting active learning and maintaining user motivation.
5. **Multimedia:** The use of combined media—images, audio, text, video, and animation—enriches the delivery of information and increases the multisensory appeal of the

game. While not mandatory, multimedia elements enhance clarity, engagement, and cognitive retention.

6. **Interfaces:** This includes both hardware and software components that facilitate interaction with the game. Intuitive, accessible, and responsive interfaces significantly impact usability, though some minimalistic games may function without advanced UI structures.
7. **World visualization:** This refers to the spatial and interactive environments that support user exploration. Although not always necessary, offering freedom of movement and environmental feedback can increase immersion and facilitate situated learning.
8. **Character design:** The inclusion of identifiable or controllable characters supports personalization, role-play, and user agency. This feature enhances engagement but may not be essential in all serious game formats, such as abstract simulations.

These core characteristics, when appropriately balanced, allow serious games to fulfill both their pedagogical and experiential objectives. While these traits may vary in form depending on the context and design goals, they collectively define the architecture of an effective serious game.

Among the diverse applications of serious games, educational video games represent one of the most prominent and impactful subtypes [31,32]. These interactive digital environments are intentionally designed to support the acquisition of academic or procedural knowledge through structured activities embedded in game-based contexts [33]. Although serious games are also used in fields such as health, defense, and social awareness, educational games focus specifically on cognitive development, skill acquisition, and knowledge reinforcement in both formal and informal learning settings [34–38]. This perspective is supported by Gros, who defines educational games as software designed to facilitate learning by combining didactic objectives with interactive, game-based mechanics [39].

Given their pedagogical focus and widespread adoption, educational video games offer a practical and representative domain for applying usability evaluation instruments. For this reason, they were selected as case studies in the present research to validate the proposed usability questionnaire.

2.3. Usability in Software and Educational Contexts

Usability is defined by the ISO 9241-11 standard [40] as:

> The extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use [40].

This definition emphasizes the importance of designing products that help users achieve their goals while providing a reliable basis for measuring and comparing the usability of different software systems. In practical terms, usability refers to the ease with which a user can perform a given task using a system or product.

According to ISO 9241-11, the three core quality components of usability are

- **Effectiveness:** The degree to which users are able to successfully complete tasks and achieve intended outcomes. A product that enables users to consistently meet their objectives is considered effective.
- **Efficiency:** The amount of effort required by users to achieve their objectives. An efficient product allows users to complete tasks with minimal effort and resource expenditure.
- **Satisfaction:** The user's subjective response to the product. Satisfaction encompasses the emotional and psychological aspects of the user experience.

Evaluating the usability of software products offers a wide range of benefits [41], including

1. Reduced learning costs and a corresponding decrease in the need for user support and assistance.
2. Optimization of design, redesign, and maintenance processes.
3. Increased conversion rates from casual users to active users or clients.
4. Enhanced reputation and credibility of the product or brand.
5. Improvement in overall quality and performance.
6. Better user well-being by reducing stress, increasing satisfaction, and increasing productivity.

In educational contexts, usability not only facilitates ease of use but also directly influences learners' motivation, comprehension, and retention. Unlike traditional productivity software, educational systems must balance functional usability with pedagogical effectiveness. Therefore, accurate and context-aware usability evaluation becomes essential when designing or adapting serious games for effective learning.

3. Systematic Literature Review

This research adopts a systematic literature review (SLR) methodology to comprehensively explore the current landscape of usability evaluation in serious games. The goal is to identify, analyze, and synthesize existing methods, techniques, and instruments that assess usability in this context, ultimately supporting the development of a dedicated evaluation tool tailored to the unique characteristics of serious games.

The review was conducted following the well-established guidelines proposed by Kitchenham and Charters [16]. This structured approach ensures the reliability and completeness of the evidence collected and strengthens the methodological foundation of the proposed instrument; see Figure 2.

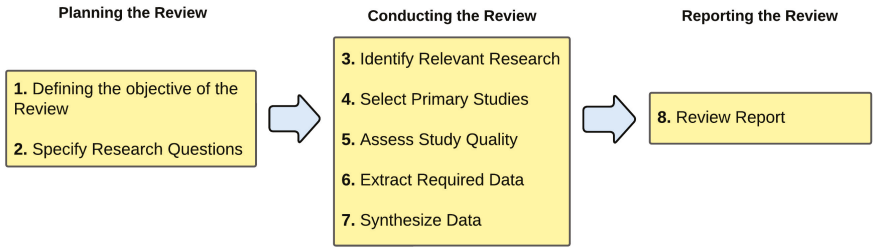


Figure 2. The systematic literature review steps.

The protocol for the systematic literature review (SLR) was structured into three main phases: (1) planning the review, (2) conducting the review, and (3) reporting the review. Each phase was carefully designed to ensure that only high-quality and relevant studies were included, thereby contributing valuable insights into usability evaluation practices in serious game environments.

3.1. Planning the Review

Usability is a critical factor in the effectiveness of serious games, especially in educational contexts where user engagement, cognitive retention, and task performance must converge. Despite its importance, the instruments currently employed to evaluate usability in serious games are often inadequate. Many of these tools originate from general software or web usability contexts and are not designed to address the distinctive features of serious games, such as interactive gameplay, narrative immersion, and pedagogical alignment.

To guide the scope and focus of the review, the following research questions (RQs) were formulated; see Table 1. These questions are intended to support the identification,

classification, and analysis of existing user-centered evaluation approaches employed in the usability assessment of both general and serious games:

Table 1. Research questions.

No.	Research Question	Motivation
RQ1	What user-centered methods, techniques, and instruments are used to evaluate the usability of games and serious games?	To identify the current landscape of usability evaluation approaches in gaming environments, including those adapted or developed specifically for serious games.
RQ2	What are the key characteristics of the user-centered methods, techniques, and instruments identified in RQ1?	To analyze the scope, advantages, and limitations of existing evaluation tools in terms of the usability dimensions they assess.
RQ3	How are the identified user-centered methods, techniques, and instruments applied in the usability evaluation of games and serious games?	To examine real-world implementations, evaluation procedures, and application contexts in order to extract practical insights that can guide the design of a new usability questionnaire.

3.2. Conducting the Review

The search was conducted across five major digital databases to ensure a complete coverage of the high-quality academic literature:

- IEEE Xplore [42];
- ACM Digital Library [43];
- Elsevier (ScienceDirect) [44];
- Google Scholar [45];
- Springer Link [46].

The primary keywords selected for the search process were “methods”, “techniques”, “instruments”, “evaluation”, “usability”, “serious games”, and “games”. These terms were derived directly from the research questions and refined through synonym analysis and semantic expansion to capture all relevant variations across sources.

The search strings were constructed following the steps recommended in the SLR guidelines, which include

- Identifying the key concepts of the research questions.
- Determining alternative spellings and synonyms for each concept.
- Verifying and refining keywords based on relevant preliminary studies.
- Using the Boolean operator “OR” to incorporate synonymous or related terms.
- Using the Boolean operator “AND” to combine the main concepts into structured queries.

To organize and facilitate the construction of search strings, a table was developed to classify the keywords into four thematic groups; see Table 2. Column “A” included terms related to evaluation along with their Spanish equivalents; Column “B” focused on terms associated with serious games and their translations; Column “C” contained keywords pertaining to usability; and Column “D” comprised terms related to methods, techniques, and instruments relevant to usability evaluation.

This structured classification helped ensure consistency and completeness in the formulation of search queries.

Table 2. Search string keywords.

A	B	C	D
A1. Evaluación	B1. Juegos Serios	C1. Usabilidad	D1. Métodos
A2. Validación	B2. Juegos	C2. Usability	D2. Técnicas
A3. Evaluation	B3. Videojuegos Educativos		D3. Instrumentos
A4. Validation	B4. Juegos educativos		D4. Methods
A5. Assessment	B5. Juegos de aprendizaje		D5. Techniques
	B6. Serious games		D6. Instruments
	B7. Games		
	B8. Educational video games		
	B9. Educational games		
	B10. Learning games		

Using this structure, the following search strings (SS) were defined:

- **SS1:** (A1 OR A2) AND (C1) AND (B1 OR B2 OR B3 OR B4 OR B5)
- **SS2:** (A3 OR A4 OR A5) AND (C2) AND (B6 OR B7 OR B8 OR B9 OR B10)
- **SS3:** (D1 OR D2 OR D3) AND (A1 OR A2) AND (C1) AND (B1 OR B2 OR B3 OR B4 OR B5)
- **SS4:** (D4 OR D5 OR D6) AND (A3 OR A4 OR A5) AND (C2) AND (B6 OR B7 OR B8 OR B9 OR B10)

The search strings were iteratively refined through pilot tests conducted in Scopus and Web of Science, aiming to balance sensitivity (maximizing the retrieval of relevant publications) and specificity (minimizing irrelevant results). During this process, terms that produced excessive noise, such as "game design" or "playability," were excluded after preliminary testing due to their low precision in retrieving relevant studies.

All retrieved articles were cataloged and organized using SR-Accelerator, an online tool that facilitated the management of metadata, inclusion/exclusion tracking, and classification [47]. Two independent reviewers conducted the screening process. Discrepancies were resolved through discussion, and in cases of persistent disagreement, a third reviewer acted as arbiter. In addition, Mendeley was employed as a reference management tool for proper citation and bibliography control throughout the research process.

The selection process for relevant studies was carried out to ensure methodological rigor and alignment with research objectives. First, a preliminary screening of all documents retrieved during the search process. In addition, all records were evaluated according to predefined inclusion and exclusion criteria (see Table 3).

In particular, one of the main exclusion criteria was applied to studies describing serious games without explicit usability evaluation data. This led to the omission of high-impact projects such as "Galaxy Zoo" [48], "Sea Hero Quest" [49], and "Borderlands Science" [50]. Despite their wide reach and scientific relevance, the lack of usability reporting made them unsuitable for this analysis, underscoring a broader gap in the literature and reinforcing the need for standardized instruments like the one proposed in this study.

Then, an in-depth analysis was conducted on the studies that had been previously shortlisted. This phase involved a thorough reading of the full text of each document to accurately assess its content and relevance. Following this, each study was assessed using a set of quality evaluation criteria, enabling us to verify its methodological rigor and its contribution to the objectives of this research.

To evaluate the methodological quality and relevance of the documents identified during the selection process, a quality assessment questionnaire was developed and applied to each study that was successfully retrieved. This questionnaire was designed to assess

how well each document aligned with the primary objectives and thematic scope of the systematic literature review.

Table 3. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Publications that propose or apply instruments, techniques, or methods to evaluate the usability of games or serious games.	Publications that report usability evaluation results but omit details of the techniques or methods used.
Studies that include case studies demonstrating the application of usability evaluation methods, techniques, or instruments in games or serious games.	Multiple reports of the same study (in such cases, the most comprehensive version was included in the review).
Peer-reviewed scientific articles written in English or Spanish.	Duplicate documents (i.e., the same study retrieved from multiple databases).
Articles that describe the process of usability evaluation applied to a game or serious game.	Articles that present a serious game but provide no information on its usability evaluation.
Engineering, master's, or doctoral theses.	Studies that focus on evaluation methods outside the scope of serious games.
	Publications for which only the abstract is available and not the full text.
	Articles whose primary objective is unrelated to usability evaluation methods in games or serious games.

The instrument consisted of seven quality evaluation questions (QEs). These questions helped determine the extent to which each study contributed to understanding usability evaluation in games and serious games. Each question addressed a specific aspect of methodological or conceptual relevance. The set of quality evaluation questions is presented in Table 4.

To assess the quality of the selected studies, a binary scoring system was applied to the quality assessment questions. For each question, a study received a score of 1 if it met the corresponding criterion and a score of 0 if it did not. This approach ensures consistency and objectivity in evaluating the methodological contribution of each document.

At the end of the assessment, only the studies that achieved a minimum score of 5 out of 7 possible points were selected for inclusion in the final analysis, ensuring that the reviewed literature contributed meaningfully to the construction of the proposed usability evaluation instrument for serious games.

Table 4. Quality Evaluation Questions.

No.	Quality Evaluation Question	Answer
QE1	Does the study define what a game or serious game is?	(+1) Yes/(+0) No
QE2	Does the study provide information about usability evaluation in games or serious games?	(+1) Yes/(+0) No
QE3	Does the study present any usability evaluation instrument, technique, or method for games or serious games?	(+1) Yes/(+0) No
QE4	Does the study describe details about the instrument, technique, or method used for evaluation?	(+1) Yes/(+0) No

Table 4. *Cont.*

No.	Quality Evaluation Question	Answer
QE5	Does the study specify the characteristics of the instrument, technique, or method of usability evaluation?	(+1) Yes/(+0) No
QE6	Does the study explain how the instrument, technique, or method is applied?	(+1) Yes/(+0) No
QE7	Does the study include a case study involving the application of a usability evaluation instrument, technique, or method?	(+1) Yes/(+0) No

To ensure methodological rigor and minimize selection bias, a structured screening and selection process was conducted in alignment with best practices for systematic literature reviews. The complete process is summarized in the flow diagram shown in Figure 3, which illustrates the identification, screening, and inclusion phases.

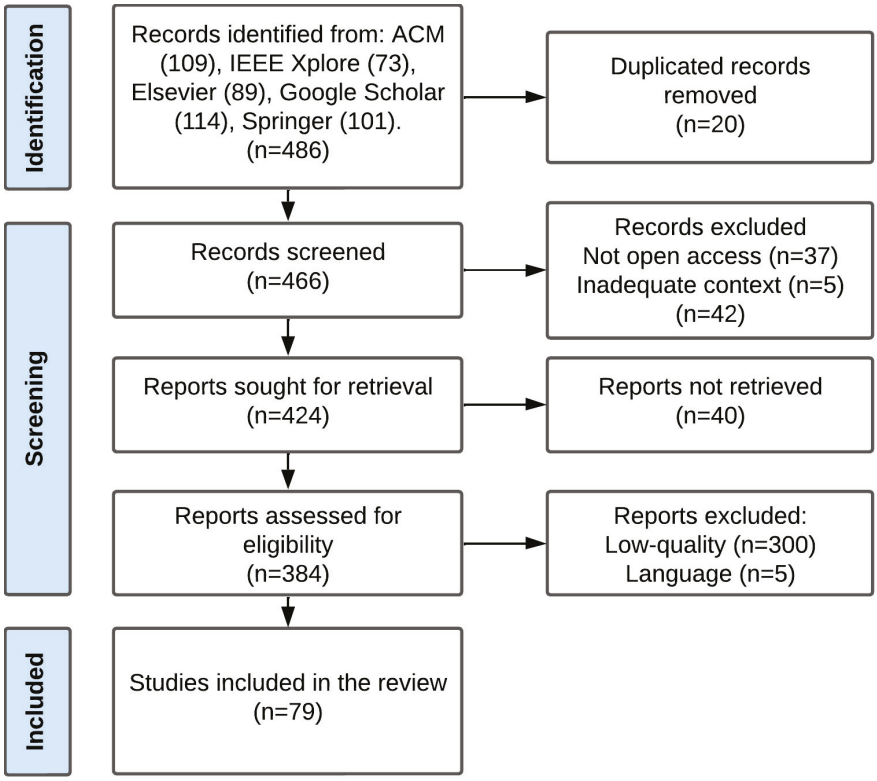


Figure 3. PRISMA flow diagram for the literature review.

Phase 1: Identification. A comprehensive search strategy was executed across five major scientific databases: ACM Digital Library (109), IEEE Xplore (73), Elsevier ScienceDirect (89), Google Scholar (114), and SpringerLink (101), yielding a total of 486 records. After applying an automated and manual duplicate removal process, 20 redundant entries were excluded, resulting in 466 unique records retained for the initial screening phase. This stage ensured a wide and representative coverage of the literature related to usability evaluation in games and serious games.

Phase 2: Screening. The 466 records were systematically reviewed based on titles, abstracts, and keywords to determine their relevance to the research scope. During this step, 42 articles were excluded, 37 due to restricted access (non-open access) and 5 for lacking contextual alignment with the focus of usability evaluation. Consequently, 424 studies were considered eligible for full-text retrieval. However, 40 of these could not be accessed

in full, reducing the articles to 384 full-text articles. These were then assessed in detail using the predefined inclusion and exclusion criteria (see Table 3).

Phase 3: Inclusion. Each of the 384 full-text articles underwent a rigorous eligibility and quality assessment process. Based on the quality evaluation criteria (see Table 4), 300 studies were excluded due to insufficient methodological quality, and 5 were excluded due to language incompatibility (i.e., not available in English or Spanish). As a result, a total of 79 high-quality studies were selected for inclusion in the final review synthesis. These papers constitute the core evidence base from which the findings of this systematic review were derived.

The complete list of the 79 included publications is presented in Table 5. The list includes the assigned paper ID (generated using the first two letters of the author's last name, the first letter of their first name, and year of publication), the article's title, the first author, and the bibliographic reference.

Table 5. List of scientific publications included in this review.

No.	Paper ID	Title	Reference
1	ToG11	Usability of Serious Games for Health	Tolentino, G. [51]
2	MaN10	Usability Testing for Educational Computer Game Using Observation Method	Diah, M. [52]
3	FaI19	Usability and Cognitive Benefits of a Serious Game to Combat Aedes Aegypti Mosquito	Falcao, I. [53]
4	HaO08	Playability Heuristics Evaluation (PHE) approach for Malaysian educational games	Hasiah, O. [54]
5	HuA14	A Usability testing on jFakih Learning Games for Hearing Impairment Children	Hussain, A. [55]
6	MaI11	Measuring Usability of Educational Computer Games Based on the User Success Rate	Ismail, M. [56]
7	DaR11	Teacher tool for Visualization and Management of a Technology-Enhanced Learning Environment	Dagdag, R. [57]
8	DeH13	Methods for Game User Research: Studying Player Behavior to Enhance Game Design	Desurvire, H. [58]
9	PaS09	Usability Assessment Heuristics in New Genre Videogames	Papaloukas, S. [59]
10	HoJ03	Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children	Höysniemi, J. [60]
11	RoU19	Usability Evaluation Using GOMS Model for Education Game "Play and Learn English"	Rosyidah, U. [61]
12	SoS14	Evaluation of Mobile Games with Playability Heuristic Evaluation System	Soomro, S. [62]
13	MaJ19	YouTube Gamers and Think-Aloud Protocols: Introducing Usability Testing	May, J. [63]
14	LiY12	Heuristic Evaluation of Digital Game Based Learning: A Case Study	Liao, Y. [64]
15	JeK08	Investigating the Applicability of Usability and Playability Heuristics for Evaluation of Pervasive Games	Jegers, K. [65]
16	HoG16	Comparing Usability and Engagement Between a Serious Game and a Traditional Online Program	Hookham, G. [66]
17	BaE06	Assessing the Applicability of the Structured Expert Evaluation Method (SEEM) for a Wider age group	Baauw, E. [67]
18	SeY09	Usability Evaluation for History Educational Games	Seng, Y. [68]
19	PaS08	Usability and Education of Games Through Combined Assessment Methods	Papaloukas, S. [69]
20	PiD08	Heuristic Evaluation for Games: Usability Principles for Video Game Design	Pinelle, D. [70]
21	RoF13	Heuristics for Video Games Evaluation: How Players Rate Their Relevance for Different Game Genres	Rodio, F. [71]
22	PiD09	Usability Heuristics for Networked Multiplayer Games	Pinelle, D. [72]
23	DeH04	Using Heuristics to Evaluate the Playability of Games	Desurvire, H. [73]
24	KoH06	Playability heuristics for mobile games	Korhonen, H. [74]
25	FaF18	Evaluating a Pervasive Game for Urban Awareness	Fabiano, F. [75]
26	KaA06	Evaluating the LEGO-RoboLab Interface with Experts	Karoulis, A. [76]
27	MoF16	Mobile Games Usability Using the Idea of Mark Overmars	Mohd, F. [77]
28	VeM18	Virtual gaming simulation of a mental health assessment: A usability study	Verkuyl, M. [78]
29	JoH16	Teaching clinical reasoning and decision-making skills to nursing students	Johnsen, H. [79]
30	SaF18	Design and implementation of a serious game on neurorehabilitation	Savazzi, F. [80]
31	SaF18b	Engaged in learning neurorehabilitation: Development validation of a serious game with user-centered design	Savazzi, F. [81]
32	BaJ12	Evaluation of simulation games for teaching engineering and manufacturing	Baalsrud, J. [82]
33	SeN19	A Review of Heuristics Evaluation Component for Mobile Educational Games	Senap, N. [83]
34	MaS19	Evaluation of user satisfaction and usability of a mobile app for smoking cessation	Malwade, S. [84]
35	AmB20	Desarrollo y evaluación de usabilidad de un juego "serio" utilizado en la enseñanza del comportamiento	Amaro, B. [85]
36	ZuG99	Usabilidad de juegos educativos	Zurita, G. [86]
37	SaY15	Metodología de Evaluación de Videojuegos Serios en Competencias Ciudadanas	Sanchez, Y. [87]
38	GoN16	Heuristic Evaluation for Serious Immersive Games and M-instruction	Gordon, N. [88]
39	OlA20	Heuristic Usability Evaluation Applied to Educational Games	Olimpio, A. [89]
40	ViE19	Heuristic Evaluation on Usability of Educational Games: A Systematic Review	Vieira, E. [90]
41	IsF14	Evaluation of the Usability of a Serious Game Aiming to Teach Facial Expressions to Schizophrenic Patients	Isleyen, F. [91]
42	PaS19	Usability Evaluation of a Serious Game Intervention on Japanese School Bullying	Paracha, S. [92]
43	WaJ16	A usability assessment for a career planning educational video game	Wang, J. [93]
44	TrG12	Lessons from an Educational Game Usability Evaluation	Trindade, G. [94]
45	MoH10	Development and potential analysis of Heuristic Evaluation for educational computer game (PHEG)	Mohamed, H. [95]
46	BaM15	HEEG: Heuristic Evaluation for Educational Games	Barbosa, M. [96]
47	GoR19	Evaluación Heurística de la usabilidad en contextos específicos	Gomez, R. [97]
48	ChJ17	Evaluación de usabilidad de juegos de video para casinos en plataformas móviles	Chacon, J. [98]
49	CaG20	Métodos de evaluación de la usabilidad en aplicaciones móviles	Castillo, G. [99]
50	ChG20	Criterios de usabilidad para el diseño e implementación de videojuegos	Chanchi, G. [100]
51	ThG19	Evaluation of Usability and Gameplay of Games on Mobile Platforms for Young People on Oncological Treatment	Theisen, G. [101]
52	ArM15	Técnicas de evaluación para videojuegos geolocalizados	Armas, M. [102]
53	BaC19	Heurísticas de Usabilidad para Mecánicas de Videojuegos	Barboza, C. [103]
54	InA09	Evaluación de Usabilidad de un Ambiente Virtual en 3D	Iniguez, A. [104]
55	Sif16	Usability evaluation methods for gesture-based games: A systematic review	Simor, F. [105]
56	LaS05	Better games through usability evaluation and testing	Laitinen, S. [106]
57	BrM10	Usability evaluation for health video games: A library of inspection heuristics	Britain, M. [107]

Table 5. *Cont.*

No.	Paper ID	Title	Reference
58	ChN15	Usability evaluation of digital Malaysian traditional games	ChePa, N. [108]
59	YaT11	A Usability Analysis of Video Games: The Development of Assessment Standards	Takeisha, T. [109]
60	MiC07	Usabilidad en un juego de memorama con reconocimiento de voz para niños	Miranda, C. [110]
61	XeM09	Games Usability and Learning–The Civilization IV Paradigm	Xenos, M. [111]
62	HoG16	A concurrent Think Aloud study of engagement and usability in a serious game	Hookham, G. [112]
63	GrT18	A Usability Review of the Learning Master Serious Game in Support of the US Army Jumpmaster's Course	Griffith, T. [113]
64	ArM16	Usability Testing for Culturally-Enhanced Serious Game	Ariffin, M. [114]
65	YaR17	Academic methods for usability evaluation of serious games: a systematic review	Yanez, R. [115]
66	JeN14	Serious games and heuristic evaluation- The cross-comparison of existing heuristic evaluation methods for games	Jerzak, N. [116]
67	EbJ17	Usability evaluation methods for mobile serious games applied to health: a systematic review	Ebertz, J. [117]
68	XiG19	A systematic literature review of usability evaluation guidelines on mobile educational games for school students	Xiao, G. [118]
69	AmB19	A usability analysis of a serious game for teaching stock market concepts in secondary schools	Amaro, B. [119]
70	SaM19	Game design frameworks and evaluating techniques for educational games: A review	Sasupilli, M. [120]
71	MoH11	Usability of educational computer game (Usa-ECG): applying analytic hierarchy process	Mohamed, H. [121]
72	BeM08	A comparison of two analytical evaluation methods for educational computer games for young children	Bekker, M. [122]
73	SaY20	Usability investigation of an educational mobile application for individuals with intellectual disabilities	Sabina, Y. [123]
74	SaA13	Usability evaluation of an application designed for the older adults	Satoshi, A. [124]
75	YaR19	Heuristic usability evaluation on games: a modular approach	Yanez, R. [125]
76	DaC17	Improving the usability in a video game through continuous usability evaluations	Daly, C. [126]
77	DoA02	A comparison of Think-Aloud, questionnaires and interviews for testing usability with children	Donker, A. [127]
78	BoM16	Usability, accessibility and gameplay heuristics to evaluate audiogames for users who are blind	Borba, M. [128]
79	SoS07	A new framework of usability evaluation for massively multi-player online game: Case study of "WOW" game	Song, S. [129]

3.3. Reporting of the Review

The analysis presented here emphasizes not only what methods, techniques, and instruments have been used in prior research but also how they have been applied and what limitations they present in the context of game-based learning environments.

RQ1: What user-centered methods, techniques, and instruments are used to evaluate the usability of games and serious games? Based on the studies reviewed in the SLR, it was observed that most of the methods used in the context of serious games are adaptations or direct applications of approaches originally developed for general software or entertainment games. However, while the overall structure of evaluation strategies remains similar, their application in serious games often demands additional pedagogical considerations.

Two primary user-centered methods were consistently identified in the reviewed literature:

- Empirical Methods: These involve observing users as they interact with the game to collect behavioral data and detect potential usability issues.
- Inquiry Methods: These rely on obtaining user feedback through instruments such as questionnaires, interviews, or focus groups, capturing users' subjective experiences and perceptions.

These methods provide complementary perspectives: empirical approaches capture real-time interaction patterns and behaviors, while inquiry approaches gather subjective insights on usability, satisfaction, and experience.

Each method is associated with specific techniques that facilitate the usability assessment process, as summarized in Table 6. Most of these techniques are applicable in both entertainment and serious game contexts, although their focus may differ depending on the educational objectives of the game.

Table 6. Primary studies on user-centered methods and techniques for usability evaluation.

Method	Techniques	Primary Studies
Empirical	Think-Aloud Protocol	[57–59,63,66,75,78,91,93,98,110,111,115,117,127]
	Direct Observation	[52,55,56,79,82,86,87,92,94,98,101,105,109,114,117,123]
	Peer Tutoring	[60]
	Eye Tracking	[98]
	Co-Discovery Learning	[69,111]
	Rapid Iterative Testing and Evaluation (RITE)	[58]
Inquiry	Questionnaires	[14,51,53,55,59,65,66,68,75,77–79,81,82,84,93,94,98,101,102,104–106,109,111–115,117–119,124,125,127,129]
	Interviews	[55,59,68,78,79,86,94,101,105,119,127]
	Focus Groups	[65,92,102]

Among the identified techniques, questionnaires are the only ones that consistently use standardized instruments for usability assessment. Other techniques, such as observation or interviews, typically rely on qualitative protocols rather than validated measurement tools.

As shown in Table 7, the SLR identified several widely used questionnaire-based instruments. These are frequently adapted to suit the context of serious games but often lack comprehensive coverage of all the features specific to these environments (narrative, gameplay dynamics, gamification, etc).

Table 7. Primary studies employing user-centered questionnaire instruments for usability evaluation.

Instrument	Primary Studies
SUS (System Usability Scale)	[14,51,53,66,75,80,93,101,104,112,113,125,129]
NASA-TLX (NASA-Task Load Index)	[125]
GEQ (Game Experience Questionnaire)	[102]
SUMI (Software Usability Measurement Inventory)	[14,55]
Ad Hoc Instruments	[80]
PSSUQ (Post-Study System Usability Questionnaire)	[79]
QUIS (Questionnaire for User Interface Satisfaction)	[14]
CSUQ (Computer System Usability Questionnaire)	[130]

RQ2: What are the key characteristics of the user-centered methods, techniques, and instruments identified in RQ1?

This research question aims to describe the core attributes of the user-centered methods, techniques, and instruments identified in the previous stage. Specifically, it seeks to clarify how these approaches are applied and what kind of insights they are capable of generating in the context of usability evaluation for games and serious games.

The literature review confirms that two primary user-centered methods are widely used in this domain:

- Empirical methods are based on observing behavior and performance of users as they interact with the game. These methods rely heavily on user participation to generate real-time usability data. Evaluators typically include a moderator and observers who guide and monitor test sessions. Participants are asked to complete specific tasks while their actions and reactions are documented.
- Inquiry methods focus on gathering data from users through verbal or written responses. These methods involve structured interaction between the evaluator and the

users, either individually or in groups, to understand their experiences, perceptions, and satisfaction with the system.

Table 8 summarizes the main techniques associated with each method:

Table 8. User-centered evaluation techniques for serious games.

Method	Technique	Description
Empirical	Think Aloud	Users verbalize their thoughts, emotions, and reasoning while interacting with the system.
	Observation	Evaluators record user behavior to detect usability issues. Provides both qualitative and quantitative insights.
	Peer Tutoring	Common in child-centered evaluations; one child teaches another to use the product, revealing intuitive understandings.
	Eye Tracking	Measures eye movement and pupil responses to evaluate interface design and visual attention.
	Co-Discovery	Two users complete tasks collaboratively, with the interaction offering deeper behavioral insights than individual sessions.
	RITE	Rapid Iterative Testing and Evaluation combines think-aloud sessions with immediate design improvements followed by continued testing.
Inquiry	Questionnaires	Structured instruments used to gather user feedback after game interaction.
	Interviews	One-on-one discussions to collect deeper insights into user satisfaction and issues.
	Focus Groups	Group discussions moderated by evaluators to explore shared experiences and perceptions.

Among all techniques, only questionnaires are supported by standardized usability instruments; see Table 9. In contrast, techniques like interviews or focus groups are often semi-structured or ad hoc and lack formal scoring schemes. Below is a summary of the most commonly used questionnaires identified through the SLR:

Table 9. Characteristics of usability questionnaires for games and serious games.

Instrument	Description
SUS	Includes 10 items rated on a 5-point Likert scale. It provides a quick and general usability score and has become a standard across software domains.
NASA-TLX	Measures perceived workload based on six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration.
GEQ	Assesses player experience across modules: core gameplay, social presence, and post-game reflection. Metrics include immersion, affect, flow, and challenge.
SUMI	Consists of 50 items categorized into subscales such as efficiency, control, affect, friendliness, and learnability.
Ad Hoc	Non-standardized instruments designed to collect preliminary data or context-specific feedback, often before user interaction with the software.
PSSUQ	Aligned with ISO 9241-11, includes 16 items and measures system usefulness, information quality, and interface quality.
QUIS	Assesses subjective satisfaction with specific interface components.
CSUQ	Similar to the PSSUQ, it consists of 19 questions and uses a 7-point Likert scale to assess satisfaction after system use.

RQ3. How are the identified user-centered methods, techniques, and instruments applied in the usability evaluation of games and serious games?

This research question aims to describe the practical application of usability evaluation strategies in user-centered studies of games and serious games. It synthesizes how empirical and inquiry-based methods are operationalized, how techniques are implemented during usability assessments, and how standardized instruments are integrated into the evaluation process.

- Empirical methods involve direct observation of users performing predefined tasks within the game environment. These methods focus on capturing behavioral data, interaction patterns, and real-time problem-solving processes. The presence of an evaluator (moderator) who guides the session and observes participants is common.
- Inquiry methods focus on gathering data from users through verbal or written responses. These methods involve structured interaction between the evaluator and the users, either individually or in groups, to understand their experiences, perceptions, and satisfaction with the system. The quality of the data collected depends on clear definitions of the evaluation objectives and usability indicators.

The following table (Table 10) summarizes how each empirical technique is applied in practice:

Table 10. Application of usability techniques for games and serious games.

Technique	Process
Think Aloud	Five participants are asked to perform tasks in the game while verbalizing their thoughts and decision-making processes, then evaluators record and analyze user feedback.
Observation	Five users with varying demographics complete a task list while being recorded. The evaluator documents user behavior and completes a follow-up questionnaire.
Peer Tutoring	Groups of 6–10 children are observed while one child teaches the game to another, then observers analyze interactions to infer usability insights.
Eye Tracking	Specialized equipment captures eye movement as five users perform tasks. Data is analyzed to assess interface clarity and visual focus.
Co-Discovery	Pairs of users, minimum 5 pairs, jointly complete tasks while being recorded. Evaluators document dialogue and actions, followed by a post-task questionnaire.
RITE	Five users engage in iterative usability testing with think-aloud protocols. Issues are fixed in real-time, and testing continues with updated versions.
Questionnaires	Recommended for 15–30 users depending on whether applied alone or with another method. Users complete gameplay tasks before filling out the questionnaire.
Interviews	Conducted post-task. If used alone, a sample of 30 users is advised; with other methods, 5 may be enough. Interviews explore user experience in depth.
Focus Groups	After gameplay tasks, group discussions are held with 5–30 participants to collectively explore usability issues and experiences.

In serious game evaluations, standardized questionnaires are widely employed to assess usability dimensions such as effectiveness, satisfaction, workload, and player experience. These instruments are typically administered immediately after users complete a predefined set of gameplay tasks, ensuring reliable and context-relevant responses.

The process involves assigning users a task list aligned with core game interactions, followed by the application of one or more questionnaires. When combined with other methods (observation or think-aloud), a minimum of 15 participants is recommended; when applied independently, at least 30 users are advised to ensure robustness [125].

Commonly used instruments include the SUS, NASA-TLX, GEQ, SUMI, PSSUQ, QUIS, CSUQ, and ad hoc questionnaires. These tools provide structured, often multidimensional data that support consistent usability assessment and facilitate comparisons across studies and contexts. Their use contributes to greater methodological rigor in the evaluation of serious games [14].

In summary, the systematic literature review provided a comprehensive and structured overview of the current state of user-centered usability evaluation in serious games. It identified the main methods, techniques, and standardized instruments employed in practice, as well as their specific characteristics and application protocols. The findings highlight a predominant reliance on empirical and inquiry-based approaches, with varied levels of methodological rigor and contextual adaptation. These insights not only confirm the need for evaluation frameworks tailored to the unique features of serious games but also serve as a foundational reference for the development of a context-sensitive usability evaluation instrument aligned with educational goals and user engagement dynamics.

4. Development of the Usability Evaluation Instrument

This section describes the design process of the proposed usability evaluation SGU (Serious Games Usability) tailored for serious games. Based on the findings of the SLR, existing questionnaires used to assess usability in digital systems were analyzed to identify limitations when applied to serious games. A structured GAP analysis was then conducted to systematically compare these instruments against the multidimensional requirements specific to serious game environments. The final result is a questionnaire-based tool that integrates both adapted and newly created items, ensuring a comprehensive assessment of usability.

4.1. GAP Analysis

To systematically support the development of the proposed instrument, a GAP analysis was carried out. This analysis aimed to critically compare the scope of widely used usability evaluation instruments with the specific multidimensional requirements of serious games. The process began by identifying a set of core usability dimensions that emerged from the systematic literature review, which are essential for capturing the full spectrum of user experience in serious games. These dimensions include serious objectives, storytelling, gamification techniques, gameplay, multimedia integration, interface design, world visualization, and character visualization.

Each of these dimensions reflects a critical component of how users engage with and perceive serious games, encompassing both functional usability and experiential qualities tied to learning, immersion, and emotional response. The selected instruments were then mapped against these dimensions to evaluate their coverage and to identify areas where existing tools fall short. The results of this GAP analysis provide the empirical foundation for designing a new, context-aware usability questionnaire that addresses the limitations of existing approaches and better aligns with the unique characteristics of serious game environments.

The findings of this analysis are summarized in Table 11, which illustrates the coverage of each reviewed instrument across the main dimensions of serious games.

Table 11. GAP analysis of existing instruments.

Questionnaire	Serious Objectives	Storytelling	World Visualization	Character Design	Gamification Techniques	Multimedia	Gameplay	Interfaces
SUS	NA	NA	NA	NA	NA	NA	+	+
SUMI	NA	NA	NA	NA	NA	NA	+	+
Ad hoc	NA	NA	NA	NA	NA	NA	NA	NA
PSSUQ	NA	NA	-	-	-	NA	+	+
CSUQ	NA	NA	-	-	-	NA	+	+
QUIS	NA	NA	NA	NA	NA	-	-	+
NASA-TLX	NA	NA	NA	NA	NA	NA	+	-
GEQ	NA	+	NA	NA	+	NA	+	-

The symbols have the following meaning:

- **-** means that the instruments consider the dimensions minimally.
- **+** means that the instruments consider the dimensions substantially.
- **NA** means that the instruments do not take the dimensions into consideration.

Based on the comparative results presented in Table 11, it was possible to identify significant gaps in the coverage provided by existing usability tests when evaluated against the specific requirements of serious games. Notably, most instruments demonstrated limited or no integration of critical dimensions such as serious objectives, narrative and story visualization, world and character visualization, gamification techniques, and multimedia integration.

These omissions suggest that current tools, although effective in general software or traditional game contexts, are insufficiently equipped to evaluate the experiential and pedagogical elements that are central to serious games.

4.2. Final Structure of the Instrument

The absence of specific evaluation items related to core components of serious games—such as learning outcomes, storyline integration, world and character design, and gamification strategies—can result in assessments that overlook fundamental usability aspects. Consequently, evaluations that rely solely on general-purpose instruments may yield incomplete or misleading insights about the game's effectiveness and user experience in educational or training scenarios.

To address this shortcoming, we developed a tailored usability evaluation instrument in the form of a structured questionnaire. The design of this tool was informed by the evidence and gaps identified in the SLR and grounded in user-centered inquiry methodology, which prioritizes the collection of direct user feedback regarding interaction, engagement, and perceived usability [131].

The construction of SGU was guided by three complementary principles. First, existing items from the SUS and PSSUQ were adapted to the educational context to ensure relevance while retaining comparability with widely used benchmarks [131]. Second, validated constructs, such as items from MEEGA+, were selectively incorporated to address serious objectives and gamification strategies [132], and questions from the Game Engagement Questionnaire (GEQ) were included to capture narrative immersion and visualization aspects [133]. Third, new items were created when no existing tool adequately covered critical dimensions identified in the GAP analysis, particularly those related to the explicit linkage between gameplay and pedagogy or multimedia integration. At the same time, exclusions were applied to avoid redundancy or conceptual misalignment, for example, by omitting workload measures from NASA-TLX and overlapping satisfaction constructs from QUIS. This structured process ensured that the resulting questionnaire integrates functional usability and pedagogical effectiveness in a coherent and context-sensitive manner.

The resulting instrument is designed to assess a comprehensive range of usability factors relevant to serious games, including gameplay mechanics, user engagement, educational integration, narrative immersion, and interface design. By systematically addressing these dimensions, the questionnaire offers a holistic and reliable evaluation framework that captures both the functional and pedagogical aspects of usability.

Structurally, the proposed tool comprises 21 items distributed across eight dimensions, each aligned with the theoretical framework previously established in this study. All items are formulated using a Likert-type scale, enabling consistent data collection, user perception quantification, and comparative analysis across different serious game implementations. The complete structure of the SGU is detailed in Table 12.

Table 12. Proposed usability evaluation questionnaire for serious games.

Category	Questions	Resource
Serious Objectives	Q1. Is the game content relevant to my interests?	MEEGA+
	Q2. Is it clear how the game content aligns with its intended educational goals?	MEEGA+
	Q3. Do I prefer learning with this game over traditional methods (e.g., lectures, readings)?	MEEGA+
	Q4. Did the game contribute to my understanding of a subject related to my studies?	MEEGA+
Story Visualization	Q5. Did the game's story or narrative capture my attention and interest?	GEQ
	Q6. Are the gameplay mechanics suitable for the narrative and game structure?	GEQ
World Visualization	Q7. Did the game environment include visual elements or metaphors I could relate to?	GEQ
Character Visualization	Q8. Did I feel in control of the character and capable of using effective strategies to solve challenges?	GEQ
Gamification Techniques	Q9. Was the level of challenge appropriate for my interest or knowledge?	MEEGA+
	Q10. Did the game introduce new obstacles or situations at an appropriate pace?	MEEGA+
	Q11. Did the game avoid becoming repetitive or tedious over time?	MEEGA+
Gameplay	Q12. Was the game clearly designed with the needs of its target users in mind?	SUMI, SUS, CSUQ
	Q13. At most times during the game, did you feel calm and at ease?	SUMI, SUS, CSUQ
	Q14. Would I like to use this game regularly?	SUMI, SUS, CSUQ
	Q15. Did I find it easy to learn how to use new features in the game?	SUMI, SUS, CSUQ
Interface	Q16. Did the interface elements (text, icons, buttons) support correct task execution and ease of interaction?	QUIS, SUS
	Q17. Was the organization of menus and buttons logical?	SUMI, SUS
	Q18. Was it easy and quick to recover from mistakes while using the game?	PSSUQ, CSUQ
	Q19. Was the information (e.g., on-screen messages, guidance) clear and useful?	PSSUQ, CSUQ
Multimedia	Q20. Did the use of multimedia improve how the information was presented?	QUIS
	Q21. Was the quality of the multimedia elements (text, images, audio, video) satisfactory?	QUIS

To complete the questionnaire, each participant carefully read each item and indicated their opinion by selecting a response that best reflected their level of satisfaction. A 5-point Likert scale was used for this purpose (Table 13), with the following equivalence:

Table 13. Likert scale.

Scale	Description
1	Totally disagree
2	Disagree
3	Neutral
4	Agreed
5	Totally agreement
NA	The question does Not Apply (NA) to the serious game evaluated

To facilitate comparative analysis and enhance interpretability, the responses collected through the 5-point Likert scale were transformed to a standardized 0–100 scale, consistent with the approach used in the System Usability Scale (SUS). This conversion allows for

direct comparison with existing usability benchmarks while preserving the ordinal nature of the original responses.

This study adopts a linear scaling approach more suitable for unidirectional questionnaires. Specifically, each response score is scaled proportionally to a 100-point range using a direct rule-of-three formula. For instruments based on a 5-point Likert scale, the transformation is expressed as

$$\text{SUS Score} = \frac{\sum_{i=1}^{n} Q_i \times 100}{5} \tag{1}$$

where

- Q_i represents the response to item on the 5-point Likert scale;
- The value 5 denotes the maximum point on the Likert scale.

This method provides an intuitive and interpretable representation of usability perceptions, particularly when all questionnaire items are phrased in a consistent direction (i.e., higher scores indicate better usability). It also simplifies the aggregation and comparison of results across different case studies and game-based evaluations.

To interpret usability results, we adopted the scoring system of the System Usability Scale (SUS) developed by John Brooke [134]. According to this model, a score below 68 suggests the presence of significant usability issues, while a score above 68 indicates acceptable usability (Figure 4). In addition, the SUS provides three complementary classification models that offer deeper insight into system usability:

- **Acceptance Range:** Scores are categorized as acceptable (above 70), marginal-high (60–69), marginal-low (50–59), or not acceptable (below 50).
- **Classification Scale:** A score above 90 earns an A (excellent), between 80 and 89 corresponds to a B (outstanding), and between 70 and 79 is rated C (good). Scores between 60 and 69 are assigned a D (fair), while any score below 60 receives an F (poor), signaling a critical need for usability improvements.
- **Adjectives classification:** Provides a qualitative interpretation of the score by assigning a descriptive label based on user perception. For example, scores below 25 are categorized as worst imaginable, those below 39 as poor, and scores around 52 as OK. A score near 73 is described as good, while scores close to 85 are considered excellent. The highest range, between 86 and 100, corresponds to the label best imaginable, indicating an exceptional user experience.

These classification models provide a comprehensive framework for evaluating and interpreting the usability of serious games, enabling developers and researchers to identify both strengths and areas for improvement based on user feedback.

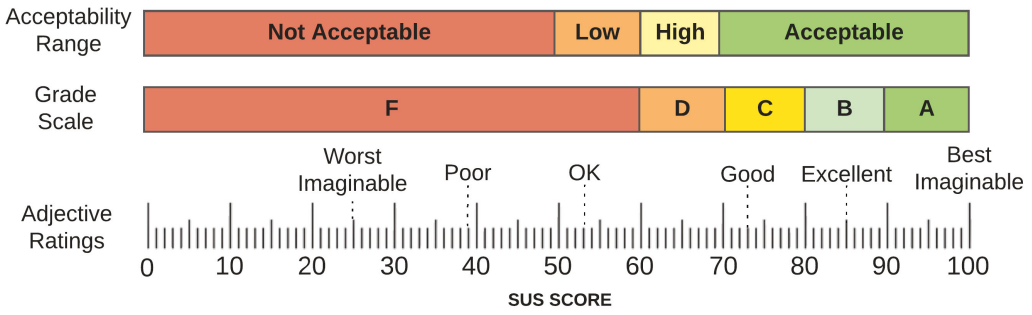


Figure 4. Bangor scale.

5. Application of the Usability Instrument in Case Studies

SGU was applied to three serious game projects: El Tesoro del Ser Politécnico [1], Flagschool [135], and Funcarac [136]. Each game targets distinct educational and cultural objectives.

"El Tesoro del Ser Politécnico" aims to foster ethical values within the polytechnic community through an exploratory game set on campus. Players navigate virtual scenarios, face interactive challenges, and uncover historical facts and ethical principles. The game is accessible through web browsers on both mobile and desktop platforms through the following link: https://tesoropolitecnico.epn.edu.ec/ (accessed on 29 August 2025).

"Flagschool" is designed to enhance critical thinking, problem-solving, creativity, and cognitive speed through the study of vexillology. The project uses interactive elements to create an engaging and educational experience. The game is available for download through the following repository: https://github.com/David-Morales-M/Vexillology-Serious-Game.git (accessed on 29 August 2025)

Finally, "Funcarac" focuses on preserving and promoting the intangible cultural heritage of indigenous agricultural communities in Ecuador. Through gameplay, users explore ancestral practices and worldviews rooted in agro-ecological knowledge. The game is accessible through web browsers on both mobile and desktop platforms through the following link: https://saminay.epn.edu.ec/ (accessed on 29 August 2025).

To evaluate usability, the protocol proposed by Abhay Rautela [137] was followed (see Figure 5). Questionnaires were selected as the primary evaluation technique due to their efficiency in capturing users' perceptions and experiential feedback.

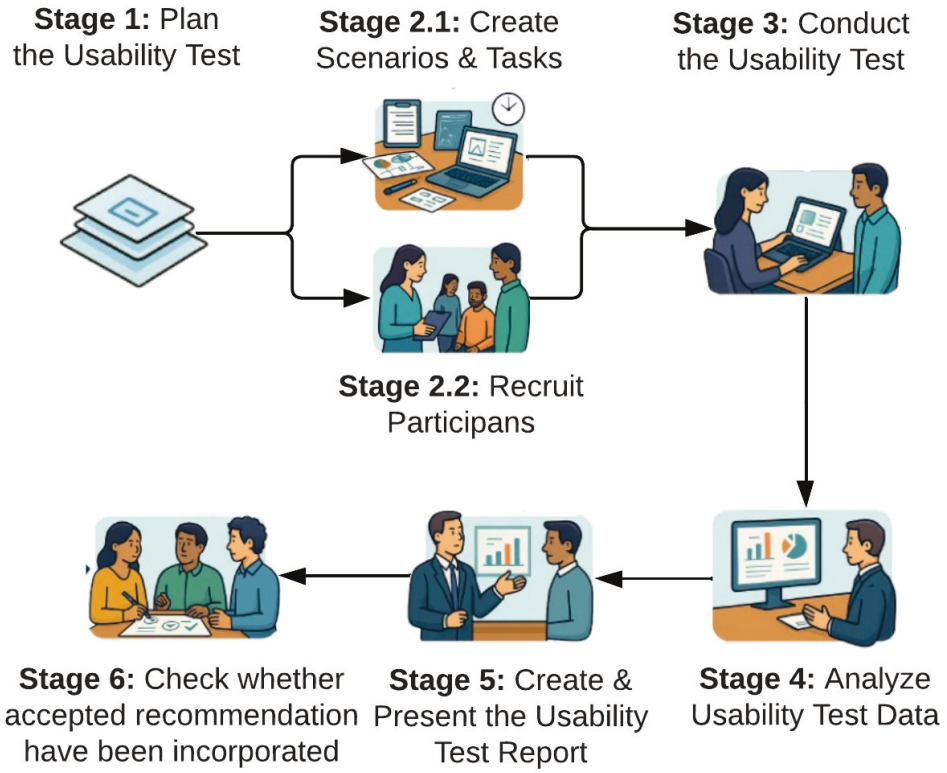


Figure 5. Usability protocol.

A minimum of 15 participants were recruited for each evaluation. As supported by Laura Faulkner's findings [138], testing with five users can identify up to 85% of

usability issues, while larger samples increase the likelihood of uncovering a broader range of problems.

Prior to testing, participants received the necessary documentation, including informed consent and a detailed task list. They then interacted with each game by completing predefined tasks and freely exploring the interface. Upon completion, participants responded to the usability evaluation questionnaire, providing feedback based on their experience.

The proposed usability evaluation instrument was applied across three serious games El Tesoro del Ser Politécnico, Flagschool, and Funcarac to assess its ability to capture the multidimensional nature of user experience in diverse educational game contexts; see Figure 6.

Overall, the questionnaire revealed high levels of user satisfaction across the three case studies, with most items exceeding a 70% positive response threshold. This indicates that the instrument successfully identified areas where each game delivered meaningful experiences, as well as those requiring refinement.

For example, items related to serious objectives (Q1–Q4) consistently scored highly, with "Funcarac" achieving 88.5% in Q1 and "Flagschool" reaching 92.0% in Q2. These results confirm that the questionnaire effectively captures users' perceptions of content relevance and learning effectiveness for evaluating the "serious" component of serious games.

Items addressing interface usability (Q16–Q19) also showed strong performance, particularly in "El Tesoro de ser Politécnico" and "Flaschool", where scores remained above 78%, suggesting that the questionnaire reliably measures key aspects of system usability such as navigability, information clarity, and error recovery.

In terms of engagement and gamification (Q9–Q11), Vexillology Game notably outperformed the others in Q9 (91.52%), evidencing the instrument's sensitivity in capturing motivational elements central to user engagement. Similarly, Q8, focused on the player's perceived control over the game character, revealed moderate ratings in Ser Politécnico (76.4%) and lower results in Funcarac (72.3%), which underscores the diagnostic capacity of this item in identifying limitations in character interaction and decision-making.

However, it is important to note that dimensions such as world visualization and character design are not mandatory components in all serious games. Their relevance depends largely on the pedagogical goals, game mechanics, and intended user experience of the game in question. For instance, some serious games may prioritize abstract cognitive tasks, textual information delivery, or simulation-based logic without employing rich visual environments or controllable avatars. In such cases, questions targeting world or character representation (e.g., Q7 and Q8) may be considered optional or omitted entirely, depending on the nature of the application. This flexibility reinforces the adaptability of the proposed questionnaire to a wide range of serious game designs, maintaining its validity while ensuring contextual relevance.

In summary, the application of the SGU across three distinct case studies validates its utility and robustness. It demonstrates not only a high level of user comprehensibility and applicability but also the capacity to generate actionable insights that reflect both functional usability and the educational effectiveness of serious games. The results reaffirm the instrument's relevance as a holistic and scalable evaluation tool for projects aiming to integrate learning objectives with engaging user experiences.

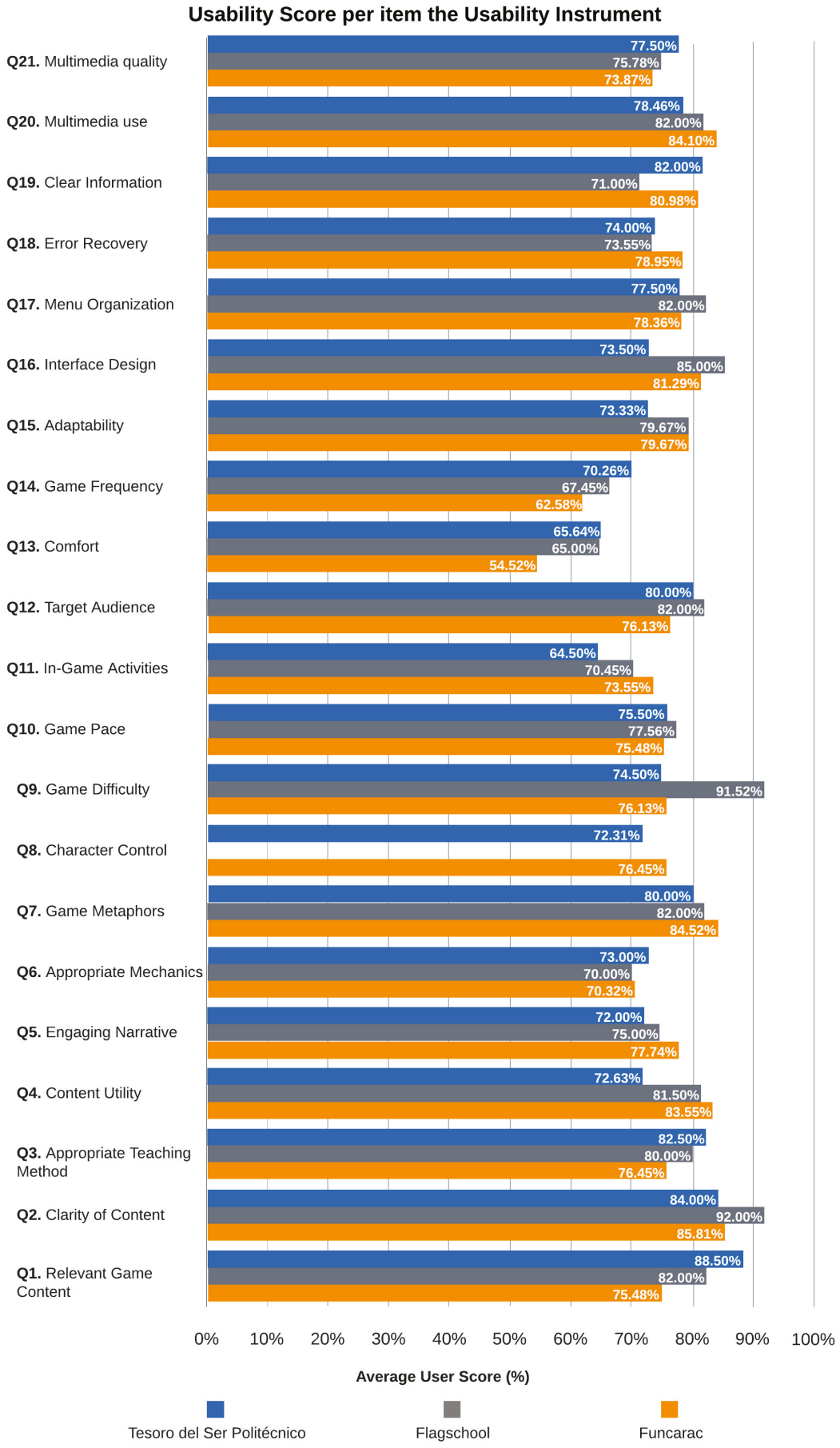


Figure 6. Average user score per item in the usability instrument.

6. Discussion

The development and application of SGU, the proposed usability evaluation instrument, represent a substantive advancement in addressing the methodological gaps identified in the systematic literature review (SLR). By integrating domains such as serious objectives, narrative immersion, character and world visualization, gamification strategies, multimedia design, and educational alignment, the tool offers a holistic framework specifically tailored to the complexities of serious games, going far beyond traditional usability tools like the SUS or CSUQ, which primarily focus on general system performance and interface satisfaction.

A critical analysis of the instrument's structure reveals that its multidimensional approach provides a more comprehensive and relevant evaluation framework for serious games. Unlike conventional tools such as the SUS or CSUQ, which are limited in capturing pedagogical or narrative aspects, the proposed questionnaire incorporates constructs from MEEGA+ and GEQ to better reflect the unique characteristics of serious games. This customization allows evaluators to measure not only ease of use and engagement but also the alignment of the game with learning objectives and its ability to foster immersive experiences.

Importantly, the instrument's design accommodates flexibility by including an N/A (not applicable) response option, which acknowledges that not all serious games require certain features, such as character or world visualization, to achieve their educational objectives. This adaptability enhances the questionnaire applicability across diverse genres of serious games, where such elements may be absent or intentionally minimized without compromising the seriousness of the game or its pedagogical intent.

Compared to instruments identified in the SLR, the proposed tool demonstrates significantly greater coverage and contextual specificity. While existing instruments like MEEGA+, GEQ, or SUMI offer partial solutions, none effectively integrate both pedagogical relevance and usability performance in a unified structure. The proposed tool fills this gap by merging validated items from various sources and organizing them into coherent dimensions that align with the core attributes of serious games. This structure enables evaluators to assess not only functionality and satisfaction but also narrative consistency, learning alignment, and engagement depth.

Furthermore, SGU facilitates comparative analysis through its standardized Likert-scale format and the subsequent transformation into a SUS-equivalent percentage scale. This allows for intuitive interpretation and benchmarking, which is particularly valuable for iterative design cycles and cross-game usability assessments. For example, games like Flagschool achieved over 90% on items related to educational clarity and challenge balance, suggesting high usability in cognitively demanding contexts. In contrast, Funcarac exhibited slightly lower results in emotional comfort and novelty, highlighting areas for design refinement.

These results are consistent with previous research showing that serious games with well-defined pedagogical goals often achieve higher usability ratings, whereas games emphasizing cultural narratives or novel experiences face greater challenges in maintaining emotional comfort and engagement. The observed differences across the case studies highlight that usability is not solely determined by interface quality but also by the integration of narrative depth, cultural context, and affective design. This underscores the importance of considering contextual and cultural variables in usability evaluations, as also noted in the recent literature. From a practical standpoint, these findings suggest that designers should prioritize both clarity of educational objectives and strategies to foster emotional immersion, ensuring that serious games are both instructive and engaging across diverse user populations.

Despite SGU demonstrating utility, certain limitations must be acknowledged. The instrument has not yet undergone large-scale psychometric validation, and its reliability and sensitivity across different user demographics and game types require further exploration. Additionally, while the structured questionnaire offers practical advantages in terms of cost and scalability, it may benefit from being complemented with qualitative methods to capture more nuanced user feedback.

In conclusion, the proposed questionnaire presents a significant contribution to the domain of serious game evaluation. It provides a targeted, user-centered, and context-aware tool capable of measuring dimensions that were previously neglected in conventional usability frameworks. Its flexibility, clarity, and comprehensive scope position it as a robust instrument for both formative and summative assessments in educational, health, and training-oriented game environments. Future work will focus on empirical validation, refinement based on user feedback, and cross-cultural adaptation to support broader adoption.

7. Conclusions, Limitations, and Future Directions

This study introduced SGU (Serious Games Usability), a novel usability evaluation instrument that addresses methodological gaps in evaluating serious games. By integrating pedagogical alignment, narrative immersion, gamification, multimedia quality, and usability principles, SGU offers a structured and comprehensive framework that goes beyond the scope of traditional tools such as the SUS or CSUQ.

The main contribution of this research is the development of SGU, a domain-specific, multidimensional, and user-centered instrument that combines traditional usability criteria with game-specific aspects such as pedagogical value, engagement, and experiential design. Its application in three case studies demonstrated strong diagnostic capacity, effectively identifying strengths and usability challenges across diverse educational, cultural, and training-oriented game designs.

However, several limitations must be acknowledged. First, the sample sizes in the case studies were limited and relatively homogeneous, which may constrain the generalizability of the findings. Second, cultural and contextual differences in user perception were not systematically addressed, leaving open questions about the instrument's applicability across different regions and populations. Third, although the inclusion of an N/A option provides flexibility, further validation is required to ensure consistency when applied across games of varying genres and levels of complexity. Finally, while the questionnaire items were derived from validated instruments and contextualized for serious games, future applications may reveal the need for further refinement or domain-specific adaptation depending on genre, user profile, or learning objectives.

Despite these limitations, the proposed tool offers strong potential for broader applicability. Its low implementation cost, clear dimensional structure, and ease of integration into iterative development cycles make it a practical tool for usability professionals, educators, and developers. Moreover, its compatibility with Likert-based scales and subsequent percentage normalization facilitate cross-study comparisons and benchmarking.

Future research should therefore prioritize large-scale psychometric validation of SGU, incorporating diverse participant profiles across different educational stages, cultural contexts, and subject domains. Cross-cultural adaptations will be essential for international adoption, while comparative studies with other usability frameworks can help establish its discriminant validity. Moreover, applying SGU to high-impact serious games, such as citizen science initiatives with millions of participants, could strengthen its robustness and relevance. To complement quantitative insights, qualitative approaches, including interviews and focus groups, should also be incorporated to capture nuanced cultural and

contextual factors that may not be reflected in numerical data. Ultimately, SGU provides a practical, flexible, and user-centered approach to usability evaluation, supporting the development of more effective and inclusive serious games.

Author Contributions: Conceptualization, M.C.-T., L.B. and M.S.; methodology, M.C.-T., D.M.-M. and M.S.; validation, M.S., L.B. and P.A.-V.; formal analysis, M.C.-T., L.B. and D.M.-M.; investigation, M.C.-T., L.B. and D.M.-M.; writing—original draft preparation, M.C.-T., M.S., L.B., J.G., J.A. and D.M.-M.; writing—review and editing, M.S., D.M.-M., J.G., J.A. and P.A.-V.; supervision, M.C.-T.; project administration, M.S., M.C.-T. and P.A.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad de Las Américas—Ecuador as part of the internal research project 489. A.XIV.24.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors wish to acknowledge the contribution of the "Laboratorio de Investigación de Sistemas de Información e Inclusión Digital"-LudoLab https://ludolab.epn.edu.ec/ (accessed on 22 July 2025) for its support in the research and publication of this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Santórum, M.; Morales-Martínez, D.; Caiza, J.; Carrión-Toro, M.; Galindo, J.; Acosta-Vargas, P. Serious Games, Serious Impact: Enhancing Ethical Education Through Interactive Learning. In Proceedings of the 2024 International Conference on Computer, Information and Telecommunication Systems (CITS), Girona, Spain, 17–19 July 2024; pp. 1–8.
2. Krajčovič, M.; Gabajová, G.; Matys, M.; Grznár, P.; Dulina, L.; Kohár, R. 3D Interactive Learning Environment as a Tool for Knowledge Transfer and Retention. *Sustainability* **2021**, *13*, 7916. [CrossRef]
3. Cook-Chennault, K.; Villanueva Alarcón, I.; Jacob, G. Usefulness of Digital Serious Games in Engineering for Diverse Undergraduate Students. *Educ. Sci.* **2022**, *12*, 27. [CrossRef]
4. Alrehaili, E.A.; Osman, H.A. A virtual reality role-playing serious game for experiential learning. *Interact. Learn. Environ.* **2022**, *30*, 922–935. [CrossRef]
5. Gurbuz, S.C.; Celik, M. Serious games in future skills development: A systematic review of the design approaches. *Comput. Appl. Eng. Educ.* **2022**, *30*, 1591–1612. [CrossRef]
6. Djaouti, D.; Alvarez, J.; Jessel, J.P. Classifying serious games: The G/P/S model. In *Handbook of Research on Improving Learning and Motivation Through Educational Games*; Advances in Game-Based Learning; IGI Global: Hershey, PA, USA, 2011; pp. 118–136. [CrossRef]
7. Vásquez, L.M.L. Juegos Con Propósitos Educativos Como Herramientas Para la Enseñanza de la Ética. Ph.D. Thesis, Universidad Nacional de Colombia, Bogotá, Colombia, 2017.
8. Tavares, N. The use and impact of game-based learning on the learning experience and knowledge retention of nursing undergraduate students: A systematic literature review. *Nurse Educ. Today* **2022**, *117*, 105484. [CrossRef]
9. Lampropoulos, G.; Kinshuk. Virtual reality and gamification in education: A systematic review. *Educ. Technol. Res. Dev.* **2024**, *72*, 1691–1785. [CrossRef]
10. Aster, A.; Hütt, C.; Morton, C.; Flitton, M.; Laupichler, M.C.; Raupach, T. Development and evaluation of an emergency department serious game for undergraduate medical students. *BMC Med. Educ.* **2024**, *24*, 1061. [CrossRef]
11. Checa, D.; Miguel-Alonso, I.; Bustillo, A. Immersive virtual-reality computer-assembly serious game to enhance autonomous learning. *Virtual Real.* **2023**, *27*, 3301–3318. [CrossRef]
12. Tito Cruz, J.; Coluci, V.R.; Moraes, R. ORUN-VR2: A VR serious game on the projectile kinematics: Design, evaluation, and learning outcomes. *Virtual Real.* **2023**, *27*, 2583–2604. [CrossRef]
13. Marin, D.; Castrillo, J.; Bermúdez, G.S. Pruebas de la usabilidad en usuarios discapacitados. *Rev. Repert. Científico* **2012**, *1*, 55–65.
14. Moreno-Ger, P.; Torrente, J.; Hsieh, Y.G.; Lester, W. Usability testing for serious games: Making informed design decisions with user data. *Adv. Hum.-Comput. Interact.* **2012**, *2012*, 369637. [CrossRef]
15. Da Silveira, A.C.; Martins, R.X.; Vieira, E.A.O. E-Guess: Evaluación de usabilidad para juegos educativos. *RIED. Rev. Iberoam. Educ. Distancia* **2021**, *24*, 245–263.

16. Kitchenham, B.; Charters, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*; Technical Report EBSE-2007-01; Keele University: Keele, UK, 2007.
17. Damaševičius, R.; Maskeliūnas, R.; Blažauskas, T. Serious Games and Gamification in Healthcare: A Meta-Review. *Information* **2023**, *14*, 105. [CrossRef]
18. Natucci, G.C.; Borges, M.A.F. The Experience, Dynamics and Artifacts Framework: Towards a Holistic Model for Designing Serious and Entertainment Games. In Proceedings of the 2021 IEEE Conference on Games (CoG), Copenhagen, Denmark, 17–20 August 2021; pp. 1–8. [CrossRef]
19. Alshar'e, M.; Albadi, A.; Jawarneh, M.; Tahir, N.; Al Amri, M. Usability Evaluation of Educational Games: An Analysis of Culture as a Factor Affecting Children's Educational Attainment. *Adv. Hum.-Comput. Interact.* **2022**, *2022*, 9427405. [CrossRef]
20. Carlier, S.; Naessens, V.; De Backere, F.; De Turck, F. A Software Engineering Framework for Reusable Design of Personalized Serious Games for Health: Development Study. *JMIR Serious Games* **2023**, *11*, e40054. [CrossRef]
21. Frasca, G. Videogames of the Oppressed: Videogames as a Means for Critical Thinking and Debate. Master's Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2001.
22. Czernik, D.S.L. Videojuegos y alfabetización digital. *Aula Innovación Educ.* **2005**, *1*, 48–50.
23. Schell, J. *The Art of Game Design: A Book of Lenses*; CRC Press: Boca Raton, FL, USA, 2008.
24. Kars, L. What Are Serious Games? 2025. Available online: https://grendelgames.com/what-are-serious-games/#:~:text=Serious%20games%20are%20games%20with,and%20other%20businesses%20and%20industries (accessed on 20 May 2025).
25. Abt, C.C. *Serious Games*; University Press of America: Lanham, MD, USA, 1987.
26. Alvarez, J.; Michaud, L. *Serious Games: Advergaming, Edugaming, Training and More*; IDATE: Montpellier, France, 2008.
27. Ferdig, R.; Winn, B., The Design, Play, and Experience Framework. In *The Design, Play, and Experience Framework*; IGI Global: Hershey, PA, USA, 2009; pp. 1010–1024. [CrossRef]
28. Carrion, M.; Santórum, M.; Flores, H.; Aguilar, J.; Perez, M. Serious Game, gamified applications, educational software: A comparative study. In Proceedings of the 2019 International Conference on Information Systems and Software Technologies (ICI2ST), Quito, Ecuador, 13–15 November 2019; pp. 55–62.
29. Carrión Toro, M.d.C. iPlus Una Metodología Centrada en el Usuario Para el Diseño de Juegos Serios. Ph.D. Thesis, Escuela Politécnica Nacional, Quito, Ecuador, 2022.
30. Deterding, S.; Sicart, M.; Nacke, L.; O'Hara, K.; Dixon, D. Gamification. using game-design elements in non-gaming contexts. In Proceedings of the CHI '11 Extended Abstracts on Human Factors in Computing Systems, New York, NY, USA, 7–12 May 2011; CHI EA '11, pp. 2425–2428. [CrossRef]
31. Ullah, M.; Amin, S.U.; Munsif, M.; Yamin, M.M.; Safaev, U.; Khan, H.; Khan, S.; Ullah, H. Serious games in science education: A systematic literature. *Virtual Real. Intell. Hardw.* **2022**, *4*, 189–209. [CrossRef]
32. Yu, Z.; Gao, M.; Wang, L. The Effect of Educational Games on Learning Outcomes, Student Motivation, Engagement and Satisfaction. *J. Educ. Comput. Res.* **2021**, *59*, 522–546. [CrossRef]
33. Didácticos, N. ¿Qué son Los Juegos Educativos? 2018. Available online: https://www.noedidacticos.com/blog/que-son-los-juegos-educativos-0de1c230067b (accessed on 20 May 2025).
34. Hodhod, R.; Hardage, H.; Abbas, S.; Aldakheel, E.A. CyberHero: An Adaptive Serious Game to Promote Cybersecurity Awareness. *Electronics* **2023**, *12*, 3544. [CrossRef]
35. Marengo, A.; Pagano, A.; Soomro, K.A. Serious games to assess university students' soft skills: investigating the effectiveness of a gamified assessment prototype. *Interact. Learn. Environ.* **2024**, *32*, 6142–6158. [CrossRef]
36. Tan, C.K.W.; Nurul-Asna, H. Serious games for environmental education. *Integr. Conserv.* **2023**, *2*, 19–42. [CrossRef]
37. Sun, L.; Guo, Z.; Hu, L. Educational games promote the development of students' computational thinking: A meta-analytic review. *Interact. Learn. Environ.* **2023**, *31*, 3476–3490. [CrossRef]
38. Guan, Y.; Qiu, Y.; Xu, F.; Fang, J. Effect of Game Teaching Assisted by Deep Reinforcement Learning on Children's Physical Health and Cognitive Ability. *Mob. Inf. Syst.* **2022**, *2022*, 6150261. [CrossRef]
39. Gros Salvat, B. Certezas e interrogantes acerca del uso de los videojuegos para el aprendizaje. *Comunicación. Revista Internacional de Comunicación Audiovisual*; Editorial Universidad de Sevilla: Sevilla, España, 2009; Volume 1, Number 7, pp. 251–264.
40. *ISO 9241-11:2018*; Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts. ISO: Geneva, Switzerland, 2018. Available online: https://www.iso.org/standard/63500.html (accessed on 18 July 2025).
41. Sánchez, W.O. La usabilidad en Ingeniería de Software: Definición y características. *Revista Ing-Novación No. 02* **2015**, *1*, 7–21.
42. IEEE Xplore. Available online: https://ieeexplore.ieee.org/Xplore/home.jsp (accessed on 15 July 2025).
43. ACM Digital Library. Available online: https://www.acm.org/ (accessed on 15 July 2025).
44. Elsevier. Available online: https://www.elsevier.com/ (accessed on 15 July 2025).
45. Google Scholar. Available online: https://scholar.google.com/ (accessed on 15 July 2025).

46. Springer Nature Link. Available online: https://link.springer.com/ (accessed on 15 July 2025).
47. Clark, J.; Glasziou, P.; Mar, C.D.; Bannach-Brown, A.; Stehlik, P.; Scott, A.M. A full systematic review was completed in 2 weeks using automation tools: A case study. *J. Clin. Epidemiol.* **2020**, *121*, 81–90. [CrossRef]
48. Zooniverse. Galaxy. 2025. Available online: https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/ (accessed on 29 August 2025).
49. Glitchers Ltd. Sea Hero Quest. 2025. Available online: https://seaheroquest.com/ (accessed on 29 August 2025).
50. Waldispuhl, J.; Szantner, A. Video game unleashes millions of citizen scientists on microbiome research. *Nat. Biotechnol.* **2025**, *43*, 36–37.
51. Tolentino, G.P.; Battaglini, C.; Pereira, A.C.V.; Oliveria, R.J.d.; Paula, M.G.M.d. Usability of Serious Games for Health. In Proceedings of the 2011 Third International Conference on Games and Virtual Worlds for Serious Applications, Athens, Greece, 4–6 May 2011; pp. 172–175. [CrossRef]
52. Diah, N.M.; Ismail, M.; Ahmad, S.; Dahari, M.K.M. Usability testing for educational computer game using observation method. In Proceedings of the 2010 International Conference on Information Retrieval & Knowledge Management (CAMP), Shah Alam, Malaysia, 17–18 March 2010; pp. 157–161. [CrossRef]
53. Falcão, I.; da S. Souza, D.; Araujo, F.P.; dos S. Fernandes, J.G.; Pires, Y.P.; Cardoso, D.L.; Seruffo, M.C. Usability and Cognitive Benefits of a Serious Game to Combat Aedes Aegypti Mosquito. In Proceedings of the 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 5–7 June 2019; pp. 720–725. [CrossRef]
54. Omar, H.; Jaafar, A. Playability Heuristics Evaluation (PHE) approach for Malaysian educational games. In Proceedings of the 2008 International Symposium on Information Technology, Kuala Lumpur, Malaysia, 26–28 August 2008; Volume 3, pp. 1–7. [CrossRef]
55. Hussain, A.; Mutalib, N.A.; Zaino, A. A usability testing on JFakih Learning Games for hearing impairment children. In Proceedings of the The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M), Kuching, Malaysia, 17–18 November 2014; pp. 1–4. [CrossRef]
56. Ismail, M.; Diah, N.M.; Ahmad, S.; Kamal, N.A.M.; Dahari, M.K.M. Measuring usability of educational computer games based on the user success rate. In Proceedings of the 2011 International Symposium on Humanities, Science and Engineering Research, Kuala Lumpur, Malaysia, 6–7 June 2011; pp. 56–60. [CrossRef]
57. Dagdag, R.J.E.; Mariano, R.V.; Atienza, R.O. Teacher tool for visualization and management of a technology-enhanced learning environment. In Proceedings of the TENCON 2011—2011 IEEE Region 10 Conference, Bali, Indonesia, 21–24 November 2011; pp. 1426–1430. [CrossRef]
58. Desurvire, H.; El-Nasr, M.S. Methods for Game User Research: Studying Player Behavior to Enhance Game Design. *IEEE Comput. Graph. Appl.* **2013**, *33*, 82–87. [CrossRef]
59. Papaloukas, S.; Patriarcheas, K.; Xenos, M. Usability Assessment Heuristics in New Genre Videogames. In Proceedings of the 2009 13th Panhellenic Conference on Informatics, Corfu, Greece, 10–12 September 2009; pp. 202–206. [CrossRef]
60. Höysniemi, J.; Hämäläinen, P.; Turkki, L. Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interact. Comput.* **2003**, *15*, 203–225. [CrossRef]
61. Rosyidah, U.; Haryanto, H.; Kardianawati, A. Usability Evaluation Using GOMS Model for Education Game "Play and Learn English". In Proceedings of the 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 21–22 September 2019; pp. 1–5. [CrossRef]
62. Soomro, S.; Ahmad, W.F.W.; Sulaiman, S. Evaluation of mobile games with playability heuristic evaluation system. In Proceedings of the 2014 International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 3–5 June 2014; pp. 1–6. [CrossRef]
63. May, J. YouTube Gamers and Think-Aloud Protocols: Introducing Usability Testing. *IEEE Trans. Prof. Commun.* **2019**, *62*, 94–103. [CrossRef]
64. Liao, Y.H.; Shen, C.Y. Heuristic Evaluation of Digital Game Based Learning: A Case Study. In Proceedings of the 2012 IEEE Fourth International Conference on Digital Game and Intelligent Toy Enhanced Learning, Takamatsu, Japan, 27–30 March 2012; pp. 192–196. [CrossRef]
65. Jegers, K. Investigating the Applicability of Usability and Playability Heuristics for Evaluation of Pervasive Games. In Proceedings of the 2008 Third International Conference on Internet and Web Applications and Services, Athens, Greece, 8–13 June 2008; pp. 656–661. [CrossRef]
66. Hookham, G.; Nesbitt, K.; Kay-Lambkin, F. Comparing usability and engagement between a serious game and a traditional online program. In Proceedings of the Australasian Computer Science Week Multiconference, Canberra, Australia, 1–5 February 2016; ACSW '16. [CrossRef]

67. Baauw, E.; Bekker, M.M.; Markopoulos, P. Assessing the applicability of the structured expert evaluation method (SEEM) for a wider age group. In Proceedings of the 2006 Conference on Interaction Design and Children, Tampere, Finland, 7–9 June 2006; IDC ′06, pp. 73–80. [CrossRef]
68. Yue, W.S.; Zin, N.A.M. Usability evaluation for history educational games. In Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Republic of Korea, 24–26 November 2009; ICIS ′09, pp. 1019–1025. [CrossRef]
69. Papaloukas, S.; Xenos, M. Usability and education of games through combined assessment methods. In Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments, Athens, Greece, 16–18 July 2008; PETRA ′08. [CrossRef]
70. Pinelle, D.; Wong, N.; Stach, T. Heuristic evaluation for games: Usability principles for video game design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, 5–10 April 2008; CHI ′08, pp. 1453–1462. [CrossRef]
71. Rodio, F.; Bastien, J.M.C. Heuristics for Video Games Evaluation: How Players Rate Their Relevance for Different Game Genres According to Their Experience. In Proceedings of the 25th Conference on l'Interaction Homme-Machine, Talence, France, 12–15 November 2013; IHM ′13, pp. 89–93. [CrossRef]
72. Pinelle, D.; Wong, N.; Stach, T.; Gutwin, C. Usability heuristics for networked multiplayer games. In Proceedings of the 2009 ACM International Conference on Supporting Group Work, Sanibel Island, FL, USA, 10–13 May 2009; GROUP ′09, pp. 169–178. [CrossRef]
73. Desurvire, H.; Caplan, M.; Toth, J.A. Using heuristics to evaluate the playability of games. In Proceedings of the CHI ′04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria, 24–29 April 2004; CHI EA ′04, pp. 1509–1512. [CrossRef]
74. Korhonen, H.; Koivisto, E.M.I. Playability heuristics for mobile games. In Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services, Helsinki, Finland, 12–15 September 2006; MobileHCI ′06, pp. 9–16. [CrossRef]
75. Fabiano, F.; Sanchez-Francisco, M.; Díaz, P.; Aedo, I. Evaluating a pervasive game for urban awareness. In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, Barcelona, Spain, 3–6 September 2018; MobileHCI ′18, pp. 197–204. [CrossRef]
76. Karoulis, A. Evaluating the LEGO–RoboLab interface with experts. *Comput. Entertain.* **2006**, *4*, 6–29. [CrossRef]
77. Mohd, F.; Noor, M.; Daud, E.H.C.; Hasbullah, S.S. Mobile Games Usability Using the Idea of Mark Overmars. In Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, Danang, Viet Nam, 4–6 January 2016; IMCOM ′16. [CrossRef]
78. Verkuyl, M.; Romaniuk, D.; Mastrilli, P. Virtual gaming simulation of a mental health assessment: A usability study. *Nurse Educ. Pract.* **2018**, *31*, 83–87. [CrossRef]
79. Johnsen, H.M.; Fossum, M.; Vivekananda-Schmidt, P.; Fruhling, A.; Åshild Slettebø. Teaching clinical reasoning and decision-making skills to nursing students: Design, development, and usability evaluation of a serious game. *Int. J. Med. Inform.* **2016**, *94*, 39–48. [CrossRef] [PubMed]
80. Savazzi, F.; Isernia, S.; Jonsdottir, J.; Di Tella, S.; Pazzi, S.; Baglio, F. Design and implementation of a serious game on neurorehabilitation: Data on modifications of functionalities along implementation releases. *Data Brief* **2018**, *20*, 864–869. [CrossRef] [PubMed]
81. Savazzi, F.; Isernia, S.; Jonsdottir, J.; Di Tella, S.; Pazzi, S.; Baglio, F. Engaged in learning neurorehabilitation: Development and validation of a serious game with user-centered design. *Comput. Educ.* **2018**, *125*, 53–61. [CrossRef]
82. Hauge, J.B.; Riedel, J.C. Evaluation of simulation games for teaching engineering and manufacturing. *Procedia Comput. Sci.* **2012**, *15*, 210–220. [CrossRef]
83. Senap, N.M.V.; Ibrahim, R. A review of heuristics evaluation component for mobile educational games. *Procedia Comput. Sci.* **2019**, *161*, 1028–1035. [CrossRef]
84. Luna-Perejon, F.; Malwade, S.; Styliadis, C.; Civit, J.; Cascado-Caballero, D.; Konstantinidis, E.; Abdul, S.S.; Bamidis, P.D.; Civit, A.; Li, Y.C.J. Evaluation of user satisfaction and usability of a mobile app for smoking cessation. *Comput. Methods Programs Biomed.* **2019**, *182*, 105042. [CrossRef]
85. Amaro, B.A.; Mira, E.E. Desarrollo y Evaluación de Usabilidad de un Juego "serio" Utilizado en la Enseñanza del Comportamiento de la Bolsa de Valores y Cereales, 2020. Trabajo Académico. Available online: https://ridaa.unicen.edu.ar/items/6b6cea48-b205-4830-a805-4fb313ded06a/full (accessed on 13 September 2025).
86. Zurita, G.; Sánchez, J.; Nussbaum, M. Usabilidad de juegos educativos. *Taller Int. Softw. Educ. TISE* **1999**, *99*, 1–18.
87. Sánchez Royo, Y. Metodología de Evaluación de Videojuegos Serios en Competencias Ciudadanas, Exposición Caso Civia. *Dk* **2015**, *53*, 1689–1699.

88. Gordon, N.; Brayshaw, M.; Aljaber, T. Heuristic evaluation for serious immersive games and M-instruction. In Proceedings of the International Conference on Learning and Collaboration Technologies, Toronto, ON, Canada, 17–22 July 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 310–319.
89. Olimpio, A.; Notargiacomo, P. Heuristic Usability Evaluation Applied to Educational Games. In Proceedings of the International Conference on Educational Technologies 2020 (ICEduTech 2020), São Paulo, Brazil, 5–7 February 2020; pp. 25–32.
90. Vieira, E.A.O.; SILVEIRA, A.C.d.; Martins, R.X. Heuristic evaluation on usability of educational games: A systematic review. *Inform. Educ.* **2019**, *18*, 427–442. [CrossRef]
91. Isleyen, F.; Gulkesen, K.H.; Cinemre, B.; Samur, M.K.; Zayim, N.; Sen Kaya, S. Evaluation of the usability of a serious game aiming to teach facial expressions to schizophrenic patients. In *e-Health–For Continuity of Care*; IOS Press: Amsterdam, The Netherlands, 2014; pp. 662–666.
92. Paracha, S.; Clawson, K.; Mitsche, N.; Hall, L.; Jamil, F. Co-design with Children: Using Participatory Design for Design Thinking and Social and Emotional Learning. *Open Educ. Stud.* **2019**, *1*, 267–280 [CrossRef]
93. Wang, J. A Usability Assessment for a Career Planning Educational Video Game. Master's Thesis, Purdue University, Lafayette, IN, USA, 2016.
94. Perry, G.T.; Kulpa, C.C.; Pinheiro, E.; Eichler, M.L. Lessons from an educational game usability evaluation. *Int. J. Interact. Mob. Technol.* **2012**, *6*, 23–28. [CrossRef]
95. Mohamed, H.; Jaafar, A. Development and potential analysis of heuristic evaluation for educational computer game (PHEG). In Proceedings of the 5th International Conference on Computer Sciences and Convergence Information Technology, Seoul, Republic of Korea, 30 November–2 December 2010; pp. 222–227.
96. Barbosa, M.B.; Rego, A.B.; De Medeiros, I. HEEG: Heuristic evaluation for educational games. In Proceedings of the XIV Brazilian Symposium of Games and Digital Entertainment, Teresina, PI, Brazil, 11–13 November 2015.
97. Gómez, R.Y. Evaluación Heurística de la Usabilidad en Contextos Específicos. Ph.D. Thesis, Universidad de Sevilla, Seville, Spain, 2019.
98. Chacón Araya, J.M. Evaluación de Usabilidad de Juegos de Video Para Casinos en Plataformas Móviles: Conceptos y Experiencias. Ph.D. Thesis, Universidad Nacional de Costa Rica, San José, Costa Rica, 2017.
99. Castillo Aguirre, G. Métodos de Evaluación de la Usabilidad en Aplicaciones Móviles. Master's Thesis, Universidad Señor de Sipán, Chiclayo, Peru, 2020.
100. Chanchí, Gabriel E., G.; Álvarez, M.C.G.; Campo, W.Y.M. Criterios de usabilidad para el diseño e implementación de videojuegos. In *Revista Ibérica de Sistemas e Tecnologias de Informação*; Universidad de Medellin: Medellín, Colombia, 2020; pp. 461–474.
101. Schneider, G.; Barbosa, D.; Santos, P.; Barbosa, J. Evaluation of Usability and Gameplay of Games on Mobile Platforms for Young People on Oncological Treatment. *RENOTE* **2019**, *17*, 122–131. [CrossRef]
102. Armas Torres, M.B. Técnicas de Evaluación Para Videojuegos Geolocalizados. Master's Thesis, Universidad de la Laguna, San Cristóbal de La Laguna, Spain, 2015.
103. Barboza Mendoza, C.P.V. Heurísticas de Usabilidad Para Mecánicas de Videojuegos. Master's Thesis, Pontificia Universidad Católica del Perú, Lima, Peru, 2019.
104. Carrillo, A.I.; Garcia-Ruiz, M. Evaluación de Usabilidad de un Ambiente Virtual en 3D. *Rev. Faz* **2009**, *3*, 56–67.
105. Simor, F.W.; Brum, M.R.; Schmidt, J.D.E.; Rieder, R.; De Marchi, A.C.B. Usability evaluation methods for gesture-based games: A systematic review. *JMIR Serious Games* **2016**, *4*, e5860. [CrossRef]
106. Laitinen, S. Better Games Through Usability Evaluation and Testing. Gamasutra. Available online: https://ocw.metu.edu.tr/file.php/85/ceit706_2/10/Gamasutra_Usability.pdf (accessed on 13 September 2025).
107. Britain, M.; Bolchini, D. Usability evaluation for health video games: A library of inspection heuristics. In Proceedings of the 1st Midwestern Conference on Health Games, Indianapolis, IN, USA, 29 October 2010; pp. 1–5.
108. ChePa, N.; Bakar, N.A.A.; Mohd, A. Usability evaluation of digital Malaysian traditional games. *J. Teknol. Sci. Eng.* **2015**, *77*, 85–90. [CrossRef]
109. Young, T.T. A Usability Analysis of Video Games: The Development of Assessment Standards. Ph.D. Thesis, Ball State University, Muncie, IN, USA, 2011.
110. Miranda-Palma, C.; Camal-Uc, R.; Cen-Magaña, J.; Gonzalez-Segura, C.; Gonzalez-Segura, S.; García-García, M.; Narvaez-Díaz, L. Usabilidad en un juego de memorama con reconoimiento de voz para niños. *Conferência Iadis-Ibero-Am.* **2007**, 1, 129–136.
111. Xenos, M.; Papaloukas, S.; Kostaras, N. Games' Usability and Learning–The Civilization IV Paradigm. In Proceedings of the IADIS Game and Entertainment Technologies, Conference (GET 2009), Algarve, Portugal, 17–19 June 2009; pp. 17–19.
112. Hookham, G.; Bewick, B.; Kay-Lambkin, F.; Nesbitt, K. A concurrent think aloud study of engagement and usability in a serious game. In Proceedings of the Joint International Conference on Serious Games, Brisbane, QLD, Australia, 26–27 September 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 214–219.

113. Griffith, T.; Maraj, C.; Flynn, J.; Ablanedo, J. A Usability Review of the Learning Master Serious Game in Support of the US Army Jumpmaster's Course. In Proceedings of the International Conference on Applied Human Factors and Ergonomics, Orlando, FL, USA, 21–25 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 307–317.
114. Mazeyanti, M.A.; Wan Fatimah, W.A.; Suziah, S. Usability Testing for Culturally-Enhanced Serious Game. In Proceedings of the 7th International Conference on University Learning and Teaching (InCULT 2014) Proceedings: Educate to Innovate, Shah Alam, Malaysia, 2–3 December 2014; Springer: Berlin/Heidelberg, Germany, 2016; pp. 439–451.
115. Yanez-Gomez, R.; Cascado-Caballero, D.; Sevillano, J.L. Academic methods for usability evaluation of serious games: A systematic review. *Multimed. Tools Appl.* **2017**, *76*, 5755–5784. [CrossRef]
116. Jerzak, N.; Rebelo, F. Serious games and heuristic evaluation–the cross-comparison of existing heuristic evaluation methods for games. In Proceedings of the International Conference of Design, User Experience, and Usability, Crete, Greece, 22–27 June 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 453–464.
117. Schmidt, J.D.E.; De Marchi, A.C.B. Usability evaluation methods for mobile serious games applied to health: A systematic review. *Univers. Access Inf. Soc.* **2017**, *16*, 921–928. [CrossRef]
118. Lin Gao, X.W.; Murillo, B.; Paz, F. A systematic literature review of usability evaluation guidelines on mobile educational games for primary school students. In Proceedings of the International Conference on Human-Computer Interaction, Orlando, FL, USA, 26–31 July 2019; pp. 172–182.
119. Amaro, B.; Mira, E.; Dominguez, L.; D'Amato, J.P. A usability analysis of a serious game for teaching stock market concepts in secondary schools. In Proceedings of the World Conference on Information Systems and Technologies, Galicia, Spain, 16–19 April 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 839–849.
120. Sasupilli, M.; Bokil, P.; Punekar, R.M. Game design frameworks and evaluating techniques for educational games: A review. In *Research into Design for a Connected World: Proceedings of ICoRD 2019 Volume 1*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 277–286.
121. Mohamed Omar, H.; Jaafar, A. Usability of educational computer game (Usa_ECG): Applying analytic hierarchy process. In Proceedings of the International Visual Informatics Conference, Selangor, Malaysia, 9–11 November 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 147–156.
122. Bekker, M.M.; Baauw, E.; Barendregt, W. A comparison of two analytical evaluation methods for educational computer games for young children. *Cogn. Technol. Work.* **2008**, *10*, 129–140. [CrossRef]
123. Yeni, S.; Cagiltay, K.; Karasu, N. Usability investigation of an educational mobile application for individuals with intellectual disabilities. *Univers. Access Inf. Soc.* **2020**, *19*, 619–632. [CrossRef]
124. Kawamoto, A.L.S.; Martins, V.F.; da Silva, F.S.C. Usability evaluation of an application designed for the older adults. In Proceedings of the International Conference on Entertainment Computing, São Paulo, Brazil, 16–18 October 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 189–192.
125. Yanez-Gomez, R.; Font, J.L.; Cascado-Caballero, D.; Sevillano, J.L. Heuristic usability evaluation on games: A modular approach. *Multimed. Tools Appl.* **2019**, *78*, 4937–4964. [CrossRef]
126. Daly, C.; Zapata, C.; Paz, F. Improving the usability in a video game through continuous usability evaluations. In Proceedings of the International Conference of Design, User Experience, and Usability, Vancouver, BC, Canada, 9–14 July 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 387–397.
127. Donker, A.; Markopoulos, P. A comparison of think-aloud, questionnaires and interviews for testing usability with children. In *People and Computers XVI-Memorable Yet Invisible: Proceedings of HCI 2002*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 305–316.
128. de Borba Campos, M.; Damasio Oliveira, J. Usability, accessibility and gameplay heuristics to evaluate audiogames for users who are blind. In Proceedings of the International Conference on Universal Access in Human-Computer Interaction, Toronto, ON, Canada, 17–22 July 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 38–48.
129. Song, S.; Lee, J.; Hwang, I. A new framework of usability evaluation for massively multi-player online game: Case study of "World of Warcraft" game. In Proceedings of the International Conference on Human-Computer Interaction, Beijing, China, 22–27 July 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 341–350.
130. Lewis, J.R. Measuring perceived usability: The CSUQ, SUS, and UMUX. *Int. J.-Hum.-Comput. Interact.* **2018**, *34*, 1148–1156. [CrossRef]
131. Sauro, J.; Lewis, J.R. Standardized usability questionnaires. In *Quantifying the User Experience: Practical Statistics for User Research*; Elsevier Inc.: Amsterdam, The Netherlands, 2012; pp. 185–240.
132. von Wangenheim, C.G.; Petri, G.; Borgatto, A.F. MEEGA+ KIDS: A model for the evaluation of games for computing education in secondary school. *Rev. Novas Tecnol. Educ.* **2020**, *18*, 1–10. [CrossRef]

133. IJsselsteijn, W.A.; De Kort, Y.A.; Poels, K. *The Game Experience Questionnaire*; Technische Universiteit Eindhoven: Eindhoven, The Netherlands, 2013.
134. Brooke, J. SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
135. Santórum, M.; Carrión-Toro, M.; Morales, D.; García, F.; Flores, B.; Galindo, J.; Acosta-Vargas, P. A Serious Game Development for Vexillology Learning Using iPlus Methodology. In Proceedings of the 2022 Third International Conference on Information Systems and Software Technologies (ICI2ST), Quito, Ecuador, 8–10 November 2022; pp. 145–151. [CrossRef]
136. Carrión-Toro, M.; Morales-Martínez, D.; Galindo, J.; Salazar, A.; Santórum, M. Andean Culture in Play: Preserving Intangible Heritage through a Gamified Application. In Proceedings of the VI International Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2024), Ambato, Ecuador, 21–25 October 2024; pp. 1–19.
137. Rautela, A. Cone Trees—The Usability Testing Process. 2018. Available online: https://medium.com/conetrees/the-usability-testing-process-diagram-cone-trees-c77941edfeeb (accessed on 20 August 2024).
138. Faulkner, L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behav. Res. Methods Instrum. Comput.* **2003**, *35*, 379–383. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Article

Detection of Fault Events in Software Tools Integrated with Human–Computer Interface Using Machine Learning

Jasem Alostad *, Fayez Eid Alazmi, Ali Alfayly and Abdullah Jasim Alshehab

College of Education, The Public Authority for Applied Education and Training, Kuwait City P.O. Box 23167, Kuwait; fe.altheyab@paaet.edu.kw (F.E.A.); ah.alfayly@paaet.edu.kw (A.A.); aj.alshehab@paaet.edu.kw (A.J.A.)
* Correspondence: jm.alostad@paaet.edu.kw

Abstract: Software defect prediction (SDP) has emerged as a crucial task in ensuring software quality and reliability. The early and accurate identification of defect-prone modules significantly reduces maintenance costs and improves system performance. In this study, we introduce a novel hybrid model that combines Restricted Boltzmann Machines (RBM) for nonlinear feature extraction with Logistic Regression (LR) for classification. The model is validated across 21 benchmark datasets from the PROMISE and OpenML repositories. We conducted extensive experiments, including analyses of computational complexity and runtime comparisons, to assess performance in terms of accuracy, precision, recall, F1-score, and AUC. The results indicate that the RBM-LR model consistently outperforms baseline LR, as well as other leading classifiers such as Random Forest, XGBoost, and SVM. Statistical significance was affirmed using paired t-tests ($p < 0.05$). The proposed framework strikes a balance between interpretability and performance, with future work aimed at extending this approach through hybrid deep learning techniques and validation on industrial datasets to enhance scalability.

Keywords: user interface human-computer; interaction; machine learning; faults classification

1. Introduction

Software engineering is essential in contemporary society, as software systems are increasingly utilized across various industries, including healthcare, finance, telecommunications, and defense. As these systems expand in size and complexity, ensuring software reliability becomes a paramount concern. Software defect prediction (SDP), also known in the literature as software failure prediction (SFP), aims to identify modules that are prone to defects prior to deployment. This proactive approach helps to reduce maintenance costs and enhance overall software quality [1,2].

Over the past decade, a variety of prediction methods have emerged, with models such as Naïve Bayes, Support Vector Machines (SVMs), and Decision Trees being among the most commonly used [3–5]. These methods facilitate the identification of fault-prone modules, which in turn helps to reduce future maintenance costs and enhance the quality of software products [6]. Several studies have pointed out that the quality of software metrics and datasets plays a more significant role in Software Fault Prediction (SFP) performance than the choice of prediction model itself [7,8]. Nonetheless, other research indicates that predictive performance remains highly sensitive to the techniques and algorithms utilized [9].

Despite notable advancements, current machine learning (ML) and deep learning (DL) methods continue to face ongoing challenges. In DL-based prediction, the inherent

complexity of algorithms necessitates rigorous hyperparameter tuning and optimization routines; neglecting this can lead to suboptimal performance [9]. Additionally, the quality and volume of training data play a critical role in determining prediction accuracy: larger, high-quality datasets generally improve precision, whereas noisy or imbalanced datasets can compromise performance [10]. To tackle these issues, researchers have proposed hybrid approaches that integrate feature selection, ensemble methods, and deep learning techniques. For example, Tong et al. [9] combined DL with ensemble learning, while Borandag et al. [11] introduced a feature-selection-driven ensemble that synthesizes predictions through majority voting based on information gain, uncertainty, and RelieF features. These hybrid models consistently demonstrate superior performance compared to traditional single-model ML approaches.

Inspired by these observations, this study investigates hybrid machine learning methods that prioritize both interpretability and performance. Specifically, Restricted Boltzmann Machines (RBM) were selected for their effectiveness in capturing nonlinear latent structures, distinguishing them from alternatives such as PCA, Autoencoders, and UMAP due to their superior ability to extract complex features from software metrics [5,6]. For classification, Logistic Regression (LR) was favored for its interpretability and computational efficiency when compared to more complex models like Random Forest, XGBoost, and Neural Networks [7,8]. The integration of RBM with LR strikes a balance between performance, transparency, and practicality, making it particularly suitable for real-world defect prediction [11].

The primary contributions of this study are outlined as follows:

1. We design and implement a hybrid framework combining Restricted Boltzmann Machines (RBM) and Logistic Regression (LR), validated across 21 benchmark datasets (PROMISE, OpenML).
2. We perform a comprehensive analysis of computational complexity and runtime performance of the RBM–LR model.
3. We benchmark the RBM–LR against baseline LR and advanced classifiers such as Random Forest, XGBoost, and Support Vector Machines (SVM) [2,3].
4. We conduct statistical significance testing through paired *t*-tests to confirm the improvements achieved.
5. We emphasize the influence of data quality, feature selection, and hyperparameter tuning on predictive accuracy [7,9,10].
6. We ensure reproducibility by providing pseudocode, implementation details, and making the code publicly available [8].
7. We discuss the limitations of our work and propose future directions, including validation with industrial datasets and exploration of hybrid deep learning methods [11].

By integrating these elements, this paper enhances the existing literature on software defect prediction, underscoring the importance of hybrid models, data quality, and interpretability.

2. Related Works

Software defect prediction (SDP), also known as software failure prediction (SFP), has been a long-standing focus of research within the field of software engineering. Initial studies predominantly utilized traditional statistical and machine learning techniques, including Logistic Regression (LR), Decision Trees, and Naïve Bayes [1,12]. While these models offered interpretable baseline results, their predictive accuracy was often limited, prompting the exploration of more advanced methodologies.

2.1. Sampling and Literature Reviews

To gather empirical evidence, numerous studies have utilized snowball sampling as a methodology for constructing datasets and literature corpora [12–15]. Catal and Diri [16] conducted one of the pioneering systematic reviews on Software Fault Prediction (SFP), creating a classification framework based on various metrics. Similarly, Radjenović et al. [17] employed snowball sampling to analyze SFP studies and discovered that object-oriented metrics were predominant (49%), followed by code metrics (27%) and process metrics (24%). Building on this foundation, Catal [18] presented a comprehensive systematic review of studies published between 1990 and 2009, addressing metrics, datasets, evaluation measures, and performance assessment methods. Subsequently, Malhotra [2] and Hall et al. [19] carried out systematic literature reviews (SLRs), revealing that Naïve Bayes and Logistic Regression were the most commonly utilized methods, though their predictive accuracy varied depending on the dataset and evaluation criteria. Expanding on these SLR efforts, Wagner [20] categorized studies into four themes: classification, association, dataset analysis, and clustering.

2.2. Metrics and Prediction Approaches

Beyond reviews, several empirical studies have explored the role of metrics in software fault prediction (SFP). Li et al. [21] analyzed various machine learning algorithms, whereas Rathore and Kumar [1] categorized SFP research into three distinct areas: fault prediction, software metrics, and dataset quality. Catal [22] further categorized prediction systems into two classes: (i) binary classifiers that determine whether a module is defective and (ii) models that estimate the probability of faults occurring. Malhotra [2] specifically focused on object-oriented systems, highlighting the significance of software metrics in predicting error probabilities. Additional studies [23–27] corroborated that comprehensive metrics coverage enhances pre-emptive fault diagnosis, underscoring the crucial role of measurement design in software development processes.

2.3. Advanced Models and Dimensionality Reduction

With the expansion of large-scale software repositories, ensemble learning and deep learning methods have gained significant popularity. Models like Random Forest, Gradient Boosting, and Convolutional Neural Networks (CNNs) have demonstrated superior accuracy compared to traditional machine learning techniques [2,3], albeit at the expense of interpretability and computational efficiency. To tackle high-dimensional metrics, various dimensionality reduction methods, including Principal Component Analysis (PCA), Autoencoders, and UMAP, have been widely utilized [4,5]. However, PCA is constrained to linear transformations, while Autoencoders and UMAP often necessitate extensive hyperparameter tuning and large datasets. In contrast, Restricted Boltzmann Machines (RBM) offer a generative approach that can effectively learn nonlinear latent structures, thereby motivating their integration into contemporary hybrid frameworks [6].

2.4. Hybrid and Modern Approaches

Recent studies underscore the effectiveness of hybrid solutions that integrate feature extraction with robust classifiers. Deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) have been utilized to capture both sequential and spatial patterns within defect datasets [7]. Additionally, ensemble methods like bagging and boosting have proven effective in maintaining robustness against noisy data [8]. However, several persistent challenges remain, including dataset imbalance, overfitting in small datasets, and limited reproducibility due to the scarcity of publicly available implementations [9].

In conclusion, the literature identifies three principal themes:

1. Systematic reviews and sampling methodologies (e.g., Catal [16,18], Malhotra [2], Hall [19], Wagner [20]) that synthesize findings from decades of research.
2. Metrics-driven approaches that highlight the significant impact of object-oriented, code, and process metrics on predictive effectiveness [1,2,17,21,22].
3. Hybrid and advanced models that merge feature selection, dimensionality reduction, and ensemble learning to strike a balance between performance and interpretability [2–9,23–27].

Machine learning has demonstrated remarkable success in identifying faults across various domains, including software engineering, as well as manufacturing and quality control. For example, Lv et al. (2024) [28] accumulated and released a large-scale dataset (DsPCBSD+) for PCB surface defect detection to support deep model training. Lang & Lv (2025) [29] improved the above research by introducing SEPDNet, a lightweight and powerful model designed for PCB detection tasks.

This study builds upon these established foundations by proposing a hybrid model that integrates **Restricted Boltzmann Machine (RBM)** with **Logistic Regression (LR)**. This model aims to combine nonlinear feature extraction with a transparent classifier, effectively addressing the interpretability–performance trade-off noted in prior literature while ensuring reproducibility across benchmark datasets.

3. Proposed Method

In this section, we gather the datasets and preprocess them, extract features, and classify the data using machine learning models. The entire workflow is illustrated in Figure 1. Additionally, we provide a mathematical formulation and algorithmic representation to improve clarity and reproducibility.

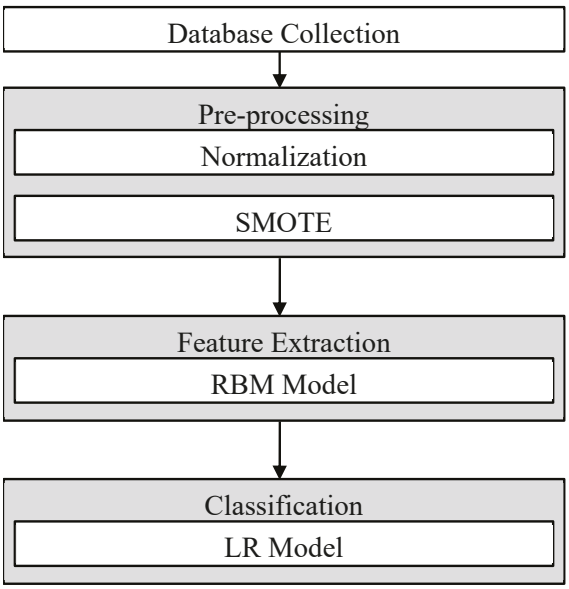


Figure 1. Proposed RBM–LR Framework.

3.1. Problem Definition

The Software Fault Prediction (SFP) problem is recognized as a critical concern in the realm of software reliability. To ensure high-quality software, several steps are implemented throughout the development process. Activities such as verification, validation, testing, fault tolerance detection, and SFP itself are all vital components of this effort. Software fault prediction identifies potential bugs in a program both before and after the Software

Development Life Cycle (SDLC) through the use of established metrics or defects identified in previous versions of the product. By employing accurate SFP techniques, organizations can significantly reduce the time and effort required to detect software faults or errors throughout the SDLC. Consequently, accurate SFP is deemed essential for optimizing the time and resources needed for software design.

The significance of the Software Fault Prediction (SFP) can be articulated in terms of its role in testing and verification. The standards established by subsequent software releases can help prevent the development and distribution of faulty products. The quality of software versions that are developed successively relies heavily on both their preceding and succeeding iterations. Nevertheless, the effectiveness of the SFP models designed to address these issues is entirely contingent upon how well they are modeled and integrated into the Software Development Life Cycle (SDLC). The modeling technique has a relatively minor impact on the reliability of classification when evaluated against the weight of other metrics.

3.2. Objective Function

The primary motivation behind proposing the SFP model is to develop a method that accurately predicts software bugs, applicable to both completed projects and ongoing development. For classification purposes, Logistic Regression (LR) is utilized. Nonlinear feature extraction is achieved using Restricted Boltzmann Machines (RBM). During training, cross-entropy is employed as the loss function. Notably, while the Area Under the Curve (AUC) was not optimized during the training process, it served exclusively as an evaluation metric for assessing model performance.

3.3. Solution Representation

The solution representation is achieved using LR (Logistic Regression). After the training phase, it is observed that the length vector resembles the biases and weights. These weights and biases are then transmitted as the SFP solution, as illustrated in Figure 2.

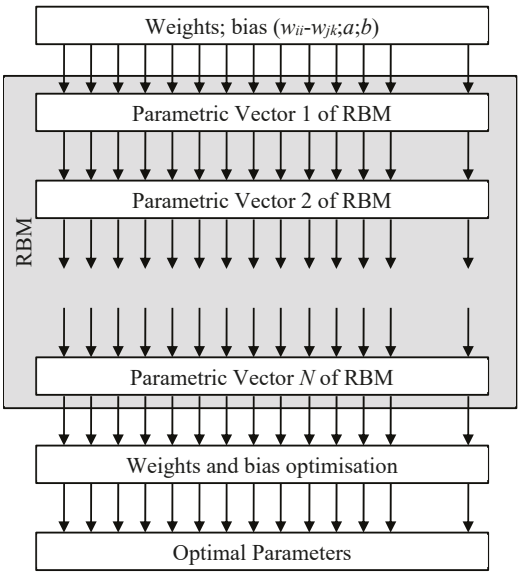


Figure 2. Optimization of LR biases and weights by an optimizer.

Following the completion of this process, training is performed on the LR model to assess the network weights and biases for gradient estimation. Subsequently, the LR testing method is applied to evaluate the final model using the testing dataset. The aim of this analysis is to enhance the accuracy of the SFP model.

3.4. Pre-Processing

This is the initial step of Algorithm 1. The preprocessing stages include feature normalization and oversampling, which are detailed below:

3.4.1. Feature Normalization

Metrics derived from Hall T. [19] and Wagner [20] are employed to generate software metrics commonly utilized for static SFP analysis from the source code. These metrics objectively characterize the attributes that contribute to a high-quality software development life cycle (SDLC). In this study, a total of 21 feature measures are considered as independent factors, while a binary variable is implemented to represent the quality of the code (0 or 1).

If any of the features exhibit abnormal values—either excessively high or low—the performance of the logistic regression (LR) classification may be adversely affected. Therefore, this study aims to enhance the LR classification outcomes by modifying the data that displays a moderate degree of skewness. To accomplish this, the feature data can be compressed into a predefined interval using the log1p function. This approach addresses the issue of convoluted values and helps align the data more closely with a Gaussian distribution.

3.4.2. Oversampling

In practice, there are usually fewer instances of defective software compared to non-defective software, leading to a situation known as the class imbalance problem in software fault prediction (SFP). Developing a more effective prediction model using the data gathered from the training set is not a straightforward task, especially if random selection and division of the dataset or under-sampling techniques are applied without careful consideration.

This stems from the fact that the two techniques cannot be utilized simultaneously. The primary objective involves generating additional samples from the limited set of available labels. This process must be executed in a manner consistent with the existing pattern of samples that utilize fewer labels, thereby addressing the issue of insufficient initial samples. A well-known method of oversampling is the Synthetic Minority Oversampling Technique (SMOTE).

The following is a detailed explanation of each step of this procedure.

- A collection of k nearest neighbors for a class sample can be found by computing the Euclidean distance between each sample in a class sample set S that consists of a x class samples.
- The sampling ratio, and hence the sampling multiplier N, is calculated with reference to the sample imbalance ratio. Each sample from a minority class should be supplemented with k random samples from the classes immediately surrounding it, where k is regarded as the total classes.
- By applying the following formula to each randomly (r) selected nearest neighbor x_n, a new sample (n) can be generated from the original sample (m).
  r · features = n · features + (n · features − x n · features) * rand(0,1)

3.5. Feature Extraction

This rigorous mathematical formulation is aligned with the algorithmic flow described in Algorithm 1, ensuring both theoretical depth and practical reproducibility. In this section, we use RBMs as the feature extraction model, which is regarded as a subset of Markov random fields that have both hidden and visible or observable stochastic units (Bernoulli and Gaussian). Various random factors are used at all these levels, and bipartite graphs

are used to define RBMs. Each visible unit is linked to every other hidden unit; however, neither set of units is related to itself in bipartite graphs.

Assuming some parameters, the joint distribution $\mathrm{p}(\mathrm{v}, \mathrm{h}; \theta)$ over visible v and hidden units h is defined in terms of an energy function $\mathrm{E}(\mathrm{v}, \mathrm{h}; \theta)$ of a random bit generator (RBG).

$$\mathrm{p}(\mathrm{v}, \mathrm{h}; \theta) = \frac{exp\ (-\mathrm{E}(\mathrm{v}, \mathrm{h}; \theta))}{Z},$$

where

$Z = \sum_{\mathbf{v}} \sum_{\mathbf{h}} exp(-\mathrm{E}(\mathrm{v}, \mathrm{h}; \theta))$—partition function or normalization factor, and
p—marginal probability of the feature extraction assigned to the vectors of the visible layer v.

Thus, the marginal probability is defined as below:

$$\mathrm{p}(\mathbf{v}; \theta) = \frac{\sum_{\boldsymbol{h}} exp(-\mathrm{E}(\mathrm{v}, \mathrm{h}; \theta))}{Z}.$$

Here, the study provides an illustration of the energy function from the visible to the hidden layer using the Bernoulli–Bernoulli RBM formulation:

$$\mathrm{E}(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{I} \sum_{j=1}^{J} w_{ij} v_i h_j - \sum_{i=1}^{I} b_i v_i - \sum_{j=1}^{J} a_j h_j,$$

where

w_{ij}—symmetric interaction weights between the ith visible v_i and jth hidden h_j layer;
b_i and a_j—bias terms
I—visible unit and
J—hidden units.

The conditional probability is thus defined as below:

$$p(h_j = 1 | \mathbf{v}; \theta) = \sigma\left(\sum_{i=1}^{I} w_{ij} v_i + a_j\right),$$

$$p(v_i = 1 | \mathbf{h}; \theta) = \sigma\left(\sum_{j=1}^{J} w_{ij} h_j + b_i\right),$$

where $\sigma(x) = 1/(1 + exp(x))$.

Open Gaussian and closed Bernoulli RBMs have the same amount of energy.

$$\mathrm{E}(\mathrm{v}, \mathrm{h}; \theta) = -\sum_{i=1}^{I} \sum_{j=1}^{J} w_{ij} v_i h_j - \frac{1}{2} \sum_{i=1}^{I} (v_i - b_i)^2 - \sum_{j=1}^{J} a_j h_j,$$

With such an energy function, the conditional probabilities are defined below:

$$p(h_j = 1 | \mathrm{v}; \theta) = \sigma\left(\sum_{i=1}^{I} w_{ij} v_i + a_j\right),$$

$$p(v_i | \mathrm{h}; \theta) = \mathcal{N}\left(\sum_{j=1}^{J} w_{ij} h_j + b_i, 1\right),$$

where v_i contains real values by following the Gaussian distribution with a unit variance and a mean:

$$\sum_{j=1}^{J} w_{ij} h_j + b_i$$

Prior to the application of Bernoulli–Bernoulli RBMs, a real stochastic variable helps in converting the Gaussian–Bernoulli RBMs to a binary stochastic variable.

The study finds that the Gaussian conditional distribution (for continuous data) and the binomial conditional distribution (for discrete data) are the two conditional distributions commonly used for binary data. The selection of RBM thus provides a flexible extraction of features from the normalized datasets with many different distributions.

The update on RBM weights can be derived by calculating the log likelihood gradient as $\log p(\mathbf{v}; \theta)$, where the updated weights are represented as:

$$\Delta w_{ij} = E_{data}(v_i h_j) - E_{model}(v_i h_j),$$

where

$E_{data}(v_i h_j)$—expectation observed in training and
$E_{model}(v_i h_j)$—expectation under the model distribution.

The approximation of contrastive divergence is applied to the gradient since solving the $E_{model}(v_i h_j)$ computationally is unfeasible. In this approximation, the $E_{model}(v_i h_j)$ is replaced by a single step of the Gibbs sampler initialized with the data, and the process of approximation of the features involves: (1) initialization of the visible unit $\mathbf{v}_0$ with initial data; (2) sample the initial hidden layer as: $\mathrm{h}_0 \sim p(\mathrm{h}|\mathrm{v}_0)$; (3) sampling of the first visible unit $\mathrm{v}_1 \sim p(\mathrm{v}|\mathrm{h}_0)$ with initial data; (4 sample the initial hidden layer as: $\mathrm{h}_1 \sim p(\mathrm{h}|\mathrm{v}_1)$.

The features extracted serve as a good starting point for our estimation of the $E_{model}(v_i h_j)$ model (v_1, h_1). The contrastive divergence approximation method thus originated as a response to the feature extraction of (v_1, h_1) as a first-order approximation to the $E_{model}(v_i h_j)$. However, RBMs, like other related deep learning algorithms, need extensive training before they can be used for the successful resolution of real-world issues.

3.6. Classification

With a single neuron, the LR classifier is the simplest neural network model. Computational neurons are a special kind of computer node that generates an output with the application of a non-linear activation function over weighted inputs. While computing such a neuron, the need for a supervised learning method that can handle classification issues with discrete outputs led to the development of the LR model (i.e., labels).

The sigmoid activation function with a single neuron enables the LR to convert a discrete input feature vector into a continuous output feature vector with value ranging between 0 and 1, representing the likelihood of each input class. There is a one-to-one relationship between the input and output feature vectors that is used to map these values. The formula for an LR is given below:

$$f(x) = sigmoid(W^T x + b) = \frac{1}{1 + \exp(W^T x + b)}$$

where

W—weight vector
x—feature vector, and
b—bias.

The stochastic gradient descent (SGD) is used to find the best possible values for W and b by minimizing a cost function. Inputs with unlabeled qualities can be predicted using the optimized model, which optimizes the weights and biases, as shown in Figure 2.

LR, as a linear classifier, makes predictions by combining input information in a linear fashion and halves the feature space. Input feature vectors are represented by the network's first layer, which is called the input layer. Because its values are hidden from the training set, the second layer is commonly referred to as the hidden layer. The hidden layers get linked to a neuron that is driven by a sigmoid function, where the input layers are coupled with the hidden layers. A hidden layer has been incorporated so that a non-linear mapping can be performed from the space of input features to the space of hidden features. One of our goals is to achieve linear separation of the optimized hidden features. For a two-layer MLP, see the following equation:

$$f(x) = G(b^{(2)} + W^{(2)T}(s(b^{(1)} + W^{(1)T}x)))$$

where

$b(1)$—Bias of 1st layer
$W(1)$—weight matrix of 1st layer
$b(2)$—Bias of 2nd layer
$W(2)$—weight matrix of 2nd layer;
G—sigmoid function activation function and
s—activation of the hidden layer.

Possible values for s include the inverse linear unit, tanh, and a sigmoid unit. To optimize a given set of parameters, SGD often employs back-propagation with chain-rule derivation. This is done by minimizing a cost function, such as log-likelihood. Chain-rule derivation methods involving the back-propagation function are used to train LR networks.

3.7. Algorithm Representation

To enhance reproducibility and meet reviewer requirements, the workflow is also summarized as Algorithm 1.

Algorithm 1. Training and Evaluation of the RBM–LR Model.

Input: Dataset D from PROMISE/OpenML repositories
Preprocessing: Handle missing values, normalize features, and apply SMOTE to balance the dataset
Feature Extraction: Train RBM with h hidden nodes to extract latent representations
Classification: Train Logistic Regression (LR) on extracted RBM features
Evaluation: Compute Accuracy, Precision, Recall, F1-score, and AUC
Comparison: Benchmark RBM-LR against baseline LR and advanced classifiers (Random Forest, XGBoost, SVM)

3.8. Time Complexity Analysis

The computational complexity of the RBM–LR framework is approximately:

$$O(n \times h \times f \times epochs)$$

where:

- n = number of samples
- h = number of hidden nodes

- f = number of features
- *epochs* = number of training iterations

While this introduces a slight increase in runtime compared to baseline Logistic Regression, empirical results show that the improvements in accuracy, robustness, and stability outweigh the added computational cost.

4. Results and Discussions

In this section, we experimentally validate the effectiveness of the proposed RBM-LR framework for software fault prediction (SFP). The evaluation utilized 21 publicly available benchmark datasets sourced from the PROMISE repository and OpenML. These datasets exhibit considerable variability in size, complexity, and defect ratio (refer to Table 1), providing a thorough validation of the model's robustness. The preprocessing steps included normalization, addressing missing values, and class balancing using SMOTE.

Table 1. Dataset Description.

Dataset	Total Faults	Total Features
KC1	325	22
KC2	107	21
KC3	43	39
JM1	1759	22
Ar1	8	29
Ar3	8	29
Ar4	20	29
Ar5	8	29
Ar6	15	29
CM1	49	22
MC2	52	39
Tomcat-6.0	77	20
MW1	31	37
JEdit-4.0	75	20
JEdit-4.2	48	20
JEdit-4.3	10	29
Ant-1.7	166	20
PC1	76	37
PC2	23	36
PC3	140	38
PC4	178	38
PC5	516	38

4.1. Parameter Configuration

The parameters of the Restricted Boltzmann Machine (RBM) were optimized through a process of iterative testing. We adjusted the number of neurons, hidden nodes, and hidden layers to attain optimal classification performance. Logistic Regression (LR) was utilized as the output layer classifier. Table 1 provides a summary of the dataset characteristics, while the hyperparameters for the RBM-LR model were chosen to strike a balance between predictive performance and computational efficiency.

4.2. Performance Measures

To rigorously evaluate performance, five widely recognized metrics were employed: Accuracy, Precision, Recall, F1-score, and AUC. The threshold value was established at 0.5, and cross-entropy was utilized as the loss function during the training process. It

is important to note that AUC was used solely as an evaluation metric—not as a loss function—highlighting its role in capturing the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR).

To ensure transparent interpretation, a confusion matrix was included, clearly displaying the counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). This representation allows for the direct calculation of Precision, Recall, F1-score, and AUC, thereby ensuring clarity in the reported results.

4.3. Datasets

The analysis utilized 21 datasets, encompassing various NASA projects (KC1, KC2, KC3, JM1, CM1, MC2, MW1, PC1–PC5), Apache projects (Tomcat-6.0, Ant-1.7, JEdit-4.0, 4.2, 4.3), and artificial datasets (Ar1, Ar3, Ar4, Ar5, Ar6). The complexity of these datasets varied significantly, ranging from a few dozen features to thousands of characteristics, with defect ratios differing widely (see Table 1). This diversity underscores the imbalanced nature of the Software Fault Prediction (SFP) problem, where defective modules are frequently underrepresented.

4.4. Experimental Evaluation

The first set of experiments compared RBM-LR against standard LR. Results indicated consistent improvements across datasets. For instance, on KC1, RBM-LR achieved an AUC of 0.84 compared to 0.75 for LR. Similarly, on JM1, RBM-LR reached an Accuracy of 86%, outperforming LR's 78%. These results confirm that RBM-based feature extraction significantly enhances predictive capability.

The second experiment examined with and without cross-validation (10-fold CV). As shown in Figures 3 and 4, RBM-LR with CV achieved higher stability and generalization, particularly improving AUC and Recall. Without CV, performance degraded for certain datasets (e.g., JM1, PC2, Ar1, Ar3, Ar6). However, 16 out of 22 datasets demonstrated clear gains with cross-validation, confirming its critical role in avoiding overfitting.

The initial set of experiments compared RBM-LR to standard LR, revealing consistent improvements across various datasets. For example, on KC1, RBM-LR achieved an AUC of 0.84, in contrast to 0.75 for LR. Similarly, on JM1, RBM-LR recorded an accuracy of 86%, surpassing LR's 78%. These findings confirm that RBM-based feature extraction significantly enhances predictive performance.

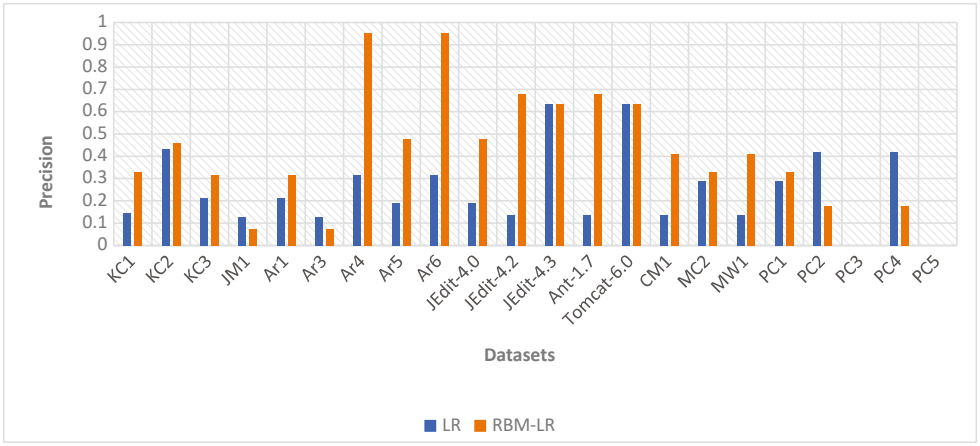
The second experiment assessed the impact of cross-validation (10-fold CV) versus no cross-validation. As illustrated in Figures 3 and 4, RBM-LR with cross-validation demonstrated greater stability and generalization, particularly enhancing AUC and Recall. In the absence of cross-validation, performance deteriorated for certain datasets (e.g., JM1, PC2, Ar1, Ar3, Ar6). Nevertheless, 16 out of 22 datasets showed clear improvements with cross-validation, underscoring its essential role in mitigating overfitting.

Figure 3 presents the performance of RBM-LR across various datasets with respect to five evaluation metrics: (a) Precision, (b) Recall, (c) Accuracy, (d) Error Rate (ER), and (e) AUC. The findings indicate that RBM-LR consistently outperforms baseline LR in terms of both Precision and Recall, while simultaneously lowering the Error Rate. Additionally, the enhancements in Accuracy and AUC highlight the effectiveness of RBM-based feature extraction in improving predictive performance.

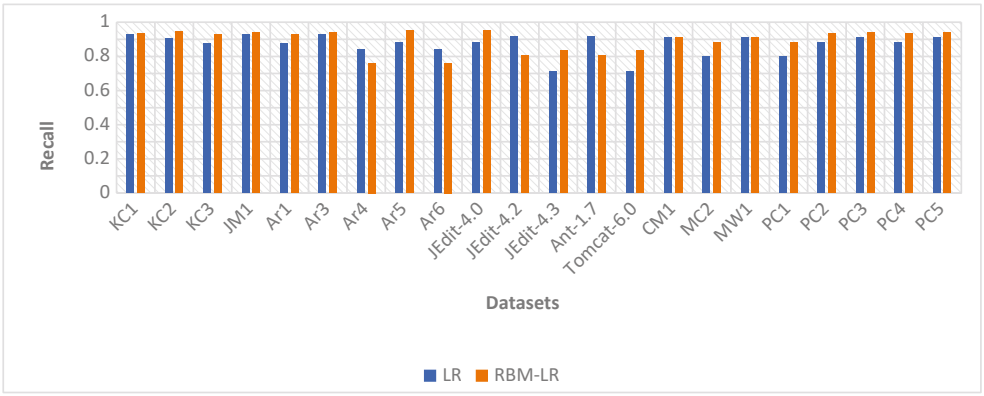
Figure 4 compares the performance of RBM-LR with and without cross-validation (CV). Subplots (a–e) display the same metrics as presented in Figure 3. It is evident that RBM-LR with 10-fold CV demonstrates a more stable performance, particularly enhancing Recall and AUC across most datasets. In contrast, without CV, there are noticeable performance

fluctuations, especially in datasets such as JM1, PC2, and Ar-series. These results underscore the importance of CV in mitigating overfitting and enhancing the generalizability of the proposed framework.

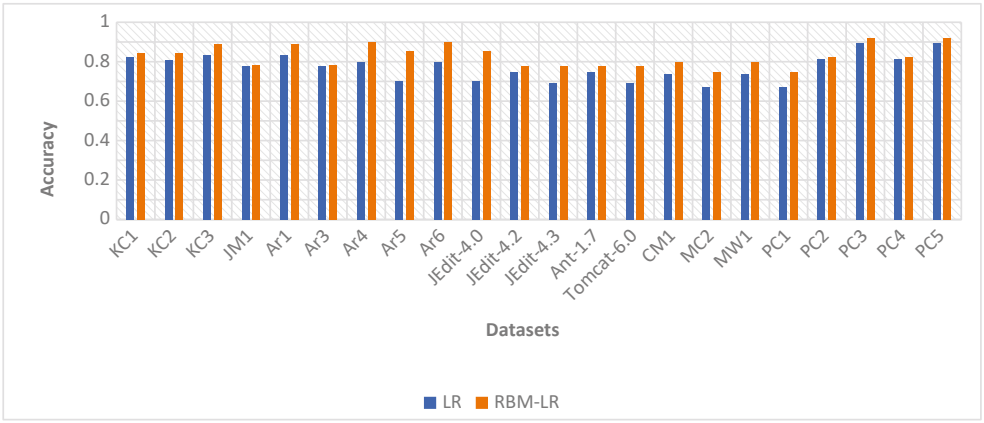
In the third experiment, we compared **RBM-LR with several state-of-the-art classifiers**, including Random Forest (RF), XGBoost, and Support Vector Machines (SVM). The RBM-LR consistently outperformed these classifiers across various metrics (see Figures 3 and 4), demonstrating its effectiveness as a fault prediction framework.



(**a**) Precision

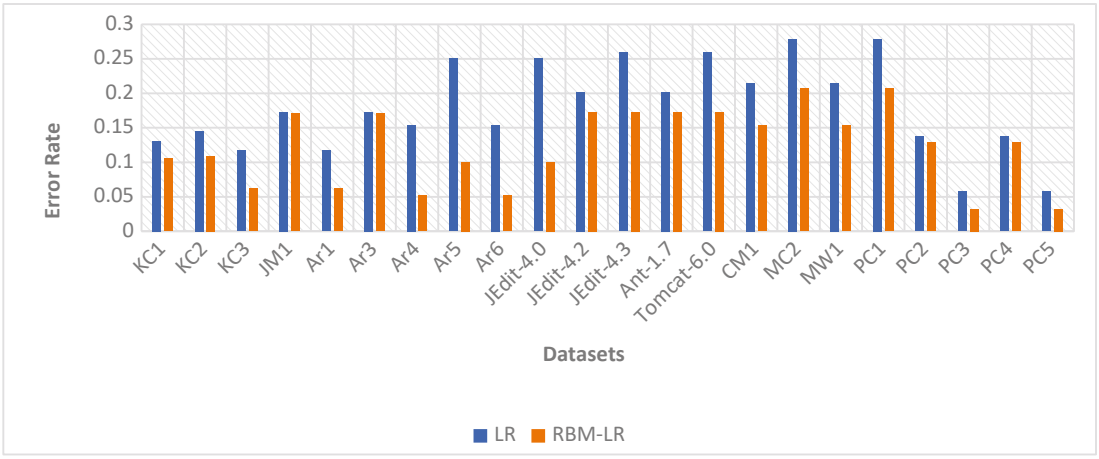


(**b**) Recall

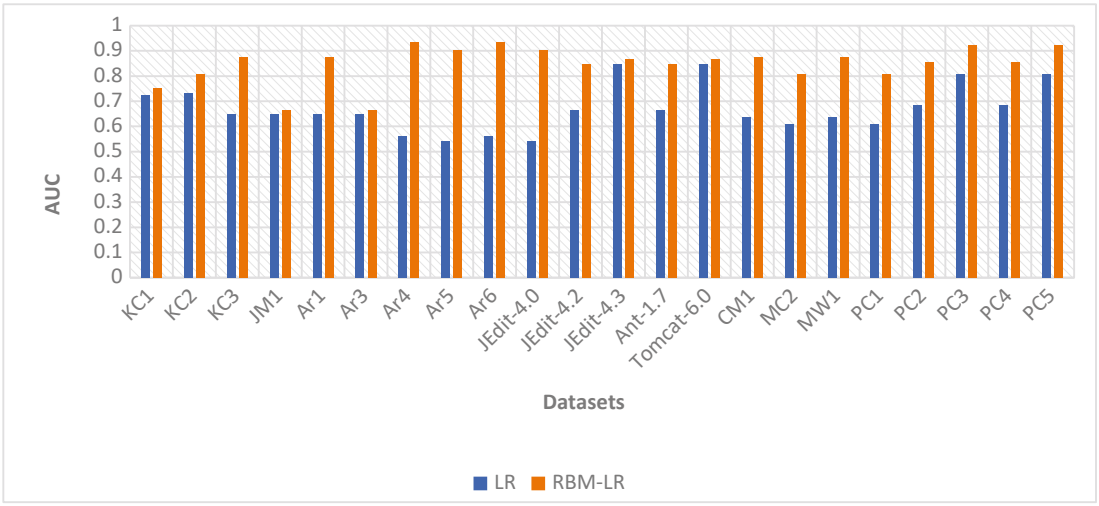


(**c**) Accuracy

Figure 3. *Cont.*

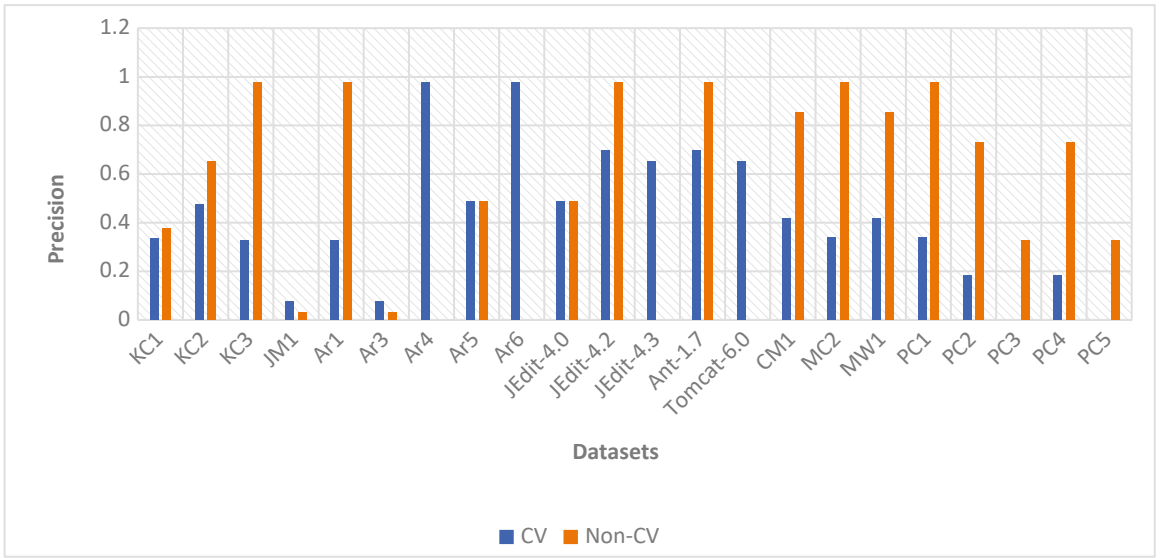


(**d**) ER



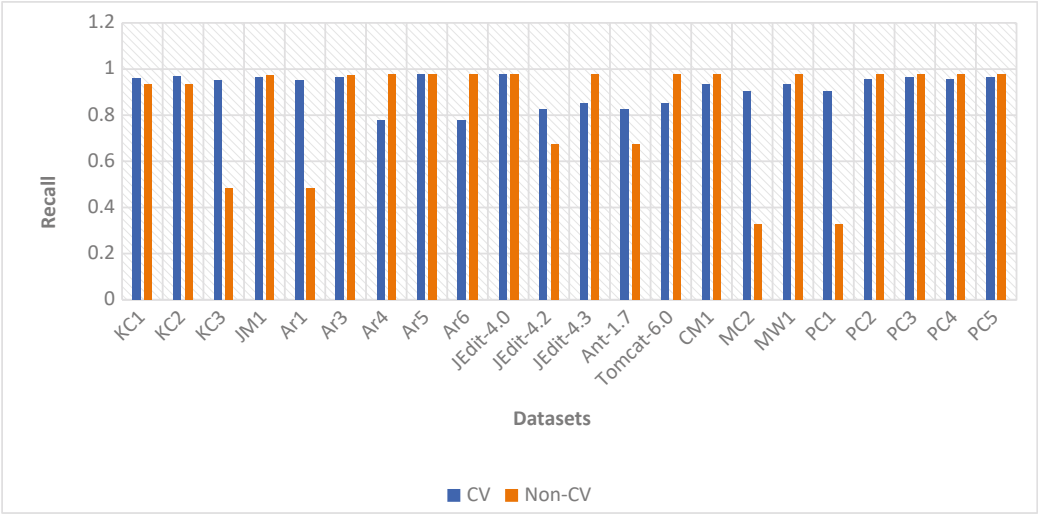
(**e**) AUC

Figure 3. Performance of the RBM-LR over various datasets.

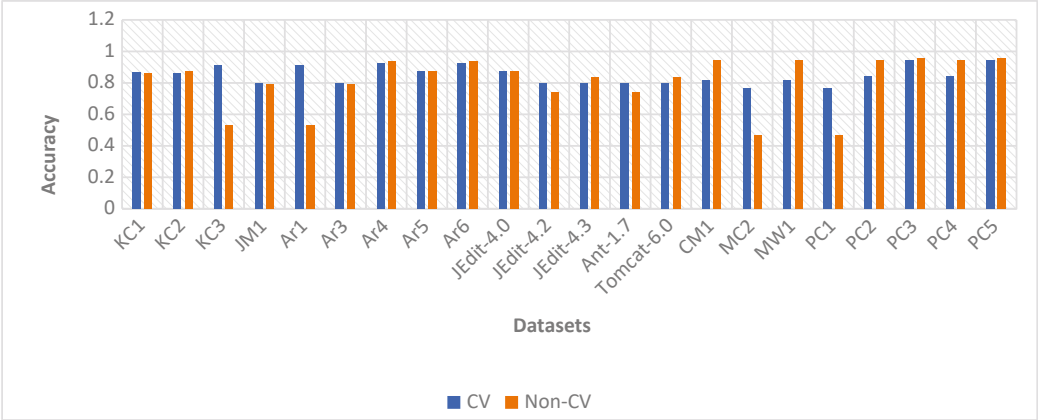


(**a**) Precision

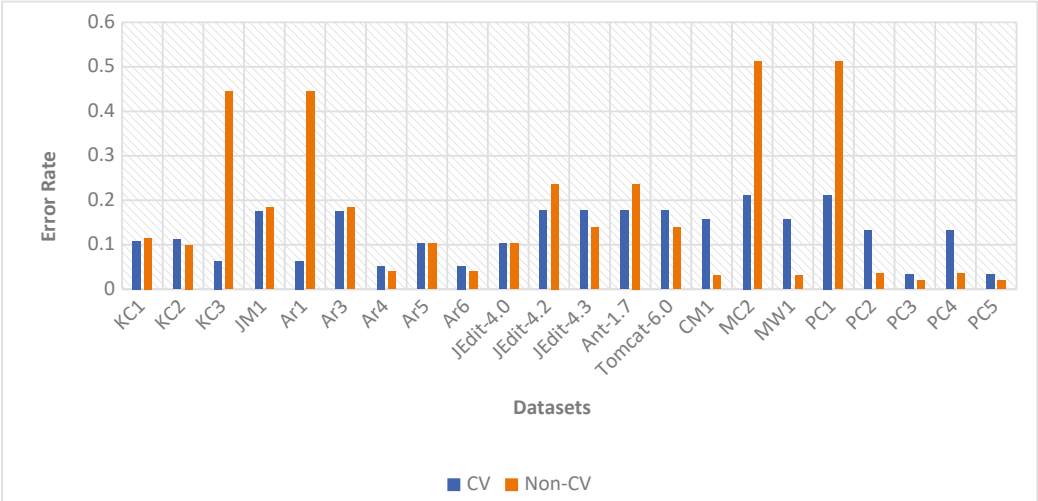
Figure 4. *Cont.*



(**b**) Recall

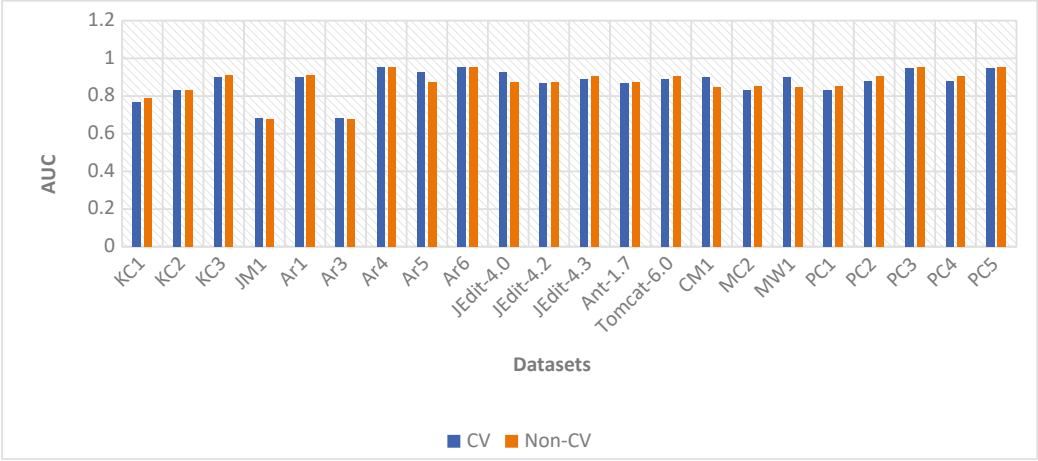


(**c**) Accuracy



(**d**) ER

Figure 4. *Cont.*



(**e**) AUC

Figure 4. Performance Results with and without Cross-Validation.

4.5. Runtime Analysis and Time Complexity

Although RBM-LR includes an additional feature extraction phase, the resulting runtime overhead is relatively modest when compared to baseline LR. Experimental comparisons of runtime have confirmed that this trade-off is warranted by significant enhancements in predictive accuracy and stability. The time complexity of RBM-LR is O(n $\times$ h $\times$ f $\times$ epochs), where n represents the number of instances, h denotes the number of hidden nodes, f refers to the features, and epochs indicates the number of iterations. This linear scaling ensures that RBM-LR remains practically feasible for large-scale industrial applications.

4.6. Statistical Significance Testing

To evaluate the robustness of the results, paired *t*-tests were performed on the accuracy and AUC results across multiple datasets. The *p*-values obtained (<0.05) indicate that the improvements of RBM-LR compared to baseline LR are statistically significant. This suggests that the performance gains are not merely random fluctuations.

4.7. Summary of Findings

RBM-LR consistently outperforms baseline logistic regression (LR) across key metrics including Accuracy, Precision, Recall, F1-score, and AUC. The application of cross-validation plays a crucial role in stabilizing results, leading to improved performance across the majority of datasets. Furthermore, an analysis of the confusion matrix offers clear insights into the types of errors, thereby enhancing interpretability.

Runtime analysis indicates that the computational costs remain manageable in relation to the performance gains achieved. Statistical significance tests confirm that these improvements are not merely incidental, but rather robust and replicable.

Together, these findings position RBM-LR as a trustworthy and scalable solution for software fault prediction, effectively balancing predictive accuracy, statistical rigor, and computational feasibility—all critical factors for successful real-world industrial adoption.

5. Conclusions

In this study, we introduced a hybrid RBM-LR framework for software fault prediction (SFP). This model combines Restricted Boltzmann Machines (RBM) for nonlinear feature extraction with Logistic Regression (LR) for classification, effectively addressing the

challenges posed by high-dimensional and imbalanced datasets while maintaining interpretability. The framework was validated using 21 benchmark datasets from PROMISE and OpenML. The experimental results demonstrate that RBM-LR consistently outperforms baseline LR and several leading classifiers, including Random Forest (RF), XGBoost, and Support Vector Machines (SVM). On average, the model achieved a 7% improvement in AUC and a 6% enhancement in accuracy compared to baseline LR. Furthermore, Precision and Recall improved by 5–8%, while the F1-score showed steady gains across the datasets. Notably, these improvements were confirmed through statistical significance testing ($p < 0.05$), underscoring the robustness of the proposed approach.

Additionally, the implementation of 10-fold cross-validation greatly stabilized performance across most datasets (16 out of 22), further validating the reliability of the RBM-LR framework. In comparisons with state-of-the-art classifiers such as RF, XGBoost, and SVM, RBM-LR consistently demonstrated superior predictive accuracy and stability.

Despite the promising results, several limitations persist. The performance of the Restricted Boltzmann Machine (RBM) is significantly influenced by hyperparameter tuning (e.g., the number of hidden nodes, learning rate), and the model incurs a moderate runtime overhead compared to standard Logistic Regression (LR). Additionally, while RBM-LR has shown strong performance on public benchmark datasets, its scalability in large-scale industrial environments warrants further exploration.

Looking ahead, several research directions are proposed. First, enhancing the framework with deep hybrid architectures such as Autoencoders, LSTM, or Transformer-based models could improve the ability to capture temporal and sequential dependencies in software evolution data. Second, applying the framework to large-scale industrial and real-time datasets would help evaluate its scalability and practical utility. Finally, integrating RBM-LR with explainability techniques (e.g., SHAP, LIME) could further increase transparency and support its adoption in safety-critical domains.

In conclusion, the RBM-LR model offers a statistically validated, interpretable, and computationally viable solution for software defect prediction. By addressing the balance between accuracy, efficiency, and interpretability, this research significantly contributes to the development of trustworthy and industrial-grade machine learning frameworks within the field of software engineering.

Author Contributions: Writing—original draft, J.A.; Writing—review and editing, F.E.A., A.A. and A.J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rathore, S.S.; Kumar, S. Towards an ensemble based system for predicting the number of software faults. *Expert Syst. Appl.* **2017**, *82*, 357–382. [CrossRef]
2. Malhotra, R. A systematic review of machine learning techniques for software fault prediction. *Appl. Soft Comput.* **2015**, *27*, 504–518. [CrossRef]
3. Bowes, D.; Hall, T.; Petrić, J. Software defect prediction: Do different classifiers find the same defects? *Softw. Qual. J.* **2018**, *26*, 525–552. [CrossRef]

4. Menzies, T.; Greenwald, J.; Frank, A. Data mining static code attributes to learn defect predictors. *IEEE Trans. Softw. Eng.* **2006**, *33*, 2–13. [CrossRef]
5. Khoshgoftaar, T.M.; Seliya, N. Software quality classification modeling using the SPRINT decision tree algorithm. *Int. J. Artif. Intell. Tools* **2003**, *12*, 207–225. [CrossRef]
6. Seliya, N.; Khoshgoftaar, T.M.; Van Hulse, J. Predicting faults in high assurance software. In Proceedings of the 2010 IEEE 12th International Symposium on High Assurance Systems Engineering, San Jose, CA, USA, 3–4 November 2010; pp. 26–34.
7. Lessmann, S.; Baesens, B.; Mues, C.; Pietsch, S. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Trans. Softw. Eng.* **2008**, *34*, 485–496. [CrossRef]
8. Arisholm, E.; Briand, L.C.; Johannessen, E.B. A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. *J. Syst. Softw.* **2010**, *83*, 2–17. [CrossRef]
9. Tong, H.; Liu, B.; Wang, S. Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning. *Inf. Softw. Technol.* **2018**, *96*, 94–111. [CrossRef]
10. Shepperd, M.; Song, Q.; Sun, Z.; Mair, C. Data quality: Some comments on the nasa software defect datasets. *IEEE Trans. Softw. Eng.* **2013**, *39*, 1208–1215. [CrossRef]
11. Borandag, E.; Ozcift, A.; Kilinc, D.; Yucalar, F. Majority vote feature selection algorithm in software fault prediction. *Comput. Sci. Inf. Syst.* **2019**, *16*, 515–539. [CrossRef]
12. Atkinson, R.; Flint, J. Accessing hidden and hard-to-reach populations: Snowball research strategies. *Soc. Res. Update* **2001**, *33*, 1–4.
13. Nam, J. *Survey on Software Defect Prediction*; Technical Report; Department of Compter Science and Engineerning, The Hong Kong University of Science and Technology: Hong Kong, China, 2014.
14. Abaei, G.; Selamat, A. A survey on software fault detection based on different prediction approaches. *Vietnam. J. Comput. Sci.* **2014**, *1*, 79–95. [CrossRef]
15. Illes-Seifert, T.; Paech, B. Exploring the relationship of a file's history and its fault-proneness: An empirical method and its application to open source programs. *Inf. Softw. Technol.* **2010**, *52*, 539–558. [CrossRef]
16. Catal, C.; Diri, B. Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. *Inf. Sci.* **2009**, *179*, 1040–1058. [CrossRef]
17. Radjenović, D.; Heričko, M.; Torkar, R.; Živkovič, A. Software fault prediction metrics: A systematic literature review. *Inf. Softw. Technol.* **2013**, *55*, 1397–1418. [CrossRef]
18. Catal, C. Software fault prediction: A literature review and current trends. *Expert Syst. Appl.* **2011**, *38*, 4626–4636. [CrossRef]
19. Hall, T.; Beecham, S.; Bowes, D.; Gray, D.; Counsell, S. A systematic literature review on fault prediction performance in software engineering. *IEEE Trans. Softw. Eng.* **2011**, *38*, 1276–1304. [CrossRef]
20. Wagner, S. A literature survey of the quality economics of defect-detection techniques. In Proceedings of the 2006 ACM/IEEE International Symposium on Empirical Software Engineering, Rio de Janeiro, Brazil, 21–22 September 2006; pp. 194–203.
21. Li, L.; Lessmann, S.; Baesens, B. Evaluating software defect prediction performance: An updated benchmarking study. *arXiv* **2019**, arXiv:1901.01726. [CrossRef]
22. Catal, C.; Diri, B. A systematic review of software fault prediction studies. *Expert Syst. Appl.* **2009**, *36*, 7346–7354. [CrossRef]
23. Du, Q.; Subramanian, M.; Pan, D. Human-Computer Interaction Corporate Law Education for Directors: A Machine Learning Approach. *Int. J. Hum.-Comput. Interact.* **2024**, *41*, 1705–1717. [CrossRef]
24. Yuan, X.; Jiang, X.; Shi, Z.; Sun, L. Human reliability analysis of intelligent mine human–computer interaction based on improved SPAR-H. *Nucl. Eng. Des.* **2024**, *423*, 113196. [CrossRef]
25. Dayanand, D.; Sutejas, K.; Shishir, V.; Padmanabhan, S.; Chavali, M. Neural Networks for Human–Machine Interface. In *Human-Machine Interface Technology Advancements and Applications*; CRC Press: Boca Raton, FL, USA, 2024; pp. 1–24.
26. Sadeghi Milani, A.; Cecil-Xavier, A.; Gupta, A.; Cecil, J.; Kennison, S. A systematic review of human–computer interaction (HCI) research in medical and other engineering fields. *Int. J. Hum.-Comput. Interact.* **2024**, *40*, 515–536. [CrossRef]
27. Sun, Y.; Guo, Q.; Zhao, S.; Chandran, K.; Fathima, G. Context-aware augmented reality using human–computer interaction models. *J. Control Decis.* **2024**, *11*, 1–14. [CrossRef]
28. Lv, S.; Ouyang, B.; Deng, Z.; Liang, T.; Jiang, S.; Zhang, K.; Chen, J.; Li, Z. A dataset for deep learning based detection of printed circuit board surface defect. *Sci. Data* **2024**, *11*, 811. [CrossRef]
29. Lang, D.; Lv, Z. SEPDNet: Simple and effective PCB surface defect detection method. *Sci. Rep.* **2025**, *15*, 10919. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Article

Exploring Avatar Utilization in Workplace and Educational Environments: A Study on User Acceptance, Preferences, and Technostress

Cristina Gasch [1,†], Alireza Javanmardi [2,*,†], Ameer Khan [2,†], Azucena Garcia-Palacios [1,†] and Alain Pagani [2,†]

[1] Department of Basic and Clinical Psychology and Psychobiology, Universitat Jaume I, 12006 Castelló de la Plana, Spain; cgasch@uji.es (C.G.); lazucena@uji.es (A.G.-P.)
[2] German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany; ameer.khan@dfki.de (A.K.); alain.pagani@dfki.de (A.P.)
* Correspondence: alireza.javanmardi@dfki.de
† These authors contributed equally to this work.

Abstract: With the rise of virtual avatars in professional, educational, and recreational settings, this study investigates how different avatar types—varying in realism, gender, and identity—affect user perceptions of embodiment, acceptability, technostress, privacy, and preferences. Two studies were conducted with 42 participants in Study 1 and 40 in Study 2, including professionals and students with varying VR experiences. In Study 1, participants used pre-assigned avatars they could control during interactions. In Study 2, an interviewer used different avatars to interact with participants and assess their impact. Questionnaires and correlation analyses measured embodiment, technostress, privacy, and preference variations across contexts. Results showed that hyper-realistic avatars resembling the user enhanced perceived embodiment and credibility in professional and educational settings, while non-realistic avatars were preferred in recreational contexts, particularly when interacting with strangers. Technostress was generally low, though younger users were more sensitive to avatar appearance, and privacy concerns increased when avatars were controlled by others. Gender differences emerged, with women expressing more concern about appearance and men preferring same-gender avatars in professional environments. These findings highlight the need for VR platform designers to balance realism with user comfort and address privacy concerns to encourage broader adoption in professional and educational applications.

Keywords: avatar; virtual reality; human–computer interaction

1. Introduction

In recent years, remote work has experienced a significant increase, driven by the need to balance personal and professional life. This shift has been made possible by technological advances that allow employees to perform their tasks from anywhere [1]. However, remote work also presents major challenges, such as lack of face-to-face interaction, distraction, difficulty in creating a suitable work environment, and blurred boundaries between personal and professional life [2]. To address these challenges, virtual reality (VR) has emerged as a promising technology. VR not only improves understanding of the environment in remote meetings [3], but also facilitates access to otherwise difficult-to-obtain resources [4]

and allows for the creation of immersive training scenarios [5,6]. In addition, VR increases immersion and the sense of embodiment [7,8], which can improve efficiency and reduce distractions during virtual meetings. However, integrating VR also has drawbacks, such as virtual motion sickness and other issues associated with prolonged use of VR hardware, including eye strain, stress, and mental overload [9,10]. These symptoms may depend on factors such as demographics, movement illusion, and viewing mode [11–15]. These issues can also increase technostress, which refers to stress and anxiety caused by the adoption and use of new technologies. In the context of VR, technostress can manifest in various ways, such as frustration with unintuitive interfaces, the cognitive effort required to adapt to complex digital environments, or sensory overload from prolonged immersion [16,17].

A relevant approach to increasing immersion in virtual environments is the use of avatars, digital representations of users, which facilitate interaction [18]. An avatar is a virtual representation of a user within a digital environment. It can range from highly realistic models to stylized or abstract figures [19]. Its design and appearance shape how users perceive both their own presence and that of others in virtual spaces, influencing communication, interaction, and overall comfort. However, using avatars can introduce additional challenges that affect the user experience and modify behavior and attitudes [20,21]. Hyper-realistic avatars, while more closely resembling real people, can create insecurities in users [22], while avatar personalization can influence behavior or even trigger psychological issues, such as body dysmorphia [23–25]. Body dysmorphia in VR refers to how much a user feels their avatar represents them, influencing interaction realism [26]. It includes self-identification (feeling the avatar as part of oneself), motor control (intuitive movement), and presence perception (sense of being in the virtual world). Opting for realistic avatars without reaching hyper-realism can mitigate some of these issues while preserving an acceptable sense of embodiment and control over the avatar [27].

Despite advances in VR and avatar design, there is limited research on how different types of avatars affect user experience in professional environments using VR. While previous studies [28–31] have examined embodiment and technostress in general VR use, few have specifically analyzed these factors in the context of virtual meetings, where social and professional interactions are crucial. Understanding how avatar realism, gender representation, and customization influence perceived embodiment, technostress, usability, and privacy is essential for improving VR-based communication tools. This study aims to bridge this research gap by investigating the impact of avatar characteristics on user experience in virtual meetings.

To better understand these effects, two studies on the use of avatars have been conducted:

- Study 1 examines how users relate to different types of avatars that they control themselves. The study evaluates four avatar conditions: (1) hyper-realistic, (2) non-realistic, (3) an avatar representing another person of the same gender, and (4) another person using the user's avatar. The study measures embodiment, technostress, usability, and privacy concerns.
- Study 2 investigates how users perceive interactions when the interviewer, rather than the participant, uses different types of avatars. This study includes hyper-realistic, non-realistic, and different-gender avatars, as well as an avatar that mirrors the participant. The goal is to analyze how these representations affect interviewer credibility, participant comfort, technostress, and privacy perceptions.

To guide the research, the following questions are addressed:

1. How do different types of avatars influence users' perceived embodiment in VR meetings?
2. What is the impact of avatar realism and representation on technostress and usability?
3. How do avatars affect perceptions of privacy and social comfort in virtual meetings?

4. Does the avatar used by the interviewer influence the credibility and engagement of the participants?

The study is based on the following hypotheses:

H1. *Users will experience greater perceived embodiment with realistic avatars compared to non-realistic avatars.*

H2. *Hyper-realistic avatars will increase technostress compared to stylized avatars.*

H3. *Users will report fewer privacy concerns when controlling their own avatars compared to being represented by others.*

H4. *Participants will perceive the interviewer as more credible when using realistic avatars compared to non-realistic avatars.*

H5. *Gender-incongruent avatars (i.e., an avatar of a different gender than the user) will negatively impact perceived embodiment and comfort.*

By addressing these research questions and hypotheses, this study aims to provide a preliminary understanding of the benefits and limitations of avatars in VR-based professional interactions. Given the exploratory nature of this research and the limited sample size, the findings should be interpreted as a foundation for future studies rather than definitive conclusions. The results will help identify key factors that warrant further investigation in larger and long-term studies. Ultimately, this study seeks to contribute to the development of more effective and user-friendly avatar designs for virtual meetings, balancing usability, perceived embodiment, and user comfort.

2. Study Design

The study was structured into two parts, each with a different focus on evaluating the use of avatars in virtual meetings. In Study 1 (SE1), participants controlled their own avatars, while in Study 2 (SE2), the interviewer used different types of avatars. Both studies were conducted in a controlled environment that simulated virtual meeting scenarios. Avatars were always introduced in the same order, instructions remained identical across all participants, and interviewers followed a structured set of questions to maintain uniformity in interactions. Investigators supervised only the questionnaire responses to minimize bias.

2.1. Participants

Inclusion criteria included individuals between 18 and 65 years of age from Spain and Germany with previous experience in online meetings or virtual reality systems. The study was conducted in both countries to increase participant variability. Participants were required to have sufficient cognitive, auditory, and/or visual abilities to read, write, or engage in conversations in the language of the study. Exclusion criteria included no or very limited computer skills and sensory disorders that would prevent participation in the study (e.g., blindness or deafness). Individuals with no or very limited computer skills were excluded due to the need for participants to interact with the technology independently, preventing basic technical difficulties from affecting the study results. The final sample consisted of the following.

- Study 1: 42 participants (22 men and 20 women) between the ages of 22 and 62, with an average age of 32.4 years (SD = 9.3). Most participants had a high educational level, with 71.5% (30 participants) holding a university degree or higher.

- Study 2: 40 participants, equally divided into 25 men and 15 women, aged between 20 and 38 years old, with an average age of 28.2 (SD = 4.7). All participants had a high level of education, with 83% having a university degree or pursuing a degree.

Given the exploratory nature of this study, participation was open to the public and advertised on the bulletin boards of research labs in Spain and Germany. Due to this recruitment method, the final participants in both countries belonged to a university setting, including students, professors, and researchers. The total sample size was determined by the number of volunteers who signed up. This approach allowed for an initial investigation into avatar perception in virtual meetings, providing preliminary insights that can guide future research with larger and more diverse samples. Both studies were approved by the university ethical committee (Ref. CEISH/25/2022). All data were collected anonymously, no photographs of participants were stored, and all participants signed an informed consent form before taking part in the study. While the sample size is limited, the findings help identify key factors that warrant further investigation in more extensive and long-term studies.

2.2. Materials

To simulate the use of avatars with varying appearances in an online meeting, the Geometry-Guided GAN for Face Animation (G3FA) was utilized [32]. This model, which has shown superior performance compared to other state-of-the-art real-time facial animation methods, enables the integration of 3D information into face animation using only 2D images. This significantly enhances the image generation capabilities of the talking head synthesis model.

In Study 1, participants used three types of avatars: a hyper-realistic avatar of the participant; a non-realistic avatar (cartoon of the participant); a hyper-realistic avatar of another person of the same gender; and a final test where another person used the participant's avatar. In Study 2, the interviewer used the following avatars: a hyper-realistic avatar of a person of the same gender as the participant; a hyper-realistic avatar ostf a person of the opposite gender; a non-realistic (cartoon) avatar; a hyper-realistic avatar of another person; and a hyper-realistic avatar representing the participant. For the hyper-realistic images, real photographs were used, while for the non-realistic images, the VToonify tool [33] was used, allowing the creation of animated-style images from real photographs. To generate this image, the cartoon (beautiful) option was selected based on a clear photograph of the participant, while the remaining settings were left at their default values.

2.3. Variables and Measurement Instruments

The evaluation protocol consists of three parts: user information collected before the study; information collected during each test; and data collected at the end of the study.

2.3.1. User Profile

Before starting the test, demographic information (gender, age, educational background) and previous experience with online meetings and virtual reality technology in various contexts (work, education, leisure) were collected.

2.3.2. Measures Obtained at the End of Each Test

In Study 1, a 5-point Likert scale questionnaire based on the response format of the System Usability Scale (SUS) [34], was used to collect user feedback at the end of each test. In addition, a psychometric approach assessed the embodiment [35] toward the avatar's face, focusing on ownership, agency, and change [36]. Questions on privacy were also

included, adapted from a validated technostress test [37]. All questions were tailored for each study to refer to the user or the interviewer.

In Study 2, it was not possible to assess Change in Embodiment, as the user did not have control over the avatar. However, Agency and Ownership were evaluated using the questions from Study 1, with the exclusion of those that could not be adapted. To assess the impact of the avatar on the interviewer's credibility and the participant's trust in the interaction, seven questions were introduced. These questions were adapted from previous studies on trust in virtual characters and online communication. Research has shown that the appearance and expressiveness of avatars significantly influence user trust [38,39]. Furthermore, studies have examined how the presence of avatars in virtual interactions affects perceived credibility and trust [40].

Additionally, two questions on privacy concerns, adapted from Study 1, were included to explore how the avatar influenced participants' perceptions of privacy and data security. Based on these studies, the questions were designed to evaluate distraction, conversational difficulty, and perceived privacy when interacting with the avatar. In the final test of Study 2, these questions also focused on the use of an avatar of a different gender.

Appendix A presents the questions used to collect data on embodiment, privacy, and technostress in Study 1 (Table A1), as well as embodiment, technostress, and gender in Study 2 (Table A2).

2.3.3. Final Study Measures

At the end of all tests, participants completed a final assessment. In Study 1, Technostress caused by avatar use was evaluated using a 5-point Likert scale, based on the response format of the System Usability Scale (SUS) and adapted from a validated test [36]. In Study 2, Technostress was not assessed, as it is more closely related to self-representation, which was not applicable in this context. Additionally, four open-ended questions were included to evaluate participants' general experience, any discomfort caused, and their perception of avatars of different genders.

The acceptability of using these types of avatars in different contexts—work, education, and leisure—was assessed in both studies. Given that the focus was on evaluating the use of the avatar itself rather than the system used, the questions were reduced to measuring future intention to use and perceived usefulness of the avatar. To achieve this, they were adapted from a validated acceptability questionnaire [41–43]. Responses were collected using a 5-point Likert scale and analyzed independently.

Finally, preferences for avatar type were evaluated in six scenarios: team meetings or interactions with external participants, educational settings as a student or teacher, and recreational activities with strangers or friends.

2.4. Method of Analysis

For the analysis, the arithmetic mean was used to aggregate responses. When analyzing individual questions, the mean was calculated across all participants to determine overall trends. For constructs measured by multiple items (e.g., Agency), a composite score was obtained by averaging the participant's responses to the corresponding questions. This method ensured comparability across participants and facilitated correlation analyses with other variables.

The statistical analysis was conducted using SPSS software (version 29.0.1.0). To evaluate the relationships between the study variables, Spearman's correlation was employed, with a significance threshold set at $p < 0.05$. The choice of Spearman's correlation was based on the observation that the relationship between the variables did not appear to be

strictly linear, making this test more suitable for capturing associations without assuming a specific linear structure. Due to the exploratory nature of the study and the limited sample size, no corrections for multiple comparisons were applied. The primary objective of this work is to identify preliminary patterns and key factors that can be further investigated in future studies with larger samples and more robust statistical analyses.

2.5. Hardware

The studies were carried out on computers equipped with NVIDIA RTX3060 graphics cards in Spain and RTX3080 in Germany, ensuring smooth performance and low latency in avatar rendering.

2.6. Procedure

The study for each participant was conducted through the following phases:

2.6.1. Pre-Study

Participants were informed about the study's purpose and signed an informed consent form. Then, they completed an initial questionnaire to gather demographic data and previous experience.

2.6.2. Study

In both studies, participants interacted with the interviewer in several different tests, each featuring a different type of avatar. The interactions lasted between 2 and 3 min, simulating conversations in virtual meetings.

- Avatar Familiarization: Before each interaction, participants were briefly introduced to the avatar that either the interviewer or the participant would use in the session.
  - S1: Hyper-realistic, non-realistic, hyper-realistic avatar of another person, and a test as a spectator while another user used their avatar.
  - S2: Hyper-realistic, non-realistic, hyper-realistic avatar of another person of the same gender, hyper-realistic avatar of a person of a different gender, and a hyper-realistic avatar of the participant.
- Simulated Interview: Participants engaged in a conversation with the interviewer, designed to make them focus on the interaction rather than on the avatar. They were allowed to change the topic or avoid questions if they wished. In S1, the interviewer positioned themselves behind the users to minimize distraction, allowing users to maintain focus on the avatar. The interview aimed to stimulate conversation and divert attention from the avatar by asking personal questions where participants could expand freely. In S2, the interview took place via video conference in separate rooms, with questions structured to simulate a job interview.
- Post-interaction Questionnaire: At the end of each interaction, participants completed a brief questionnaire to evaluate each avatar type.

2.6.3. Post-Study

After completing all tests, participants completed the final questionnaire.

3. Results

3.1. User's Prior Experience

Regarding prior experience with video calls and virtual reality systems, the following results were obtained:

- Study 1: Regarding video conferencing experience, 69% reported frequent use, 23% had regularly used VR
- Study 2: Systems, while 14.3% had never used VR before this study. Regarding video conferencing use, 58% reported frequent use and 43.5% had occasionally used a virtual reality system.

3.2. User's Perception of Embodiment

In both studies, embodiment perception was a key dimension in measuring the effectiveness of avatars. The questionnaire elements were classified into three properties that influence this perception: ownership, agency, and change (the latter not applicable to Study 2). In terms of ownership, Figure 1 illustrates that both hyper-realistic avatars in both studies achieved a greater resemblance to human faces than the non-realistic avatar (Q2). Participants particularly attributed a stronger sense of facial ownership to their hyper-realistic avatars in questions about facial features (Q1 and Q3).

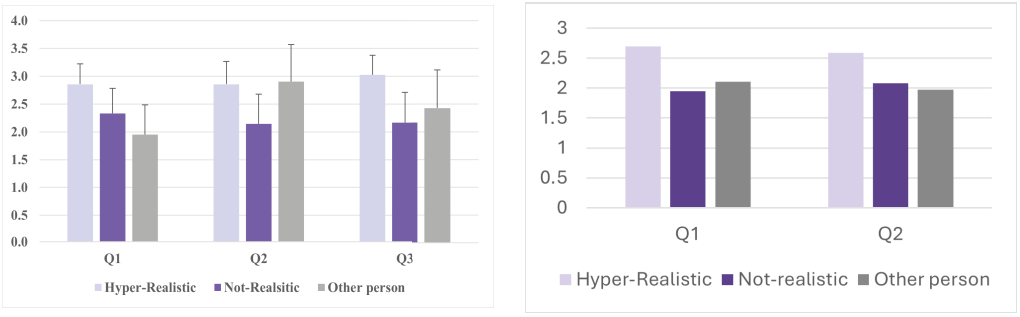


Figure 1. Property. **Left**: Study 1; **Right**: Study 2.

For agency, Figure 2 shows that, in Study 1, users' own avatars received higher scores across all questions. The hyper-realistic avatar scored highest on questions related to avatar movement (Q4, Q7, and Q8), while the non-realistic avatar scored higher on questions related to enjoyment (Q5) and comfort (Q6). Conversely, the avatar representing another person did not score significantly high in any question, except for comfort. In Study 2, the difference between avatar types was much more balanced, although the participant's own hyper-realistic avatar retained slightly higher values. It is notable that the question about feeling like they were speaking with a real person rather than a representation (Q4) received a clearly lower value.

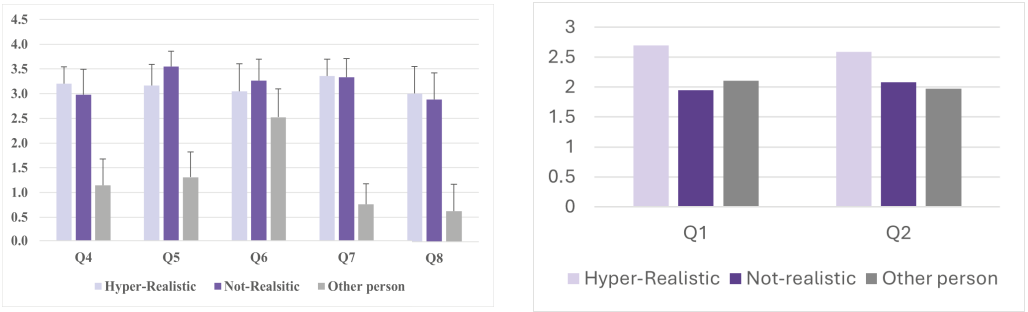


Figure 2. Agency. **Left**: Study1; **Right**: Study 2.

Lastly, the results for change in Study 1 are presented in Figure 3. In all three cases, the values remained relatively low (below 2.5). However, the non-realistic avatar showed a pronounced sense of change. Interestingly, the hyper-realistic avatar led more users to

question changes in their own face (Q10). Additionally, the avatar that did not belong to the user received the lowest scores across the three questions.

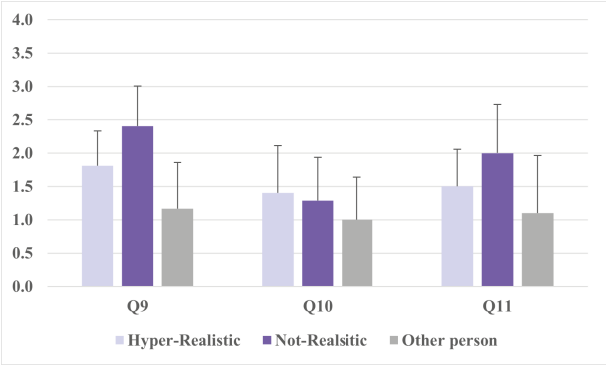


Figure 3. Change. Study 1.

A correlation analysis with sociodemographic data and prior experience was performed using Spearman's correlation. Results indicated no significant correlation with age, gender, education level, or previous videoconferencing experience. However, prior experience with virtual reality showed a correlation, as displayed in Table 1, where the Agency factor in Study 1 demonstrated improvement with prior VR experience, suggesting a greater sense of control over the avatar with increased VR exposure. No significant correlations were found in Study 2.

Table 1. Correlation between virtual reality experience and video conference experience. The values in bold indicate significant values.

		Study 1		**Study 2**	
		r	***p***	**r**	***p***
Property	H	0.252	0.107	0.101	0.638
	N	0.191	0.227	0.022	0.92
	O	−0.076	0.631	0.096	0.655
Agency	H	**0.469**	**0.002**	0.034	0.874
	N	**0.329**	**0.003**	0.067	0.757
	O	**0.411**	**0.007**	0.128	0.398
Change	H	0.141	0.373	0.018	0.132
	N	−0.062	0.697	0.181	0.41
	O	0.104	0.51	0.087	0.551

3.3. Credibility and Confidence

As shown in Figure 4, in Study 2, participants generally felt that the avatar did not cause a notable lack of credibility or trust. However, the participant's own hyper-realistic avatar scored below average on all questions. On the other hand, the lowest-scoring avatar was the non-realistic one. In a Spearman correlation analysis between results and prior experience, a negative correlation was again found between prior VR experience and question Q5 (r = −498; $p = 0.008$), indicating that users with VR experience had greater confidence when interacting with an avatar. Regarding the relationship between credibility and trust with embodiment properties, a significant negative correlation was found with the sense of ownership. The resulting values are shown in Table 2, suggesting that the greater the sense that the avatar belongs to the interlocutor, the higher the credibility and confidence.

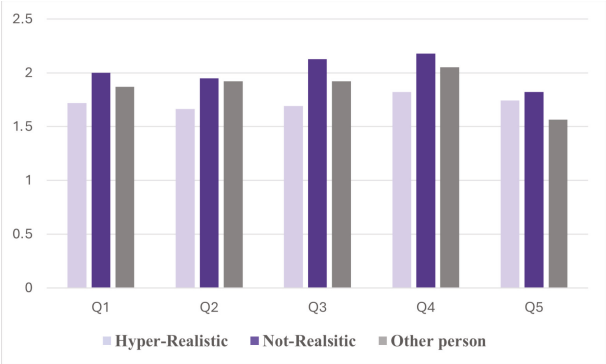


Figure 4. Values obtained in credibility and confidence from Study 2.

Table 2. Correlation between presence properties and credibility and confidence. The values in bold indicate significant values.

	Property	Agency
Hyper-Realistic	−**0.404; 0.011**	−0.214; 0.190
Not Realistic	−0.01; 0.95	−0.430; 0.795
Other person	−**0.371; 0.009**	−0.524; 0.620
User	−**0.367; 0.021**	−**0.330; 0.04**

3.4. Privacy

In Study 1, users did not express concern about privacy for the first three avatars (with values below 0.5 on a scale of 0 to 4). However, this increased to 1.5 when they saw that others could use their avatar. In Study 2, higher privacy concerns were evident from the first avatar. Figure 5 (left), which presents the results of the second study, shows that hyper-realistic avatars generated more concern than the non-realistic ones, with concerns increasing with the use of the participant's own avatar and gender changes.

To further explore the correlation between privacy concerns and embodiment perception, a Spearman correlation analysis was conducted. In Study 1, there was a notable correlation between hyper-realistic avatars and the sense of change (r = 0.575; p = 0.001) and, to a lesser extent, the sense of control (r = 0.397; p = 0.014). In Study 2, a stronger correlation was found between the sense of control and avatars representing another person (r = 0.377; p = 0.018) and the participant's own avatar (r = 0.460; p = 0.003). Despite being a low correlation, both types of avatars negatively correlated with the sense of ownership. However, the most notable correlation was with the use of an avatar of a different gender, showing discomfort (r = 0.745; p = 0.0001).

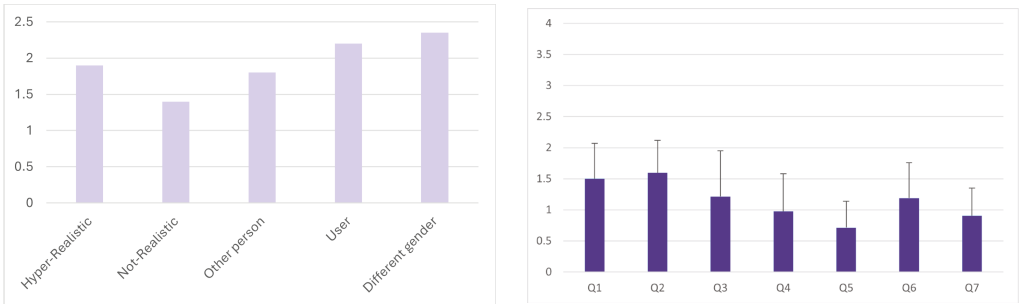


Figure 5. Left: Privacity results in Study 2; **Right**: Visual comparison of technostress questionnaire responses in Study 1.

3.5. Technostress

Figure 5 (right) provides a visual comparison of questionnaire responses, as shown in Table A1, offering insights into levels of technostress. Specifically, Questions 4, 5, and 7, related to privacy, fati gue, and self-esteem while using avatars, showed the lowest values, all below 1. In contrast, Questions 1 and 2, focused on the avatar's appearance, scored higher, exceeding 1.5. These findings suggest that merely using avatars does not inherently induce high levels of technostress. However, additional analysis using Spearman's correlation revealed significant associations:

- Gender correlated with Questions 6 (r = 0.359; $p = 0.05$) and 7 (r = 0.298; $p = 0.018$), indicating a greater concern among women regarding possible judgment or discrimination based on avatar appearance, aligning with previous research on appearance-related anxiety in video conferencing.
- Age correlated with Questions 1 through 5 (Average: r = 0.319; $p = 0.045$), indicating higher concerns among younger participants regarding avatar appearance, privacy, and fatigue, consistent with existing literature.
- Videoconferencing experience correlated with apprehension about expressing one's true personality (r = 0.377; $p = 0.014$), suggesting that more experience amplifies concerns in this domain.

3.6. Acceptability

As shown in Figure 6, users perceived avatar technology as useful in both educational and leisure settings in both studies, with ratings above 2.5. However, uncertainty remains about their effectiveness in the workplace, although users still recognize their utility (S1: 2.3; S2: 2). For the intention to use, the effect is similar: scores are higher than for utility, but the same trend persists, with leisure scoring the highest and work the lowest.

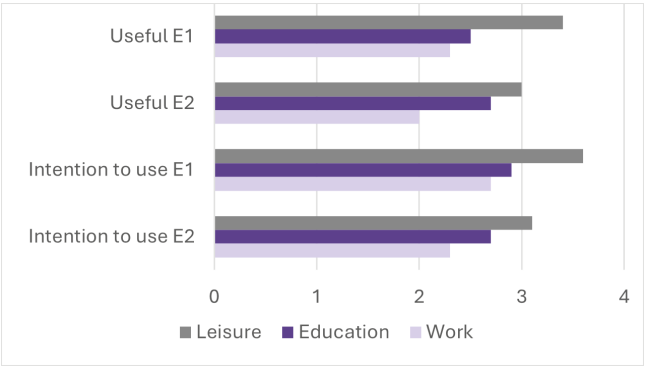


Figure 6. Perceived utility and intention to use.

To discern factors influencing users' willingness to adopt avatar systems, a Spearman correlation analysis was conducted. Initially, the analysis considered prior experience and demographic data. In the first study, a positive correlation with previous VR experience was observed in the first study (Work: r = 0.397, $p = 0.009$; Education: r = 0.399, $p = 0.009$; Leisure: r = 0.339, $p = 0.009$). In contrast, in the second study, this correlation was only significant for education and leisure (Education: r = 0.459 $p = 0.024$; Leisure: r = 0.417, $p = 0.004$). For embodiment factors, ownership emerged as significant in determining perceived usefulness of avatars in both work and educational contexts, thus influencing the intention to use them. Detailed results are provided in Table 3.

Table 3. Correlation between usability and property factor.

	S1 r	S1 *p*	S2 r	S2 *p*
Useful in work	0.446	0.003	0.351	0.008
Useful in education	0.312	0.048	0.281	0.033
Intention to use in work	0640	0.001	0.343	0.012
Intention to use in education	0.336	0.029	0.294	0.041

3.6.1. Avatar Preferences

Figure 7 (left) presents a comparative analysis of avatar preferences across different settings in Study 1. In workplace scenarios, users preferred avatars representing themselves, ideally with the highest possible image quality. This preference increased with the seriousness of meetings. Similarly, in educational contexts, particularly regarding educators, users favored a hyper-realistic avatar of themselves (47.6%). For students, self-representation was also preferred as they progressed. In leisure activities, preferences varied depending on the activity. When interacting with friends, users prioritized their hyper-realistic avatars. Conversely, in settings requiring interaction with both acquaintances and strangers, such as gaming environments, users preferred non-realistic avatars (52.4%).

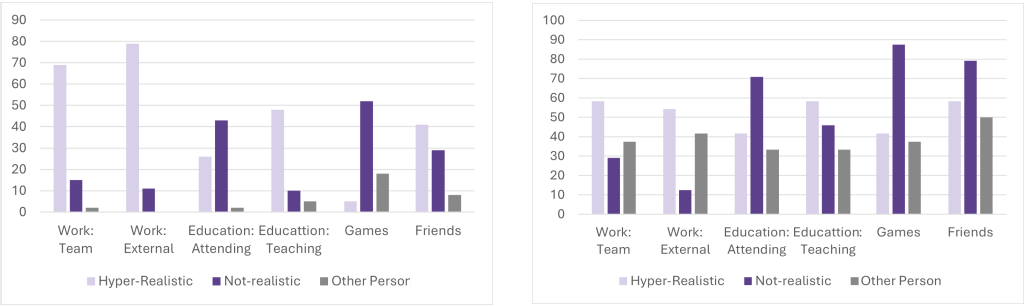


Figure 7. Avatar preferences comparative. **Left**: Study 1; **Right**: Study 2.

Similarly, in Study 2, users preferred interacting with a hyper-realistic avatar in work settings, ideally a real-life likeness. In educational settings, as students, they preferred non-realistic avatars, while as teachers, they chose hyper-realistic avatars of themselves. In leisure settings, users preferred interacting with non-realistic avatars, although the difference between a non-realistic avatar and hyper-realistic avatars was smaller when interacting with friends. This comparison is shown in Figure 7 (right).

3.6.2. Use of Different-Gender Avatars

In Study 2, an additional test involved using an avatar of a different gender. Results showed that this test scored lower than the others on both embodiment properties (below 1.5 on a scale of 0 to 4), particularly the question about feeling like they were speaking with a real person (0.85). Regarding questions on credibility and trust, results are displayed in Figure 8, showing that users experienced more issues interacting with the interviewer (2.6), felt increased discomfort (2.2), and had higher privacy concerns (2.3).

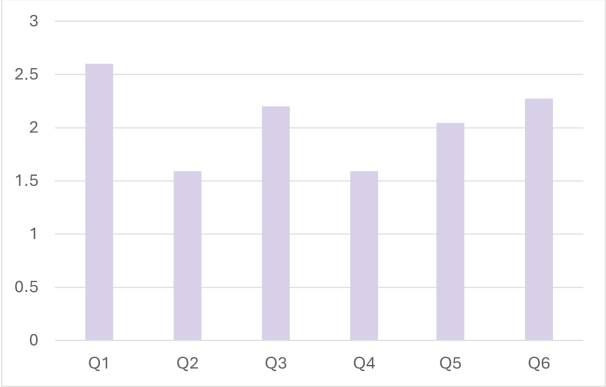


Figure 8. Credibility and confidence with other gender avatar.

Additionally, to better understand user preferences when interacting with another person, participants could select whether they preferred interacting with a male or female avatar. Results based on user gender are found in Table 4. For men: in the workplace, they preferred interacting with a male avatar; in education, they mostly preferred male avatars; and in leisure, where the preference for another person's avatar increased, the preference was for both genders, with a greater tendency toward female avatars when interacting with strangers. For women, the number of users who preferred another person's avatar was considerably lower. Notably, in leisure, there was no gender preference for the avatar they interacted with, but in other settings, they preferred either both genders or male avatars.

Additionally, most users mentioned in open-ended questions that interacting with an avatar whose voice did not match the appearance felt less serious and initially distracting until they grew accustomed to it.

Table 4. User preferences depending on gender.

		Total	Both	Man	Woman
Work Team	Men	9	2	7	
	Women	3	3		
Work External	Men	12	1	8	
	Women	3	1	2	
Education Attending	Men	10	7	2	1
	Women	3	1	2	
Education Teaching	Men	6		4	2
	Women	3	1	2	
Games	Men	15	8	2	4
	Women	6	6		
Friends	Men	12	8	3	1
	Women	8	8		

4. Discussion

This study examined the use of different types of avatars in work, educational, and recreational settings, evaluating user perceptions of embodiment, acceptability, technostress, privacy, and preferences. Two studies were conducted: in the first, participants controlled their own avatars, while in the second, an interviewer used various avatars. Both studies included hyper-realistic and non-realistic avatars with variations in gender and identity.

Findings indicate that hyper-realistic avatars enhance the sense of embodiment, especially when they closely resemble the user. This perceived embodiment positively

influenced the perceived usefulness and intention to use avatars in work and educational settings, where users preferred avatars that resembled themselves. However, in recreational contexts, participants favored non-realistic avatars, particularly for informal interactions or engagements with strangers.

The study reinforced the importance of embodiment in avatar use, showing that a stronger sense of embodiment improves perceived utility and intention to use. This result aligns with previous studies on embodiment in VR [12,13], which suggest that similarity to the user enhances identification and the sense of presence in virtual environments. Therefore, implementing strategies that enhance ownership and control while minimizing perceptions of bodily distortion in avatars is crucial. Avatars that closely match the user's appearance improved the sense of control and reduced discomfort related to appearance changes.

Although technostress related to avatar use was relatively low, concerns about avatar appearance emerged, especially among younger users, reflecting potential insecurities about body image, which is consistent with previous studies [34,41]. Privacy concerns also arose, particularly when avatars were used by others. While initial privacy concerns were minimal, they may increase over time due to the perception that data used to create avatars could be sold or stolen. In environments where confidentiality is critical, such as the workplace, ensuring transparency in data usage and maintaining security will be essential for the widespread adoption of avatar technology.

Results also showed that the type of avatar used by the interviewer influenced participants' perceptions of credibility and trust. Hyper-realistic avatars—whether resembling the user or another person of the same gender—were perceived as more credible, while non-realistic avatars generated discomfort and a lower perception of professionalism. This suggests that, in professional settings such as job interviews or formal presentations, selecting avatars that enhance credibility and trust is crucial.

Although a general preference for hyper-realistic avatars was observed in professional settings, the preference for non-realistic avatars in recreational environments aligns with previous research on virtual identity, such as the studies by Yee and Bailenson [44], who found that in informal contexts, users tend to choose more stylized or abstract representations to foster creativity or maintain anonymity.

This phenomenon can be explained through self-presentation theory [45], which suggests that individuals adjust their appearance based on the social context. In professional environments, where credibility and professionalism are essential, users prefer avatars that reflect their own image. In contrast, in recreational settings, aesthetic flexibility allows for the exploration of different aspects of identity without the constraints of the physical world.

Additionally, the results suggest gender differences in participant preferences. Women showed greater concern about the avatar's appearance and its impact on how they were perceived by others, whereas men felt more comfortable interacting with same-gender avatars in work and educational settings. These differences should be considered in avatar design to ensure that all users feel comfortable and confident when engaging with this technology.

Furthermore, cultural background and prior experience with virtual reality may have influenced participant perceptions of avatars. Since the study was conducted in Spain and Germany, differences in technology adoption, self-representation, and privacy concerns could have played a role in avatar preferences. While cultural influences were not explicitly analyzed, future studies should explore how regional differences and attitudes toward virtual interactions impact avatar selection. Additionally, prior VR experience may have shaped participants' comfort levels and perceived embodiment. Those familiar with VR might have adapted more easily to hyper-realistic avatars, while less experienced users may have felt greater discomfort or uncertainty. Expanding future research to include

individuals with minimal or no VR exposure would help clarify how familiarity with immersive technologies affects user perception and adoption.

Personalizing one's representation can negatively affect self-perception. Jawad et al. [46] found that AI-driven personalization algorithms can influence self-image, group identity, and online social interactions.

An important concern emerging from the findings is the psychological impact of excessive avatar customization, particularly the risk of body dysmorphia. The ability to modify avatar appearance may heighten body image insecurities, especially among younger users, suggesting the need to establish limits on customization to prevent unrealistic expectations. This underscores the importance of designing balanced avatars that faithfully represent users without negatively affecting self-esteem.

Given the exploratory nature of this study, its small sample size, and the fact that both studies were conducted in a single session, the findings should be interpreted as preliminary insights rather than definitive conclusions. As a result, the study has several limitations.

First, since all participants had at least basic technological knowledge, it remains unclear how individuals with minimal or no experience with technology would engage with avatars. Additionally, the limited sample size and the study's focus on only two countries restricted the ability to fully assess potential cultural differences in avatar preferences. Expanding the sample to include a broader and more diverse range of participants and cultural backgrounds would improve the generalizability of the results.

Second, the long-term effects of avatar-based interactions remain uncertain. Investigating the prolonged use of this technology and its impact on group dynamics would provide valuable insights into its sustainability and effectiveness over time.

Another limitation is the lack of consideration for psychological and social factors that may influence avatar preferences, which could have affected the findings. Incorporating physiological measures, such as heart rate or skin conductance, would allow for a more objective assessment of stress levels. Additionally, qualitative approaches, such as post-study interviews, could help identify specific stressors, including the uncanny effect of certain avatars or the pressure to perform in virtual interactions.

Furthermore, a dedicated study on the potential risks associated with avatars could offer deeper insights into privacy concerns and inform the development of effective mitigation strategies. Addressing issues such as data misuse, identity theft, and user control over personal information would be essential for the widespread adoption of avatar-based technologies.

Finally, these studies have been conducted exclusively in a professional setting. Future research should explore how these findings translate to other contexts, such as training simulations or customer service interactions, where avatar perception and perceived embodiment could influence user experience and performance in distinct ways.

5. Conclusions

Avatars hold significant potential for enhancing virtual interactions, particularly in professional and educational settings. The findings suggest that perceived embodiment plays a crucial role in avatar utility and acceptance. To maximize their benefits, it is essential to prioritize strategies that enhance the sense of ownership and control while minimizing perceptions of bodily alteration.

Addressing concerns regarding privacy and appearance will also be key to widespread avatar adoption. Ensuring transparency in data usage and providing balanced customization options can help reduce risks related to body dysmorphia and self-presentation anxiety. Additionally, future research should investigate specific privacy risks, such as data misuse

and identity theft, and propose mitigation strategies. User-controlled data permissions, anonymization techniques, and increased transparency in platform policies—such as clear guidelines on data storage and usage—could help alleviate user concerns and improve trust in avatar-based technologies.

Furthermore, ethical risks, particularly those related to excessive avatar customization and its potential to contribute to body dysmorphia, require further attention. Over-customization may lead to unrealistic body standards, negatively impacting users' self-esteem. To address this, future avatar design should incorporate guidelines that limit unrealistic body modifications, promote inclusive avatar options, and encourage representations that reflect diverse and authentic identities. Developers bear a responsibility to protect user well-being, ensuring that avatar customization tools enhance self-expression without reinforcing harmful beauty standards.

Finally, further research is needed to assess the impact of avatars on gender perception, trust in interactions, and group dynamics in virtual environments. Exploring long-term effects and considering individual and cultural differences will contribute to a more inclusive and effective development of avatar technologies in the future.

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: C.G. and A.G.-P.; data collection: C.G., A.J. and A.K.; software: A.J.; analysis and interpretation of results: C.G.; draft manuscript preparation: C.G. and A.J.; review and editing, A.G.-P. and A.P.; project administration, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the EU project CORTEX2 under the European Union's Horizon Europe research and innovation programme (grant agreement: N° 101070192).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Ethics Committee of Universitat Jaume I (protocol code CEISH/25/2022, 21 December 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Study 1 questions.

	Questions used in the Study 1
	Property
Q1	It felt like the virtual face was my face.
Q2	The virtual face felt like a human face.
Q3	I had the feeling that the virtual face belonged to me.
	Agency
Q4	The movements of the virtual face seemed to be my own movements.
Q5	I enjoyed controlling the virtual face.
Q6	I have felt comfortable using the virtual face.
Q7	I felt as if I was causing the movement of the virtual face.
Q8	The movements of the virtual face were synchronous with my own movements.

Table A1. *Cont.*

	Change
Q9	I had the illusion of owning a different face from my own.
Q10	I felt the need to check if my face really still looked like what I had in mind.
Q11	I felt as if the form or appearance of my face had changed.
	Privacy
Q12	I feel that the use of this type of avatar is an intrusion into my privacy.
Q13	I feel that this kind of avatar reveals private personal information without my consent.
	Technostress
Q1	Do you feel or would you feel pressured or stressed about maintaining a "perfect" image or appearance through your avatar?
Q2	Do you feel or think you would feel anxiety or stress when comparing your avatar to other people's avatars in virtual environments?
Q3	Do you experience or think that you would experience difficulties in expressing your true identity or personality through your avatar?
Q4	Do you or would you feel uncomfortable or stressed about the lack of privacy or the potential exposure of your real identity while using an avatar?
Q5	Do you feel or think you would feel that using avatars in virtual environments exhausts you emotionally or mentally?
Q6	Do you experience or think you would experience worries or stress related to the possibility of your avatar being judged or discriminated against by other users?
Q7	Do you feel or think that you would feel that the use of avatars in virtual environments negatively affects your self-esteem or self-confidence?

Table A2. Study 2 questions.

	Questions used in the Study 2
	Property
Q1	It felt like the virtual face was the other person.
Q2	The virtual face felt like a human face.
	Agency
Q3	The movements and expressions of the virtual face felt real.
Q4	I felt that the virtual face was not a representation but the real person.
Q5	Overall, I felt like I was talking to the other person.
	Credibility and confidence
Q1	The avatar has distracted me from the conversation.
Q2	I found it more difficult to maintain the conversation with the avatar than with the real person's face.
Q3	I felt that it was more difficult to speak with the virtual face than with the real person.
Q4	I felt that it was more difficult for me to look at the virtual face during the conversation than with the real person.
Q5	Overall, I felt stressed or uncomfortable talking with the virtual face.
Q6	I feel that the use of this type of avatar is an intrusion into privacy.
Q7	I feel that this kind of avatar reveals private personal information without consent.
	Gender
Q1	I feel it affected my interaction with the interviewer.
Q2	It made me feel uncomfortable during the conversation.
Q3	I found it more challenging to maintain the conversation.
Q4	I feel it affected my confidence with the interviewer to answer their questions.
Q5	I feel the credibility and/or authority of the interviewer changed.
Q6	I was concerned about sharing private information while using avatars.

References

1. Popovici, V.; Popovici, A.L. Remote Work Revolution: Current Opportunities and Challenges for Organizations. *Ovidius Univ. Ann. Econ. Sci. Ser.* **2020**, *20*, 468–472.
2. Fereydooni, N.; Walker, B.N. Virtual Reality as a Remote Workspace Platform: Opportunities and Challenges. In Proceedings of the New Future of Work 2020, Virtual, 3–5 August 2020.
3. Sun, D.; Kiselev, A.; Liao, Q.; Stoyanov, T.; Loutfi, A. A New Mixed-Reality-Based Teleoperation System for Telepresence and Maneuverability Enhancement. *IEEE Trans.-Hum.-Mach. Syst.* **2020**, *50*, 55–67. [CrossRef]
4. Ramkumar, N.; Fereydooni, N.; Shaer, O.; Kun, A.L. Visual Behavior During Engagement with Tangible and Virtual Representations of Archaeological Artifacts. In Proceedings of the 8th ACM International Symposium on Pervasive Displays, Palermo, Italy, 12–14 June 2019. [CrossRef]
5. Thomsen, A.S.S.; Bach-Holm, D.; Kjærbo, H.; Højgaard-Olsen, K.; Subhi, Y.; Saleh, G.M.; Park, Y.S.; La Cour, M.; Konge, L. Operating Room Performance Improves after Proficiency-Based Virtual Reality Cataract Surgery Training. *Ophthalmology* **2017**, *124*, 524–531. [CrossRef]
6. Seymour, N.E.; Gallagher, A.G.; Roman, S.A.; O'brien, M.K.; Bansal, V.K.; Andersen, D.K.; Satava, R.M. Virtual Reality Training Improves Operating Room Performance Results of a Randomized. *Ann. Surg.* **2002**, *236*, 458–464. [CrossRef]
7. Cummings, J.J.; Bailenson, J.N. How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychol.* **2016**, *19*, 272–309. [CrossRef]
8. McGlynn, S.A.; Sundaresan, R.M.; Rogers, W.A. Investigating AgeRelated Differences in Spatial Presence in Virtual Reality. *Hum. Factors Ergon. Soc. Annu. Meet.* **2018**, *62*, 1782–1786. [CrossRef]
9. Souchet, A.D.; Lourdeaux, D.; Oagani, A.; Rebenitsch, L. A narrative review of immersive virtual reality's ergonomics and risks at the workplace: Cybersickness, visual fatigue, muscular fatigue, acute stress, and mental overload. *Virtual Real.* **2023**, *27*, 19–50. [CrossRef]
10. Stanney, K.M.; Hash, P. Locus of User-Initiated Control in Virtual Environments. *Teleoperators Virtual Environ.* **1998**, *7*, 447–459. [CrossRef]
11. McCauley, M.E.; Sharkey, T.J. Cybersickness: Perception of Self-Motion in Virtual Environments. *Teleoperators Virtual Environ.* **1992**, *1*, 311–318. [CrossRef]
12. Reason, J.T.; Brand, J.J. *Motion Sickness*; Academic Press: Cambridge, MA, USA, 1975; 310p.
13. Moss, J.D.; Muth, E.R. Characteristics of Head-Mounted Displays and Their Effects on Simulator Sickness. *Hum. Factors* **2011**, *53*, 308–319. [CrossRef]
14. Rebenitsch, L.; Owen, C. Review on cybersickness in applications and visual displays. *Virtual Real.* **2016**, *20*, 101–125. [CrossRef]
15. Knight, M.M.; Arns, L.L. The relationship among age and other factors on incidence of cybersickness in immersive environment users. In Proceedings of the APGV 2006: Symposium on Applied Perception in Graphics and Visualization, Boston, MA, USA, 28–29 July 2006; p. 162.
16. Khan, B.A.; Pirkkalainen, H.; Salo, M. Technostress in Work-Related Use of Social Virtual Reality. *ECIS* **2024**, *15*, 2304.
17. Yang, X.; Chen, X.; Gao, D.; Wang, S.; Han, X.; Wang, B. Have-fun: Human avatar reconstruction from few-shot unconstrained images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 742–752.
18. Slater, M.; Linakis, V.; Usoh, M.; Kooper, R. Immersion, Presence and Performance in Virtual Environments: An Experiment with Tri-Dimensional Chess. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST, Hong Kong, 1–4 July 1996; pp. 163–172.
19. Qian, Z.; Wang, S.; Mihajlovic, M.; Geiger, A.; Tang, S. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 5020–5030.
20. Aymerich-Franch, L. Avatar embodiment experiences to enhance mental health. In *Technology and Health*; Academic Press: Cambridge, MA, USA, 2020; pp. 49–66. [CrossRef]
21. Porras, B.; Ferrer, M.; Olszewska, A.; Yilmaz, L.; González, C.; Gracia, M.; Gültekin, G.; Serrano, E.; Gutiérrez, J. Is This My Own Body? Changing the Perceptual and Affective Body Image Experience Among College Students Using a New Virtual Reality Embodiment-Based Technique. *J. Clin. Med.* **2019**, *8*, 925. [CrossRef]
22. Park, J. The effect of virtual avatar experience on body image discrepancy, body satisfaction and weightregulation intention. *Cyberpsychol. J. Psychosoc. Res. Cyberspace* **2018**, *12*, 3. [CrossRef]
23. O'Donnell, C. Getting Played: Gamification and the Rise of Algorithmic Surveillance. *Surveill. Soc.* **2014**, *12*, 349–359. [CrossRef]
24. Raman, U. Body Image Perception: Adolescent Boys and Avatar Depiction in Video Games. *Auctus J. Undergrad. Res. Creat. Scholarsh.* **2016**, 1–11. https://scholarscompass.vcu.edu/auctus/48.

25. Khan, B.A. *Emergence of Technostress in Multi-User VR Environments for Work-Related Purposes*; Tampere University: Tampere, Finland, 2023.
26. Mancin, P.; Ghisi, M.; Spoto, A.; Cerea, S. The relation between body dysmorphic disorder symptoms and photo-based behaviors: Is body appreciation a protective factor? *Body Image* **2024**, *51*, 101764. [CrossRef]
27. Westerman, D.; Tamborini, R.; Bowman, N.D. The effects of static avatars on impression formation across different contexts on social networking sites. *Comput. Hum. Behav.* **2015**, *53*, 111–117. [CrossRef]
28. Omar, K.; Fakhouri, H.; Zraqou, J.; Marx, J. Usability Heuristics for Metaverse. *Computers* **2024**, *13*, 222. [CrossRef]
29. Tursunov, J.; Rozinaj, G.; Minarik, I.; Vanco, M. Creating a Metaverse-Based Meeting Room with Live Avatar Control and Web Communication. In Proceedings of the 2024 International Symposium ELMAR, Zadar, Croatia, 16–18 September 2024; pp. 57–60.
30. Combe, T.; Fribourg, R.; Detto, L.; Normand, J.M. Exploring the Influence of Virtual Avatar Heads in Mixed Reality on Social Presence, Performance and User Experience in Collaborative Tasks. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 2206–2216. [CrossRef]
31. Alcántara, J.C.; Tasic, I.; Cano, M.D. Enhancing digital identity: Evaluating avatar creation tools and privacy challenges for the metaverse. *Information* **2024**, *15*, 624. [CrossRef]
32. Javanmardi, A.; Pagani, A.; Stricker, D. G3FA: Geometry-guided GAN for Face Animation. In Proceedings of the 35th British Machine Vision Conference, BMVC 2024, Glasgow, UK, 25–28 November 2024.
33. Yang, S.; Jiang, L.; Liu, Z.; Loy, C.C. VToonify: Controllable High-Resolution Portrait Video Style Transfer. *ACM Trans. Graph. (TOG)* **2022**, *41*, 5153. [CrossRef]
34. Brooke, J. SUS: A "quick and dirty" Usability Scale. In *Usability Evaluation in Industry*; Jordan, P.W., Thomas, B., Weerdmeester, I.L.M., Eds.; Taylor and Francis: Abingdon, UK, 1996; pp. 189–194.
35. Roth, D.; Latoschik, M.E. Construction of the Virtual Embodiment Questionnaire (VEQ). *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 3546–3556. [CrossRef] [PubMed]
36. Longo, M.R.; Schüür, F.; Kammers, M.P.M.; Tsakiris, M.; Haggard, P. What is embodiment? A psychometric approach. *Cognition* **2008**, *107*, 978–998. [CrossRef]
37. Nimrod, G. Technostress: Measuring a new threat to well-being in later life. *Aging Ment. Health* **2018**, *22*, 1080–1087. [CrossRef]
38. Langer, M.; König, C.J.; Krause, K.; Fitili, A. Examining Digital Interviews for Personnel Selection: The Effects of Interview Mode on Performance Ratings and Social Presence. *Comput. Hum. Behav.* **2022**, *126*, 107012. [CrossRef]
39. McKnight, D.H.; Carter, M.; Thatcher, J.B.; Clay, P.F. Trust in a Specific Technology: An Investigation of Its Components and Measures. *ACM Trans. Manag. Inf. Syst. (TMIS)* **2011**, *2*, 1–25. [CrossRef]
40. Bente, G.; Rüggenberg, S.; Krämer, N.C.; Eschenburg, F. Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations. *Hum. Commun. Res.* **2008**, *34*, 287–318. [CrossRef]
41. Castilla, D.; Botella, C.; Miralles, I.; Bretón-López, J.; Dragomir-Davis, A.M.; Zaragoza, I.; Garcia-Palacios, A. Teaching digital literacy skills to the elderly using a social network with linear navigation: A case study in a rural area. *Int. J.-Hum.-Comput. Stud.* **2018**, *118*, 24–37. [CrossRef]
42. Castilla, D.; Garcia-Palacios, A.; Miralles, I.; Breton-Lopez, J.; Parra, E.; Rodriguez-Berges, S.; Botella, C. Effect of Web navigation style in elderly users. *Comput. Hum. Behav.* **2016**, *55*, 909–920. [CrossRef]
43. Castilla, D.; Suso-Ribera, C.; Zaragoza, I.; Garcia-Palacios, A.; Botella, C. Designing ICTs for Users with Mild Cognitive Impairment: A Usability Study. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5153. [CrossRef] [PubMed]
44. Yee, N.; Bailenson, J. The Proteus Effect: The effect of transformed self-representation on behavior. *Hum. Commun. Res.* **2007**, *33*, 271–290. [CrossRef]
45. Goffman, E. The Moral Career of the Mental Patient. *Psychiatry* **1959**, *22*, 123–142. [CrossRef]
46. Jawad, M.; Talreja, K.; Bhutto, S.A.; Faizan, K. Investigating how AI Personalization Algorithms Influence Self-Perception, Group Identity, and Social Interactions Online. *Rev. Appl. Manag. Soc. Sci.* **2024**, *7*, 533–550. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Article

Wormhole Whispers: Reflecting User Privacy Data Boundaries Through Algorithm Visualization

Xiaoxiao Wang †, Jingjing Zhang †, Huize Wan and Yuan Yao *

School of Architecture and Design, Beijing Jiaotong University, Beijing 100044, China;
21312070@bjtu.edu.cn (X.W.); 21312052@bjtu.edu.cn (J.Z.); 21312071@bjtu.edu.cn (H.W.)
* Correspondence: yuanyao@bjtu.edu.cn
† These authors contributed equally to this work.

Abstract: As user interactions on online social platforms increase, so does the public's concern over the exposure of user privacy data. However, ordinary users often lack a clear and intuitive understanding of how their personal online information flows and how algorithms are involved in this process. We have developed a design to address this issue that visualizes data algorithms and transmission based on data collection and analysis. The primary contribution of this work is an innovative approach that uses metaphorical visual language to represent the flow of user data, algorithms, and information. It also aims to present these concepts interactively, allowing users to gain a more vivid understanding of the dynamic changes in their personal data. This participatory and interactive design seeks to provide users, as well as various stakeholders, an opportunity to reflect on the relationship between online society, privacy data boundaries, and users' rights to be informed.

Keywords: user privacy data; data visualization; data privacy and security

1. Introduction

As smartphone penetration continues to rise, internet usage has become increasingly integrated into daily life. By July 2024, there were 5.45 billion internet users worldwide, representing 67.1% of the global population. Of these, 5.17 billion, or 63.7%, were social media users 2024 Number of internet and social media users worldwide as of July 2024, https://www.statista.com/statistics/617136/digital-population-worldwide/, accessed on 26 November 2024.

In this context, concerns regarding online privacy and information security are growing, with users becoming more aware of the risks of data exposure [1]. Research indicates that some users are wary of the accuracy of big data algorithms, fearing their personal information could be accessed and used without their consent [2]. Meanwhile, many users remain unclear about data usage policies, algorithms, and processes, especially when these data are embedded in complex machine learning systems, which are often opaque and protected by trade secrets. These factors have amplified concerns about personal data exposure [3], particularly in "contracting" scenarios where data services are required. The associated terms are often lengthy, filled with complex legal jargon, and difficult to navigate [4].

For most internet users, despite their concerns about privacy breaches, there is little opportunity to take effective action [5]. Moreover, a less obvious risk is that while sharing

information on social media platforms is often considered "voluntary", users may not fully understand the implications of providing data to third parties. As a result, they may unknowingly accept terms that compromise their privacy [6], either actively or under duress [7].

Furthermore, several high-profile data privacy breaches, such as the June 2024 case in which Italy's antitrust authority fined Meta for failing to adequately inform Instagram users that their data would be used for commercial purposes (2024 Italian watchdog fines Meta for unfair commercial practices, https://www.reuters.com/technology/italys-antitrust-fines-meta-35-mln-euros-unfair-commercial-practices-2024-06-05/), highlight the negative impact of these issues, accessed on 25 December 2024.

China's vast online population reached nearly 1.1 billion internet users (1.09967 billion) as of June 2024. Additionally, the internet penetration rate in rural areas has been steadily increasing, and the construction of network infrastructure continues to advance (2024 The 54th Statistical Report on China's Internet Development, https://www.cac.gov.cn/2024-03/25/c_1713038218396702.htm, accessed on 26 November 2024), which encompasses a wide range of demographic groups across urban and rural settings. This extensive coverage ensures a high level of representativeness, making China the primary data source and focus of our analysis. We conducted a user survey on the popular Chinese social media platform Weibo (2025 Weibo: Share anything you want, let the whole world hear your voice, https://www.weibo.com, accessed on 18 January 2024), collecting 685 posts using the hashtags "BigData" and "Eavesdropping" as filters. The survey revealed that users' primary data privacy concerns include the specific methods platforms use to collect and handle user data, the security of data in various usage scenarios, and the potential for misuse. However, most algorithm visualizations are intended for professionals, aiming to convey and optimize specialized knowledge [8]. Many existing studies on data privacy education aim to raise users' awareness and teach security and privacy concepts effectively through serious games [9], card-based methods [10], and similar approaches. In our experience, designers rarely attempt to make complex algorithms more accessible, and users seldom have the chance to understand how their online data flows clearly. The high technical barriers create a dilemma between concern and action for the average user. This limits independent thinking about data privacy and contributes to a lack of design insights in this field, presenting an opportunity for further design and research.

Therefore, we aim to conduct an empirical study on Visualization of data flow process and related algorithms, focusing on the perspective of ordinary users to summarize and represent the circulation and processing of user data on social media platforms. This approach seeks to raise user awareness of the flow of their private data. First, the study designed a web scraping and analysis process for the privacy policies of five social media applications and the data from 13 third-party SDKs. Following that, we combined the results of data analysis and related algorithms (clustering analysis, collaborative filtering, and association rule learning) with a poetic metaphorical data visualization method, creating a system of metaphorical visual symbols. Finally, we invited 10 users to participate in an experiment to experience and discuss the visualization design and its themes.

Building on the aforementioned context, our research lies at the intersection of three fields: data visualization, user experience design, and privacy data science education. We focus on how innovative visualization and interactive methods can help everyday users understand and navigate complex data privacy concepts. The primary contribution of our work is not in technical expression, but in providing a novel methodological framework for data privacy education through design innovation. Our target audience primarily consists of general internet users interested in protecting personal data privacy but lacking

specialized knowledge. Additionally, our research offers valuable insights for privacy educators, data visualization researchers, and platform designers. The main contributions of this study are as follows:

1. We propose a metaphor-based visualization framework that translates complex data privacy mechanisms into accessible visual narratives, specifically targeting non-expert users' understanding of algorithmic data processing on social media platforms.
2. We develop a systematic transformation framework that converts technical privacy concepts into visual metaphors through three interconnected stages: (1) data flow analysis of privacy policies and SDK sharing patterns, (2) algorithmic principle deconstruction focusing on clustering, filtering and association learning, and (3) interactive visualization design validated through user studies.
3. Through empirical evaluation with users, we demonstrate how metaphor-based visualization can (1) increase user comprehension of privacy-related algorithms, (2) motivate users to reconsider their privacy decisions, and (3) provide insights for improving privacy policy presentation in social media platforms. Additionally, we identify key challenges and opportunities in designing privacy-aware interfaces that balance user understanding with technical accuracy.

This paper is organized as follows. We begin by reviewing related work in three areas: the disclosure of privacy policies, users' concerns about data security, and research on algorithms and algorithm visualization. In Section 3, we present our design approach, which includes data collection and analysis, visual transformation using poetic metaphors, and an interactive design based on user research. In Section 4, we discuss our findings and reflections, focusing on user experience, users' reflections and actions regarding the flow of privacy data, the accessibility benefits of metaphorical algorithm visualization, and the issue of hidden privacy data. Finally, we examine the strengths and limitations of our design approach and suggest directions for future work.

2. Related Work

2.1. Disclosure of Privacy Policy

Currently, some studies are actively attempting to disclose data security and transmission methods to users while analyzing their awareness of privacy policy content [11,12]. These studies highlight issues with privacy policies and emphasize that in "free" internet services, data is often treated as a form of currency (2022 Why Source Data Is The New Currency For Retailers, https://www.forbes.com/sites/forbestechcouncil/2021/11/03/why-source-data-is-the-new-currency-for-retailers/?sh=7a02117a5e11, accessed on 16 February 2024) In the digital environment of social networks, personal data is particularly interconnected. Users are reminded that agreeing to allow access to their profiles carries the risk of exposing their personal data from network connections to unauthorized visitors [13]. Additionally, these studies evaluate the potential privacy risks of various system designs [14] and work on visualizing examples of longitudinal information that third-party trackers can infer from user browsing [15]. As noted by researchers such as Solove D. J., "Data-breach harm might often be intangible, but it still is very real" [16]. These studies provide valuable support for the data security and policy aspects of our research on visualization and public awareness design.

2.2. Users Concerned About Data Security

We found that although users are very concerned about data security and personal privacy, they still face a dilemma between worry and action. The most common concerns

are related to data collection methods. For example, users worry about data being collected without a clear explanation [17]. When purchasing IoT devices, users consider privacy and security factors [18]. Despite concerns about privacy exposure, users have almost no ability to negotiate with online service providers due to a lack of sufficient focus groups and expertise. They are caught in the dilemma of disclosing too much personal information versus the pressure of social isolation [19].

These studies help us better understand the real psychology of users and reveal that users lack an intuitive understanding of the flow of their personal online information. In our design research, we attempt to provide users with easy-to-understand visualizations of privacy algorithms.

2.3. Algorithm Research and Visualization

The application of algorithm analysis in various software and platforms is becoming increasingly widespread. Current research mainly focuses on optimizing various algorithm models [20,21] and their applications in network and privacy security [22]. There are also many studies on potential issues caused by algorithms, such as the link between algorithms and harm on social media platforms [23] and how algorithms shape collective action [24].

However, most existing studies primarily focus on textual analysis and the deconstruction of algorithmic principles, or they offer in-depth references for professionals and technicians. Many studies have confirmed the educational benefits of algorithm visualization, enabling students to better understand algorithms [25,26]. Yet, there is relatively little research dedicated to the popularization and visualization of privacy algorithms for general users. As designers, we aim to bridge this gap by creating direct and impactful visualization tools for users. Examples include Joern Roeder and Jonathan Pirnay's fbFaces device (2011 fbFaces [Installation]. University of Visual Arts and Design Kassel, https://www.alphavillefestival.co.uk/fbfaces, accessed on 14 February 2024), Dima Yarovinsky's "I Agree" (2018 This Guy Printed out the Terms of Service for the World's Most Popular Apps, https://twistedsifter.com/2018/05/i-agree-by-dima-yarovinsky/, accessed on 14 February 2024), and Laurie Frick's "Felt Personality" (2018 Felt Personality, https://www.lauriefrick.com/felt-personality, accessed on 14 February 2024). The extensive work of previous researchers provides valuable technical theories and design insights, which support the development of our own visualization systems.

3. Design Method

This study begins with a meta-data user survey, using the analysis of user attitudes to explore the underlying causes of the current situation. Data collection was based on privacy policies from five of China's most popular social media platforms, along with web-scraped data from 123 SDKs mentioned in the policies of 13 popular apps. Collection sources included official websites and app privacy policy terms. Based on these data, literature research, and big data analysis methods, we focused on the personalized recommendation features of each platform. We selected three key algorithms—clustering analysis, association rule learning, and collaborative filtering—for detailed explanation, analysis, and visualization. Additionally, user experiments and design tests were conducted using third-person interactive design, aimed at helping ordinary users better understand data privacy and boundaries.

3.1. Data Collection

China has the largest internet user base in the world, with over 1.1 billion users as of June 2024. This figure not only highlights the vast scale of the user base but also

positions China as a leading ground for global digital transformation. In this context, users' awareness of privacy is shifting from passive acceptance to active consideration, reflecting a broader global trend among digital users. As such, the study of Chinese users provides valuable insights for anticipating future global user behavior.

To understand users' psychological attitudes toward privacy data collection on social media platforms, the research team first surveyed key Chinese platforms, such as Weibo and Bilibili, and gathered relevant data. Many users expressed concerns about eavesdropping on these platforms. In China, "big data" has become a widely accepted technical term, while "eavesdropping" is commonly used to describe concerns about privacy violations. The frequent discussions surrounding these keywords on social media often lead to in-depth debates and experience-sharing among users. Using the keywords "#bigdata" and "#eavesdropping", the research team collected 685 posts through web crawlers. After excluding irrelevant data, such as advertisements, marketing posts, and off-topic comments, 635 valid posts remained. Data collection occurred on 30 October 2023, and the posts were published between October 2022 and October 2023.

We also selected the five most popular social media platforms in China by download volume (WeChat (2025 WeChat: Connecting a billion people with calls, chats, and more, accessed on 16 January 2025. https://www.wechat.com/en/), Weibo, REDnote (2025 REDnote: Inspire Lives Shares and discovers the wonders of the world, accessed on 16 January 2025. https://www.xiaohongshu.com/explore), Bilibili (2025 Bilibili: Southeast Asia's leading anime, comics, and games (ACG) community where people can create, watch and share engaging videos, accessed on 16 January 2025. https://www.bilibili.com/), and Taobao (2025 Taobao: Online shopping platform, accessed on 16 January 2025. https://www.taobao.com/)) and conducted a comprehensive analysis of their user privacy policies, as shown in Figure 1. This analysis included collecting, organizing, and performing word frequency analysis on the policy texts.

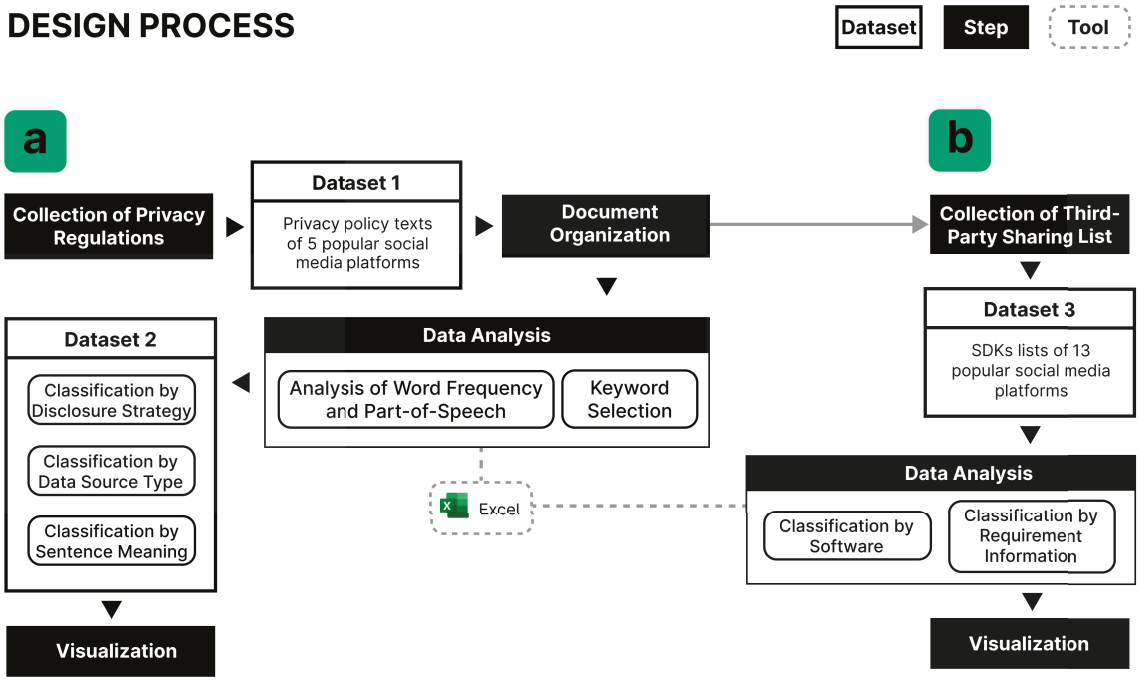


Figure 1. Privacy policy and SDK data collection analysis and visualization flowchart. (**a**) Collecting and organizing user privacy policies of the five most popular social media platforms. (**b**) Collecting and organizing a third-party SDK data-sharing checklist.

At the same time, we found that all five popular applications mentioned above have third-party SDK data-sharing lists, as shown in Figure 1. To analyze the sharing relationships of personal data between different software and third-party SDKs, we collected the quantity and types of user information provided by 13 popular software applications to third-party SDKs. Our data sources included official websites and in-app terms. We organized and compared the similarities and differences between different software, extracting the frequency of various third-party SDKs appearing under different information requirement categories.

3.2. Data Analysis

Statistical data indicate that 94.33% of comments expressed concern/fear, as shown in Table 1. Privacy invasion and data security are the primary emotional concerns. The most frequent negative keywords are "terrifying", "annoying", and "suspicious", highlighting users' discomfort with how platforms currently define privacy boundaries. This suggests that many users believe they are being eavesdropped on and monitored by big data. These insights provide essential data and user research references for future studies on privacy data transmission and visualization design.

Table 1. Classify the reactions of Weibo users using the filtering term "BigData" and "eavesdropping".

Reaction	Definition	Example	Percentage (N)
Concern/Fear	Describes the user expressing anxiety or concern about data security, privacy, or negative consequences of technology.	*Big data is terrifying. It eavesdropped on me saying my birthday is coming up.*	94.33% (599)
Frustration/Anger	Expresses dissatisfaction or frustration with the system, perceived invasions of privacy, or general irritation with technology.	*Every time I talk to a friend, my phone pushes related ads. It feels like I'm being listened to, and it's really infuriating.*	3.78% (24)
Confusion/Curiosity	The user is unsure about what is happening or is asking for clarification.	*How does this app know I've been looking for these things recently? I haven't even searched for them online.*	1.26% (8)
Suspicion	The user expresses doubt or suspicion, possibly related to conspiracy theories or fears about being monitored.	*I'm afraid of using certain social media platforms. They seem to be constantly monitoring me, tracking all of my online activities.*	0.63% (4)

In analyzing privacy policy data, we focused on keywords related to the collection and processing of personal data to ensure relevance, with special attention to adjectives, adverbs, and other terms indicating attitude and degree, thereby supporting subsequent classification and visualization, as shown in Figure 2. We applied the "Classification by Disclosure Strategy" to examine implicit strategies within these policies—such as mandatory, suggestive, or guiding approaches—to highlight their stance on user consent (e.g., "mandatory" or "optional"). Additionally, we classified data collection sources, distinguishing between internal and external origins to clarify specific data collection paths (Figure 1). We found that most platform privacy policies persuade users to accept or comply with terms by stating that certain features cannot be used otherwise, thereby obtaining user consent. Additionally, platforms consider de-identified personal data, which cannot identify individuals, as non-private user data. These data are widely used in platform algorithms and shared with multiple parties. In situations of information asymmetry, users may place excessive trust in these platforms. Furthermore, most platform privacy policies are complex

in both text and concept, making them difficult to read. Some network service providers may also write privacy policies in very general terms, only briefly explaining what data will be collected and how it will be used.

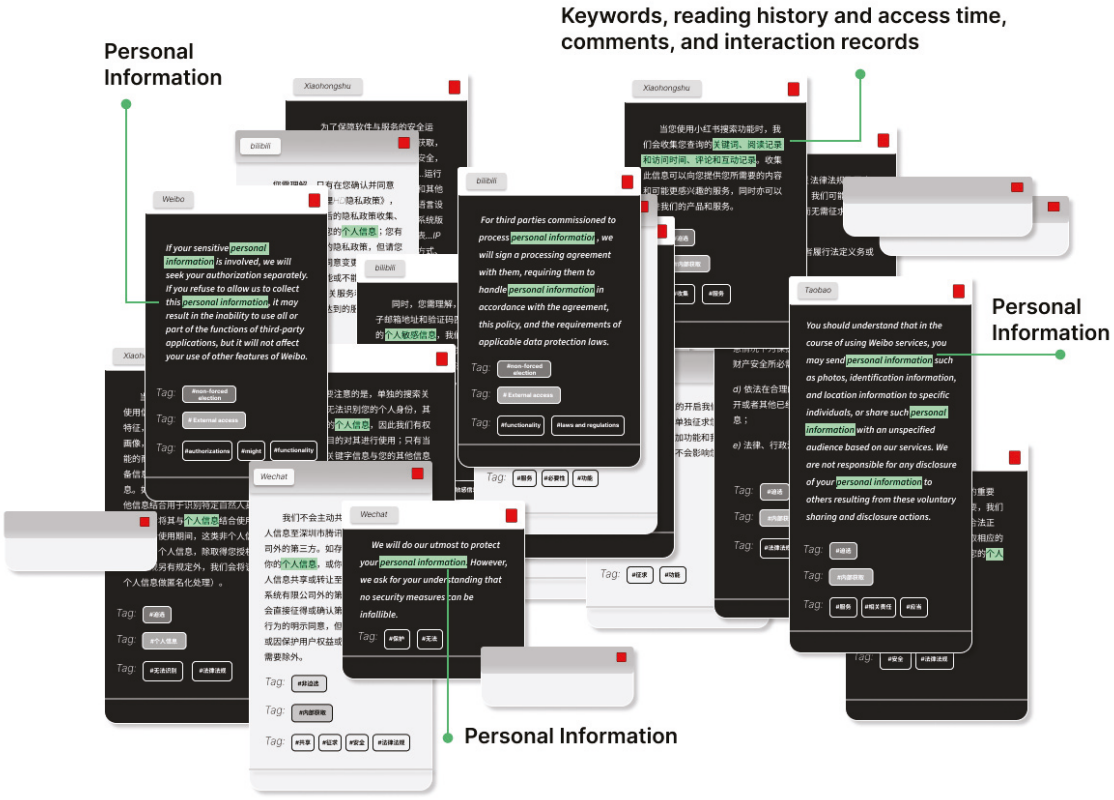


Figure 2. Organization of user privacy policies from the five most popular social media platforms.

Additionally, app agreements often include clauses such as "may include third-party products and services" and "share necessary information with third-party vendors", but the purposes and scope of the shared information are not clearly defined. The services provided by third parties are mostly invisible to users, so the embedded SDKs in apps may also obtain the same data access and usage permissions without the user's knowledge. Our data analysis revealed that the information shared between apps and third parties is extensive, including device identifiers, installation permissions, and even information from other apps. For example, among the 13 apps and 15 types of collected information, the SDK embedded in the "REDnote" app appeared most frequently ("Huawei Push SDK" appeared 15 times and "Mobile SDK" appeared 34 times). As users, we primarily rely on the information disclosed by developers in privacy policies. This makes it difficult to know where our data go and to determine whether the permissions we grant are necessary for the app's services or are being marketed and sold as currency between platforms.

Through the investigation of previous studies and the analysis of collected data, we noticed that many users and researchers mentioned "algorithms". Users suspect they are being "eavesdropped on" also because platforms and operators use behavior-targeting strategies after obtaining user data. These strategies involve using algorithms to process users' behaviors (browsing, searching, purchasing, etc.) on websites or applications to push related advertisements.

Reasons for Choosing the Algorithms

In the data processing of social media platforms, the choice of specific algorithms directly impacts how user privacy data are utilized. Analyzing previously collected social media posts revealed that users' main privacy concerns focus on how platforms collect, analyze, and utilize their personal data. Based on these findings, we selected three representative algorithms, closely related to user privacy, for visualization research.

Cluster Analysis: The Foundation of User Profiling. Cluster analysis algorithms are commonly used on social media platforms for user segmentation and profiling, directly addressing the issue of "being categorized", which concerns many users. This algorithm is a valuable tool in data science [27,28] and plays a crucial role as a core element of data analysis. By analyzing data such as browsing habits, search history, and interaction behaviors, the algorithm groups users with similar characteristics. The primary reason for selecting this algorithm is that it provides an intuitive way to demonstrate how platforms interpret and categorize user behavior. It is also closely linked to the "information cocoon" phenomenon often mentioned in user feedback. Additionally, it forms the foundation for personalized services, significantly shaping how users access information.

Collaborative Filtering: The Core of Recommendation Systems. Collaborative filtering algorithms are central to personalized recommendations and are closely linked to users' common sense of being "monitored". There are two primary types of collaborative filtering algorithms: user-based and item-based. As one of the most widely used recommendation algorithms [29], it explains why users receive suggestions based on the behavior of similar users. This method illustrates how platforms can deliver personalized recommendations without directly accessing users' private data.

Association Rule Learning: Predicting Behavioral Patterns. Association rule learning, also known as market basket analysis (MBA), is a data mining technique that helps retailers understand buyers' purchasing behaviors, enabling them to make more informed decisions [30]. By analyzing user behavior patterns and predicting potential interests, this technique directly addresses users' concerns about platforms "predicting their needs". It shows how platforms use historical data to forecast user interests, how user data drive business decisions, and how users can better understand the data analysis behind personalized services.

The selection of these algorithms for this study is based on two main considerations. First, they directly influence the personalized content users encounter, and their mechanisms are closely tied to user privacy. Second, they represent three key aspects of data utilization: data classification, association analysis, and personalized recommendations. This approach allows for a comprehensive exploration of the data processing workflow. The combination of these three algorithms addresses the privacy concerns most frequently raised by users. It also creates a complete data processing chain, from user segmentation (cluster analysis), to personalized recommendations (collaborative filtering), and finally to behavior prediction (association rule learning). By analyzing and visualizing these algorithms, users can gain a clearer understanding of how platforms handle their personal data, empowering them to make more informed privacy decisions.

After organizing and analyzing metadata, we identified issues in privacy policies, particularly regarding the use and circulation of user information by third-party SDKs. Further analysis revealed the role of algorithms in this data circulation. User privacy data are extensively shared and used by third-party SDKs after users agree to the terms and policies. Algorithms act as a black box, processing these data and pushing content that makes users feel "eavesdropped on", as shown in Figure 3. These three components are

integral to the operational systems of most platforms. Therefore, in the subsequent design phase, algorithm visualization will be a key foundation of interactive design.

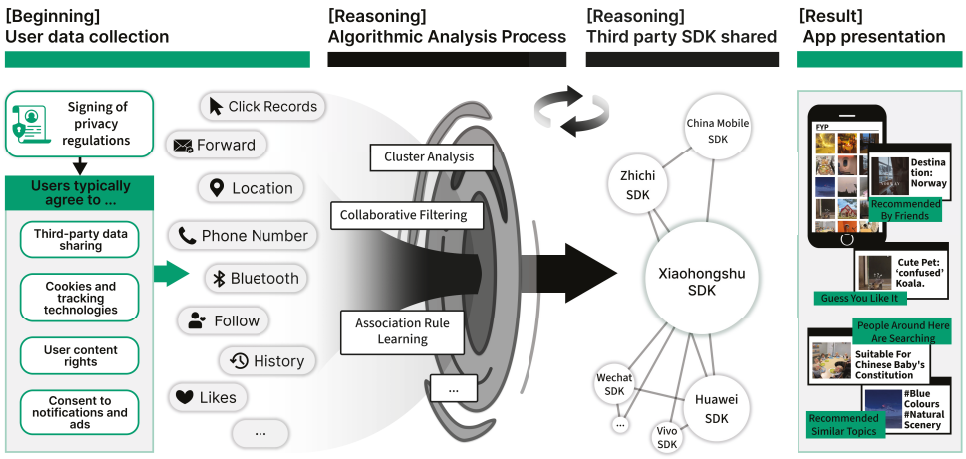


Figure 3. The basic process of design methodology, when combined with the wormhole metaphor, establishes the foundation for this visualization's metaphorical logic.

3.3. Visual Transformation

3.3.1. Visual Method: Infopoetry

Clearly and rigorously expressing the flow process and underlying mechanisms of visualized privacy data across platforms is a challenging task that requires substantial professional knowledge. In design, translation is not about understanding "the meaning expressed in another language", but about conveying something that originally had no clear form or textual structure in a novel or unexplored way. Therefore, this study introduces the design concept of "Infopoetry" [31]. Infopoetry goes beyond traditional information design; it aims to establish a representative scenario rather than merely transforming measurable numbers into visual charts.

In this project, the design practice attempts to construct relevant scenarios to allow visual symbols to reconstruct the data privacy dilemmas users encounter in reality (Figure 3). The approach follows four steps—extracting data features, emotional reflection, building metaphors, and determining the appropriate medium—to establish information visualization and corresponding interactive design strategies. Although the algorithms involved in this process are not limited to the three mentioned above, here, we will only deconstruct and visualize the principles of the three most popular algorithms.

3.3.2. Visualization

Through previous data collection and analysis, we found that personal information is like interstellar matter in the universe, either attracted or altered. Its data transmission process is akin to entering "wormholes" and "black boxes", where no one knows what changes might occur or where the data might go. The black-box operation of social media platform algorithms exacerbates this opacity. These wormholes and black boxes constitute a significant part of the current electronic information ecosystem. Therefore, to better disclose the principles of algorithms to users, we use the metaphor of roaming the interstellar universe to visualize key algorithms in the privacy data transmission process, namely clustering analysis, collaborative filtering, and association rule learning.

Clustering analysis groups data objects based on their descriptions and relationships found in the data. The goal is to make objects within the same group similar (related) to

each other, while objects in different groups are dissimilar (unrelated). Clusters obtained through clustering can be represented by their centers, sizes, densities, and descriptions.

In the planetary metaphor, clustering analysis is like grouping similar types of planets into one galaxy. For example, during a user's interstellar adventure, if they see planets with rings, monochrome planets, and two-tone planets, they might group all ringed planets into one galaxy, all monochrome planets into another, and all two-tone planets into a third. Each galaxy represents a "cluster". A visualization of this metaphor can be seen in Figure 4. This metaphor clusters based on the type of planets (i.e., "features"). In user behavior data analysis, users are categorized based on their "features", which could be their shopping habits, the ads they click on, or the time they spend on a website.

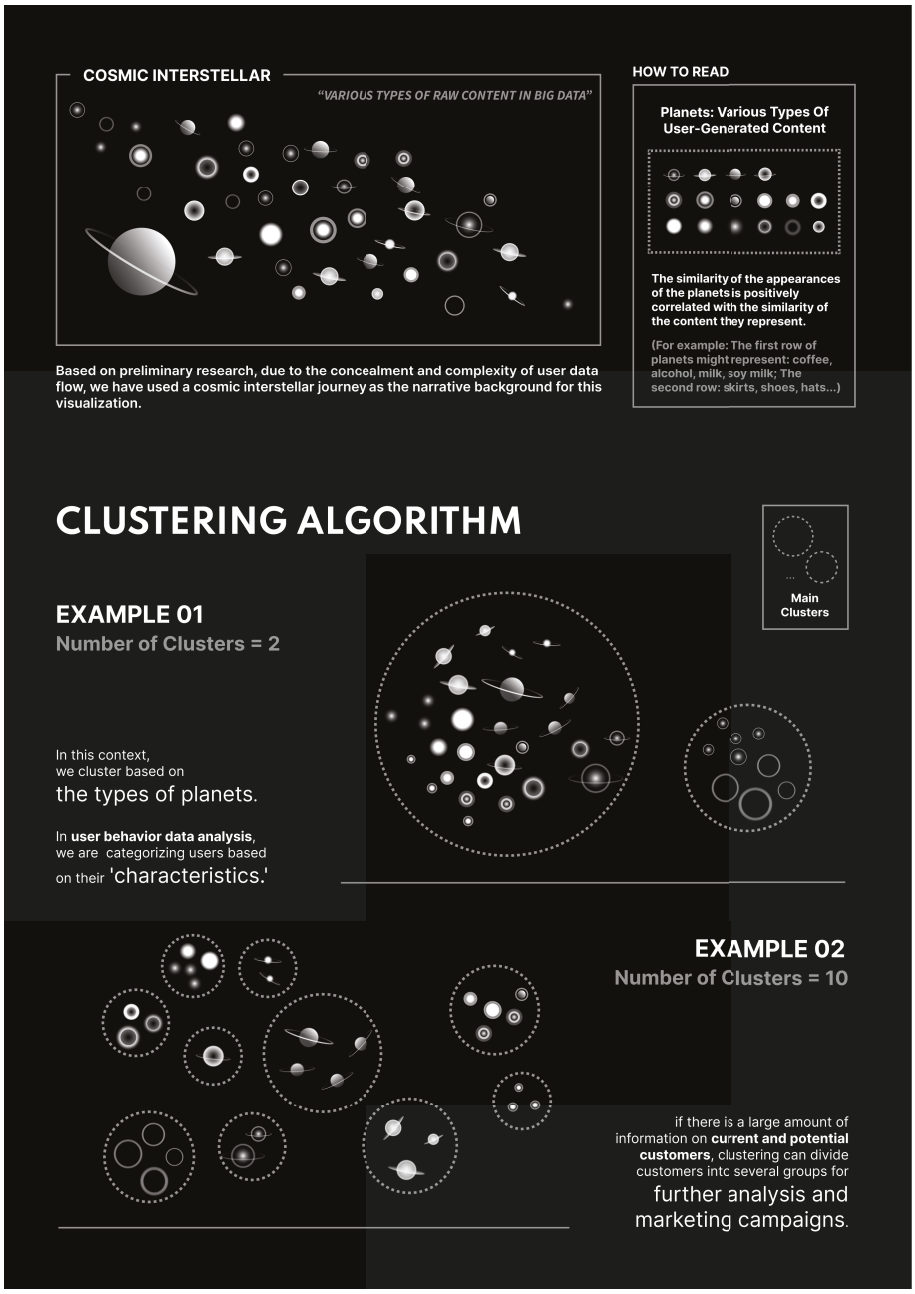


Figure 4. Visualization diagram of clustering algorithm: dividing planets into different galaxies based on their characteristics.

Collaborative filtering is divided into two types: user-based collaborative filtering (UserCF) and item-based collaborative filtering (ItemCF). UserCF analyzes the similarity between users to find a group of users most similar to the target user. It then predicts the target user's preferences for unknown items based on the behavior of these similar users. ItemCF, on the other hand, analyzes the similarity between items to find a group of items most similar to the target item. It then predicts the target user's preferences for unknown items based on their preferences for these similar items.

In the visualization of collaborative filtering algorithms, planets liked by friends, planets liked by similar space explorers, and planets browsed are more likely to pass through layers of filters to appear before the user, as shown in Figure 5.

Association rule learning is a key technique in unsupervised learning algorithms used to discover associations between itemsets in a dataset. It identifies association rules by analyzing the frequency and relationships of itemsets within the dataset.

In this visualization design, during their cosmic exploration journey, as shown in Figure 6, users may notice that whenever they find a central gaseous planet, a large gaseous planet is often nearby. Similarly, whenever users see a glowing planet, there is likely another planet with a similar structure nearby. This is less common with other types of planets; for example, ringed planets are usually found next to distinctly layered planets. These association rules are valuable as they help us explore the universe more effectively and find rare resources. It is similar to how a great detective follows traces or how one finds patterns and reasons in games like Sudoku or Minesweeper.

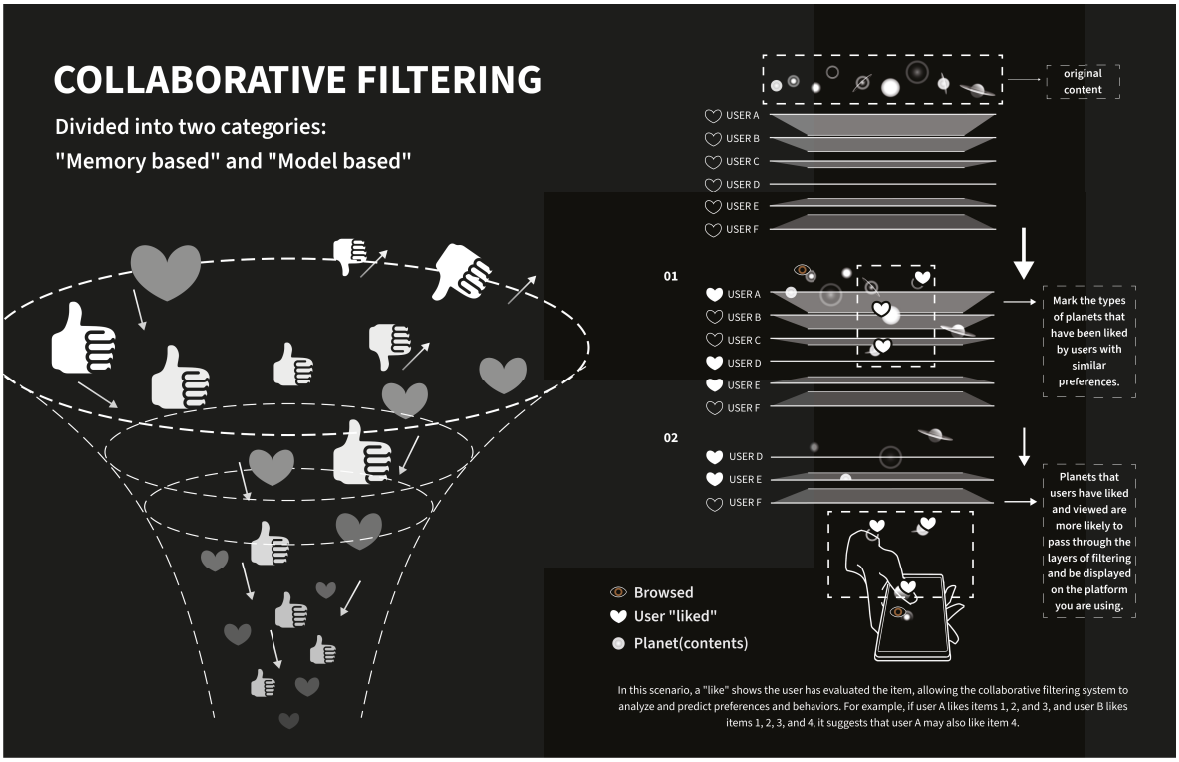


Figure 5. Visual diagram of collaborative filtering algorithm: user-centered filtering of content.

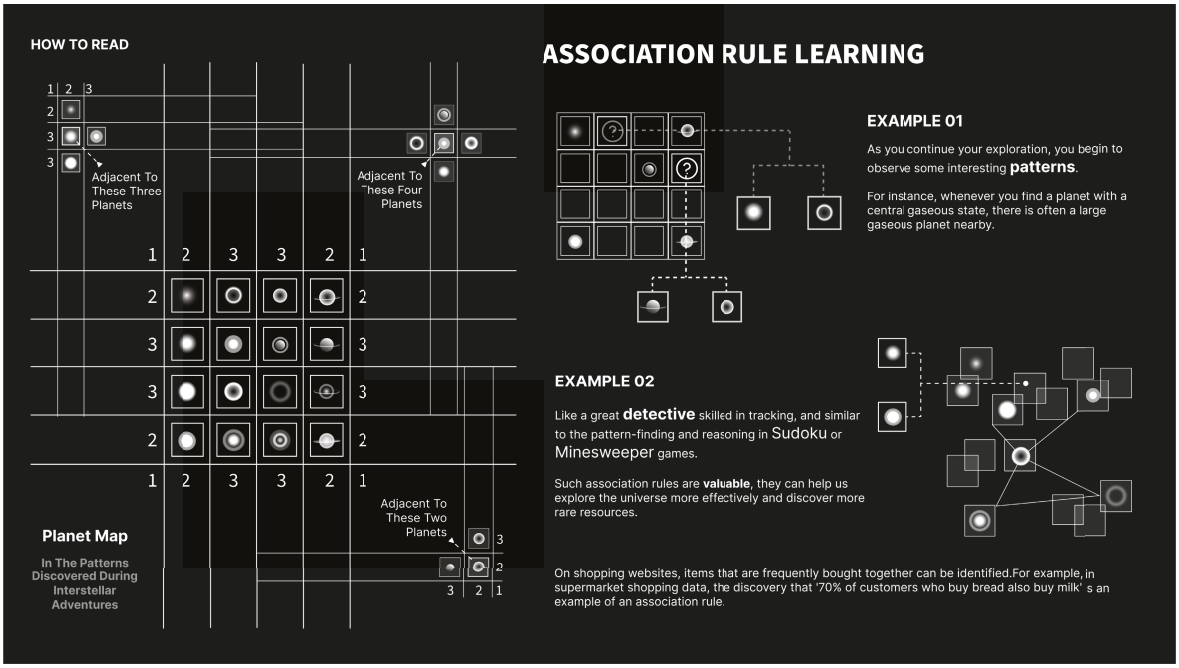


Figure 6. Visual diagram of association rule learning algorithm: discovering patterns in a large number of planets (contents).

3.4. Interaction Design & User Study

To enhance the accessibility of the visualization results, we added interactive features to the previous data visualizations. Users can click, drag, and pan to simulate the process of navigating the platform's webpage, as shown in Figure 7. This gamified interactive scenario amplifies previously overlooked privacy boundary issues, making the mechanisms of algorithm and privacy data conversion more intuitive.

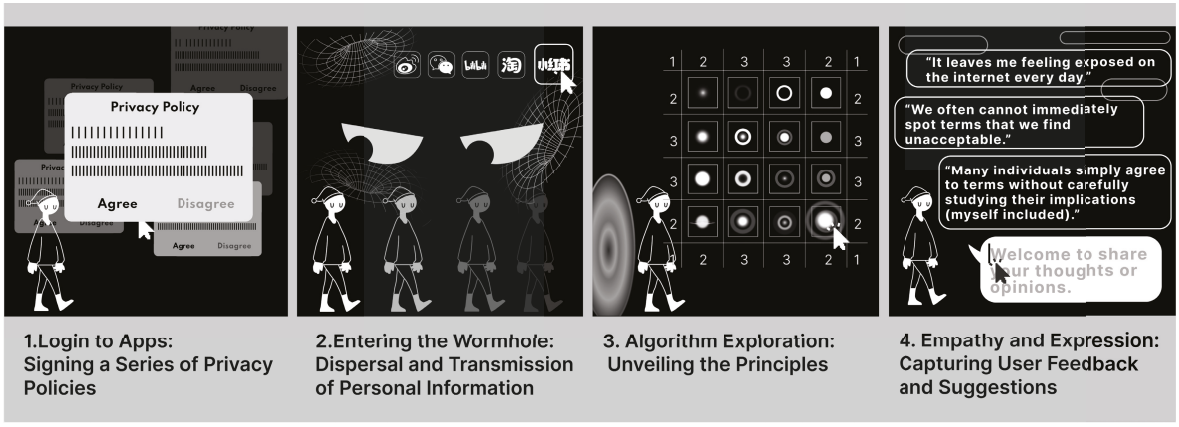


Figure 7. The four components of interaction design transformation.

To gather feedback from general users, we invited 10 participants (six females, four males, mean age = 23.1, *SD* = 1.58) to experience the interactive process and share their thoughts, as shown in Table 2. The average interaction time was 6 min. By observing and recording the participants' behaviors and feedback, we gained insights into how users perceive the flow and use of personal privacy data across multiple social platforms. All participants indicated that the project gave them a more intuitive understanding of privacy policies and algorithms. Several participants mentioned that the experience helped them better understand their previous experiences and concerns.

Table 2. Feedback from general users.

Participants	Age	Reaction
P1 (F)	20	*I definitely have a more intuitive understanding of the privacy policies, including some of the algorithms.*
P2 (F)	22	*Honestly, I've always thought agreeing to terms and conditions is kind of funny.*
P3 (M)	22	*Regarding the latter part about cluster analysis, when I first started using TikTok, I thought, "Wow, so this is how most people are". Until I stumbled upon the concept of the "information bubble".*
P4 (M)	23	*Whenever I show interest in something, suddenly everything related to it, just flood in from everywhere.*
P5 (F)	23	*Our privacy data is rapidly transmitted to various unknown places, indeed like traveling through a wormhole*
P6 (F)	23	*I really hope that the designers and tech teams can do more to educate regular people like me and push those capitalists to stop spying on us and make some real changes.*
P7 (M)	23	*This kind of game makes me feel like the internet has gone too far. Now I get why I always get calls from people trying to sell me a car or a house.*
p8 (F)	24	*Most apps tend to overreach when it comes to permissions, and some even block you from using them if you don't agree to everything.*
P9 (M)	25	*It's just like those forced pop-ups in games—nowadays, so many apps bombard you with permission requests when you open them.*
P10 (F)	26	*If apps could just stick to privacy agreements and keep users' information safe, I think that would be amazing!*

Most participants indicated that this project gave them a more intuitive understanding of privacy policies and algorithms. In particular, regarding the demonstration of the "black-box" nature of technology, participants' feedback validated the accuracy of the design. For example, P1 expressed during the experience "The flow of user information on the platform should involve many complex algorithms and technologies, all of which are invisible to the outside world. This metaphor of a 'black box' precisely matches my imagination of this phenomenon". P7 also remarked "Our private data is rapidly and efficiently transmitted to various unknown places, mysterious and complex. It indeed feels like walking through a wormhole".

Some participants also shared their feelings after experiencing this study. P6 mentioned "After searching for skincare products for my mom on an online shopping platform, I received related product advertisements on Douyin or another shopping platform the next second. It really scared me. Now I roughly understand what's going on".

P8 raised concerns about current authorization mechanisms: "Major apps often require excessive permissions, and in some cases, refusing to grant them renders the app unusable. In these situations, personal privacy is compromised, and it feels like being exposed on the internet every day".

The visualization design also enhances users' understanding of the data collection and usage mechanisms. For example, P10 noted "Now I have a better understanding of the methods they use to collect my information, so I can avoid them when necessary, without unnecessary suspicion or concern". In contrast, P9 emphasized the importance of prior informed consent for data authorization, stating "I always find out afterward that I've agreed to share my information, which makes me worried that one day my private data could be exposed". This feedback highlights the impact of the hidden nature of privacy issues on users.

These comments reaffirm the negative impact on users due to the hidden nature of these issues. More than half of the participants indicated they would reconsider their privacy decisions. However, some participants were pessimistic about their ability to control privacy decisions and expressed a desire for the right to negotiate with platforms. Additionally, we received several suggestions regarding privacy policies. Participants recommended simplifying the service agreements that pop up when opening an app to show only the conditions needed for main functions, with secondary functions being negotiated with users later. They also suggested adding design elements to long agreements, such as icons and cartoons, to aid understanding.

4. Findings and Reflection

4.1. User Experience and Reflection on the Flow of Privacy Data by Users

Through interactive visualization of these principles, participants reported that their emotions shifted from initial unease and confusion to broader associations with previous experiences, such as shopping or trending topic recommendations. After comparing these with their own real-life experiences, participants moved toward self-reflection. This suggests that the visualization and interaction design can prompt users to reflect on negative experiences on social platforms, encouraging greater awareness of related topics and knowledge. It also raises the question of whether current social platforms, in collecting and processing user data, have crossed privacy boundaries from a user experience perspective. This, in turn, increases users' awareness and understanding of privacy boundaries, encouraging them to advocate for improved 'privacy data protection' and user-centered optimization strategies from platform operators. For example, some platforms should avoid strategies that hide comments based on gender, as this could lead to a homogenization of viewpoints, stifling diverse perspectives and potentially simplifying complex discourse.

4.2. Metaphorical Visualization for Enhancing Algorithm Understanding and User Engagement

This study introduces a more universal and accessible design approach to engage a broader audience, beyond just professionals, in understanding and discussing user data issues. The process began with metadata analysis, followed by a deconstruction of algorithmic mechanisms at the principle level. To make this complex process both comprehensible and artistic for the public, we used wormholes and planets as metaphors to create a poetic visual language system. We also explored ways to broaden the impact of the design, focusing on enhancing the user experience through interactive elements like characters, scenes, and sound effects to boost immersion. Finally, user testing was conducted to optimize and validate the design exploration.

The results suggest that representing the transmission of user privacy data as a 'cosmic wormhole' symbol can enhance users' understanding of the underlying algorithms and mechanisms in a more engaging, immersive way. This approach encourages users to reconsider their relationship with the digital avatars formed by their personal data.

4.3. The Concealment and Boundary Crossing Issues of Privacy Data Use

This design study highlighted several issues, including the hidden nature of user privacy policies, the collection and circulation of user data by third parties, and the complexity of algorithmic involvement. Drawing on the classification research by Mathur et al. [32], it examined the increasing use of 'dark pattern' interfaces in modern interaction design—interfaces that are misleading and manipulative. The study found that platforms, through their service agreements, not only require users to authorize access to private information such as phone numbers, locations, and photo albums but also collect data like

search history, network status, and clipboard content. This practice exploits users' belief that these types of data are unimportant or unrelated to traditional notions of 'privacy'. As a result, many users are unaware of the extent to which their information can be utilized and transformed by big data algorithms and analysis. Has this already crossed the privacy boundaries users once thought were secure? This issue not only highlights how platforms redefine privacy through algorithmic processes but also triggers negative emotions in users when interacting with these platforms.

4.4. Strengths and Limitations

This study offers several advantages. First, our research focuses on the perspectives and concerns of the general public, rather than experts, regarding data collection and processing. Although ongoing discussions in fields like computer science and programming address data and privacy security, which are of significant theoretical and practical value, our work shifts the focus to enhancing user awareness. By exploring the presentation of data privacy-related algorithms, we aim to empower users with the knowledge necessary for informed privacy decisions.

Second, the study's novelty lies in integrating poetic metaphors with the complex processes of data flow. We designed a series of simple, visually appealing visualizations that reflect the characteristics and goals of data flow. These visualizations create a symbolic scenario that helps users better understand data security and privacy through interactive, metaphor-driven design.

Third, our empirical research highlights users' experiences and potential shifts in their decision-making. The results show that introducing new visual symbols to represent real-world data privacy challenges can effectively help participants visualize and interact with their understanding of algorithms and mechanisms in a more engaging and immersive way.

Our study has several limitations. First, we conducted open-ended interviews instead of in-depth ones, which may have led us to overlook complex relationships from the participants' perspectives. Additionally, we did not track long-term changes in privacy-related decision-making behaviors. Second, the current interactive system is still in its early prototype phase, requiring multiple iterations to optimize the system's interaction design. Future research will aim to include more participants to gather comprehensive feedback and refine the visualization design. Furthermore, we plan to expand the study to a wider range of regions and apply the analytical framework to major global social platforms. This will allow us to explore the adaptability and effectiveness of privacy algorithm visualizations in different cultural contexts. Through these efforts, we hope to better understand and address the limitations identified in the current study, providing valuable insights for future research and design in related fields.

5. Conclusions

This study, from the perspective of a broad user base, employs the method of "infopoetry" to capture the characteristics of the "eavesdropping" process and the roles of its components, guiding the development of visualization and interactive design. Compared to the fields of technical implementation and optimization, our research emphasizes the intersection of data visualization, user experience design, and data privacy education. We adopt a unique approach to investigate how innovative visualization and interactive methods can improve non-expert users' comprehension of complex data privacy concepts, offering a fresh perspective for privacy education. Participant feedback suggests that this approach effectively enhances general users' awareness of data security and privacy. The results also uncover hidden issues in the current online environment and social media

platforms. The overlapping interests of multiple stakeholders drive the "eavesdropping" phenomenon, while the involvement of algorithms further increases the "value" of user information.

At the same time, we explore the effects of presenting the originally opaque data flow process to the public through artistic processing. We believe that this study can alleviate users' psychological burdens when using the internet, reduce negative speculations about platform technology caused by excessive anxiety, and enhance users' self-protection. Additionally, it can foster broader dialogue between developers and users.

In future work, we aim to conduct more in-depth and extensive data and text analyses on this topic, inviting a larger user group to discuss the boundaries of data privacy on social media platforms. Additionally, we will explore how users perceive the security of their personal privacy data across multiple social platforms, also providing more comprehensive visualizations of algorithm principles. The study also explores how designers can use aesthetic and highly accessible data visualization techniques to demystify complex algorithmic principles and data privacy policies for the general public.

Furthermore, we aim to convert the results of this work into strategies for optimizing user privacy data management from a design-driven technology perspective. We will seek ethically sound solutions to enhance transparency in interactive information, guiding designers to propose optimized policies and design methods. Ultimately, we hope to promote mutual trust between users and platforms, improving the online environment.

Author Contributions: Conceptualization, X.W., J.Z. and H.W.; methodology, X.W. and J.Z.; data analysis, X.W., J.Z. and H.W.; interaction design, X.W., J.Z. and H.W.; user experiments design and execution, X.W., J.Z. and H.W.; writing—original draft preparation, X.W., J.Z. and Y.Y.; writing—review and editing, X.W., J.Z. and Y.Y.; visualization, X.W. and J.Z.; supervision, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China: 62407003.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors are grateful to the colleagues and students who supported this study. Special thanks go to all participants in the user experiments for their valuable contributions. Their involvement and insights helped advance the study, providing foundational data that enriched our understanding of user perspectives on data privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Brudy, F.; Ledo, D.; Greenberg, S.; Butz, A. Is anyone looking? mitigating shoulder surfing on public displays through awareness and protection. In Proceedings of the International Symposium On Pervasive Displays, Copenhagen, Denmark, 3–4 June 2014; pp. 1–6.
2. Royakkers, L.; Timmer, J.; Kool, L.; Van Est, R. Societal and ethical issues of digitization. *Ethics Inf. Technol.* **2018**, *20*, 127–142. [CrossRef]
3. Yeung, K. 'Hypernudge': Big Data as a mode of regulation by design. *Inf. Commun. Soc.* **2017**, *20*, 118–136. [CrossRef]
4. Luger, E.; Moran, S.; Rodden, T. Consent for all: Revealing the hidden complexity of terms and conditions. In Proceedings of the SIGCHI Conference On Human Factors In Computing Systems, Paris, France, 27 April–2 May 2013; pp. 2687–2696. [CrossRef]
5. Barth, S.; De Jong, M. The privacy paradox–Investigating discrepancies between expressed privacy concerns and actual online behavior–A systematic literature review. *Telemat. Inform.* **2017**, *34*, 1038–1058. [CrossRef]

6. Trottier, D. Social Media as Surveillance: Rethinking Visibility in a Converging World. In *Social Media as Surveillance: Rethinking Visibility in a Converging World*; Amsterdam University Press: Amsterdam, The Netherlands, 2012; pp. 1–213.
7. Gibbs, S. Samsung's voice-recording smart TVs breach privacy law, campaigners claim. *Guardian* **2015**, *27*. Available online: https://www.theguardian.com/technology/2015/feb/27/samsung-voice-recording-smart-tv-breach-privacy-law-campaigners-claim (accessed on 13 June 2024).
8. Schez-Sobrino, S.; Garcıa, M.; Lacave, C.; Molina, A.; Glez-Morcillo, C.; Vallejo, D.; Redondo, M. A modern approach to supporting program visualization: From a 2D notation to 3D representations using augmented reality. *Multimed. Tools Appl.* **2021**, *80*, 543–574. [CrossRef]
9. Yang, J.; Niculaescu, O.; Ghinita, G. A Game-Oriented Educational Tool for Location Privacy Topics. In Proceedings of the 25th ACM SIGSPATIAL International Conference On Advances In Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017. [CrossRef]
10. Denning, T.; Lerner, A.; Shostack, A.; Kohno, T. Control-Alt-Hack: The design and evaluation of a card game for computer security awareness and education. In Proceedings of the 2013 ACM SIGSAC Conference On Computer & Communications Security, Berlin, Germany, 4–8 November 2013; pp. 915–928. [CrossRef]
11. Hosseini, M.; Breaux, T.; Slavin, R.; Niu, J.; Wang, X. Analyzing privacy policies through syntax-driven semantic analysis of information types. *Inf. Softw. Technol.* **2021**, *138*, 106608. [CrossRef]
12. Nokhbeh Zaeem, R.; Barber, K. A study of web privacy policies across industries. *J. Inf. Priv. Secur.* **2017**, *13*, 169–185. [CrossRef]
13. Lovato, J.; Allard, A.; Harp, R.; Onaolapo, J.; Hébert-Dufresne, L. Limits of individual consent and models of distributed consent in online social networks. In Proceedings of the 2022 ACM Conference On Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 2251–2262.
14. Dimova, Y.; Kode, M.; Kalantari, S.; Wuyts, K.; Joosen, W.; Mühlberg, J. From Privacy Policies to Privacy Threats: A Case Study in Policy-Based Threat Modeling. In Proceedings of the 22nd Workshop On Privacy In The Electronic Society, Copenhagen, Denmark, 26 November 2023; pp. 17–29.
15. Weinshel, B.; Wei, M.; Mondal, M.; Choi, E.; Shan, S.; Dolin, C.; Mazurek, M.; Ur, B. Oh, the places you've been! User reactions to longitudinal transparency about third-party web tracking and inferencing. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 149–166.
16. Solove, D.; Citron, D. Risk and anxiety: A theory of data-breach harms. *Tex. Law Rev.* **2017**, *96*, 737. [CrossRef]
17. Frik, A.; Kim, J.; Sanchez, J.; Ma, J. Users' expectations about and use of smartphone privacy and security settings. In Proceedings of the 2022 CHI Conference On Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022; pp. 1–24.
18. Emami-Naeini, P.; Dixon, H.; Agarwal, Y.; Cranor, L. Exploring how privacy and security factor into IoT device purchase behavior. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–12.
19. Altman, I. Privacy regulation: Culturally universal or culturally specific? *J. Soc. Issues* **1977**, *33*, 66–84. [CrossRef]
20. Abualigah, L.; Diabat, A.; Mirjalili, S.; Abd Elaziz, M.; Gandomi, A.H. The arithmetic optimization algorithm. *Comput. Methods Appl. Mech. Eng.* **2021**, *376*, 113609. [CrossRef]
21. Corbellini, A.; Mateos, C.; Godoy, D.; Zunino, A.; Schiaffino, S. An architecture and platform for developing distributed recommendation algorithms on large-scale social networks. *J. Inf. Sci.* **2015**, *41*, 686–704. [CrossRef]
22. Jacob, I.; Darney, P. Design of deep learning algorithm for IoT application by image based recognition. *J. ISMAC* **2021**, *3*, 276–290. [CrossRef]
23. Saurwein, F.; Spencer-Smith, C. Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media Commun.* **2021**, *9*, 222–233. [CrossRef]
24. Milan, S. When algorithms shape collective action: Social media and the dynamics of cloud protesting. *Soc. Media+ Soc.* **2015**, *1*, 2056305115622481. [CrossRef]
25. Naps, T.; Eagan, J.; Norton, L. JHAVÉ—An environment to actively engage students in Web-based algorithm visualizations. In Proceedings of the Thirty-first SIGCSE Technical Symposium On Computer Science Education, Austin, TX, USA, 7–12 March 2000; pp. 109–113.
26. Stasko, J. Using student-built algorithm animations as learning aids. In Proceedings of the Twenty-eighth SIGCSE Technical Symposium on Computer Science Education, San Jose, CA, USA, 27 February–1 March 1997; pp. 25–29.
27. Xu, D.; Tian, Y. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
28. Sinaga, K.; Yang, M. Unsupervised K-means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
29. Cao, C.; Ni, Q.; Zhai, Y. An improved collaborative filtering recommendation algorithm based on community detection in social networks. In Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation, Madrid, Spain, 11–15 July 2015; pp. 1–8.

30. Kumar, D.; Kashyap, R.; Gayathri, N. Market Basket Analysis: Identify the changing trends of market data using association rule mining. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 51–59.
31. Zingale, S. Infopoesia: L'enunciazione poetica dei dati. Una sperimentazione fra arte e design della comunicazione. *E/C* **2020**, *30*, 366–373.
32. Mathur, A.; Kshirsagar, M.; Mayer, J. What makes a dark pattern... dark? design attributes, normative considerations, and measurement methods. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–18.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Article

CMDAF: Cross-Modality Dual-Attention Fusion Network for Multimodal Sentiment Analysis

Wang Guo [1,2], Kun Su [1,2], Bin Jiang [1,2], Kang Xie [3,*] and Jie Liu [2,*]

[1] Shenzhen Research Institute, Shandong University, Shenzhen 518057, China; 17863138533@126.com (W.G.); nukus@mail.sdu.edu.cn (K.S.); jiangbin@mail.sdu.edu.cn (B.J.)
[2] School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai 264209, China
[3] Key Lab of Information Network Security, Ministry of Public Security, Shanghai 201204, China
* Correspondence: xiekang@stars.org.cn (K.X.); liujie_sdwh@163.com (J.L.)

Abstract: Multimodal sentiment analysis (MSA) seeks to predict subjective human sentiments by utilizing information from multiple modalities. It has been applied in diverse scenarios. Recent studies suggest that MSA benefits from integrating diverse modalities, emphasizing the fusion of multimodal information at different levels and the joint learning of modality-consistent and modality-inconsistent representations. In this work, we put forward a cross-modality dual-attention fusion network (CMDAF) that incorporates a cross-modality dual-attention fusion module. This module facilitates the integration of representations across modalities, enabling effective fusion from shallow to deep levels. A Transformer-based projection module transforms the fused representations into modality-consistent and modality-inconsistent forms, further enhanced through multi-scale contrast learning. Experiments on three MSA datasets (MOSI, MOSEI, and CH-SIMS) demonstrate the superiority of CMDAF, achieving state-of-the-art performance on most metrics.

Keywords: multimodal sentiment analysis; multi loss; dual attention; multimodal fusion; contrastive learning

1. Introduction

Multimodal fusion learning aims to create fine-grained fusion representations from different modalities, which in turn enhances the overall performance of the model. In recent years, online social media platforms like Facebook, YouTube, TikTok, and Twitter have experienced rapid growth. This growth has led to an explosion of multimodal data (e.g., textual, acoustic, and visual) that captures users' emotional expressions. This influx of data not only motivates advancements in multimodal fusion learning but also presents significant challenges. Currently, multimodal fusion learning has garnered extensive research interest across various fields, including multimodal sentiment analysis [1,2], cross-modal retrieval [3,4], visual questioning [5], and emotion recognition in conversation [6,7].

Multimodal sentiment analysis (MSA) aims to analyze information in multiple modalities, including textual elements, acoustic attributes, and visual indicators. This approach enhances the prediction of both the polarity and intensity of human sentiments. Unlike traditional unimodal sentiment analysis, MSA effectively alleviates semantic and sentimental ambiguities by integrating and leveraging the rich, complementary information from multiple sources. Recently, MSA has been increasingly applied to various downstream fields, especially in areas such as healthcare [8], e-business [9], and recommended systems [10].

Multimodal sentiment analysis techniques rely on the comprehensive extraction of features from different data modalities. Much prior research has employed custom algorithms to extract fixed features, including acoustic attributes (such as pitch, bandwidth, and speech rate) and visual elements (like facial structure, head angle, and pose). Although these fixed features incorporate prior knowledge, they often lack sensitivity across different datasets. Recent studies, including MMML and UniMSE, highlight the significant potential

of pre-trained models during the feature extraction phase, providing valuable insights for our work.

Multimodal sentiment analysis must account for the inherent heterogeneity in multimodal data, necessitating the accurate capture of inter-modality correlations. These correlations can be categorized into modality-consistent and modality-inconsistent information. Many current studies tend to prioritize modality-consistent information, which enhances the model's recognition capabilities but risks fostering an over-reliance on modal details that contribute significantly. Previous research [11] indicated that textual modality is generally more informative than acoustic and visual modalities. Consequently, when discrepancies arise in affective polarity between textual and other modalities, the model's performance can be substantially impacted. For instance, if a girl states, "I really like the sun", models will likely interpret her mood as positive. However, if her tone conveys sadness or she is depicted getting wet in the rain, the models might erroneously classify her sentiment as negative. Thus, leveraging the variability among modalities is crucial for improving model performance.

Inspired by the observations above, we propose a novel Cross-Modality Dual-Attention Fusion Network (CMDAF). It conducts multi-level interactions to learn and understand inter-modality correlation, which improves the performance of MSA.

In summary, our contribution can be summarised as follows:

- We propose the Cross-Modality Dual-Attention Fusion Network. The network uses pre-trained models to extract unimodal raw data from the three modalities and convert them into a unified dimensional embedding representation. It facilitates inter-modal information interaction.
- We achieve multi-level inter-modal information interaction by feeding the representations from three modalities into the cross-modality dual-attention fusion module.
- We propose the Transformer-based Features Projector. It projects unimodal features into modality-consistent and modality-inconsistent representations. Building on the above projections, we introduce unimodal classification loss and multi-scale contrast learning frameworks to further enhance the learning of the two representations for each modality.
- We validated the performance of the network on the MOSI, MOSEI, and CH-SIMS datasets. Experimental results demonstrate that CMDAF outperforms other baselines across all metrics on the MOSI dataset. Additionally, it also outperforms the baselines on the MOSEI and CH-SIMS datasets for most metrics.

2. Related Work

2.1. Multimodal Sentiment Analysis

Many previous researches on multimodal sentiment analysis have experimented with various multimodal fusion methods. Zadeh et al. [12] proposed a multimodal dictionary representation that learns four types of interactions between facial expressions and speech. TFN [13] emphasizes learning both inter-modality and intra-modality representations through an end-to-end approach. CAT-LSTM [14] developed a recurrent model to capture contextual information among different discourses. Rahman et al. [15] proposed the multimodal adaptive gating mechanism (MAG) built on BERT [16], which captures information in vector form.

Yu et al. [17] developed a Chinese multimodal sentiment analysis dataset (CH-SIMS) featuring fine-grained modal annotations. They designed a joint training module using self-supervised learning to identify inter-modality consistency and inconsistency, emphasizing samples with significant discrepancies. In addition, CH-SIMS directs the learning subtasks to focus on samples with substantial discrepancies between modal supervision. MISA [18] aims to bridge the representation gap between different modalities and facilitate effective fusion by learning modality-invariant and modality-specific representations. The model contains two subspaces to be learned, one for capturing commonalities and the other for capturing modality uniqueness. Arjmand et al. [19] proposed the Transformer-based

Speech Prefix Language Model (TEASEL), which utilizes a lightweight spatial encoder for speech. MM-DFN [20] developed a novel graph-based dynamic fusion module to minimize redundancy and enhance the complementarity of schemas. This module achieves these goals by capturing the dynamics of contextual information across various potential spaces.

Paraskevopoulos et al. [21] proposed a top–down cross-modal structure. His framework efficiently extracts multi-level representations among modalities and facilitates effective interactions. Mult [11] introduced a directed pairwise cross-modal attention mechanism designed to operate in an end-to-end fashion. This approach effectively addresses the challenges of misalignment in multimodal data as well as the complexities associated with long-term dependencies. ViT [22] divides images into blocks (patches) and uses a linear embedding sequence of these image blocks as input to the Transformer. This work demonstrates the independence and efficiency of attention applications in computer vision. BLIP2 [23] is a lightweight pre-training approach using frozen visual pre-training models and LLMs with generalization and efficiency. MMML [24] introduces a cross-modality module designed to capture the interdependencies between different modalities.

2.2. Contrastive Learning

UniMSE [1] performs inter-modality contrast learning to maximize inter-class variance across samples. To address the problem of large-scale unsupervised visual data, Vatt [25] trained models through Noise Contrast Estimation (NCE) in different granularity spaces of semantic hierarchies. Inspired by recent contrast learning algorithms, SimCLR [26] integrates multiple data augmentation techniques and introduces nonlinear transformations to enhance the effectiveness of representation learning. MVGRL [27] focuses on learning multi-level representations from points to graphs. ConFEDE [28] integrates inter-sample contrast learning and intra-sample modal decomposition into a simple unified loss function. ConFEDE constructs contrast relations between all decomposed features.

3. Methodology

3.1. Overall Architecture

The overall architecture of CMDAF is illustrated in Figure 1. First, a feature extractor module independently extracts unimodal features from text, visual, and acoustic data. After obtaining the unimodal featureembeddings, these embeddings are processed through a cross-modality attention module, which facilitates dynamic fusion at multiple levels—both within each modality and across modalities. The fused features are then fed into a transformer-based feature projection module, where each fused representation is transformed into both modality-consistent and modality-inconsistent forms. Additionally, we adopt a multi-loss training strategy, which includes three key components, namely multimodal classification loss, unimodal classification loss, and contrastive learning loss.

3.2. Task Definition

Given a conversation $C = \{u_1, u_2, \cdots, u_{|c|}\}$ containing $|C|$ utterances $u_1, \cdots, u_{|c|}$, the goal of MSA is to predict the emotional polarity and intensity for each utterance u_i in C. The input of each utterance can be represented as $I = \{I_t, I_a, I_v\}$, where I^m denotes the unimodal raw feature from the utterance i, and $m \in \{t, v, a\}$, where $\{t, v, a\}$ denotes the textual, acoustic, and visual modalities. To better learn human emotions, we introduce contextual information into the conversation. We integrate each utterance with the previous k utterances as textual input. We also integrate the current and previous audio and video clips. Thus, each utterance corresponds to k+1 rounds of text, two rounds of speech, and video.

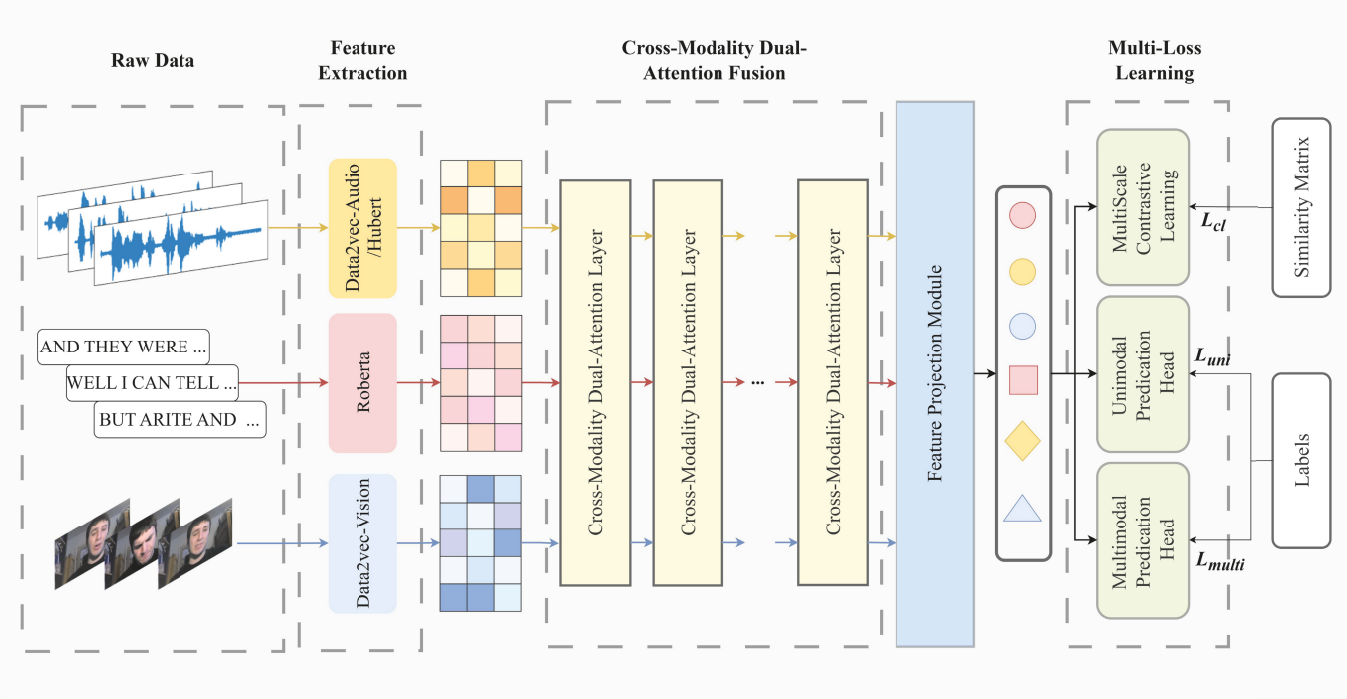


Figure 1. The overall architecture of the CMDAF. For the raw data, the network obtains unimodal features by pre-training feature extractors. The architecture comprises the following three essential components: the cross-modality dual-attention fusion module, the feature projection module, and the multi-loss learning module.

3.3. Feature Extraction

In contrast to traditional fixed feature extraction algorithms, many recent studies have demonstrated that pre-trained models perform effectively in unimodal feature extraction [1,7,24]. Inspired by these studies, we adopt and fine-tune the pre-trained feature extractor to obtain rich unimodal features.

We utilize RoBERTa [29] as a text encoder to extract and encode features from the textual modality. RoBERTa is a Bert-based pre-trained language model. It improves the performance of BERT through large-scale data training and model parameter optimization. Roberta outperforms the original BERT on several natural language processing tasks. Data2vec [30] proposes a powerful self-supervised algorithm for multiple modal inputs. It also represents a new, comprehensive self-supervised learning paradigm demonstrating strong multimodal learning representation capabilities. Due to the excellent performance of Data2vec, we use Data2vec-audio and Data2vec-vision as feature encoders for audio and vision, respectively. In addition, we employ HuBERT to extract features for Chinese speech in CH-SIMS.

We directly extracted features from the raw text data using RoBERTa. For visual features, we utilized Data2Vec-Vision to process the T-frame images of each clip. Additionally, extracting audio files from the video is a prerequisite for acoustic feature extraction. Notably, all three modal features (text, visual, and acoustic) have a dimensionality of 768, allowing them to be used directly for subsequent processing.

3.4. Cross-Modality Dual-Attention Module

The main goal of MSA focuses on rich modal fusion representations from multimodal features. As shown in Figure 2, we propose the cross-modality multi-head dual-attention module, which comprises both multi-head dual-attention and self-attention components. It addresses the heterogeneity problem in multimodal data and captures the intra-modality and inter-modal correlations. The module contains N layers, enabling the network to perform multi-level feature fusion. Specifically, primary multimodal feature fusion is

conducted in the shallow layers, whereas advanced multimodal fusion takes place in the deeper layers. Since textual modality is more instructive for multimodal sentiment analysis [1,11], we do not apply the dual-attention mechanism equally to the three modalities but mainly use textual modal features to guide the other modality's features for cross-modality interaction.

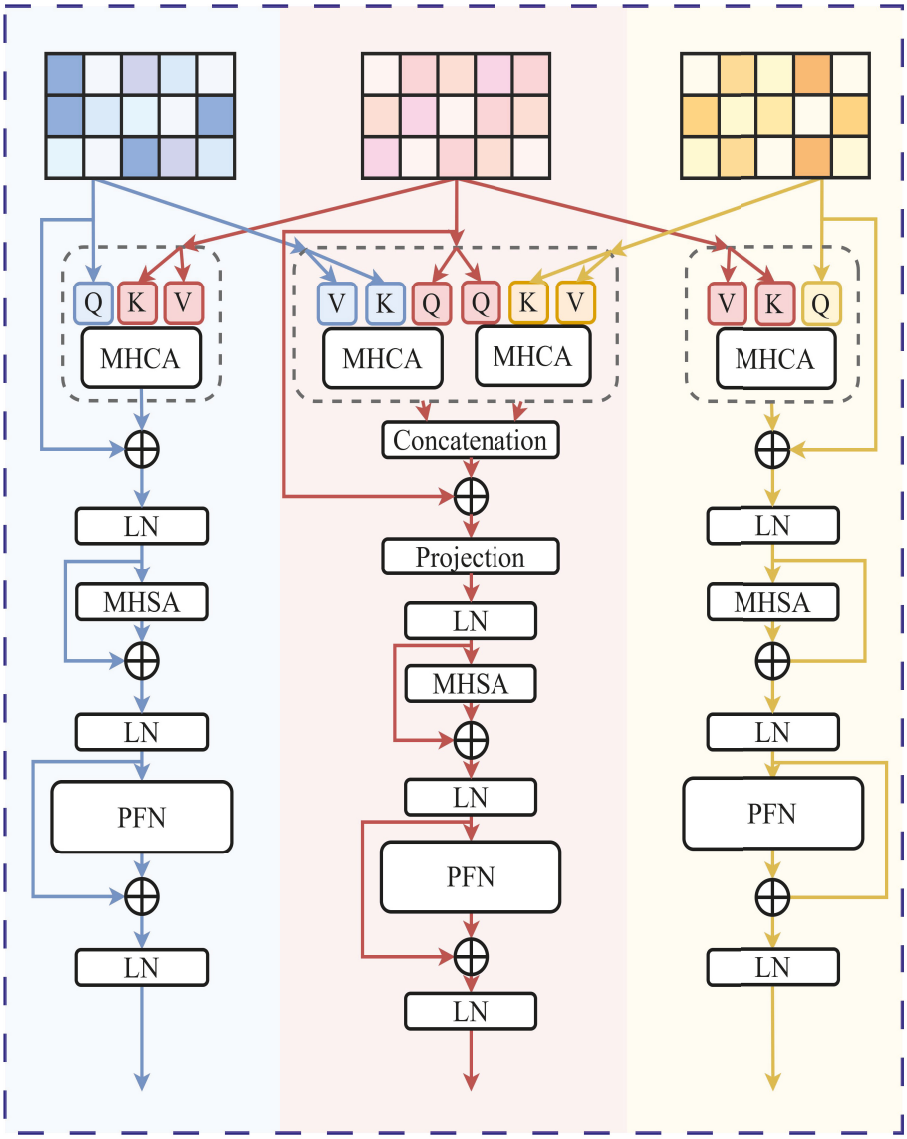


Figure 2. The network structure of the cross-modality dual-attention module.

In the first stage, the multi-head cross-modal dual-attention encoders focus on capturing and integrating inter-modal interactions across different modalities. This mechanism enables the model to identify complex dependencies among multimodal data. It provides a more comprehensive understanding of the data environment. This process can be computed as follows:

$$X_i^{T,V} = MHCA(X_{i-1}^T, X_{i-1}^V, X_{i-1}^V), \tag{1}$$

$$X_i^{T,A} = MHCA(X_{i-1}^T, X_{i-1}^A, X_{i-1}^A), \tag{2}$$

$$X_i^T = CA(X_i^{T,V}, X_i^{T,A}) + X_{i-1}^T, \tag{3}$$

$$X_i^T = LN(Proj(X_i^T), \tag{4}$$

where the $MHCA$ $(\cdot)$ denotes the multi-head dual-attention network, $CA(\cdot)$ denotes the concatenation operation, $Proj(\cdot)$ denotes the projection layer, and $LN(\cdot)$ denotes the layer normalization operation.

$$X_i^{V,T} = MHCA(X_{i-1}^V, X_{i-1}^T, X_{i-1}^T), \tag{5}$$

$$X_i^{A,T} = MHCA(X_{i-1}^A, X_{i-1}^T, X_{i-1}^T), \tag{6}$$

$$X_i^V = LN(Proj(X_i^{V,T}), \tag{7}$$

$$X_i^A = LN(Proj(X_i^{A,T}), \tag{8}$$

where the $MHCA(\cdot)$ denotes the multi-head dual-attention network, $CA(\cdot)$ denotes the concatention operation, $Proj(\cdot)$ denotes the projection layer, $LN(\cdot)$ denotes the layer normalization operation.

In the second stage, the multi-head self-attention encoders focus on digging deeper and extracting the rich interaction information within each modality. This process helps the model better understand and represent the intrinsic structure of single-modality data. The stage can be formalized as follows:

$$X_{is}^m = MHSA(X_i^m, X_i^m, X_i^m), \tag{9}$$

$$X_{is}^m = LN(X_{is}^m + X_i^m), \tag{10}$$

$$X_{is}^m = LN(PFN(X_{is}^m) + X_{is}^m), \tag{11}$$

where $MHSA(\cdot)$ denotes the multi-head self-attention network, $PFN(\cdot)$ denotes the pointwise feed-forward layer, $m \in \{t, v, a\}$.

The first phase of intermodal information interaction involves four main MHCA (multi-head cross-attention) networks. These networks are designed to handle the varying tendencies between different modalities when there are significant differences in their characteristics.

3.5. Features Projection Module

Inspired by the structure of ConFEDE [28], we proposed a transformer-based features projection module, which is illustrated in Figure 3. The module consists of four projectors: the modality-consistent projector, the text-inconsistent projector, the audio-inconsistent projector, and the visual-inconsistent projector. The modality-consistent projector projects each of the three modalities into a modality-consistent feature X_{cs}^m, while each of the other three projectors is a modality-inconsistent projectors, each transforming the corresponding modality feature into a modality-inconsistent feature X_{ics}^m. Afterward, we establish the contrastive relationships among all features by implementing multi-scale contrastive learning. This process can be expressed as follows:

$$X_{cs}^m = ConsTrans(X^m), \tag{12}$$

$$X_{ics}^m = InconsTrans(X^m), \tag{13}$$

where $m \in \{t, a, v\}$, l$ConsTrans(\cdot)$ and $InconsTrans(\cdot)$ denote the modal-consistency feature Transformer and modal-inconsistency feature Transformer, respectively.

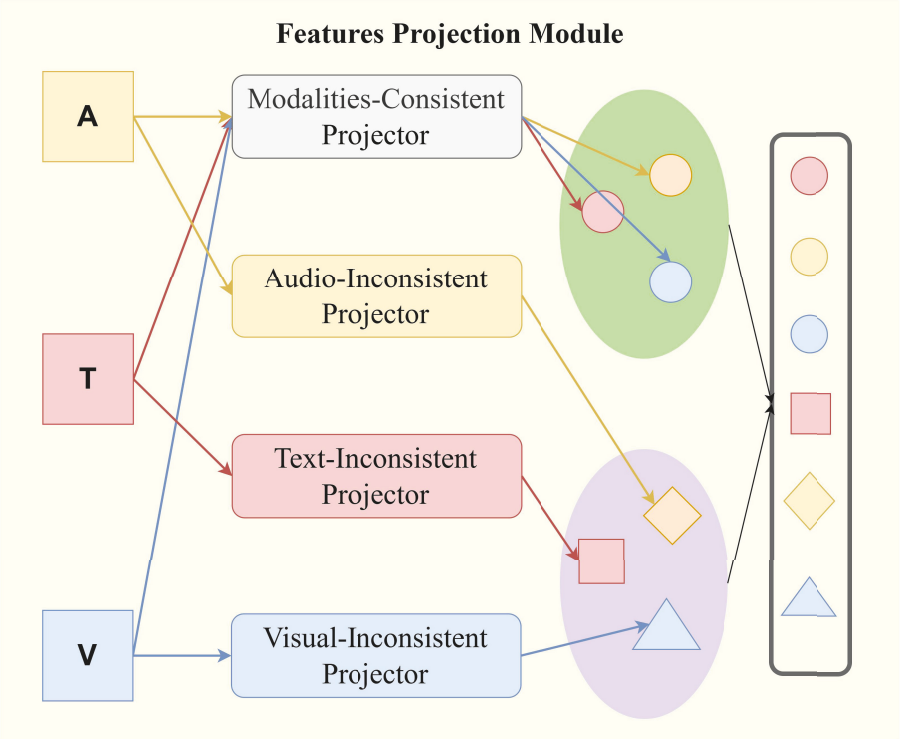


Figure 3. The network structure of the features projection module.

3.6. Multi-Loss Learning

Many recent studies have used the multi-loss training strategy for multimodal fusion learning, which incorporates different subtask losses in the multimodal fusion task to help improve model performance. Following [1,24,28], we use multi-loss learning further to improve the performance of understanding multimodal fusion representations. The multi-loss objective function is described as follows:

$$L = L_{multi} + \alpha L_{uni} + \beta L_{cl}, \tag{14}$$

where L_{multi} is the multimodal classification loss, L_{uni} represenst the unimodal classification loss and L_{cl} represents the multi-scale contrastive loss. Additionally, α and β are hyper-parameters, constrained between 0 and 1, that regulate the weighting of the components' contributions.

3.6.1. Multimodal Classification Loss

In contrast to the sharp turning point of ReLU, GELU implements an adaptive gating mechanism through its internal Gaussian CDF component. It provides a smooth activation profile that helps the deep learning model to be more stable during the training process. We concatenate the consistent and inconsistent features from the three modalities as inputs to the MLP, which generates the multimodal sentiment prediction results. These results are then optimized using smooth L1 loss.

3.6.2. Unimodal Classification Loss

For each of the six single features [T^is, V^is, A^is, T^id, V^id, and A^id] obtained after the consistent/inconsistent Transformer module, we feed them into an MLP-based classifier to predict the unimodal sentiments and also optimize them using smooth L1Loss.

Compared to other datasets, CH-SIMS contains both multimodal and unimodal annotations, which is a Chinese multimodal sentiment analysis dataset with finer modal annotations. For datasets without unimodal annotations, we apply multimodal annotations for unimodal sentiment prediction.

3.6.3. Multi-Scale Contrastive Loss

Following [28], we present a multi-scale contrastive learning framework called NT-Xent. It is designed to perform intra-modality and inter-modality contrastive learning together. On the one hand, inter-sample contrast learning can magnify the distance between dissimilar samples and reduce the distance between similar samples. On the other hand, intra-sample contrast learning can highlight modality-consistent and modality-inconsistent representations. For each anchor sample, the multi-scale contrastive loss function can be represented as follows:

$$\ell_{\mathrm{cl}}^{i} = \sum_{p \in \mathcal{N}^i} -\log \frac{\exp(\mathrm{sim}(p)/\tau)}{\sum_{k \in \mathcal{N}^i \cup \mathcal{P}^i} \exp(\mathrm{sim}(k)/\tau)}, \tag{15}$$

where $\mathcal{P}^i$ denotes all the positive pairs, and $\mathcal{N}^i$ denotes all the negative pairs, and sim is a cosine similarity calculation for positive and negative sample pairs, respectively.

4. Experiments

4.1. Datasets

As shown in Table 1, we employ three widely used datasets for MSA to evaluate the performance of the cross-modality dual-attention fusion network. These datasets are the Multimodal Opinion-level Sentiment Intensity dataset (MOSI) (Zadeh et al., 2016) [31], Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) (Zadeh et al., 2018) [32], and Chinese Multimodal Sentiment Analysis Dataset (CH-SIMS) (Yu et al., 2020) [17].

Table 1. The details of MOSI, MOSEI, CH-SIMS.

Dataset	Train	Valid	Test	All	Language
MOSI [31]	1284	229	686	2199	English
MOSEI [32]	16,326	1871	4659	22,856	English
CH-SIMS [17]	1368	456	4457	6281	Chinese

MOSI contains 2199 discourse video clips collected from YouTube monologue film reviews. Each clip has been manually assigned sentiment scores that range from -3 to +3, indicating both the sentiment polarity and its relative intensity.

MOSEI is an extended version of MOSI annotated with sentiment and emotion. It includes 22,856 film review clips sourced from YouTube. Importantly, there is no overlap between MOSI and MOSEI, as their data collection and labeling processes are conducted independently.

CH-SIMS is a Mandarin language dataset containing the same modalities, audio, text, and video, collected from 2281 annotated video clips. It includes data from TV programs and films.

4.2. Evaluation Metrics

Following previous studies (CH-SIMS [17], UniMSE [1], and ConFEDE [28]), we choose the mean absolute error (MAE), Pearson correlation (Corr), seven-class classification accuracy (ACC-7), binary classification accuracy (ACC-2), and F1 scores as evaluation metrics. Specifically, ACC-2 and the F1 score are calculated for negative/positive (NP) and non-negative/negative (NN) classifications using the MOSI and MOSEI datasets [28]. All indicators reflect better performance with higher values, except for the mean absolute error (MAE), where a lower value indicates better performance.

4.3. Baselines

To thoroughly validate the effectiveness of our CMDAF, we perform a fair comparison with several SOTA methods for MSA tasks. We do not select methods that only report results from a single round of experiments nor methods that do not have an officially

valid code for reproduction. For MSA, the methods we compare can be broadly classified into two groups: (1) early multimodal fusion methods, such as the Tensor Fusion Network (TFN) (Zadeh et al., 2017) [13] and Low-rank Multimodal Fusion (LMF) (Liu et al., 2018) [33], and (2) methods exploring modality interaction fusion, including Multimodal Transformer (MulT) (Tsai et al., 2019) [11], modality-invariant and specific representations (MISAs) (Hazarika et al., 2020) [18], large pre-trained transformers (MAG-BERT) (Rahman et al., 2020) [15], Multi-Task Learning (Self-MM) (Yu et al., 2021) [34], hybrid contrastive learning (HyCon) (Mai et al., 2021) [35], a unified learning framework (ConFEDE) (Yang et al., 2023) [28], and Multimodality Multi-Loss Fusion Network (MMML) (Wu et al., 2023) [24].

4.4. Experimental Settings

We use Roberta-Base, We used Roberta-base, Data2Vec-Audio-base, HuBERT-base, and Data2Vec-Vision-base as unimodal feature extractors. Feature extractors were fine-tuned on MSA datasets Roberta-Base to better extract MSA-specific unimodal features. The number of previous utterances k is 2. The length of the extracted unimodal feature is 768. In addition, we adopt Transformer-based encoders as the backbone of CMDAF. The batch size is 16, the main learning rate is 5×10^{-5}, and the AdamW optimizer is used with loss functions, including smooth L1Loss and NT-Xent. The whole process is run on a single RTX 3090 GPU, manufactured by NVIDIA, Santa Clara, CA, USA, and the experimental results are averaged over three rounds. α and β are set to 0.1 and 0.02, respectively.

4.5. Results

We evaluated the experimental performance of CMDAF on MOSI, MOSEI, and CH-SIMS. In comparison to the previous SOTA, CMDAF showed a significant improvement across all three publicly available datasets. The experimental scores of CMDAF were averaged over three rounds of experimental results.

The comparison results with baseline methods are presented in Tables 2–4. The performances of all other baselines, except for MAG-BERT, are sourced from published papers [1,2,24,28]. $\uparrow$ signifies that the higher values reflect better performance, while $\downarrow$ indicates that lower values for MAE are preferred. The best results are highlighted in boldface, while the underlined values indicate the second-best results. For ACC-2 and the F1 score, the left and right of the "/" represents the negative/non-negative and negative/positive, respectively.

The proposed method outperforms all baselines across all metrics on the MOSI dataset as shown in Table 2. Compared to the SOTA model MMML, our method improves by 1.03% in NP for ACC-2, by 1.4% in NN for ACC-2, by 0.98% in NP for the F1 score, by 1.54% in NN for the F1 score, by 2.4% in the Pearson correlation (Corr), and by 3.7% in MAE. In addition, the proposed model shows exceptional ability in multi-class classification, outperforming the optimal baseline by 1.04% on ACC-7.

As seen in Table 3, CMDAF outperforms all baselines across all metrics except ACC-7 for the MOSEI dataset. Furthermore, the ACC-2 and F1 scores of NN are significantly higher than all baselines at 0.99% and 1.07%, respectively. Table 4 demonstrates that our proposed method outperforms all baselines across most metrics on the CH-SIMS dataset. Specifically, our method improves ACC-2 by 0.89%, F1 scores by 0.95%, and ACC-5 by 0.73%.

As shown in the results, our method has a significant improvement in classification ability. The superior network performance implies that CMDAF can understand modality-consistent and modality-inconsistent information more comprehensively. Our approach provides richer and more accurate information for downstream tasks, demonstrating capabilities in multimodal sentiment analysis.

Table 2. Comparison SOTA: the MOSI dataset demonstrates exceptional performance. Note: [a]: results from [1]; [b]: results from [28]; [c]: results from [24].

Model	MAE↓	CORR↑	ACC-7↑	ACC-2↑	F1↑
LMF [a] [33]	0.917	0.695	33.20	-/82.50	-/82.40
TFN [a] [13]	0.901	0.698	34.90	-/80.80	-/80.70
MulT [a] [11]	0.861	0.711	-	81.50/84.10	80.60/83.90
MISA [a] [18]	0.804	0.764	-	80.79/82.10	80.77/82.03
MAG-BERT [a] [15]	0.712	0.796	-	84.20/86.10	84.10/86.00
HyCon [a] [35]	0.713	0.790	46.60	-/85.20	-/85.10
Self-MM [b] [34]	0.708	0.796	46.67	83.44/85.46	83.36/85.43
ConFEDE [b] [28]	0.742	0.784	42.27	84.17/85.52	84.13/85.52
MMML [c] [24]	0.643	0.838	48.25	85.91/88.16	85.85/88.15
CMDAF (ours)	0.606	0.862	49.29	87.31/89.19	87.39/89.13

Table 3. Comparison with SOTA: the MOSEI dataset demonstrates exceptional performance. Note: [a]: results from [1]; [b]: results from [28]; [c]: results from [24].

Model	MAE↓	CORR↑	ACC-7↑	ACC-2↑	F1↑
LMF [a] [33]	0.623	0.7	48.00	-/82.0	-/82.1
TFN [a] [13]	0.593	0.677	50.20	-/82.50	-/82.10
MulT [a] [11]	0.580	0.713	-	-/82.50	-/82.30
MISA [a] [20]	0.568	0.717	-	82.59/84.23	82.67/83.97
MAG-BERT [a] [15]	-	-	-	84.70/-	84.50/-
HyCon [b] [35]	0.601	0.776	52.80	-/85.40	-/85.60
Self-MM [b] [34]	0.531	0.765	53.87	83.76/85.15	82.17/85.83
ConFEDE [b] [28]	0.522	0.78	54.86	81.65/85.82	82.17/85.83
MMML [c] [24]	0.517	0.791	54.95	86.32/86.73	86.23/86.49
CMDAF (ours)	0.509	0.806	54.11	86.43/87.72	86.41/87.56

Table 4. Comparison with SOTA: the CH-SIMS dataset demonstrates exceptional performance. Note: [a]: results from [2]; [b]: results from [28]; [c]: results from [24].

Model	ACC-2↑	ACC-3↑	ACC-5↑	F1↑	MAE↓	Corr↑
ALMT [a] [2]	81.19	68.93	45.73	81.57	0.404	0.619
ConFEDE [b] [28]	82.23	70.15	46.30	82.08	0.392	0.637
MMML [c] [24]	82.93	69.37	49.38	82.90	0.332	0.732
CMDAF (ours)	83.82	70.42	50.11	83.85	0.325	0.723

4.6. Ablation Study

To verify the effect of different components on overall performance, we perform a range of ablation experiments on the MOSI dataset, with the results shown in Tables 5–7. Note that "w/o" indicates that CMDAF is evaluated without a specific setting.

4.6.1. Effects of Different Modalities

We investigate the effect of incorporating acoustic and visual modalities on the proposed model. As shown in Table 5, the removal of both acoustic and visual modalities leads to a noticeable decline in the overall performance of the model. Model performance degradation is more pronounced in the absence of inputs from acoustic and visual modalities. This phenomenon suggests that MSA can analyze and use multimodal information in a complementary way, removing semantic and affective ambiguity. The results above demonstrate that the textual modality provides richer and more accurate affective information than the other two modalities.

Table 5. Ablation study of CMDAF on MOSI: different modalities. A and V denote the acoustic and visual modalities, respectively. Note that the best result is highlighted in bold.

	MAE↓	**CORR↑**	**ACC-2↑**	**F1↑**
CMDAF	**0.606**	**0.862**	**87.31/89.19**	**87.39/89.13**
w/o A	0.619	0.844	85.71/87.91	85.68/87.94
w/o V	0.611	0.851	85.86/88.11	85.79/88.10
w/o A,V	0.620	0.841	85.28/87.53	85.30/87.59

An interesting phenomenon is that removing acoustic modalities weakens the performance of the network more than removing visual modalities. This phenomenon also appears in recent work [1]. A plausible explanation for this observation is that the network can more easily extract key emotional information from the acoustic than from the visual modality. In other words, images typically contain more complex noise.

4.6.2. Effects of Cross-Modality Dual-Attention Fusion Module

As seen in Table 6, we examine the effect of the cross-modality dual-attention (CMDA) module on model performance. We can observe that eliminating CMDA has a greater impact on modal performance. These results suggest that CMDA can address the heterogeneity of multimodal data and effectively capture intra-modality and inter-modality correlations. In addition, CMDA allows each of the three modalities to be fully learned and enhances the interaction and fusion between the modalities. The results clearly demonstrate that the CMDA plays a crucial role in emotion prediction.

Table 6. Ablation study of CMDAF on MOSI: CMDA vs. MHSA. CMDA represents the cross-modality dual-attention module. Note that the best result is highlighted in bold.

	MAE↓	**CORR↑**	**ACC-2↑**	**F1↑**
CMDAF	**0.606**	**0.862**	**87.31/89.19**	**87.39/89.13**
w/o CMDA	0.618	0.841	85.99/88.39	85.96/88.31

4.6.3. Effects on Multi-Loss

Moreover, we can observe from Table 7 the instructive significance of multi-scale contrastive learning and unimodal sentiment analysis for multimodal sentiment analysis. Through the application of multi-scale contrastive learning, the model effectively captures both similarities and dissimilarities across modalities. This capability allows for a nuanced analysis of the underlying sentiments in the data, particularly when there are significant emotional variations among different modal features. As shown in Table 7, 'w/o intra' exhibits higher performance than 'w/o inter' on MAE and lower accuracy on NN ACC-2. On the one hand, intra-sample contrast learning improves the binary classification accuracy of the model by comparing textual modality with other modal features. On the other hand, inter-sample contrast learning improves the multi-classification stability by distinguishing between consistency and inconsistency across samples. Unimodal sentiment analysis also emphasizes the consistency information embedded in modal similarity features and the discrepancy information retained by dissimilarity features.

Table 7. Ablation study of CMDAF on MOSI: Multi-Loss. CL and Inter represent multi-scale and inter-sample contrastive learning modules, respectively. Uni represents unimodal sentiment classification. Note that the best result is highlighted in bold.

	MAE↓	CORR↑	ACC-2↑	F1↑
CMDAF	**0.606**	**0.862**	**87.31/89.19**	**87.39/89.13**
w/o CL	0.614	0.848	85.71/87.76	85.70/87.81
w/o Intra	0.608	0.858	86.03/88.25	0.8599/88.29
w/o Inter	0.612	0.851	86.30/88.72	86.27/88.70
w/o Uni	0.611	0.852	86.25/88.57	86.28/88.56

4.7. Visualization of Features Extraction

Extracting richer modal features from the raw data significantly enhances the overall performance of our approach. To assess the efficacy of feature extraction using the pre-trained model, we visualize the representations obtained through various extraction methods. Using the MOSI dataset as a case study, we select all samples in the test set, excluding those with neutral annotations. Figure 4a illustrates the modal representation from the standard MOSI preprocessing. In contrast, Figure 4b presents the feature representations generated by our proposed method, highlighting the improvements achieved. It is evident that the clustering of features from the samples extracted using our method is more pronounced. This observation underscores the enhanced effectiveness of our approach in the extraction of sample features.

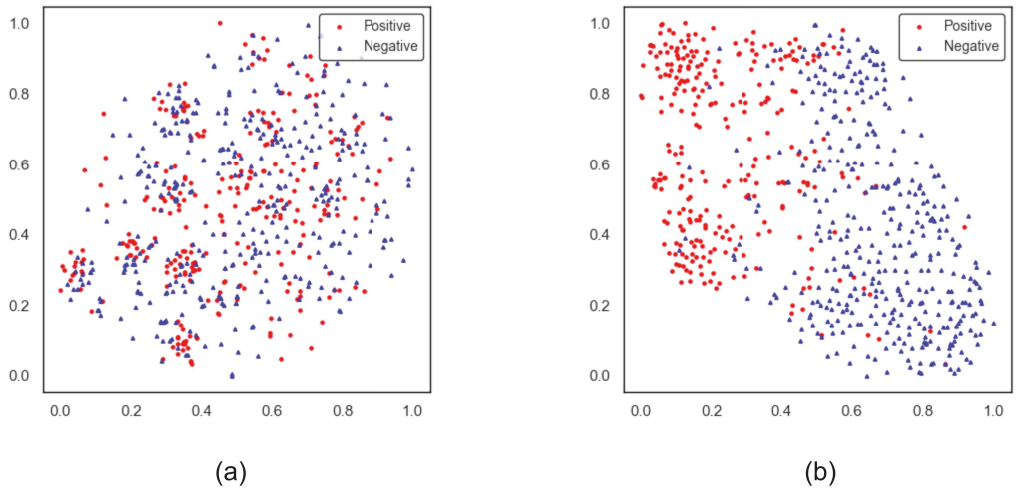


Figure 4. The T-SNE visualization compares the feature representations of three modalities between (**a**) samples extracted using the MOSI method and (**b**) samples obtained through the pre-trained extractors of our proposed method.

5. Conclusions

In summary, we propose a cross-modal dual-attention fusion network (CMDAF). First, we fine-tune pre-trained models to extract rich features from the raw data for CMDAF. Next, CMDAF simultaneously learns both modality-consistent and modality-inconsistent information, enabling multi-level interaction and fusion of inter-modal information. Additionally, we introduce multi-scale contrastive learning to enhance both intra-modal and inter-modal comparisons. The results demonstrate that CMDAF achieves state-of-the-art performance across most metrics on the three datasets. While our proposed CMDAF method demonstrates several advantages, there are some limitations that should be addressed. First, our method is currently designed to work in environments where all three modalities are present. The performance of the method in scenarios where one modality is

missing requires further investigation. Additionally, we plan to explore the application of CMDAF in downstream tasks, such as healthcare, in future work.

Author Contributions: Conceptualization, W.G., K.S. and J.L.; methodology, W.G.; software, W.G.; validation W.G. and K.S.; formal analysis W.G.; investigation, W.G.; resources, W.G.; data curation W.G.; writing-original draft W.G.; visualization, W.G.; writing-review and editing K.S., B.J. and J.L.; supervision, K.S., B.J. and J.L.; Funding acquisition K.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Key Lab of Information Network Security, Ministry of Public Security and the Shenzhen Fundamental Research Program under Grant JCYJ20230807094104009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK (accessed on 21 September 2023), https://thuiar.github.io/sims.github.io/chsims (accessed on 21 September 2023). These data were downloaded from the following public resources: https://drive.google.com/drive/folders/1A2S4pqCHryGmiqnNSPLv7rEg63WvjCSk (accessed on 21 September 2023).

Conflicts of Interest: The authors declare no conflicts of interest regarding the publication of this paper.

References

1. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv* **2022**, arXiv:2211.11256.
2. Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; Yu, T. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv* **2023**, arXiv:2310.05804.
3. Teng, J.; Lu, X.; Gong, Y.; Liu, X.; Nie, X.; Yin, Y. Regularized two granularity loss function for weakly supervised video moment retrieval. *IEEE Trans. Multimed.* **2021**, *24*, 1141–1151. [CrossRef]
4. Tan, R.; Xu, H.; Saenko, K.; Plummer, B.A. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 2083–2092.
5. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
6. Chudasama, V.; Kar, P.; Gudmalwar, A.; Shah, N.; Wasnik, P.; Onoe, N. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4652–4661.
7. Li, J.; Wang, X.; Liu, Y.; Zeng, Z. CFN-ESA: A Cross-Modal Fusion Network With Emotion-Shift Awareness for Dialogue Emotion Recognition. *IEEE Trans. Affect. Comput.* **2024**, *15*, 1919–1933. [CrossRef]
8. Pujol, F.A.; Mora, H.; Martínez, A. Emotion recognition to improve e-healthcare systems in smart cities. In Proceedings of the Research & Innovation Forum 2019: Technology, Innovation, Education, and their Social Impact, Rome, Italy, 24–26 April 2019; pp. 245–254.
9. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
10. Dang, C.N.; Moreno-García, M.N.; Prieta, F.D.l. An approach to integrating sentiment analysis into recommender systems. *Sensors* **2021**, *21*, 5666. [CrossRef] [PubMed]
11. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Volume 2019, p. 6558.
12. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
13. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
14. Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; Morency, L.P. Multi-level multiple attentions for contextual multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 1033–1038.
15. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; Volume 2020, p. 2359.

16. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 July 2019; Volume 1, p. 2.
17. Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; Yang, K. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 3718–3727.
18. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
19. Arjmand, M.; Dousti, M.J.; Moradi, H. Teasel: A transformer-based speech-prefixed language model. *arXiv* **2021**, arXiv:2109.05522.
20. Hu, D.; Hou, X.; Wei, L.; Jiang, L.; Mo, Y. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7037–7041.
21. Paraskevopoulos, G.; Georgiou, E.; Potamianos, A. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 4573–4577.
22. Dosovitskiy, A. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
24. Wu, Z.; Gong, Z.; Koo, J.; Hirschberg, J. Multimodal multi-loss fusion network for sentiment analysis. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024; pp. 3588–3602.
25. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.H.; Chang, S.F.; Cui, Y.; Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24206–24221.
26. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
27. Hassani, K.; Khasahmadi, A.H. Contrastive multi-view representation learning on graphs. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 4116–4126.
28. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 9–14 July 2023; pp. 7617–7630.
29. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
30. Baevski, A.; Hsu, W.N.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 1298–1312.
31. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259.
32. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
33. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
34. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 10790–10797.
35. Mai, S.; Zeng, Y.; Zheng, S.; Hu, H. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2276–2289. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Article

Affective Stroking: Design Thermal Mid-Air Tactile for Assisting People in Stress Regulation

Sheng He [1], Hao Zeng [2], Mengru Xue [3,*], Guanghui Huang [4], Cheng Yao [1,*] and Fangtian Ying [1]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China; hesheng@zju.edu.cn (S.H.)
[2] School of Industrial Design, Nanjing University of the Arts, Nanjing 210003, China; zenghaonua@163.com
[3] Ningbo Innovation Center, Zhejiang University, Ningbo 315000, China
[4] Faculty of Humanities and Arts, Macau University of Science and Technology, Macau 999078, China; ghhuang1@must.edu.mo
* Correspondence: mengruxue@zju.edu.cn (M.X.); yaoch@zju.edu.cn (C.Y.)

Abstract: Haptics for stress regulation is well developed these years. Using vibrotactile to present biofeedback, guiding breathing or heartbeat regulation is a dominant technical approach. However, designing computer-mediated affective touch for stress regulation is also a promising way and has not been fully explored. In this paper, a haptic device was developed to test whether the computer-mediated affective stroking on the forearm could help to assist people in reducing stress. In our method, we used mid-air technology to generate subtle pressure force by blowing air and generating thermal feedback by using Peltier elements simultaneously. Firstly, we found intensity and velocity parameters to present comfort and pleasant stroking sensations. Afterward, an experiment was conducted to find out whether this approach could help people mediate their perceived and physiological stress. A total of 49 participants were randomly assigned to either a Stroking Group (SG) or a Control Group (CG). Results showed that participants from SG felt more relaxed than those from CG. The physiological stress index, RMSSD, increased and LF/HF decreased in SG although these changes were not statistically significant. Our exploration created subtle, non-invasive, noiseless haptic sensations. It could be a promising alternative for assisting people in stress regulation. Design implications and future applicable scenarios were discussed.

Keywords: mid-air; thermal; haptic; tactile; affective touch; computer-mediated stroking; heart rate variability; stress reduction; digital health; human–computer interaction

1. Introduction

1.1. Background

In everyday life, people encounter various stressors, eliciting people's anxiety feelings. In the field of human–computer interaction (HCI), researchers have developed technological devices and systems, such as guided breathing [1], biofeedback for reflection [2], relaxation training through biofeedback [3,4], and tools for aiding meditation [5] to help individuals reduce stress and regulate emotions. The mainstream methods of stress regulation typically employ visual [1–4] and auditory [5,6] stimuli. In recent years, a growing body of research has been examining the role of haptics. In the context of haptic-based stress regulation, most cases involve vibrations or pressure sensations used to guide breathing rhythms [7–9], reflect heartbeat rhythms [10–13], or use tactile sensation to present biofeedback to relax [14] and ease anxiety [15]. However, the current haptic sensation is relatively simple, resulting in a monotonous tactile display form, and there is no way to create a more detailed tactile experience. Meanwhile, these haptic actuators generated mechanical noise, affecting the overall user experience.

Touch is a primary non-verbal communication channel for conveying intimate emotions and it is essential for our physical and mental well-being [16,17]. In the appropriate

context, non-harmful touch can not only reduce physiological stress indicators, such as heart rate, blood pressure, and cortisol levels, but also enhance heart rate variability. It can also increase levels of oxytocin, serotonin, and dopamine, which are linked to positive emotional states and overall well-being. Moreover, this type of touch can reduce sympathetic nervous activity linked to stress and anxiety [18–23]. Humans, compared to other species, have specialized nerve fibers (CT afferents) that are highly sensitive to gentle, warm, and caress-like touch, yielding positive emotional and physiological effects [24]. Prior research has shown that touch can support emotional regulation through affective touch, which includes gentle stroking characterized by slow, warm movements. This form of touch has been proven to promote secure attachment [25–27]. Other studies also proved that affective touch led to a reduction in autonomic responses and anxiety levels, indicating a calming effect [28]. Touch could be taken as "comforting or caring". Offering comfort is a vital prosocial behavior, which is frequently manifested through affiliative touch directed at individuals in distress, providing alleviation from their troubled state [29]. Usually, individuals who receive affective touch, experience three stages of emotion regulation [30]: situation modification, attentional deployment, and cognitive reappraisal (feeling closeness, connection, and support). This process helps to alleviate anxiety and stress through both external assistance and personal effort.

1.2. Studies About Affective Touch

Affective touch occurs when in contact with the skin [31]. Research in the field of neuroscience has found that CT afferent fibers are closely linked to encoding the emotional aspects of pain, itch, and tickle, as well as sensations of roughness and pleasantness [32]. Additionally, the emotional responses to touch differ between hairy and glabrous skin areas [33]. Two key factors influencing CT afferents are the intensity and speed of the stimulus. Generally, gentle, slow, and light movements are associated with more pleasant experiences compared to firmer strokes [34], and rough surfaces or rapid movements are perceived as less pleasant compared to smooth surfaces [35]. The rating of pleasantness for brush stroking gestures is greater when the stroking velocity is optimal for activating CT afferents (e.g., 3 m/s), compared to other velocities [36]. Temperature sensations are also linked to the mechanisms underlying the perception of pleasantness. People are more likely to accept caress-like touches that are similar to their own body temperature, approximately around 32 °C [37].

1.2.1. Different Actuation Technologies to Imitate Pleasant Stroking

In the field of Human–Computer Interaction, many studies focus on using different actuation technologies to reproduce CT stroking. The common approach is using vibrotactile stimulation. With varying intensity, frequency, velocity, and direction of motion, researchers generate pleasant sensations, similar to real touch [38–40]. Jukka Raisman et al. also found the more continuous the stimulation, the more pleasant it was and the lower the arousal level [41].

In addition, studies have used SMA or voice coil to generate mechanical movement on the skin, thereby producing a feeling of pressure or friction and providing pleasant and natural sensations [42–44]. Lawrence H. Kim et al. studied the VPS (Vibration, Pressure, and Shear) display and found that participants rated the shear mode the highest pleasantness and continuity [45]. Weicheng Wu et al. designed a wearable device consisting of a fabric sleeve and thermoplastic pneumatic actuators arranged in a linear array and found pressure changes during the inflation process did not affect continuity or pleasure [46]. Yiran Zhao et al. developed a wearable device, using synthetic fur which simulated a gentle stroking sensation on the user's forearm. Results showed that participants who received affective touch experienced lower state anxiety and the same physiological stress response level compared to the control group participants. They also found that affective touch facilitated stress regulation by rendering pleasantness, providing emotional support, and shifting attention [47].

1.2.2. Using Mid-Air and Thermal Technology to Influence Affections

Existing technologies have limits for displaying haptic renderings, such as affective touch. The tactile sensations produced by vibration devices are relatively simplistic, often requiring visual assistance to understand the specific haptic gestures [48,49]. Additionally, both vibration and mechanical devices generate noticeable noise [43], which affects the ability of mediated touch to provide a similar experience to human-to-human touch and thus could result in less authentic and skewed emotional expressions. Emerging trends in contactless haptic interfaces are gaining attraction, utilizing mid-air technologies, such as air-jet [50,51] and acoustic radiation pressure [52–55], to generate tactile stimulation without physical contact with the user. Mohamed Yassine Tsalamlal et al. emphasized a significant impact of the air jet's intensity and movement speed on assessments of valence, arousal, and dominance [50]. Marianna Obrist et al. suggested that spatial, directional, haptic characteristics, such as frequency, intensity, and duration of the haptic stimulation, were important factors in generating haptic sensations by employing mid-air tactile stimulation [52]. Sean Chew et al. found that users could directly experience the applications without the need to wear any devices, enhancing the speed and hygiene of the experience [53]. What is more, mid-air tactile is a new promising technology used for stress regulation. Yuka Sato et al. assessed how users perceive haptic stimuli created by air vortex rings applied to the cheek and examined their impact on physiological responses. The findings indicate that specific stimuli may help alleviate stress [51]. In the field of VR, researchers have explored the effect of combining mid-air technology with thermal haptics to create a better immersive experience [55].

Researchers have also found a fundamental link between temperature and intimacy [56], social connections [57], social closeness [58], trust [59], and a warm personality [60]. Furthermore, psychotherapy research shows that warm hands in clients are linked to emotional security [61]. Participants thought these thermal sensations were comfortable and felt like human touch. In the HCI community, thermal touch technologies created thermal sensations of heat or cold. Many studies have shown that temperature affects both the valence and arousal of emotions [62–64]. Warm stimuli are perceived as more pleasant/positive compared to cold stimuli, and both emotional valence and arousal are influenced by modifying the rate or degree of temperature change [62]. Using thermal tactile for stress regulation has become more and more popular. Some works have focused on using heat to guide participants to pay attention to themselves [65] and engage in physical meditation [66] or relaxation [67]. Muhammad Umair et al. designed a personalized thermal tactile module for affect regulation following stress induction in participants [68]. The results showed that thermal tactile feedback effectively helped regulate stress and anxiety, playing a crucial role in the process. Individuals use the perceptual attributes of thermal patterns, along with emotional interpretations and metaphors, to attribute familiar meanings to thermal patterns. However, presenting temperature strokes is technically challenging. Yuhu Liu introduced a haptic device designed to generate a stroking sensation on the forearm by simultaneously applying pressure and thermal feedback. The compact and soft device employs micro blowers and inflatable pouches to create a pressure sensation while utilizing water to provide thermal feedback. According to a user study, it effectively simulated a thermal stroking sensation, and it was observed that cooler temperatures enhanced the perceived pleasantness of the stroking [69].

1.3. Scope of Our Research

A substantial number of haptic technologies have been developed to simulate the sensation of stroking, yet the exploration of combining mid-air and thermal tactile technologies remains relatively limited. Also, few cases use the combination to simulate affective touch to assist stress regulation. A large number of blank areas have yet to be verified. We emphasize mid-air and thermal tactile technologies for stress regulation due to their experiential and material qualities, such as convenience to proceed, comfort, subtlety, ability to direct attention, gradual nature, and privacy. We believe that combining mid-air and thermal feedback can create a comfortable and pleasant stroking experience, further helping users

alleviate anxiety and support stress regulation. Compared to other affective touch gestures, such as hugging or patting, which require more complex hardware, stroking demands lower technical complexity. Additionally, stroking can maximize the potential of mid-air technology by simulating a gentle and soothing tactile experience, which is more effective for alleviating stress and anxiety. For the body part selection, we chose the forearm. The forearm is the most extensively studied body part, providing a solid foundation of previous research [38–47,50,53]. Moreover, it is one of the most commonly and easily touched body parts, offering significant potential for developing various corresponding devices. This paper explores affective touch technologies based on mid-air technology, combining thermal sensation for personal stress regulation. We defined research questions as follows:

- How can designs utilize mid-air and thermal tactile technology to create a sensation of being caressed and evoke positive feelings in users?
- What are the effects and impacts of this technology on participants' stress regulation?
- What are other potential applications or scenarios that can adopt this technology?

To address these questions, we developed a haptic stroking simulation device. Initially, we evaluated the device to determine the suitable intensity and velocity settings for delivering comforting and pleasant stroking sensations. Subsequently, we conducted a between-group experiment and semi-open interviews with 49 participants. They were randomly divided into either the Stroking Group (SC) or the Control Group (CC). We hypothesized that the computer-mediated stroking device for stress regulation could help to assist people in reducing stress, compared to those who did not use our designed device. After experiencing the stress induction phase, participants who interacted with our haptic device showed a significant reduction in anxiety based on the State-Trait Anxiety Inventory (STAI). The heart rate variability (HRV) feature of those participants also altered, indicating reduced stress and increased relaxation, although the difference was not statistically significant. Our contributions are as follows:

1. We introduce a novel method that creates subtle stroking sensations by combining mid-air and thermal feedback technologies, thereby exploring new design possibilities offered by haptic sub-modalities;
2. We explore design insights of computer-mediated affective stroking, which contributes to the expansion of the design space for stress regulation technologies;
3. We summarize promising design solutions for practical application scenarios, which refine future research and design in mid-air and thermal tactile technologies.

2. Materials and Methods

2.1. Hardware Design and Components

The Stroking Device is composed of three components: the Stroking Component, Thermal Control Component, and Hardware Driver Component (Figure 1). The first component is the Stroking Component, consisting of 8 fans, each measuring 4 cm $\times$ 4 cm. These fans are relatively small and can be combined freely. Therefore, it can be better designed according to the needs of the situation. They are arranged in two columns. Each column contains 4 fans and has a total length of 20.5 cm. The distance between each fan is 1.5 cm. One column is placed on the top of the forearm, while the other is positioned on the outer side (Figure 2). Each corresponding pair of fans from the two columns forms a row, resulting in four rows (P1, P2, P3, P4).

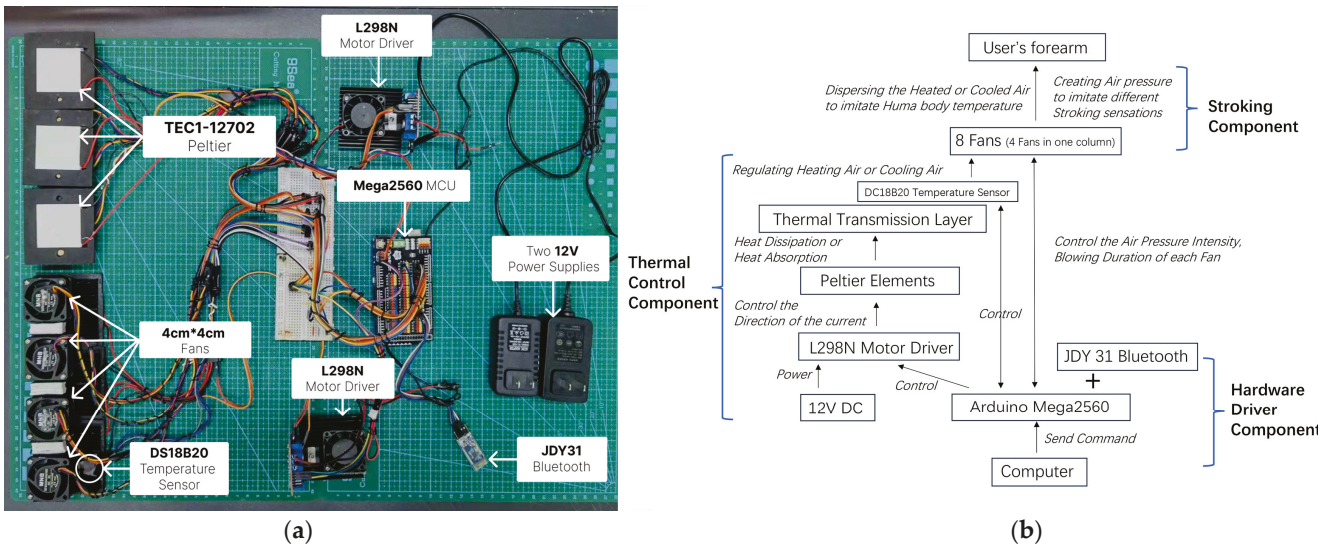


Figure 1. Hardware design: (**a**) hardware of the stroking device and (**b**) framework of the whole system.

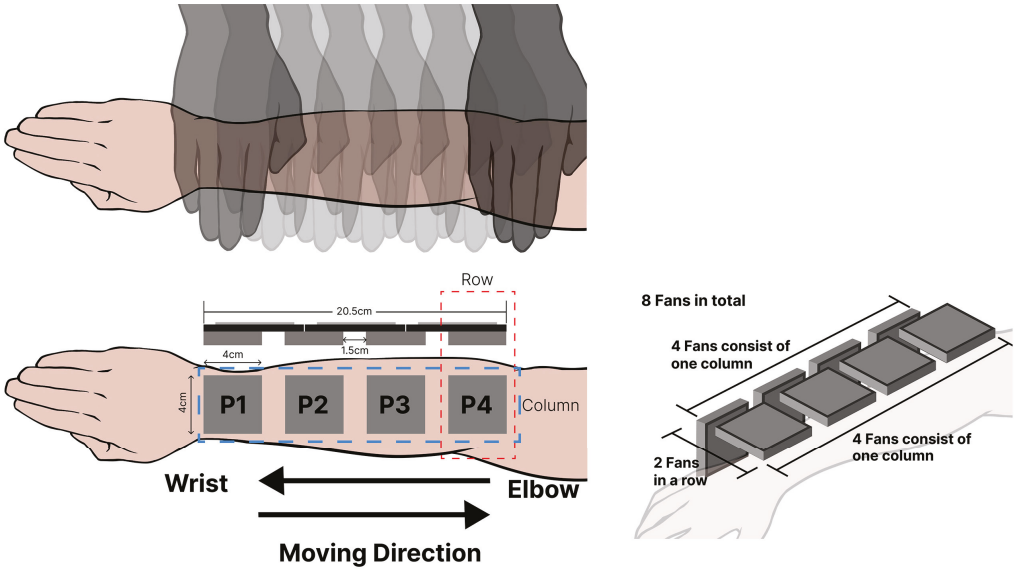


Figure 2. Arrangements of 8 fans and blowing-air pressure on the skin to imitate stroking sensation.

The second component is the Thermal Control Component (Figure 1). Each column of four fans is fixed onto a metal heat conduction layer, which has three Peltier elements (TEC1-12702) attached to its back. Peltier elements have been frequently used as thermal driving devices in previous studies [62–64]. The heat generation and transfer by the Peltier elements are quick and easy. A motor driver (L298N Dual H Bridge) regulates the temperature hot/cold changes in the Peltier elements by controlling the current. Heat is transferred through the conductive layer to the back of each fan. The fan blades disperse the heated air, allowing participants to feel the temperature, thereby simulating the human warmth that should be included in stroking. In this setup, we set the constant temperature to 32 °C [37]. It is within a comfortable temperature range for humans. The DS18B20 temperature sensor detects real-time temperature changes. A PID (Proportional, Integral, Derivative) algorithm is implemented in the software to maintain the temperature at 32 ± 0.5 °C. L298N motor drivers are powered by 12 V power supplies.

The third component is the Hardware Driver Component (Figure 1). Because there are many electronic components, the Mega2560 MCU can provide enough interfaces to ensure the normal operation of the system. We use it to control the fan's blowing speed, the variation in the simulated stroking speed, and the output current of the two motor drivers to manage the thermal changes produced by Peltier elements. Additionally, we develop an Android application that connects to the MCU via a JDY31 Bluetooth module. This allows us to trigger various stroking effects, facilitating participant relaxation based on affective touch in subsequent experiments.

2.2. Computer-Mediated Stroking Design

2.2.1. Parameters of Generating Stroking

Based on previous research [38–40], stroking is a dynamic movement influenced by three factors: the force applied during stroking, the speed of stroking, and the direction of stroking.

Since we use an Arduino Mega 2560 MCU, all control programs are written in Arduino. We translate the stroking force into the intensity of the fan's airflow. In Arduino, the output signal has only two states: high (5 V) and low (0 V). However, many applications require analog signals (signals that can take any value between 0 V and 5 V). PWM (Pulse Width Modulation) simulates an analog signal by adjusting the duty cycle of the high state. By adjusting the duty cycle, we can change the proportion of time the fan motor receives power, thereby controlling the fan motor's speed and, consequently, the intensity of the fan's airflow.

The stroking speed is controlled by the rhythm of the fan's airflow in our device. We group the fans into four rows (P1, P2, P3, and P4). In the Arduino program, only one row of fans blows air at a time, while the other three rows remain off. The blowing sequence is either P1-P2-P3-P4 or P4-P3-P2-P1 (Figure 2). By setting the time intervals for each row's blowing, we control the overall speed of the simulated stroking produced by the device.

$$V = \frac{L}{T}, \tag{1}$$

where L is the length of the device, T is the total duration when our device starts blowing from one end to when it finishes blowing at the other end, and V is the stroking velocity. Since we set the intervals to 1.5 s, 2 s, and 2.5 s. The total duration (T) we set in the Arduino program will be 7.5 s, 10 s, and 12.5 s, respectively. Our device has a total length (L) of 20.5 cm. Thus, the corresponding velocities (V) calculated will be 2.73 cm/s, 2.05 cm/s, and 1.64 cm/s, respectively. Research has shown that the optimal speed for a comfortable stroking sensation on hairy skin is between 1 and 10 cm/s [36,69]. The user's forearm skin will feel the slight pressure from the airflow, and the movement of this pressure across different parts of the forearm creates a continuous sensation similar to stroking (Figure 2). Since the blowing sequence can proceed in two directions, and the device's size generally covers most people's forearms, participants can experience stroking in two directions, either from wrist to elbow or from elbow to wrist.

2.2.2. Pilot Study

The purpose of this study is to test which combination of independent parameter variables can best simulate the effect of stroking, providing users with comfort and a sense of pleasantness. Our experimental device consists of an array of 8 fans arranged in four rows, surrounding the outer and upper sides of the arm (Figure 3). To avoid any potential influence from the experimental device, we used KT boards to make a box that covers the device.

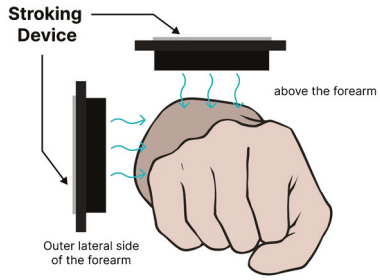


Figure 3. Fans around the outer and upper sides of the arm, blowing air pressure on the forearm.

We designed a 2 ("Moving Direction") × 3 ("Duration") × 2 ("Intensity") repeated measures experiment. The "Moving Direction" variable includes two levels: from the wrist to the elbow, and from the elbow to the wrist. The "Duration" variable refers to the time interval between the activation of adjacent fans, with intervals of 1.5 s, 2 s, and 2.5 s. The "Intensity" variable controls the air force output of the fans using Arduino's PWM, with levels set at "200" and "255."

We recruited 10 participants, including 7 males and 3 females, aged between 23 and 25. All participants were students from a local university. Each participant signed a consent form, agreeing to participate in the experiment and allowing their data to be used for research purposes.

The experiment included 12 parameter combinations based on a 2 ("Moving Direction") × 3 ("Duration") × 2 ("Intensity") design, plus two additional sets without parameters, resulting in 14 sets of wind haptic parameters. These 14 sets were counterbalanced by randomly ordered for each participant to experience.

Upon entering the experimental environment, participants were asked to choose their dominant hand, and the experimenter adjusted the device's position accordingly. Each participant then inserted their arm into the device's box and closed their eyes. During the experience, the participants needed to report whether they felt a stroking sensation and identify the direction of the stroke. They then evaluated each parameter set on four dimensions: perceived continuity, perceived authenticity, perceived comfort, and perceived pleasantness, using a 7-point Likert scale for their ratings. These are the four metrics commonly measured in previous studies on mediated stroking design [41–46]. Participants reported their scores, and the experimenter recorded the data. At the end of the experiment, we conducted a brief semi-structured interview to gather qualitative data from the participants.

2.2.3. Mediated Stroking Design Decisions

Shapiro–Wilk normality tests showed that most participants' ratings are normally distributed ($p > 0.05$). Ratings also met Mauchly's Test of sphericity ($p > 0.05$).

Firstly, Intensity had statistically significant effects on the ratings of continuity and authenticity. It had no statistically significant effects on the ratings of comfort and pleasantness. The ratings for continuity were 4.87 ± 0.29 for PWM_200 and 5.59 ± 0.28 for PWM_255. The ratings for Authenticity were 4.28 ± 0.37 for PWM_200 and 4.88 ± 0.36 for PWM_255. Compared to PWM_200, PWM_255 provides greater power output. Therefore, with increased airflow, the perception becomes more pronounced, resulting in better scores across all evaluation metrics. Secondly, Duration had statistically significant effects on the ratings of continuity, comfort, and pleasantness. It had no statistically significant effect on the ratings of authenticity. Post hoc LSD comparisons showed that there was a significant difference between 1.5 s and 2.5 s on continuity ($p = 0.03$), comfort ($p = 0.01$), and pleasantness ($p = 0.003$). Since the air generated a gentle sensation, the 2.5 s interval provided a longer blowing duration, resulting in a more sustained and stable airflow. This made the perception more pronounced, leading to higher scores across all evaluation metrics. Thirdly, the two-way interaction of Intensity × Duration had statistically significant effects on the ratings of continuity ($F(2, 18) = 5.42$, $p = 0.01$), authenticity ($F(2, 18) = 8.07$,

$p = 0.001$), and comfort (F (2, 18) = 5.23, $p = 0.01$). However, it had no statistically significant effect on the ratings of pleasantness (F (2, 18) = 2.11, $p = 0.15$). Figure 4 shows the results. The two-way interaction of Moving Direction x Intensity and the two-way interaction of Moving Direction × Duration did not reveal any significant effects on any of the measured variables. The three-way interaction was not found significant on any dependent variable.

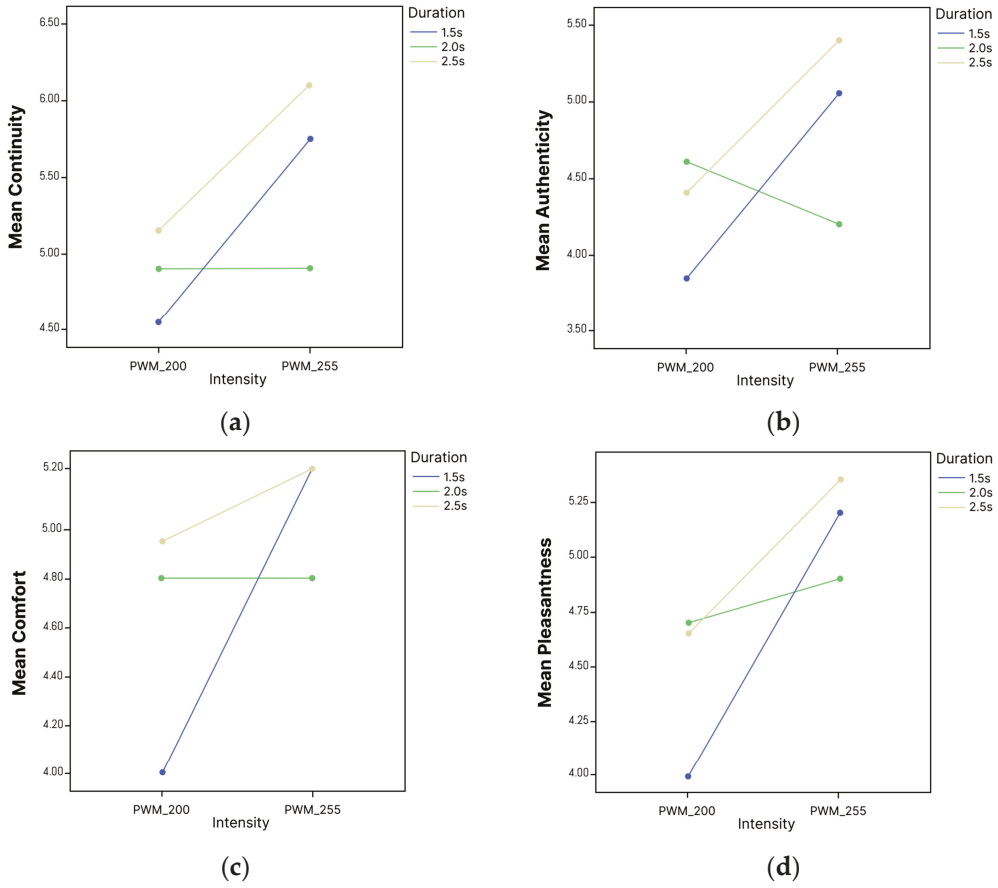


Figure 4. The two-way interaction of Intensity × Duration on four metrics: (**a**) perceived continuity; (**b**) perceived authenticity; (**c**) perceived comfort; and (**d**) perceived pleasantness.

From the study, we can see the intensity of the airflow affects the perception of the stroking force. In our experiment, users felt that the stroking was closer to a gentle human touch when the force was generated by the PWM_255 setting. Additionally, the duration of the interval influences the airflow output, altering the perceived stroking force. The longer the duration is, the stronger the perceived force is. In our study, the 2.5 s interval provided better results across all evaluation metrics. However, due to the varying speeds in real stroking, the 2.5 s interval setting was felt slower. On the other hand, the 1.5 s interval setting resembles the speed of a real human stroking more closely.

We also found that the user's perception of stroking primarily depends on the stroking force and the speed. The higher airflow intensity closely resembles the feeling of actual stroking. Blowing duration also affects the real-stroking perception, but not as important as the intensity. In our study, each fan operated sequentially to create a continuous tactile sensation, with the airflow rhythm being consistently maintained across different stroking parameter sets. The participants could tell the difference in the stroking direction. However, they told us that the Direction of the stroking did not lead to any differences in perceived continuity, perceived authenticity, perceived comfort, and perceived pleasantness for them because the blowing rhythm was fixed.

Considering the above findings, we finally chose PWM_255 as the intensity parameter, with a 1.5 s interval for a fast-stroking effect and a 2.5 s interval for a slow-stroking effect.

The comfort and pleasure provided by the combinations of PWM_255 and 1.5 s interval or 2.5 s interval scored among the highest of all the combinations. We tried the effects of fast-slow stoking cycling and slow-fast stroking cycling within 5 min on the subjects and found that there was not much difference between the two. We decided to use fast-slow stroking as a cycle in the following experiments.

2.2.4. Air Pressure Testing

We further tested the force of air pressure exerted on the skin by a fan under PWM_255 setting for intensity, 1.5 s interval, and 2.5 s interval setting for fast and slow stroking speed. First, we measured the airflow velocity of the fan using an anemometer and then calculated the dynamic pressure using the formula from fluid mechanics. The dynamic pressure is the pressure caused by the airflow when an object is exposed to wind and can be expressed by the following formula:

$$q = \frac{1}{2} \cdot p \cdot v^2, \tag{2}$$

where q is the dynamic pressure, measured in Pascals (Pa); p is the air density, measured in kilograms per cubic meter (kg/m^3), typically around 1.225 kg/m^3 under standard conditions; and v is the airflow speed, measured in meters per second (m/s). Subsequently, the dynamic pressure can exert a force F on the surface of an object, which can be calculated using the following formula:

$$F = q \cdot A, \tag{3}$$

where F is the force, measured in Newtons (N), and A is the surface area of the object exposed to the airflow, measured in square meters (m^2). Substituting the dynamic pressure equation into the formula for force yields:

$$F = \frac{1}{2} \cdot p \cdot v^2 \cdot A, \tag{4}$$

In our case, the surface area of each fan is 0.0016 m^2, which corresponds to the skin area exposed to air pressure. The time intervals were set to 1.5 s for the slow stroking and 2.5 s for the fast stroking, with each fan blowing for durations of 3 s and 5 s, respectively. Figure 5 illustrates the variation of the force generated by a single fan blowing on a fixed skin area for a predetermined duration. The data represent the mean values obtained from three separate measurements.

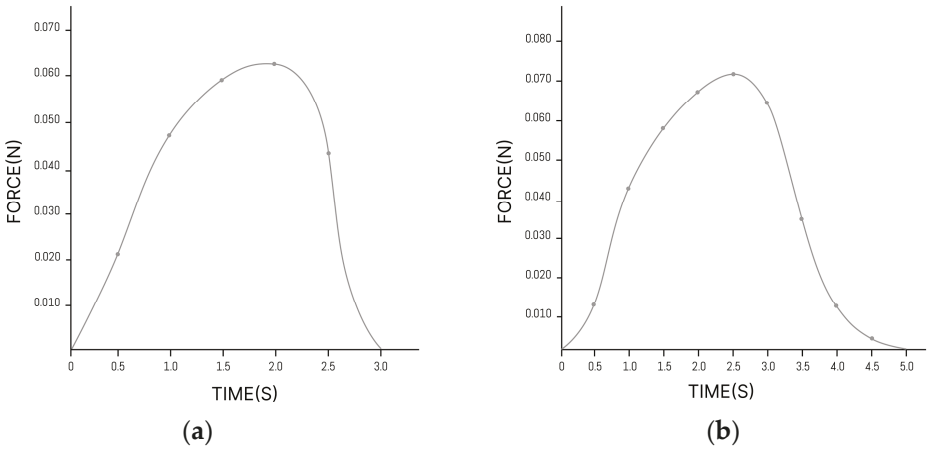


Figure 5. Force generated by a single fan blowing once for 3 s (**a**) and 5 s (**b**).

We compared the air pressure parameter with the pressure parameter caused by human touch. In this instance, we used OKWILL pressure sensors to measure the pressure values exerted on the skin of the human arm by both light and firm real human touches. We also sensed air pressure from our device. The testing durations were set to a predetermined

fast stroking time of 7.5 s and a slow stroking time of 12.5 s. As shown in Figures 6 and 7, the pressure data from real human touches fluctuated continuously, whereas the air pressure generated by our device exhibited an initial increase and a decrease at the start and end, respectively, while maintaining relatively stable output in between. The results indicated that the pressure generated by the fan differed from that of a firm human touch, but was relatively similar to that of a light human touch. These results aligned with our expectations, as our goal was to replicate the sensation of gentle stroking and further assess its potential for stress relief in users.

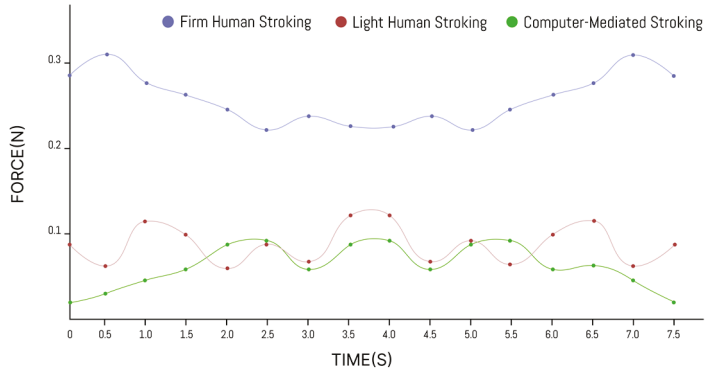


Figure 6. Comparison of pressure parameters caused by air and human touch for fast stroking.

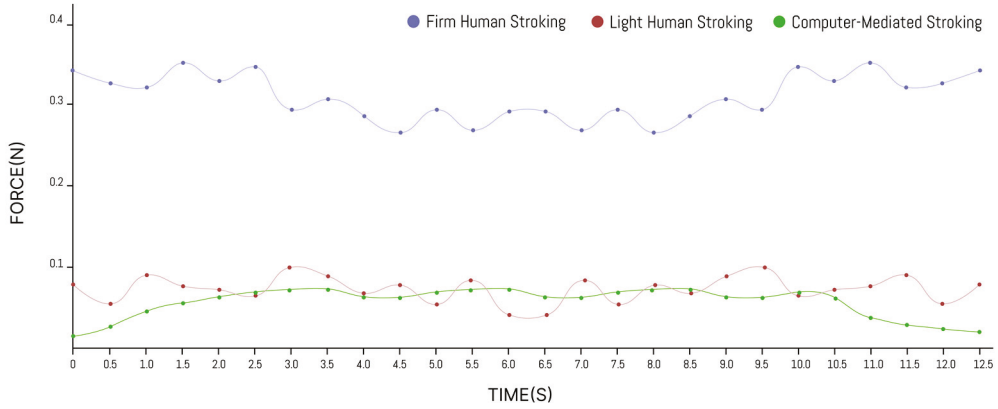


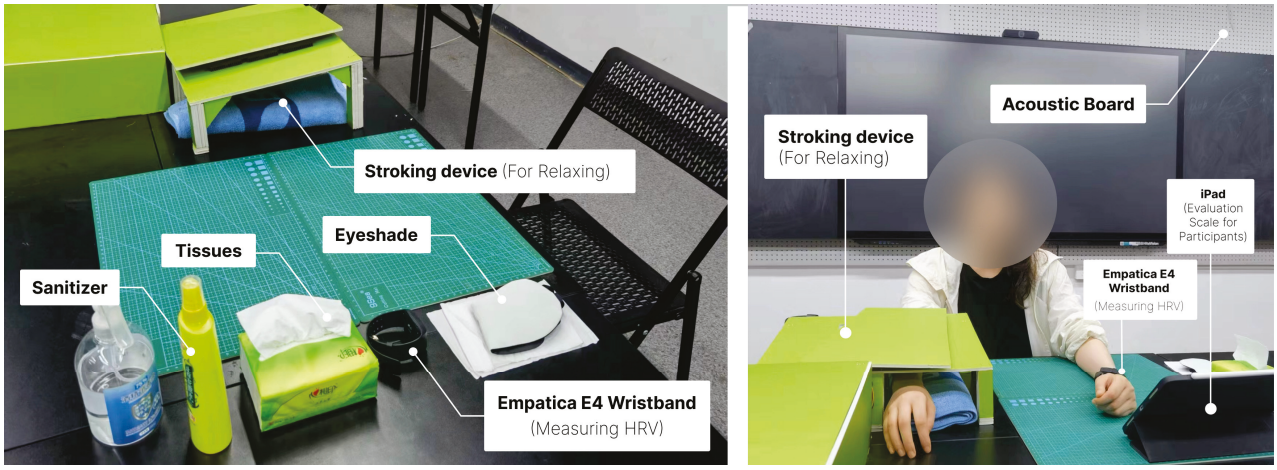
Figure 7. Comparison of pressure parameters caused by air and human touch for slow stroking.

2.3. Experiment Design and Setup

The purpose of this experiment was to evaluate the effectiveness of computer-mediated stroking in reducing stress. The participants were divided into two groups: the Stroking Group (SG) and the Control Group (CG). Both groups carried out a within-subjects experiment, including phases of baseline relaxation, stress induction, and subsequent relaxation. Initially, baseline physiological data were collected for two groups. After obtaining data from the first phase, both groups undergo the same stress induction task. In the final phase, participants from SG relaxed with their eyes closed and received computer-mediated stroking from the device at the same time (Figure 6), while CG only relaxed with their eyes closed.

The experimental room was surrounded by acoustic boards as the experimental environment to avoid noise. The room's windows were equipped with curtains to completely block outside light and prevent potential visual distractions. The experiment took place during the summer, and we used an air conditioner to maintain the room temperature between 22 and 24 degrees Celsius. The lighting in the room was cool-colored light. Inside the room, we arranged the tables and chairs for the experiment. A mat was placed in the center of the table. Since the experiment involved wearable devices, we placed sanitizer

and tissues on the left side of the mat for disinfection and cleaning. Since participants needed to close their eyes during the experiment, we also prepared an eyeshade on the left for each participant. The Empatica E4 Wristband (Measuring HRV) was also placed on the left side. On the right side of the mat, we placed the mediated stroking device, covered by a KT board-made box. In front of the participant's seat, we placed an iPad for completing experimental questionnaires (Figure 8). The experiment was approved by the ethics committee from Macau University of Science and Technology (MUST-20230925001).



(**a**) (**b**)

Figure 8. Experiment design: (**a**) experiment setup and (**b**) experiment environment.

2.4. Participants

When selecting participants, we inquired about their physical and mental health. Only those with good overall health were eligible to participate in our experiment. A total of 55 participants (24 males and 31 females) took part in the experiment. Their professions included university students and professors, freelancers, and corporate employees, with ages ranging from 23 to 50 years old. The participants were recruited through Snowball Sampling [70] and they were all right-handed. All participants read and signed informed consent forms before the experiment. They were randomly assigned to either the SG or the CG. In total, 6 participants quit the experiment because they did not want to experience stress induction (doing mathematical tasks) during the experiment or had something interrupted and could not wait until the experiment finished. In the end, the SG consisted of 25 participants, while the CG had 24 participants. In the SG, there were 8 participants (4 males, 4 females) aged 20–29, 8 participants (3 males, 5 females) aged 30–39, and 9 participants (3 males, 6 females) aged 40–49. In the CG, there were 8 participants (3 males, 5 females) aged 20–29, 8 participants (4 males, 4 females) aged 30–39, and 8 participants (4 males, 4 females) aged 40–49 (Table 1). The participants we selected represented the target user group for our design in terms of occupation, age, and gender distribution.

Table 1. The participants were assigned to the experimental conditions.

	Age 20–29			**Age 30–39**			**Age 40–49**					
	Males	**Females**	**Total Participants**	**Males**	**Females**	**Total Participants**	**Males**	**Females**	**Total Participants**	**Total Males**	**Total Females**	**Total Participants**
Stroking Group (SG)	4	4	8	3	5	8	3	6	9	10	15	25
Control Group (CG)	3	5	8	4	4	8	4	4	8	11	13	24

2.5. Procedure

The experiment consists of three phases (Figure 9), lasting a total of 30–40 min. Before the first phase, participants read and sign an informed consent form, followed by a brief introduction to the experiment. They were equipped with the Empatica E4, which continuously captured their physiological data throughout the experiment.

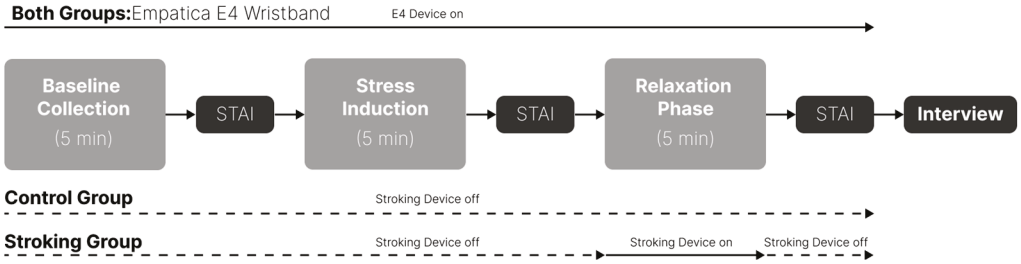


Figure 9. Experiment Procedure.

In the first phase, all participants were asked to sit quietly with their eyes closed for 5 min to collect HRV baseline data. The participants then completed the "state" part of the State-Trait Anxiety Inventory (STAI-S) [71] questionnaire. This phase will last about 10 min.

In the second phase, all participants underwent a 5 min stress induction, details can be seen in Section 2.6. Following this, they completed the STAI-S questionnaire for the second time. This phase was completed within a total of 10 min.

In the third phase, all participants were required to close their eyes and relax uniformly. Members of the SG received an affective touch from a stroking device at the same time, while members of the CG only relaxed with their eyes closed. We applied the fast and slow stroking designed in the pilot study. The stroking protocol was a fast stroking back and forth, followed by a slow stroking back and forth, repeated in a cycle. A total of 15 cycles were performed in 5 min. Both groups experienced 5 min. The participants then completed the STAI-S questionnaire for the third time. This phase was also completed within 10 min.

At the end of the experiment, we also conducted user interviews to collect and analyze qualitative data from the participants. The interpretation of the quantitative findings would be supported by the interview data.

2.6. Anxiety-Inducing Stimulus

The participants of the two groups received an anxiety-inducing event in the form of a mental arithmetic task, after the baseline measures had been taken and prior to the final relaxing intervention. The task [72] required participants to count aloud a reverse counting task starting from 9000 and subtracting 7 each time. This part lasted 5 min.

2.7. Hypotheses

Our experiment hypotheses were as follows:

- Participants who used the computer-mediated stroking device for stress regulation would reduce subjective stress perception more than the control group (H1);
- Participants who used the computer-mediated stroking device for stress reduction were expected to show a greater increase in the physiological indicator RMSSD, compared to the control group (H2);
- Participants who used the computer-mediated stroking device for stress reduction were expected to show a greater decrease in the physiological indicator LF/HF, compared to the control group (H3).

2.8. Measurements

According to previous research [13], due to individual differences, comparing the mean data between groups during the relaxation phase does not effectively reveal the

actual impact of the design intervention. Instead, comparing the mean changes in metrics (relaxation phase—stress induction phase) between the SG and CG is a more accurate way to assess the true effect of our design intervention. We used a subjective scale and two physiological indicators as follows:

- State-Trait Anxiety Inventory (STAI);

For this experiment, which focused on short-term changes in anxiety rather than long-term anxiety issues at the individual level, we used the part of "state" of the State-Trait Anxiety Inventory (STAI-S) [71] to collect subjective data from participants at the end of each experimental phase. The questionnaire consisted of 20 items and used a 4-point scale for responses, ranging from 1 ("not at all" or "almost never") to 4 ("very much" or "almost always"), based on the participant's current state.

- Root Mean Square of Successive Differences (RMSSD);

RMSSD [73–75] is a commonly used metric for assessing heart rate variability (HRV). It measures the differences in time intervals between successive heartbeats (RR intervals), reflecting instantaneous changes in heart rate and the ability to self-regulate. In HRV analysis, RMSSD is primarily used to evaluate the activity level of the parasympathetic nervous system, which plays a crucial role in heart rate regulation. An increase in RMSSD indicates heightened parasympathetic activity, often observed during recovery from high-intensity exercise or stressful environments [74,75]. Conversely, a decrease in RMSSD suggests reduced parasympathetic activity, which is associated with higher levels of stress [74,75].

- Low Frequency/High Frequency (LF/HF);

In HRV analysis, LF and HF [76,77] are commonly used frequency domain indicators. The LF/HF ratio is frequently used to assess the balance between sympathetic and parasympathetic nervous system activity. A higher LF/HF ratio may indicate predominant sympathetic activity, which is typically associated with higher stress levels [76,77]. Conversely, a lower LF/HF ratio may suggest a dominant parasympathetic activity, which is generally linked to lower stress levels and better relaxation [76,77].

We calculated HRV metrics from the cleaned RR intervals for each 5 min segment of our experiment [47]. The processing software was Kubios 4.1.1. This 5 min window is selected as it is the widely accepted minimum duration for deriving neurophysiologically significant HRV metric [74].

- Interview.

In our interviews, we addressed eight questions across three parts (Appendix A). The first part (Q1, Q2) focused on the sensory experience and related associations of the computer-mediated stroking itself (with emphasis on the perception of stroking, temperature sensation, comfort, and pleasantness). The second part (Q3–Q6) involved the participants' perceptions during the relaxation phase of the experiment and the impact on individual stress regulation (whether or not it was effective). The third part (Q7, Q8) discussed potential improvements in design features and future usage scenarios. We audio-recorded interviews with each participant. The data were transcribed following the methodology of qualitative content analysis [78]. Two researchers independently transcribed and analyzed the recordings line by line, extracting descriptions relevant to the themes. They named similar events and situations, categorized them, and conducted frequency counts, including the number of participants under each description, to highlight the significance of each theme.

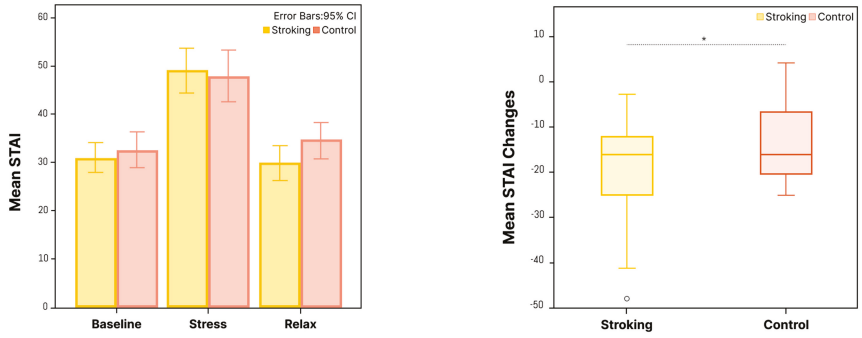
3. Results

We conducted operational checks to determine whether the stress induction phase successfully elicited anxiety states and changes in objective HRV metrics among all participants. The Shapiro–Wilk test indicated that not all subjective and objective measures were statistically normal across three phases. Therefore, we conducted the Wilcoxon test and Mann–Whitney U Test for non-parametric measures and independent samples t-test for the

parametric measurement between baseline phase and stress-induction phase. Significant differences were found between these two phases for STAI scores ($p < 0.01$), RMSSD metrics ($p < 0.01$), and LF/HF metrics ($p < 0.01$). During the stress induction phase, we successfully elicited anxiety states in all participants, as evidenced by the decrease in RMSSD and the increase in LF/HF.

3.1. Reduces Subjective Anxiety

We assessed whether our stroking device effectively helped participants relax and alleviate anxiety by comparing the change in STAI scores (relaxation phase—stress induction phase) between the SG and the CG. The Shapiro–Wilk test indicated the change in STAI scores was statistically normal for both groups. An independent-samples T test was conducted to analyze the data. The results indicated a significant difference ($p = 0.045$, two-tailed) between the SG (M = −19.6, SD = 10.50) and the CG (M = −13.38, SD = 9.04). Then, we calculated effect sizes and confidence intervals. Cohen's d was 0.63 and the 95% confidence intervals ranged from −11.43 to −0.14. The result showed the effect size was medium, and the 95% confidence intervals also proved a significant difference between groups. The participants from the Stroking Group felt more relaxed than those from the Control Group (Figure 10). Hypothesis 1 was supported.



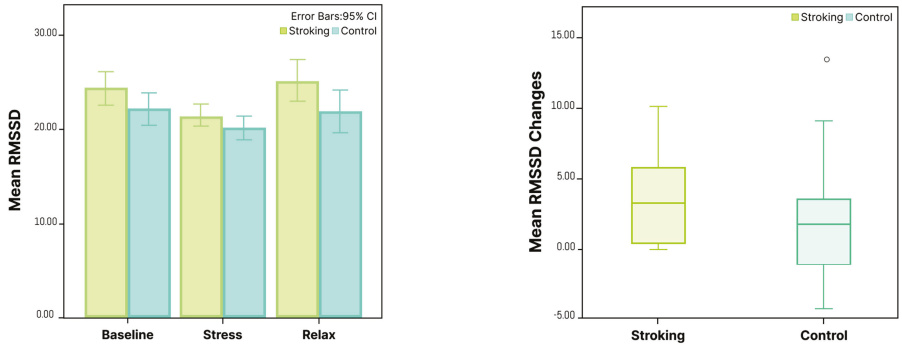
(**a**) (**b**)

Figure 10. STAI scores: (**a**) mean scores across three phases and (**b**) changes during the relaxation phase. (*: $p < 0.050$).

3.2. Enhances Heart Rate Variability

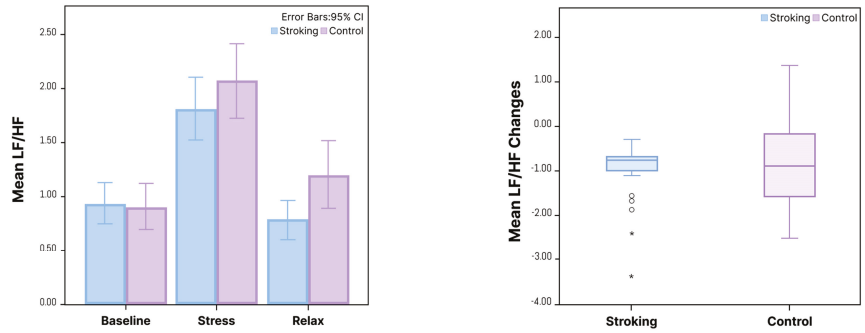
We evaluated whether our stroking device effectively helped the participants relax and alleviate anxiety by examining changes in RMSSD metrics (relaxation phase—stress induction phase) and changes in LF/HF metrics (relaxation phase—stress induction phase) between the SG and the CG. The Shapiro–Wilk test indicated changes in RMSSD metrics and LF/HF metrics were statistically normal for both groups.

An independent-samples T test was conducted on the changes in RMSSD metrics (Figure 11), revealing no significant difference ($p = 0.07$, two-tailed) between the SG (M = 3.65, SD = 3.33) and the CG (M = 1.71, SD = 4.06). Then, we calculated effect sizes and confidence intervals. Cohen's d was 0.52 and the 95% confidence intervals ranged from −0.20 to 4.07. It meant the effect size was medium, while the 95% confidence intervals presented no significant difference between groups. Similarly, an independent-samples T test on the changes in LF/HF metrics showed no significant difference ($p = 0.051$, two-tailed) between the SG (M = −1.03, SD = 0.68) and the CG (M = −0.87, SD = 1.00) (Figure 12). Then, we calculated effect sizes and confidence intervals between groups. Cohen's d was 0.19 and the 95% confidence intervals ranged from −0.65 to 0.33. It meant the effect size was small and the 95% confidence intervals also presented no significant difference between groups.



(**a**) (**b**)

Figure 11. RMSSD metrics: (**a**) mean RMSSD metrics across three phases and (**b**) changes in RMSSD metrics during the relaxation phase.



(**a**) (**b**)

Figure 12. LF/HF metrics: (**a**) mean LF/HF metrics across three phases and (**b**) changes in LF/HF metrics during the relaxation phase. (*: abnormal value).

In all, the results showed that RMSSD metrics increased, and LF/HF metrics decreased in SG, which means our haptic device had a positive effect on participants in improving Heart Rate Variability, although they were not significantly proven. Hypothesis 2 and 3 were denied. We also conducted a three-way ANOVA (gender, age, and group), and found that neither the main effects, two-way interactions, nor the three-way interaction revealed any statistically significant influence of age and gender on the changes in subjective and objective measures.

3.3. Qualitative Results

3.3.1. User's Perception of Mediated Stroking

The majority of participants (20/25) felt the sensation of gentle stroking. Seventeen participants felt it resembled a human's stroking. Three participants compared it to stroking sensations from plush toys, bristle brushes, or curtains. However, five participants thought that it lacked a tangible feeling. Among them, three participants felt it was more like wind and two participants did not think it resembled stroking. Most participants (16/25) like the temperature sensation. They felt the temperature was comfortable and similar to body temperature. Seven participants felt it was slightly warm (with two of them preferring cooler temperatures). One participant found it a bit cool, and one participant felt it was not noticeable.

Almost all participants (22/25) reported the overall feeling is comfortable. However, three participants mentioned a tingling sensation, which was a little bit disturbing. We asked the participants to imagine what kind of emotions were conveyed by computer-mediated stroking. Seven participants mentioned feelings of calmness, tranquility, and relaxation. Nine participants reported feelings of comfort, reassurance, and safety. Three

participants felt that the stroking helped them slow down and conveyed gentle emotions. Three participants felt that the machine was trying to communicate with them like a human, expressing care, or even conveying a playful feeling. Three participants did not have any particular associations.

Overall, our design provided participants with a sensation of gentle stroking. Due to the inherently gentle nature of airflow haptics and the addition of body-temperature-comfortable warmth, almost all users reported a positive and comfortable sensory experience. These laid the foundation for further stress regulation and relaxation.

3.3.2. Users Recalled Positive Experience for Stress Regulation

Regarding whether the mediated stroking evoked positive experiences, the majority of participants (20/25) reported recalling positive and pleasant experiences during relaxation. Among them, seven participants mentioned feeling relaxed before bedtime. P02: "*It felt like the state before sleep, receiving a new kind of hypnotic effect*". Seven participants felt that the simulated touch closely resembled a gentle touch from a real person, evoking experiences from intimate relationships, such as touches from parents or other caregivers during childhood, or emotional touches from romantic partners. P01: "*It reminded me of being stroked as a child and because I am a mom, I usually stroke my baby similar to this sensation*". P22: "*It felt like someone stroking me to sleep, thinking of my grandmother*". P24: "*It felt like when my mother used to hold and pat and stroke me as a child*". P34:" *it feels like my wife is flirting with me*". Two participants reminisced about past experiences with nature. P18: *"When I was a child, there was no air conditioning at home. Lying on the sofa by the window, the breeze would blow the curtains, and they would brush over me. It's a pleasant memory. The stroking was light, barely noticeable."* P42: "*It reminded me of standing on a grassy hillside, with the wind blowing, lifting my hair*". One participant felt it was similar to a spa experience, which was very pleasant. Three participants found it fun and intriguing, wanting to know what was stroking them. The remaining five participants had neutral emotions, with no significant emotional fluctuations.

Regarding whether the mediated stroking helped with stress regulation, almost all participants (23/25) mentioned that it provided emotional comfort. Concerning their stress regulation and relaxation methods during the relaxation phase, twelve participants mentioned zoning out or dozing off. Ten participants recalled pleasant memories and experiences related to intimate relationships. Four participants focused on the stroking itself. However, two participants felt it had little effect. P16 said: "*my personality tends to be unaffected by external factors*". P12 mentioned: "*He couldn't fully relax because they were preoccupied with thinking about the previous math problems*".

According to the interviews mentioned above, the positive and enjoyable sensory experiences, combined with the recalled positive memories, counterbalanced the negative feelings of stress and anxiety induced by the math test. Affective stroking helped participants to modify the situation, divert attention, and achieve cognitive reappraisal, as part of the emotion regulation strategy [30], alleviating negative emotions. It might contribute to further stress relief and relaxation.

3.3.3. Users Raised Expectations for More Design Details

Fourteen participants provided various suggestions for the further development of the technology and equipment, indicating they would use it if these conditions were met. Among them, four participants suggested integrating it with existing devices, such as massage equipment and office or home furniture. Four participants mentioned the need for more body parts and areas to be covered, such as both arms and eventually various parts of the body. Four participants emphasized user customization. P05: "*Control over the stroking rhythm*". P17: "*The interaction process should be simple*". P18: "*Users should be able to set preferences; for example, I enjoy the feeling of friction with textiles when sleeping. This includes control over force, temperature, and body area*". P37: "*The ability to adjust the air temperature would enhance comfort*". One participant mentioned the need for more detailed

usage scenarios, and another suggested designing more touch gestures and combining them with multimodal interactions. The remaining five participants expressed that they might use the technology, with four preferring their own methods, such as exercising, listening to music, or watching videos. One participant mentioned that she would not use the device to relax during stressful situations, but rather to enhance relaxation when she is already in a calm state. Another six participants mentioned that the current device already helps them relax, and they would continue using it without other suggestions. Design needs mentioned by the participants stemmed directly from their individual life experiences, providing valuable insights for our future research directions.

3.3.4. Users' Thoughts About Future Applications and Scenarios

We asked the participants to give more possibilities of using this technology and device in other applications or scenarios, according to their own experiences. Eight participants mentioned applications in gaming, entertainment, movie-watching, exhibition spaces, VR immersive experiences, and multimodal synesthetic experiences. Eight participants suggested integrating the technology with existing devices to aid relaxation and emotional regulation. Examples included scarves, clothing, home furnishings, office desks, car seats, massage chairs, beauty devices, sleep pods, and spa spaces. Six Participants expressed interest in using this technology to imitate humans' or animals' touch for remote communications with their family members and friends. Six participants mentioned it could be used as a relaxation aid during personal psychological therapy. Four participants suggested designing it as a sleep aid device. Three participants mentioned it could be used for personal relaxation and meditation. One participant mentioned that the touch felt like a playful interaction with a robot. This suggests that in the future, this technology could be used to design companion robots that engage in touch-based interactions, potentially addressing social issues such as loneliness. The participants mentioned potential usage scenarios, including homes, medical facilities, elderly care centers, physical therapy settings, workplaces, and car interiors. The participants' diverse opinions suggest that this technology has a wide range of potential applications in the future. A detailed discussion can be found in Section 4.2.2.

4. Discussion

4.1. Why Did the STAI Measurements Not Align with the Physiological Responses?

The results showed participants from the SG significantly felt more relaxed than those from the CG based on STAI measurements. The physiological stress index, RMSSD, also increased more, and LF/HF also decreased more in the SG, compared to the CG. However, the physiological data were not significantly proven. Three possible reasons are as follows:

- The novelty of the device may induce subjective preference in users: Some participants in the SG felt excited by using our novel device. They may have anticipated greater relaxation effects, potentially leading to elevated self-reported relaxation scores that do not necessarily reflect genuine physiological changes;
- The limits of the stress induction method: Real-world stressors are varied. The stress induction task mostly affects cognitive load. In real life, people are affected by lots of emotional-driven situations such as social stress caused by interpersonal relationships. Many participants noted that the anxiety induced by mathematical tasks is temporary, unlike ongoing life stressors. This made it easier for them (no matter which group they were in) to relieve anxiety during the relaxation phase, which may challenge the effectiveness of our design intervention;
- Not enough time for the affective touch to take effect on physiological response: Previous research has shown that the effect of affective touch on parasympathetic activation becomes significant only when participants remain in a stress-free environment for more than 16 min [47]. Our study only proceeded for 5 min for intervention.

Future research should aim to conduct experiments either in the lab with a new stress induction method, or in real-world settings, tracking users' stress regulation behaviors

for a period of time. We should also let participants become familiar with the device and then test them to avoid of preference or bias, creating enough time for an affective touch to impact the parasympathetic activity.

4.2. Design Implications

4.2.1. Individual Haptic Sensory Experiences Vary Significantly and User's Needs Differ

Numerous studies have highlighted that touch heavily relies on individualized contextual interpretation. Individual customization to meet specific needs is essential.

- The touch of the skin with different intensities affects the tactile sensation: Because we used mid-air tactile to imitate touch on human skin, most users appreciated the gentle sensation it provided. However, some users felt that the intensity was still insufficient compared to physical touch, lacking a sense of realism. Some participants mentioned that touch is not always at the same intensity. Further design should consider the change in stroking intensity to imitate real touch sensations. Others reported a tingling sensation akin to electric shock. This might be due to the fact that we connected eight fan devices simultaneously, and the MCU's supply voltage and current were insufficient, leading to inadequate power for each individual device. As a result, the airflow was weaker or produced a tingling sensation. Variations in hardware fan specifications might also cause inconsistencies in start-up times during operation, affecting continuity and diminishing the realism of the touch experience. Future designs should find appropriate voltage and current supply, consider the balance between airflow intensity and speed, and test the different sensory experiences they provide to users.
- People like thermal tactile at different individuals' preferences: To simulate human body temperature, we maintained the temperature at 32 ± 0.5 °C. Most users stated that a warm touch makes them feel real. They found this range comfortable and acceptable, while a few expressed contrary opinions, feeling either too hot or too cold. Previous studies have indicated a wide variability in users' temperature perceptions; some users experience pleasure from warmth [62,68], while others find cold temperatures pleasurable [68,69]. Future designs should further explore how variations in temperature intensity and rate affect sensory experiences. Consideration could be given to allowing users to customize the temperature to better suit their individual needs.
- Variations in stroking rhythm provide more real touch sensations: During the relaxation phase, we provided users with a fixed, uniform stroking rhythm, which led some users to perceive it as monotonous, particularly over the course of 5 min. Future designs should incorporate variations in stroking rhythm to enhance user experience. This could include one-directional repetitive stroking, bi-directional back-and-forth stroking, or variations in rhythm and pauses during stroking, all of which could contribute to a more authentic stroking sensation. Additionally, offering users the ability to customize their stroking preferences could be considered.
- Considering different body parts and areas for relaxation: Many users felt that applying the touch solely on the forearm was insufficient. They expressed a desire for this tactile experience to extend to more body parts, such as the head and face, and even for a full-body relaxation experience. Our mediated stroking simulation has evoked a sense of relaxation in users reminiscent of various other contexts. These similar experiences have led users to desire that our technology also provide comparable sensations in similar body areas.

4.2.2. Possible Design Directions

Several participants (P02, P05, P08, P22, P24, P29, P31, P41, P42, and P51) became extremely relaxed during the relaxation phase due to our affective touch-based assistance, leading them to either doze off or enter a state of mental blankness. This led them to believe that our technology could be used as a sleep aid. This effect might be attributed to the

inherently gentle nature of air-based haptics or possibly because our haptic simulation evokes the comforting experiences of being soothed by caregivers in childhood (as mentioned by many users). We observed a study on using auditory vibrations to aid sleep before bedtime [79], and future research could explore the potential application of using our technology in this context.

Users also expressed a desire to integrate haptics with other modalities. Previous research has suggested that the synchronous integration of haptics with other modalities can better convey emotional experiences [48]. In the experiment, many users mentioned that they preferred to use visual or auditory methods to decompress and relax. Research has shown that combining touch with other modalities, such as sound, can more accurately convey the overall characteristics of the tactile experience [80]. The study and design of haptics in conjunction with other modalities are thus expected to become increasingly important.

Our intermediary stroking simulation has users recall similar relaxation experiences in many other scenarios, leading users to believe that our technology could be effectively combined with existing relaxation and massage devices, or embedded within spa facilities or space for meditation.

Some participants (P05, P11, P17, P29, P34) felt that our rendered touch technology created a sense of interacting with a real person or a robot. Mediated remote emotion communication [81] and emotional haptic interaction with robots [82] are crucial future development areas that merit dedicated exploration and research. What is more, our haptic technology could be taken as haptic feedback and integrated into other technical research programs, such as robot control [83]. It would help facilitate robot design and assist in understanding users' perceptions about interactions with robots.

Regarding the design direction of the user interface or form factors, we see four possibilities: (1) a lightweight design to meet the needs of wearable devices, particularly integrating with traditional clothing items such as hats, scarves, and garments; (2) integration with existing massage instruments, which can be either wearable massage devices like neck massagers or embedded in larger massage equipment such as massage chairs; (3) development of embedded devices that integrate into living spaces, combining with sofas and beds, embedding into workplaces with desks and chairs, or integrating into automotive environments with car dashboards or seats; and (4) combining haptic rendering technology with robotics could enhance robots' ability to engage in social interactions with humans.

4.3. Limitation and Future Work

Firstly, we only used mid-air haptics to simulate a stroking gesture. Further research is needed to determine whether other gestures, such as hugging or tapping, might yield similar or even better stress regulation effects. Secondly, due to time constraints, we were unable to recruit more participants and thus did not compare our approach with other haptic feedback technologies, constraining the understanding of this method's performance. Future research should explore the different impacts of various haptic technologies. Thirdly, some participants mentioned that the temperature of a hand stroking might not always be gentle or warm. Further studies are required to investigate how variations in temperature and the rate of temperature change, as well as differences in the pressure of touch, affect stress regulation. Fourthly, according to the limits mentioned in Section 4.1, we should carefully design our research. Further studies could use other stress-induction methods like TSST to induce social stress [47] or be conducted in the wild based on different stress situations. It is better to conduct longer research and avoid the bias or preference caused by participants for using the novel device. We want to fully test whether our approach could help people or not in stress regulation.

5. Conclusions

In this paper, we designed stroking sensations by using Peltier elements to generate warm air and blow it by an array of fans. Computer-mediated stroking is subtle, comfortable, gradual in nature, noiseless, and private, giving participants comfortable and pleasant sensations, evoking participants positive experiences, and distracting attention from stressful work. The experiment showed that participants in the SG reduced anxiety and felt more relaxed after stress induction by using our haptic device. This novel technology shows the potential for stress regulation and needs to be developed further on detailed parameters, such as intensity and velocity of airflow, temperature variations, and the rate of temperature change, according to specific situations and personal requirements.

This work presents an innovative approach that combines mid-air and thermal feedback technologies, opening up new design possibilities within haptic sub-modalities. It also delves into the design insights of affective stroking, contributing to the expansion of the design landscape for stress regulation technologies. Meanwhile, this work outlines possible design solutions for real-world applications, which will guide future research and development in mid-air and thermal tactile technologies. At last, all discoveries are contributing to the advancement of studies in the field of digital health, promoting people's mental health and well-being.

Author Contributions: Conceptualization, S.H. and H.Z.; methodology, S.H. and M.X.; software, S.H.; validation, S.H. and H.Z.; formal analysis, S.H. and H.Z.; investigation, S.H. and H.Z.; resources, M.X. and G.H.; data curation, S.H. and M.X.; writing—original draft preparation, S.H. and H.Z.; writing—review and editing, S.H., M.X. and C.Y.; visualization, S.H. and H.Z.; supervision, G.H., C.Y. and F.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Funds for the Central Universities (Grant No. 226-2024-00164); "Leading Goose" R&D Program of Zhejiang (Project No. 2023C01216), Zhejiang Provincial Natural Science Foundation of China under Grant No. LR24F020001, Research Center of Computer Aided Product Innovation Design, Ministry of Education.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of Macau University of Science and Technology (protocol code: MUST-20230925001, 12 June 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to thank Yan Liu, Fang He, Fang Tong, Yujie Zhou, Ming Zhang, Chunxiao Huo, Yongjun He, Qian Wang, Gang Meng, Xiaoxian Wang, Fan Wang, Tian Qin, Fuye Zhang, Xi Yu, Hao Jiang, and Xiaoyou He for contributing to various aspects of the study and the experiment. We sincerely appreciate all the volunteers who dedicated their time to participating in the experiment. We also thank the reviewers for their valuable feedback and suggestions, which have greatly improved the quality of our paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Questions of User Interview

Q1: How do you feel about this stroking sensation? (Does it feel like a gentle stroking? How does the temperature feel? How comfortable is the sensation?)
Q2: How do you interpret this stroking and the emotions it conveys?
Q3: What kind of experience does this stroking provide?
Q4: Does the stroking evoke positive and pleasant feelings?
Q5: Does the stroking help to provide emotional comfort?
Q6: What is your personal stress regulation method during the relaxation phase?
Q7: If the prototype were further developed, would you be inclined to use this feature to assist with stress regulation in the future?
Q8: What other scenarios do you think this feature could be useful for?

References

1. Zepf, S.; El Haouij, N.; Lee, J.; Ghandeharioun, A.; Hernandez, J.; Picard, R.W. Studying Personalized Just-in-Time Auditory Breathing Guides and Potential Safety Implications during Simulated Driving. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2020), Genoa, Italy, 12–18 July 2020; pp. 275–283.
2. Xue, M.; Liang, R.-H.; Yu, B.; Funk, M.; Hu, J.; Feijs, L.M.G. AffectiveWall: Designing Collective Stress-Related Physiological Data Visualization for Reflection. *IEEE Access* **2019**, *7*, 131289–131303. [CrossRef]
3. Yu, B.; Hu, J.; Funk, M.; Liang, R.-H.; Xue, M.; Feijs, L.M.G. RESonance: Lightweight, Room-Scale Audio-Visual Biofeedback for Immersive Relaxation Training. *IEEE Access* **2018**, *6*, 38336–38347. [CrossRef]
4. Rudics, E.; Nagy, Á.; Dombi, J.; Hompoth, E.A.; Szabó, Z.; Horváth, R.; Balogh, M.; Lovas, A.; Bilicki, V.; Szendi, I. Photoplethysmograph Based Biofeedback for Stress Reduction under Real-Life Conditions in Healthcare Frontline. *Appl. Sci.* **2023**, *13*, 835. [CrossRef]
5. Vidyarthi, J.; Riecke, B.E.; Gromala, D. Sonic Cradle: Designing for an Immersive Experience of Meditation by Connecting Respiration to Music. In Proceedings of the Designing Interactive Systems Conference (DIS '12), Newcastle Upon Tyne, UK, 11–15 June 2012; pp. 408–417.
6. Leslie, G.F.; Ghandeharioun, A.; Zhou, D.; Picard, R.W. Engineering Music to Slow Breathing and Invite Relaxed Physiology. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3–6 September 2019; pp. 1–7.
7. Miri, P.; Flory, R.; Uusberg, A.; Culbertson, H.; Harvey, R.H.; Kelman, A.; Peper, E.; Gross, J.J.; Isbister, K.; Marzullo, K. PIV: Placement, Pattern, and Personalization of an Inconspicuous Vibrotactile Breathing Pacer. *ACM Trans. Comput.-Hum. Interact.* **2020**, *27*, 1–44. [CrossRef]
8. Yu, B.; Feijs, L.; Funk, M.; Hu, J. Breathe with Touch: A Tactile Interface for Breathing Assistance System. In Proceedings of the Human-Computer Interaction—INTERACT 2015, Bamberg, Germany, 14–18 September 2015; pp. 45–52.
9. Aslan, I.; Burkhardt, H.; Kraus, J.; André, E. Hold My Heart and Breathe with Me: Tangible Somaesthetic Designs. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI 2016), Gothenburg, Sweden, 23–27 October 2016; Article No. 92. pp. 1–6.
10. Azevedo, R.T.; Bennett, N.; Bilicki, A.; Hooper, J.; Markopoulou, F.; Tsakiris, M. The Calming Effect of a New Wearable Device during the Anticipation of Public Speech. *Sci. Rep.* **2017**, *7*, 2285. [CrossRef]
11. Costa, J.; Guimbretière, F.; Jung, M.F.; Choudhury, T. BoostMeUp: Improving Cognitive Performance in the Moment by Unobtrusively Regulating Emotions with a Smartwatch. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 1–23. [CrossRef]
12. Choi, K.Y.; Ishii, H. ambienBeat: Wrist-Worn Mobile Tactile Biofeedback for Heart Rate Rhythmic Regulation. In Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '20), Sydney, Australia, 9–12 February 2020; pp. 17–30.
13. Costa, J.; Adams, A.T.; Jung, M.F.; Guimbretière, F.; Choudhury, T. EmotionCheck: Leveraging Bodily Signals and False Feedback to Regulate Our Emotions. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September 2016; pp. 758–769.
14. Zhou, Y.; Murata, A.; Watanabe, J. The Calming Effect of Heartbeat Vibration. In Proceedings of the 2020 IEEE Haptics Symposium (HAPTICS), Crystal City, VA, USA, 24–27 March 2020; pp. 677–683.
15. Haynes, A.C.; Lywood, A.; Crowe, E.M.; Fielding, J.L.; Rossiter, J.M.; Kent, C. A Calming Hug: Design and Validation of a Tactile Aid to Ease Anxiety. *PLoS ONE* **2022**, *17*, e0259838. [CrossRef]
16. van Erp, J.B.F.; Toet, A. Social Touch in Human–Computer Interaction. *Front. Digit. Humanit.* **2015**, *2*, 2. [CrossRef]
17. Jablonski, N.G. Social and Affective Touch in Primates and Its Role in the Evolution of Social Cohesion. *Neuroscience* **2021**, *464*, 117–125. [CrossRef]
18. Holt-Lunstad, J.; Birmingham, W.A.; Light, K.C. Influence of a "Warm Touch" Support Enhancement Intervention Among Married Couples on Ambulatory Blood Pressure, Oxytocin, Alpha Amylase, and Cortisol. *Psychosom. Med.* **2008**, *70*, 976–985. [CrossRef]
19. Field, T.; Hernandez-Reif, M.; Diego, M.; Schanberg, S.; Kuhn, C. Cortisol Decreases and Serotonin and Dopamine Increase Following Massage Therapy. *Int. J. Neurosci.* **2005**, *115*, 1397–1413. [CrossRef] [PubMed]
20. Triscoli, C.; Croy, I.; Olausson, H.; Sailer, U. Touch Between Romantic Partners: Being Stroked Is More Pleasant Than Stroking and Decelerates Heart Rate. *Physiol. Behav.* **2017**, *177*, 169–175. [CrossRef] [PubMed]
21. Grewen, K.M.; Girdler, S.S.; Amico, J.; Light, K.C. Effects of Partner Support on Resting Oxytocin, Cortisol, Norepinephrine, and Blood Pressure Before and After Warm Partner Contact. *Psychosom. Med.* **2005**, *67*, 531–538. [CrossRef] [PubMed]
22. Triscoli, C.; Croy, I.; Steudte-Schmiedgen, S.; Olausson, H.; Sailer, U. Heart Rate Variability Is Enhanced by Long-Lasting Pleasant Touch at CT-Optimized Velocity. *Biol. Psychol.* **2017**, *128*, 71–81. [CrossRef]
23. Lindgren, L.; Rundgren, S.; Winsö, O.; Lehtipalo, S.; Wiklund, U.; Karlsson, M.; Stenlund, H.; Jacobsson, C.; Brulin, C. Physiological Responses to Touch Massage in Healthy Volunteers. *Auton. Neurosci.* **2010**, *158*, 105–110. [CrossRef]
24. Morrison, I. Keep Calm and Cuddle On: Social Touch as a Stress Buffer. *Adapt. Hum. Behav. Physiol.* **2016**, *2*, 344–362. [CrossRef]
25. Parker, R.A. The Effects of Physical Touch and Thermal Warmth on Interpersonal Trust. ProQuest Thesis, Indiana University of Pennsylvania, Indiana, PA, USA, 2011.

26. Umair, M.; Sas, C.; Latif, M.H. Towards Affective Chronometry: Exploring Smart Materials and Actuators for Real-Time Representations of Changes in Arousal. In Proceedings of the 2019 Designing Interactive Systems Conference (DIS '19), San Diego, CA, USA, 23–28 June 2019; pp. 1479–1494.
27. Kidd, T.; Devine, S.L.; Walker, S.C. Affective Touch and Regulation of Stress Responses. *Health Psychol. Rev.* **2023**, *17*, 60–77. [CrossRef]
28. Mazza, A.; Cariola, M.; Capiotto, F.; Diano, M.; Schintu, S.; Pia, L.; Dal Monte, O. Hedonic and Autonomic Responses in Promoting Affective Touch. *Sci. Rep.* **2023**, *13*, 11201. [CrossRef]
29. Lim, K.Y.; Hong, W. Neural Mechanisms of Comforting Prosocial Touch and Stress Buffering. *Horm. Behav.* **2023**, *153*, 105391. [CrossRef]
30. Gross, J.J. The Emerging Field of Stress regulation: An Integrative Review. *Rev. Gen. Psychol.* **1998**, *2*, 271. [CrossRef]
31. McGlone, F.; Vallbo, A.B.; Olausson, H.; Loken, L.; Wessberg, J. Discriminative Touch and Emotional Touch. *Can. J. Exp. Psychol.* **2007**, *61*, 173–183. [CrossRef] [PubMed]
32. Sailer, U. How Sensory and Affective Attributes Describe Touch Targeting C-Tactile Fibers. *Exp. Psychol.* **2020**, *67*, 224–236. [CrossRef] [PubMed]
33. McGlone, F.; Reilly, D. The Cutaneous Sensory System. *Neurosci. Biobehav. Rev.* **2010**, *34*, 148–159. [CrossRef] [PubMed]
34. Sailer, U.; Ackerley, R. Exposure Shapes the Perception of Affective Touch. *Dev. Cogn. Neurosci.* **2019**, *35*, 109–114. [CrossRef] [PubMed]
35. Greco, A.; Guidi, A.; Bianchi, M.; Lanatà, A.; Valenza, G.; Scilingo, E.P. Brain Dynamics Induced by Pleasant/Unpleasant Tactile Stimuli Conveyed by Different Fabrics. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 2417–2427. [CrossRef]
36. Loken, L.S.; Wessberg, J.; McGlone, F.; Olausson, H. Coding of Pleasant Touch by Unmyelinated Afferents in Humans. *Nature Neurosci.* **2009**, *12*, 547–548. [CrossRef]
37. Ackerley, R.; Wasling, H.B.; Liljencrantz, J.; Olausson, H.; Johnson, R.D.; Wessberg, J. Human C-Tactile Afferents Are Tuned to the Temperature of a Skin-Stroking Caress. *J. Neurosci.* **2014**, *34*, 2879–2883. [CrossRef]
38. Israr, A.; Poupyrev, I. Tactile Brush: Drawing on Skin with a Tactile Grid Display. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11), Vancouver, BC, Canada, 7–12 May 2011; pp. 2019–2028.
39. Huisman, G.; Darriba Frederiks, A.; Van Dijk, B.; Heylen, D.; Kröse, B. The TaSST: Tactile Sleeve for Social Touch. In Proceedings of the 2013 World Haptics Conference (WHC), Daejeon, Republic of Korea, 14–17 April 2013; pp. 211–216.
40. Huisman, G.; Darriba Frederiks, A.; van Erp, J.B.F.; Heylen, D.K.J. Simulating Affective Touch: Using a Vibrotactile Array to Generate Pleasant Stroking Sensations. In Proceedings of the EuroHaptics 2016, London, UK, 4–7 July 2016; pp. 240–250.
41. Raisamo, J.; Raisamo, R.; Surakka, V. Comparison of Saltation, Amplitude Modulation, and a Hybrid Method of Vibrotactile Stimulation. *IEEE Trans. Haptics* **2013**, *6*, 517–521. [CrossRef]
42. Knoop, E.; Rossiter, J. The Tickler: A Compliant Wearable Tactile Display for Stroking and Tickling. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15), Seoul, Republic of Korea, 18–23 April 2015; pp. 1133–1138.
43. Muthukumarana, S.; Elvitigala, D.S.; Forero Cortes, J.P.; Matthies, D.J.C.; Nanayakkara, S. Touch me Gently: Recreating the Perception of Touch using a Shape-Memory Alloy Matrix. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20), Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
44. Culbertson, H.; Nunez, C.M.; Israr, A.; Lau, F.; Abnousi, F.; Okamura, A.M. A Social Haptic Device to Create Continuous Lateral Motion Using Sequential Normal Indentation. In Proceedings of the 2018 IEEE Haptics Symposium (HAPTICS), San Francisco, CA, USA, 25–28 March 2018; pp. 32–39.
45. Kim, L.; Castillo, P.; Follmer, S.; Israr, A. VPS Tactile Display: Tactile Information Transfer of Vibration, Pressure, and Shear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 1–17. [CrossRef]
46. Wu, W.; Culbertson, H. Wearable Haptic Pneumatic Device for Creating the Illusion of Lateral Motion on the Arm. In Proceedings of the 2019 IEEE World Haptics Conference (WHC), Tokyo, Japan, 9–12 July 2019; pp. 193–198.
47. Zhao, Y.; Tao, Y.; Le, G.; Maki, R.; Adams, A.; Lopes, P.; Choudhury, T. Affective Touch as Immediate and Passive Wearable Intervention. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2022**, *6*, 1–23. [CrossRef]
48. Eid, M.A.; Al Osman, H. Affective Haptics: Current Research and Future Directions. *IEEE Access* **2015**, *4*, 26–40. [CrossRef]
49. Wei, Q.; Hu, J.; Li, M. Enhancing Social Messaging with Mediated Social Touch. *Int. J. Hum.-Comput. Interact.* **2022**, *40*, 1669–1688. [CrossRef]
50. Tsalamlal, M.Y.; Ouarti, N.; Martin, J.-C.; Ammi, M. Haptic Communication of Dimensions of Emotions Using Air Jet-Based Tactile Stimulation. *J. Multimodal User Interfaces* **2015**, *9*, 69–77. [CrossRef]
51. Sato, Y.; Ueoka, R. Investigating Haptic Perception of and Physiological Responses to Air Vortex Rings on a User's Cheek. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 3083–3094.
52. Obrist, M.; Subramanian, S.; Gatti, E.; Long, B.; Carter, T. Emotions Mediated Through Mid-Air Haptics. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 18–23 April 2015; pp. 2053–2062.

53. Chew, S.; Dalsgaard, T.-S.; Maunsbach, M.; Seifi, H.; Bergström, J.; Hornbæk, K.; Irisarri, J.; Ezcurdia, I.F.; Iriarte, N.; Marzo, A.; et al. TOUCHLESS: Demonstrations of Contactless Haptics for Affective Touch. In Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; Article No. 475. pp. 1–5.
54. Lan, R.; Sun, X.; Wang, Q.; Liu, B. Ultrasonic Mid-Air Haptics on the Face: Effects of Lateral Modulation Frequency and Amplitude on Users' Responses. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; Article No. 712; pp. 1–12.
55. Singhal, Y.; Wang, H.; Gil, H.; Kim, J.R. Mid-Air Thermo-Tactile Feedback using Ultrasound Haptic Display. In Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology, Osaka, Japan, 8–10 December 2021; Article No. 28. pp. 1–11.
56. Vess, M. Warm Thoughts: Attachment anxiety and sensitivity to temperature cues. *Psychol. Sci.* **2012**, *23*, 472–474. [CrossRef]
57. Fay, A.J.; Maner, J.K. Interactive Effects of Tactile Warmth and Ambient Temperature on the Search for Social Affiliation. *Soc. Psychol.* **2020**, *51*, 199–204. [CrossRef]
58. IJzerman, H.; Semin, G.R. The Thermometer of Social Relations. *Psychol. Sci.* **2009**, *20*, 1214–1220. [CrossRef]
59. Kang, Y.; Williams, L.E.; Clark, M.S.; Gray, J.R.; Bargh, J.A. Physical temperature effects on trust behavior: The role of insula. *Soc. Cogn. Affect. Neurosci.* **2011**, *6*, 507–515. [CrossRef]
60. Williams, L.E.; Bargh, J.A. Experiencing Physical Warmth Promotes Interpersonal Warmth. *Science* **2008**, *322*, 606–607. [CrossRef]
61. Mittelmann, B.; Wolff, H.G. Emotions and Skin Temperature: Observations on Patients During Psychotherapeutic (Psychoanalytic) Interviews. *Psychosom. Med.* **1943**, *5*, 211. [CrossRef]
62. Wilson, G.; Dobrev, D.; Brewster, S.A. Hot Under the Collar: Mapping Thermal Feedback to Dimensional Models of Emotion. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 4838–4849.
63. Tewell, J.; Bird, J.; Buchanan, G.R. The Heat is On: A Temperature Display for Conveying Affective Feedback. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 1756–1767.
64. Wilson, G.; Brewster, S.A. Multi-Moji: Combining Thermal, Vibrotactile & Visual Stimuli to Expand the Affective Range of Feedback. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 1743–1755.
65. Roquet, C.D.; Sas, C. Interoceptive Interaction: An Embodied Metaphor Inspired Approach to Designing for Meditation. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; Article No. 265. pp. 1–17.
66. Ståhl, A.; Jonsson, M.; Mercurio, J.; Karlsson, A.; Höök, K.; Banka Johnson, E.-C. The Soma Mat and Breathing Light. In Proceedings of the Extended Abstracts of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 305–308.
67. Williams, M.A.; Roseway, A.; O'Dowd, C.; Czerwinski, M.; Morris, M.R. SWARM: An Actuated Wearable for Mediating Affect. In Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction, Stanford, CA, USA, 15–19 January 2015; pp. 293–300.
68. Umair, M.; Sas, C.; Chalabianloo, N.; Ersoy, C. Exploring Personalized Vibrotactile and Thermal Patterns for Affect Regulation. In Proceedings of the 2021 ACM Designing Interactive Systems Conference, Virtual Event, 28 June–2 July 2021; pp. 891–906.
69. Liu, Y.; Nishikawa, S.; Seong, Y.A.; Niiyama, R.; Kuniyoshi, Y. ThermoCaress: A Wearable Haptic Device with Illusory Moving Thermal Stimulation. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; Article No. 214. pp. 1–12.
70. Yang, Y.; Wang, Y.; Easa, S.M.; Zheng, X. Analyzing Pedestrian Behavior at Unsignalized Crosswalks from the Drivers' Perspective: A Qualitative Study. *Appl. Sci.* **2022**, *12*, 4017. [CrossRef]
71. Spielberger, C.D.; Gonzalez-Reigosa, F.; Martinez-Urrutia, A.; Natalicio, L.F.; Natalicio, D.S. The State-Trait Anxiety Inventory. *Rev. Interam. Psicol./Interam. J. Psychol.* **1971**, *5*. [CrossRef]
72. Whited, A.; Larkin, K.T.; Whited, M. Effectiveness of emWave Biofeedback in Improving Heart Rate Variability Reactivity to and Recovery from Stress. *Appl. Psychophysiol. Biofeedback* **2014**, *39*, 75–88. [CrossRef] [PubMed]
73. Young, H.; Benton, D. We Should Be Using Nonlinear Indices When Relating Heart-Rate Dynamics to Cognition and Mood. *Sci. Rep.* **2015**, *5*, 16619. [CrossRef] [PubMed]
74. Shaffer, F.; Ginsberg, J.P. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health* **2017**, *5*, 258. [CrossRef] [PubMed]
75. Shaffer, F.; McCraty, R.; Zerr, C.L. A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability. *Front. Psychol.* **2014**, *5*, 1040. [CrossRef]
76. Choi, M.-H.; Kang, K.-Y.; Lee, T.-H.; Choi, J.-S. Correlations between SSQ Scores and ECG Data during Virtual Reality Walking by Display Type. *Appl. Sci.* **2024**, *14*, 2123. [CrossRef]
77. Kantor, J.; Vilímek, Z.; Vítězník, M.; Smrčka, P.; Campbell, E.A.; Bucharová, M.; Grohmannová, J.; Špinarová, G.; Janíčková, K.; Du, J.; et al. Effect of Low Frequency Sound Vibration on Acute Stress Response in University Students—Pilot Randomized Controlled Trial. *Front. Psychol.* **2022**, *13*, 980756. [CrossRef]
78. Bryman, A. *Social Research Methods*; Oxford University Press: Oxford, UK, 2016.

79. Zhang, H.; Zheng, R.; Yang, S.; Wei, W.; Shan, H.; Zhang, J. Waves Push Me to Slumberland: Reducing Pre-Sleep Stress through Spatio-Temporal Tactile Displaying of Music. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; Article No. 995. pp. 1–15.
80. Zhang, Z.; Alvina, J.; Héron, R.; Safin, S.; Détienne, F.; Lecolinet, E. Touch without Touching: Overcoming Social Distancing in Semi-Intimate Relationships with SansTouch. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; Article No. 651. pp. 1–13.
81. Raisamo, R.; Salminen, K.; Rantala, J.; Farooq, A.; Ziat, M. Interpersonal Haptic Communication: Review and Directions for the Future. *Int. J. Hum.-Comput. Stud.* **2022**, *166*, 102881. [CrossRef]
82. Okuda, M.; Takahashi, Y.; Tsuichihara, S. Human Response to Humanoid Robot That Responds to Social Touch. *Appl. Sci.* **2022**, *12*, 9193. [CrossRef]
83. Kim, H.; Miyakoshi, M.; Kim, Y.; Stapornchaisit, S.; Yoshimura, N.; Koike, Y. Electroencephalography Reflects User Satisfaction in Controlling Robot Hand through Electromyographic Signals. *Sensors* **2023**, *23*, 277. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

applied sciences

MDPI

Review

Uncovering Research Trends on Artificial Intelligence Risk Assessment in Businesses: A State-of-the-Art Perspective Using Bibliometric Analysis

Juan Carlos Muria-Tarazón [1,*], Juan Vicente Oltra-Gutiérrez [1], Raúl Oltra-Badenes [1] and Santiago Escobar-Román [2]

[1] Departamento de Organización de Empresas, Universitat Politècnica de València, 46022 Valencia, Spain; jvoltra@omp.upv.es (J.V.O.-G.); rauloltra@doe.upv.es (R.O.-B.)
[2] Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València, 46022 Valencia, Spain; sescobar@upv.es
* Correspondence: jmuria@doe.upv.es

Abstract: This paper presents a quantitative vision of the study of artificial intelligence risk assessment in business based on a bibliometric analysis of the most relevant publications. The main goal is to determine whether the risk assessment of artificial intelligence systems used in businesses is really a subject of increasing interest and to identify the most influential and productive sources of scientific research in this area. Data were collected from the Web of Science Core Collection, one of the most complete and prestigious databases. Regarding the temporal evolution of publications and citations this study evidences, this research subject shows rapid growth in the number of publications (at a compound annual rate of 31.20% from 2018 to 2024 inclusive), showing its high attraction for researchers, responding to the need to implement systematic risk assessment processes in the organizations using AI to mitigate potential harms, ensure compliance with regulations, and enhance artificial intelligence systems' trust and adoption. Especially after the surge of large language models like ChatGPT or Gemini, AI is revolutionizing the dynamics of human–computer interaction using natural language, video, and audio. However, as the scientific community initiates rigorous studies on AI risk assessment within organizational contexts, it is imperative to consider critical issues such as data privacy, ethics, bias, and hallucinations to ensure the successful integration and interaction of AI systems with human operators. Furthermore, this paper constitutes a starting point, including for any researcher who wants to be introduced to this topic, indicating new challenges that should be dealt by researchers interested in AI and hot topics, in addition to the most relevant literature, authors, and journals about this research subject.

Keywords: artificial intelligence; bibliometrics; risk assessment; risk; impact assessment; AI

1. Introduction

Artificial intelligence (AI) is a highly used concept in business nowadays, and is sometimes even used as a synonym for digital transformation. In many cases, managers and other decision-makers do not understand the pros and cons of AI initiatives inside their organizations.

The confidentiality and privacy of the corporate or personal data provided to the AI systems is sometimes assumed and not guaranteed in many cases [1].

Also, there are concerns about the ethical aspects of AI implementation [2,3], especially in some fields like healthcare [4]. That is the reason we can find recent and interesting works on explainable (or interpretable) AI [5,6].

On the other hand, AI is a promising technology, and businesses should see how this technology allows them to compete in a better way, in terms of quality improvement or cost reduction.

This paper aims to identify and study the state-of-the-art research on the assessment of these risks introduced by artificial intelligence systems on businesses because this could be an emerging theme with interesting development opportunities.

For this, we will use these research hypotheses:

1. The state-of-the-art developments in this research subject are characterized by a concentration of highly cited publications, predominantly authored by leading researchers from top-tier institutions and published in high-impact journals.
2. The number of publications on AI risk assessment in businesses has significantly increased over the last seven years, reflecting a growing global emphasis on cybersecurity and explainable AI.
3. Collaborative research involving multiple institutions is undertaken frequently and generates a scientific impact on the research subject of AI risk assessment in businesses.

Just to clarify what we are talking about, risk assessment is defined in the ISO 31000:2018 standard [7] as "the overall process of risk identification, risk analysis, and risk evaluation", so it is a more general concept than risk analysis or risk evaluation. Actually, these concepts are part of the assessment process.

This risk assessment should be done systematically, iteratively, and collaboratively, according to ISO, and so it should also be done in the AI context.

Besides that, official references from the government about risk assessment are present in the European AI Regulation, which was recently published [8], or in the NIST AI risk management framework [9].

This allows businesses to apply a risk–benefit analysis in their decisions concerning AI investments, in case they are both acquiring or developing an AI system.

Our study demonstrates that this is a very novel concept in research, although it has increased its relevance in synchrony with the explosion of research production in artificial intelligence.

This paper aims to assess the relevance of the research subject of risk assessment of artificial intelligence initiatives in businesses, to be useful as a guide for other researchers interested in this topic, indicating the most relevant references, listing the most relevant researchers and institutions, identifying the journals with a specific interest in this topic, and finally, signaling the ability of this research subject to attract funds for research.

This paper is distributed in the following way: The first section presents an introduction to the research subject and to the bibliometric analysis, explaining why and how we have decided to use these methods and indicating relevant works on bibliometric analysis in AI and the differences between them.

The next section presents the methodology we have decided to use for the study, including the search strategy, the methods of analysis, and the tools used.

The following section analyzes the results of our research, highlighting the most important findings in the most productive and influential journals, the most influential articles, the most prolific and influential authors, institutions, and countries, and the collaboration between countries. Then, we identify the landscapes and evolution of our research subject and the emerging related topics.

Finally, this paper presents the conclusions of this research and the bibliography used.

2. Literature Review

According to Noyons et al. [10], Gil-Gomez et al. [11], and Ninkov et al. [12], bibliometrics represents a statistically standard way to analyze scientific performance and dynamics in a specific research subject, based on a quantitative analysis of bibliographic data of academic publishing.

Bibliometrics is, according to them, a useful way to determine if the interest of the subject is expanding or reducing in a specific period of time and to identify the most relevant authors, publications, and institutions generating knowledge on the topic.

Nevertheless, bibliometrics has been used in this way in an important number of research subjects. Just looking at management research areas, we can find relevant works such as green supply management [13], operations research [14], sustainable manufacturing [15], entrepreneurship [16], or crowdfunding [11].

Besides that, according to Passas [17], while a meta-analysis is focused on quantitative analysis and a systematic literature review is focused on qualitative analysis, a bibliometric analysis can cover both sides, and that is the approach we used when defining this research. Additionally, given the size and the heterogeneity of our dataset, a systematic literature review was not realistic if we wanted to cover all the different approaches, and a meta-analysis was not feasible without excessively restricting the number of studies analyzed, given the mentioned heterogeneity and scarcity of quality studies.

By looking at new technologies (such as artificial intelligence), we can also find relevant bibliometrics works in Big Data Analytics [18,19], artificial intelligence in e-commerce [20], blockchain in logistics and supply chain [21], digital transformation [22], artificial intelligence in society [23], or even cybersecurity in healthcare [24].

In our preliminary search to verify that this paper would be relevant enough, we did not find previous bibliometric research on the business risk assessment of artificial intelligence. There are some works, such as the aforementioned Fosso Wamba et al. [23], and Zhang et al. [25] and Albahri and AlAmoodi [26], about the impact of artificial intelligence on society. The first two were focused on ethics in society, and the last one performed a bibliometric analysis on cybersecurity aspects of artificial intelligence, but none were specifically on the risk assessment of artificial intelligence in businesses.

Finally, and complementarily to our bibliometric analysis, we conducted additional manual searches using alternative platforms such as arXiv and ResearchGate.

ArXiv: Our search on arXiv yielded eight references, primarily communications to international conferences and preprints. Our reading of these references identified no leading journal or relevant paper to be included in our research.

ResearchGate: On ResearchGate, we discovered several journals with noteworthy publications, although they were not considered influential in our initial bibliometric analysis. For example, Novelli et al. published works in "*AI & Society*", a Q3 journal with an impact factor of 2.9, and in "*Digital Society*", which is not indexed in the Web of Science catalog with a high number of citations. These works [27,28] are coincident in proposing an alternative model for AI risk assessment based on real-world risk scenarios and frameworks developed previously for climate change, offering a different approach from the existing European AI regulation. Both works are highly cited according to ResearchGate. Additionally, the "*AGI—Artificial General Intelligence, Robotics, Safety & Alignment*" journal includes impactful articles such as "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence" by Slattery et al. [29], which proposes a living database with 777 risk entries based on various taxonomies, hosted at MIT "https://airisk.mit.edu (accessed on 10 January 2025)", one of the most comprehensive and dynamic repositories of potential AI risks for businesses, classified by both causes and by domains.

These findings suggest that some researchers publish their work on AI risk assessment in journals that may not be highly visible in mainstream repositories but offer valuable and innovative real-world approaches; so, we should take these into account, although it does not replace a bibliometric analysis at all.

Future research could explore the reasons behind this publication trend and the potential benefits of considering these alternative sources.

3. Materials and Methods

The bibliometric analysis in this paper is developed to identify the most active and influential research clusters, both in terms of authors and themes on the subject of artificial intelligence risk assessment in business.

Applying a bibliometric analysis can avoid the subjectivity of a literature review and comprehensively and objectively reflect the status of a research subject.

The analysis is based on data collected from the Web of Science Core Collection (WoSCC) database, a database with a specially high reputation as a scientific repository [30,31]. These data made it possible to analyze the number of publications and number of citations, in terms of affiliation, authors, or countries, and elaborate the h-Index, co-occurrence network, and thematic or strategic map, among other techniques, to set a high-resolution picture of the current situation of this research area.

For this research, we excluded Google Scholar due to its inclusion of non-peer-reviewed publications and the lack of a mechanism to export the entire dataset for further analysis. This limitation forces the manual selection of each publication before exporting, which is impractical for a comprehensive bibliometric analysis.

Additionally, the differences between Scopus' subject areas and the Web of Science Core Collection (WoSCC) research areas' classification posed challenges for direct comparison. For instance, equating Scopus' "Business, Management and Accounting" with WoSCC's "Business Economics" is highly subjective and debatable. Ultimately, WoSCC provided a slightly larger dataset than Scopus, and WoSCC's inclusion of high-quality, peer-reviewed journals ensures reliable and representative data, consequently enhancing the validity of the results presented.

To develop a bibliometric study, this paper follows the steps indicated in Figure 1:

1. The choice of the subject ("artificial intelligence risk assessment on business"), the time span (we chose two periods; the first one includes all production before 2018, and the second one includes the period between 2018 and 2024, both inclusive), and keywords ("risk assessment" AND "business" AND ("artificial intelligence" OR "machine learning")).
2. The choice of Web of Science Core Collection as the database to use responds to productivity criteria (92 million records since the year 1900 with more than 22,000 peer-reviewed journals) and influence criteria (2.2 billion cited references) [32]
3. In order to reduce the noise generated by the results, this paper excluded retracted publications and 2025 production (to ensure a fair comparison of annual production); and it included only articles and review articles of the following research areas: Computer Science, Engineering, Business Economics, Operations Research Management Science, Mathematics, Science Technology Other Topics, Telecommunications, Government Law, Public Administration, Social Sciences Other Topics, Information Science Library Science, Sociology, Mathematical Methods In Social Sciences, Instruments Instrumentation, Remote Sensing, and Medical Informatics.

4. Our research obtained 244 publications that met the chosen criteria (227 of them between 2018 and 2024). Of these, 18 publications (7.4%) were reviews and 226 (92.6%) were articles.
5. According to the methodology described, this paper classifies the results according to the performance analysis and the science mapping.
6. Finally, this paper discusses and presents the conclusions, identifying the most productive and influential authors, sources, articles, and countries and the most relevant relationships between them.

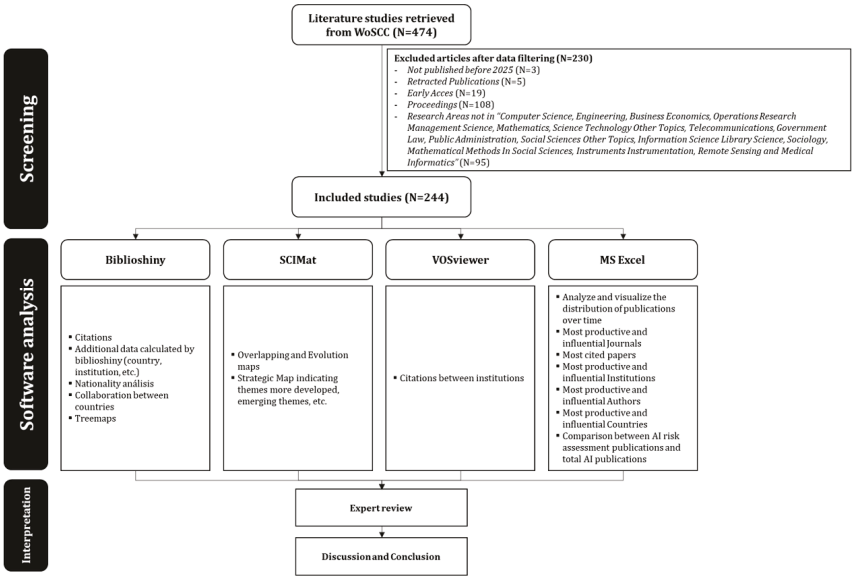


Figure 1. Flowchart of the bibliometric research strategy.

Method of Analysis

According to Noyons et al. [10], there are two main procedures in bibliometric analysis: performance analysis and science mapping.

While the former aims to evaluate the impact of the activity of authors, institutions, countries, etc., the second is focused on displaying the structure and dynamics of the research on one specific area, according to Cobo et al. [33].

A performance analysis is usually composed of the number of publications, the number of citations, and an index based on both numbers, the h-Index [34]. Although many authors have reviewed the h-Index and proposed alternatives ([35–38], among others), this index offers a high value and is easy to obtain, so it is used very frequently in bibliometric performance analysis.

We discarded the use of the g-Index in our bibliographic analysis for several reasons: we saw few differences in using this index compared with the use of the h-Index (no new entrants nor items exiting from the tables). This can be due to the generally low number of citations [39], as this topic is very new, and the g-Index is not an official number given by WoS. Thus, the process of calculations could introduce undesired errors in our analysis due to authors using different names, the absence of ORCID ID in some cases, etc.

Based on the previous considerations, this paper shows the journals ordered according to the h-Index computed based on the author's number of existing publications and citations in the WoSCC dataset (h); second, by the number of citations (TC); and third, by total scientific production in AI (TPAI).

In the case of the most productive and influential journals, this paper combines these indexes with the impact factor (IF) for 2023 (the last year published at the time we wrote

this paper), and the last 5-year impact factor (5-IF) to provide a complete picture of each included source.

Regarding the science mapping techniques, these are aimed at monitoring a research field to define its cognitive structure, that is, the structure of relationships between keywords, authors, etc., and its evolution in specific periods [10].

For the selected period, we analyzed all results from the Web of Science Core Collection (WoSCC) spanning from 1997 to 2024, the time at which this analysis was conducted. Upon examining the evolution of published material, we observed a significant increase starting in 2018. Furthermore, we chose to specifically analyze the period between the substantial rise of large language model-based chatbots (2021) and the last year analyzed. Consequently, we divided the dataset into two distinct periods (pre-2018 and post-2018) and conducted a detailed examination of the years 2021–2024 for a specific inquiry.

This research used a suite of software tools for comprehensive quantitative analysis and visualization of the collected literature. These tools included R-version 4.4.1 combined with Bibliometrix and Biblioshiny [40], VOSviewer 1.6.20, SCIMat 1.1.06 [33], and Microsoft Excel for Microsoft 365 Version 2410 Build 16.0.18129.20158.

This paper presents and discusses different charts within this science mapping, like the coupling maps chart, the co-occurrence networks chart, and the thematic evolution chart.

4. Results

The results are presented in the following order: the most productive and influential journals are shown and discussed first; then we analyze the evolution of published articles on the research subject we selected; and after that, this paper discusses the most prolific and influential authors, and the most productive and influential institutions. Then a country analysis is performed, and finally, after presenting the landscape and evolution, and the current emerging issues of our research subject, we discuss the general results.

4.1. Publishing Journals

After analyzing the 244 journals returned by the previously described query to the Web of Science Core Collection, Table 1 was prepared to show the publication trends of each of the top 25 journals:

In Table 2, if we look just at the last four years of our dataset (2021–2024), a period chosen for the appearance of relevant outcomes on large language models (LLMs) from OpenAI (GPT-3), Google (BERT), Facebook (preparing OPT), and Microsoft (Azure OpenAI service) that have enabled easier access to AI systems for the general public and social and economic actors, then we see how, except for the top 5, all the other positions change, showing differences in which journals are becoming more active on this topic (like *RISKS*, *RISK ANALYSIS* or *SENSORS*), and which ones are losing positions (like *ARTIFICIAL INTELLIGENCE REVIEW*, *STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT*, and *SCIENTIFIC REPORTS*).

Finally, if we zoom out and look at publishers, we identify the top 5 in Table 3.

We can see in Table 3 that MDPI is the most productive publisher between these years, with four journals having an h-Index greater than 1. Looking at the journals of this publisher, *SENSORS* is the most cited journal, and also contains MDPI's most cited article of the dataset with 40 citations.

Note that, due to the novelty of the research subject, the dataset is relatively small, and given the multidisciplinary nature of AI, encompassing aspects of computer science, business, ethics, and more, the combination of that results in a very heterogeneous set of journals.

Table 1. Publication trends by journal (top 25).

	1997–2017	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*TOTALS*
EXPERT SYSTEMS WITH APPLICATIONS	5	0	0	1	0	4	2	1	13
ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE	0	0	1	0	0	0	3	3	7
SUSTAINABILITY	0	0	1	0	0	0	4	2	7
IEEE ACCESS	0	1	0	0	2	2	0	2	7
APPLIED SOFT COMPUTING	0	1	0	0	2	0	2	1	6
ARTIFICIAL INTELLIGENCE REVIEW	1	0	1	0	0	0	3	1	6
STOCHASTIC ENVIRONMENTAL RESEARCH & RISK ASSESSMENT	0	0	1	0	1	1	1	1	5
SCIENTIFIC REPORTS	0	0	1	0	0	0	1	3	5
RISKS	0	0	0	0	2	1	1	1	5
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH	1	0	0	2	0	0	1	0	4
SAFETY SCIENCE	0	0	1	2	1	0	0	0	4
SENSORS	0	0	0	1	0	2	0	1	4
RISK ANALYSIS	0	0	0	0	0	1	1	2	4
REMOTE SENSING	0	0	0	1	1	0	1	0	3
INDUSTRIAL MANAGEMENT & DATA SYSTEMS	0	0	0	1	1	0	1	0	3
NANOTOXICOLOGY	0	0	1	1	0	0	0	0	2
JOURNAL OF CLEANER PRODUCTION	0	0	0	0	1	0	1	0	2
SOCIETY	0	0	0	0	2	0	0	0	2
RELIABILITY ENGINEERING & SYSTEM SAFETY	0	0	0	0	1	0	0	1	2
TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	0	0	0	1	0	0	1	0	2
COMPLEX & INTELLIGENT SYSTEMS	0	1	0	0	0	0	1	0	2
NEURAL COMPUTING & APPLICATIONS	0	0	0	0	1	1	0	0	2
PATTERNS	0	0	0	0	2	0	0	0	2
DECISION SUPPORT SYSTEMS	0	0	0	1	0	0	1	0	2
RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE	1	0	0	0	0	0	1	0	2
TOTAL PUBLICATIONS IN TOP 25 JOURNALS	8	3	7	11	17	12	26	19	103
TOTAL PUBLICATIONS IN DATASET	13	10	10	26	40	38	52	55	244

5 4 3 2 1 0 Publications

In Table 4, ordered first by h-Index, second by the total number of citations, and third by the total number of publications on this research subject, we can see clearly that the Journal "*EXPERT SYSTEMS WITH APPLICATIONS*" is the most influential one, according to its h-Index (11), owning the highest number of papers about the research subject (13) and the highest number of citations (661), and being second in terms of the ratio between citations and publications, with a value of 50.85. This journal is also one of the two unique members of the list with a paper having more than 100 citations, i.e., "Risk assessment in social lending via random forests". R denotes the rank, while Name refers to the journal or source title. The h-Index is represented by h, and TC signifies the total number of citations. TP stands for the total number of papers published. The ratio of total citations to total published papers is expressed as TC/TP. The number of papers with more than 100, 50, 25,

and 10 citations is indicated as >100, >50, >25, and >10, respectively. IF represents the 2023 impact factor, and 5-IF denotes the five-year impact factor.

Table 2. Publication trends by journal (top 25 between 2021 and 2024).

	2021	***2022***	***2023***	***2024***	***TOTALS***	***VAR (+/−)***
EXPERT SYSTEMS WITH APPLICATIONS	0	4	2	1	7	
ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE	0	0	3	3	6	
SUSTAINABILITY	0	0	4	2	6	
IEEE ACCESS	2	2	0	2	6	
APPLIED SOFT COMPUTING	2	0	2	1	5	
RISKS	2	1	1	1	5	+3
ARTIFICIAL INTELLIGENCE REVIEW	0	0	3	1	4	−1
STOCHASTIC ENVIRONMENTAL RESEARCH & RISK ASSESSMENT	1	1	1	1	4	−1
SCIENTIFIC REPORTS	0	0	1	3	4	−1
RISK ANALYSIS	0	1	1	2	4	+3
SENSORS	0	2	0	1	3	+1
JOURNAL OF CLEANER PRODUCTION	1	0	1	0	2	+5
SOCIETY	2	0	0	0	2	+5
RELIABILITY ENGINEERING & SYSTEM SAFETY	1	0	0	1	2	+5
NEURAL COMPUTING & APPLICATIONS	1	1	0	0	2	+7
REMOTE SENSING	1	0	1	0	2	−2
PATTERNS	2	0	0	0	2	+6
INDUSTRIAL MANAGEMENT & DATA SYSTEMS	1	0	1	0	2	−3
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH	0	0	1	0	1	−9
SAFETY SCIENCE	1	0	0	0	1	−9
TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	0	0	1	0	1	−1
COMPLEX & INTELLIGENT SYSTEMS	0	0	1	0	1	−1
DECISION SUPPORT SYSTEMS	0	0	1	0	1	+1
RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE	0	0	1	0	1	+1
NANOTOXICOLOGY	0	0	0	0	0	−9
TOTAL PUBLICATIONS IN TOP 25 JOURNALS	17	12	26	19	74	
TOTAL PUBLICATIONS IN DATASET	40	38	52	55	185	

6 5 4 3 2 1 0 Publications

Table 3. Publication trends by publisher (top 5 between 2021 and 2024).

	2021	**2022**	**2023**	**2024**	**TOTALS**	**Journal's h-Index > 1**
MDPI	6	5	8	12	31	4
ELSEVIER	8	2	6	10	26	1
SPRINGER	3	4	6	7	20	4
PERGAMON-ELSEVIER SCIENCE LTD	0	4	5	4	13	2
WILEY	1	3	3	2	9	1

The rest of the journals have five or fewer publications on this research subject.

If we look at the number of citations, the top five Journals are "*EXPERT SYSTEMS WITH APPLICATIONS*", "*APPLIED SOFT COMPUTING*", "*ARTIFICIAL INTELLIGENCE REVIEW*" (containing another paper with more than 100 citations, "Financial credit risk

assessment: a recent review"), "*EUROPEAN JOURNAL OF OPERATIONAL RESEARCH*", and "*SAFETY SCIENCE*", all with more than 100 citations.

Table 4. The 25 most productive and influential journals on artificial intelligence risk assessment in businesses.

R	Row Labels	h	TC	TP	TC/TP	>100	>50	>25	>10	IF (2023)	5-IF
1	*EXPERT SYSTEMS WITH APPLICATIONS*	11	661	13	50.85	1	4	1	5	7.5	7.6
2	*APPLIED SOFT COMPUTING*	5	156	4	39.00	0	1	1	0	7.2	7
3	*ARTIFICIAL INTELLIGENCE REVIEW*	4	198	4	49.50	1	0	2	1	10.7	11.7
4	*EUROPEAN JOURNAL OF OPERATIONAL RESEARCH*	4	130	3	43.33	0	1	1	2	6	5.9
5	*SAFETY SCIENCE*	4	129	3	43.00	0	0	3	0	4.7	5.3
6	*STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT*	4	78	5	15.60	0	1	0	0	3.9	3.6
7	*ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*	4	43	5	8.60	0	0	0	1	7.5	7.4
8	*SENSORS*	3	81	3	27.00	0	0	2	0	3.4	3.7
9	*SUSTAINABILITY*	3	39	5	7.80	0	0	0	2	3.3	3.6
10	*IEEE ACCESS*	3	29	5	5.80	0	0	0	1	3.4	3.7
11	*NANOTOXICOLOGY*	2	77	2	38.50	0	0	2	0	3.6	4.6
12	*SCIENTIFIC REPORTS*	2	68	5	13.60	0	1	0	0	3.8	4.3
13	*JOURNAL OF CLEANER PRODUCTION*	2	63	1	63.00	0	0	1	1	9.7	10.2
14	*SOCIETY*	2	52	2	26.00	0	0	1	1	1.4	0.9
15	*RELIABILITY ENGINEERING & SYSTEM SAFETY*	2	52	2	26.00	0	0	1	0	9.4	8.1
16	*TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE*	2	50	1	50.00	0	0	1	0	12.9	13
17	*COMPLEX & INTELLIGENT SYSTEMS*	2	49	1	49.00	0	0	1	1	5	5.2
18	*NEURAL COMPUTING & APPLICATIONS*	2	48	2	24.00	0	0	1	1	4.5	4.7
19	*REMOTE SENSING*	2	46	3	15.33	0	0	1	1	4.2	4.9
20	*PATTERNS*	2	37	2	18.50	0	0	0	1	6.7	6.6
21	*INDUSTRIAL MANAGEMENT & DATA SYSTEMS*	2	34	3	11.33	0	0	0	2	4.2	5.4
22	*DECISION SUPPORT SYSTEMS*	2	32	2	16.00	0	0	1	0	6.7	7.5
23	*RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE*	2	28	2	14.00	0	0	0	1	6.3	5.8
24	*RISKS*	2	20	4	5.00	0	0	0	0	2	1.7
25	*RISK ANALYSIS*	2	19	3	6.33	0	0	0	1	3	3.5

"*ARTIFICIAL INTELLIGENCE REVIEW*", with 198 citations, is the one with the second highest impact factor in the list (both in 2023, 10.7, and in the 5 last years, 11.7) of this top five, and "*SAFETY SCIENCE*", the one with a lower impact factor (4.7 and 5.3 for 2023 and for the 5 last years, respectively).

Based on these data, we can consider these top five journals to be the most productive and influential ones.

A special mention should be made regarding "*TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE*", in 16th position, "*JOURNAL OF CLEANER PRODUCTION*", ranked in 13th position, and "*ARTIFICIAL INTELLIGENCE REVIEW*", ranked third, as the three journals with the highest impact factor in the list. Note that this impact factor is not focused on the subject of our research, but it is general.

4.2. Evolution of Published Articles

As we could expect after the previous analysis, the number of published articles and citations is low compared to the bibliometric analysis conducted for other topics, showing that this topic has not been a main topic in scientific research until recent years.

A chart was elaborated to identify whether the research on this area was stable or increasing (or decreasing).

In Figure 2, we can see that the annual number of publications has been increasing since 2018, with a Compound Annual Growth Rate (CAGR) of 31.20% calculated from 2018 to 2024 inclusive.

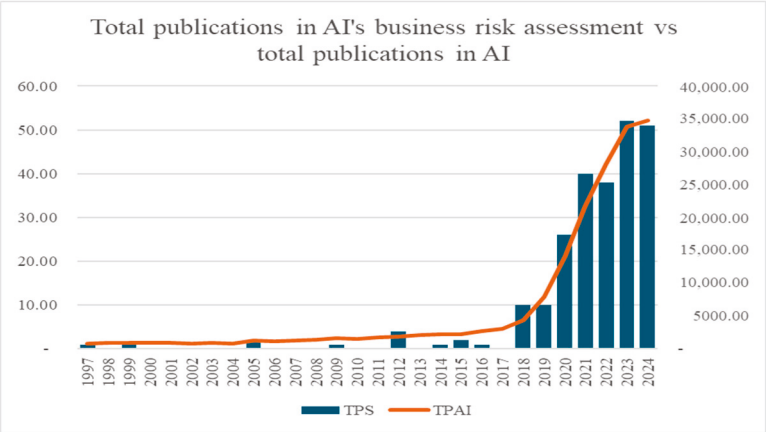


Figure 2. Number of annual publications according to WoS Core Collection query on artificial intelligence's business risk assessment (TPS) vs. annual publications in AI (TPAI).

We can benchmark this CAGR with the one from the results of a more general query "ARTIFICIAL INTELLIGENCE" in WoS Core Collection between 2018 and 2024 and with the other restrictions applied to make it comparable, i.e., 41.15%. So, a part of our research subject CAGR could be explained by the CAGR of research in the artificial intelligence subject, although we can take account of two additional considerations:

- While the CAGR of publications in the period between 1997 and 2017 of the search term "artificial intelligence" was 5.14%, the CAGR of publications in the same period of our search terms was 0%.
- On the other hand, analyzing the correlation between publications on "artificial intelligence" and publications analyzed in this paper, we can find a correlation of 0.99 with a p-value < 0.001, so we can consider it is very correlated.

In Figure 2 we can see the compared data:

In any case, we can consider that the interest in this topic has increased, especially since 2018. The first paper on this research subject was published in 1997 (42 years after the first mention of artificial intelligence, in 1955 [41]).

Note that the youth of this research topic, while justifying the importance of this paper's research, is why, in most of the tables, we can see lower productivity and influence measurements compared to other research topics with a longer life in scholar terms. That is the reason we built Table 5, allowing readers to compare the period between 2018 and 2024 with the whole dataset results.

Table 5 shows that 94.67% of the papers on this research topic were published in the period between 2018 and 2024, and 83.5% of the papers with more than 20 citations were published in 2018–2024 period as well.

This is the reason why, in specific parts of this paper, when appropriate, we made a distinction between these two periods.

Table 5. General citation structure on artificial intelligence risk assessment in businesses in WOS database.

Citations	All Time		2018–2024	
	Number of Papers	% Papers	Number of Papers	% Papers
>200 citations	3	1.230	1	0.433
>100 citations	2	0.820	1	0.433
>50 citations	13	5.328	10	4.329
>20 citations	41	16.803	37	16.017
≤20 citations	185	75.820	182	78.788
Total	244	100.000	231	100.000

4.3. The Most Influential Articles

The next step in the performance analysis is to identify the most influential articles on this subject so far. With this aim, Table 6 was built, ranking the 30 articles with the biggest number of citations, also calculating the citations per year in the whole period to include a relative indicator that allows for comparison of the influence of each article independently from the year it was published.

The most cited paper in the dataset we are analyzing was published by Wang, Yongqiao et al. [42], followed closely by papers by Malekipirbazari and Aksakalli [43] and Liao et al. [44].

These should, therefore, be considered the most influential studies in the business risk assessment of artificial intelligence.

On the other hand, if we look at the number of citations per year, the most important paper was published very recently, in 2022, and appears in fourth position in our list [45], another symptom of the youth of the research topic our paper is analyzing. The second paper in the top 30 was the aforementioned paper by Liao et al., and the third was by Malekipirbazari and Aksakalli [43], with 25.89 citations per year.

So, in relative terms of citations per year, the paper by Sun et al. should be considered the most influential paper, while in absolute terms of the total number of citations, the Wang, Yongqiao, et al. 2005 paper should be the most influential paper.

Additionally, Table 6 shows that five of the most cited papers correspond with the most influential journal in Table 4, "*EXPERT SYSTEMS WITH APPLICATIONS*".

The five journals that concentrate the majority of citations in Table 6 are, in this order, "*EXPERT SYSTEMS WITH APPLICATIONS*", "*IEEE TRANSACTIONS ON FUZZY SYSTEMS*", "*INTERNATIONAL JOURNAL OF FUZZY SYSTEMS*", "*ARTIFICIAL INTELLIGENCE REVIEW*", and "*APPLIED SOFT COMPUTING*". These thirty papers gather 2180 citations, 59.94% of the total citations in the dataset.

The three journals that concentrate the highest number of papers in Table 6 are, in this order, "*EXPERT SYSTEMS WITH APPLICATIONS*", "*SAFETY SCIENCE*", and "*APPLIED SOFT COMPUTING*", but the fact is that, actually, the thirty most influential papers of Table 6 are very scattered in up to twenty-four different journals.

Finally, if we look at the WoS h-Index of both corresponding authors, Sun has an h-Index of seven and Wang has an h-Index of one, again a symptom of the freshness of this research field (Malekipirbazari, the corresponding author of the second paper in the list, has an h-Index of ten). The exception is Liao, the corresponding author of the paper in third position, who has an h-Index of seventy-two.

Table 6. The 30 most cited papers on artificial intelligence risk assessment in businesses.

R	Journal	TC	Title	Author/s	Year	C/Y
1	*IEEE TRANSACTIONS ON FUZZY SYSTEMS*	247	A new fuzzy support vector machine to evaluate credit risk [42]	WANG YQ; WANG SY; LAI KK	2005	13.00
2	*EXPERT SYSTEMS WITH APPLICATIONS*	233	Risk assessment in social lending via random forests [43]	MALEKIPIRBAZARI M; AKSAKALLI V	2015	25.89
3	*INTERNATIONAL JOURNAL OF FUZZY SYSTEMS*	216	Hesitant fuzzy linguistic term set and its application in decision making: a state-of-the-art survey [44]	LIAO HC; XU ZS; HERRERA-VIEDMA E; HERRERA F	2018	36.00
4	*IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*	122	Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks [45]	SUN JK; ZHANG JB; LI QF; YI XW; LIANG YX; ZHENG Y	2022	61.00
5	*ARTIFICIAL INTELLIGENCE REVIEW*	106	Financial credit risk assessment: a recent review [46]	CHEN N; RIBEIRO B; CHEN A	2016	13.25
6	*EXPERT SYSTEMS WITH APPLICATIONS*	87	Improving returns on stock investment through neural network selection [47]	QUAH TS; SRINIVASAN B	1999	3.48
7	*EXPERT SYSTEMS WITH APPLICATIONS*	85	Exploring the behaviour of base classifiers in credit scoring ensembles [48]	MARQUÉS AI; GARCÍA V; SÁNCHEZ JS	2012	7.08
8	*APPLIED SOFT COMPUTING*	78	Dynamic ensemble classification for credit scoring using soft probability [49]	FENG XD; XIAO Z; ZHONG B; QIU J; DONG YX	2018	13.00
9	*EXPERT SYSTEMS WITH APPLICATIONS*	74	Two-level classifier ensembles for credit risk assessment [50]	MARQUÉS AI; GARCÍA V; SÁNCHEZ JS	2012	6.17
10	*COGNITIVE SYSTEMS RESEARCH*	72	Enterprise credit risk evaluation based on neural network algorithm [51]	HUANG XB; LIU XL; REN YQ	2018	12.00
11	*NANOMATERIALS*	63	Practices and trends of machine learning application in nanotoxicology [52]	FURXHI I; MURPHY F; MULLINS M; ARVANITIS A; POLAND CA	2020	15.75
12	*EUROPEAN JOURNAL OF OPERATIONAL RESEARCH*	63	Unsupervised quadratic surface support vector machine with application to credit risk assessment [53]	LUO J; YAN X; TIAN Y	2020	15.75
13	*EXPERT SYSTEMS WITH APPLICATIONS*	62	A novel tree-based dynamic heterogeneous ensemble method for credit scoring [54]	XIA YF; ZHAO JH; HE LY; LI YG; NIU MY	2020	15.50
14	*INTERNATIONAL JOURNAL OF MANAGING PROJECTS IN BUSINESS*	61	A review of artificial intelligence based risk assessment methods for capturing complexity-risk interdependencies cost overrun in construction projects [55]	AFZAL F; SHAO YF; NAZIR M; BHATTI SM	2021	20.33
15	*STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT*	58	Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model [56]	ADNAN RM; PETROSELLI A; HEDDAM S; SANTOS CAG; KISI O	2021	19.33

Table 6. *Cont.*

R	Journal	TC	Title	Author/s	Year	C/Y
16	*SCIENTIFIC REPORTS*	57	Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images [57]	VARGHESE B; CHEN F; HWANG D; PALMER SL; ABREU ALD; UKIMURA O; ARON M; ARON M; GILL I; DUDDALWAR V; PANDEY G	2019	11.40
17	*INTERNATIONAL JOURNAL OF MACHINE LEARNING AND CYBERNETICS*	57	Improved TODIM method for intuitionistic fuzzy MAGDM based on cumulative prospect theory and its application on stock investment selection [58]	ZHAO MW; WEI GW; WEI C; WU J	2021	19.00
18	*PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*	57	The hemispheric contrast in cloud microphysical properties constrains aerosol forcing [59]	MCCOY IL; MCCOY DT; WOOD R; REGAYRE L; WATSON-PARRIS D; GROSVENOR DP; MULCAHY AP; HU YX; BENDER FAM; FIELD PR; CARSLAW KS; GORDON H	2020	14.25
19	*APPLIED SOFT COMPUTING*	50	A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment [60]	LAPPAS PZ; YANNACOPOULOS AN	2021	16.67
20	*BUSINESS HORIZONS*	50	Cybersecurity: risk management framework and investment cost analysis [61]	LEE I	2021	16.67
21	*JOURNAL OF FORECASTING*	49	Application of machine learning methods to risk assessment of financial statement fraud: evidence from China [62]	SONG XP; HU ZH; DU JG; SHENG ZH	2014	4.90
22	*SAFETY SCIENCE*	49	Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis [63]	SARKAR S; MAITI J	2020	12.25
23	*NANOTOXICOLOGY*	49	Nanotoxicology data for *in silico* tools: a literature review [64]	FURXHI I; MURPHY F; MULLINS M; ARVANITIS A; POLAND CA	2020	12.25
24	*TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE*	46	Machine learning and credit ratings prediction in the age of fourth industrial revolution [65]	LI JP; MIRZA N; RAHAT B; XIONG DP	2020	11.50
25	*COMPUTER LAW & SECURITY REVIEW*	45	Principles and business processes for responsible AI [66]	CLARKE R	2019	9.00
26	*SAFETY SCIENCE*	43	A machine learning approach for monitoring ship safety in extreme weather events [67]	RAWSON A; BRITO M; SABEUR Z; TRAN-THANH L	2021	14.33
27	*RELIABILITY ENGINEERING & SYSTEM SAFETY*	43	The value of meteorological data in marine risk assessment [68]	ADLAND R; JIA HY; LODE T; SKONTORP J	2021	14.33
28	*NORTH AMERICAN JOURNAL OF ECONOMICS AND FINANCE*	41	A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees [69]	GOLBAYANI P; FLORESCU I; CHATTERJEE R	2020	10.25

Table 6. *Cont.*

R	Journal	TC	Title	Author/s	Year	C/Y
29	*SENSORS*	40	Secure smart wearable computing through artificial intelligence-enabled internet of things and cyber-physical systems for health monitoring [70]	RAMASAMY LK; KHAN F; SHAH MHM; PRASAD BVVS; IWENDI C; BIAMBA C	2022	20.00
30	*JOURNAL OF CLEANER PRODUCTION*	40	Development of flood hazard map and emergency relief operation system using hydrodynamic modeling and machine learning algorithm [71]	RAHMAN M; CHEN NS; ISLAM MM; MAHMUD GI; POURGHASEMI HR; ALAM M; RAHIM MA; BAIG MA; BHATTACHARJEE A; DEWAN A	2021	13.33

4.4. The Most Prolific and Influential Authors

The next ranking shows the most prolific and influential authors. It was elaborated based on the number of publications of the authors on the risk assessment of artificial intelligence in business.

Table 7 details these authors and their production, along with the number of citations, the local h-Index (elaborated with the data collected), and the global h-Index (taken directly from the WoS database).

Table 7. The 20 most productive and influential authors on artificial intelligence risk assessment in businesses.

R	Author's Name	Institution	Country	Local h-Index	Global h-Index	TC	TP	TC/TP	≥ 50	≥ 20	≥ 10
1	MULLINS, M	Univ Limerick	Ireland	4	21	155	4	38.75	1	2	1
2	GARCIA, V.	Univ Jaume 1	Spain	3	23	180	3	60.00	2	1	0
3	MARQUES, A. I.	Univ Jaume 1	Spain	3	34	180	3	60.00	2	1	0
4	SANCHEZ, J. S.	Univ Jaume 1	Spain	3	31	180	3	60.00	2	1	0
5	FURXHI, I	Univ Limerick	Ireland	3	12	140	3	46.67	1	2	0
6	MURPHY, F	Univ Limerick	Ireland	3	23	140	3	46.67	1	2	0
7	POLAND, CA	ELEGI/Colt Laboratory	Scotland	3	27	140	3	46.67	1	2	0
8	GOH, M	Natl Univ Singapore	Singapore	3	49	78	3	26.00	0	2	1
9	RAO, CJ	Wuhan Univ Technol	Peoples' R China	3	20	78	3	26.00	0	2	1
10	HERRERA-VIEDMA, E	Univ Granada	Spain	2	101	283	3	94.33	1	2	0
11	XU, ZS	Sichuan University	Peoples' R China	2	125	256	3	85.33	1	1	0
12	LIAO, HC	Sichuan University	Peoples' R China	2	72	225	2	112.50	1	0	0
13	CHEN, A	Chinese Acad Sci	Peoples' R China	2	10	144	2	72.00	1	1	0
14	CHEN, N	Beijing City Univ	Peoples' R China	2	21	144	2	72.00	1	1	0
15	ARVANITIS, A	Aristotle University of Thessaloniki	Greece	2	11	112	2	56.00	1	1	0
16	PETROSELLI, A	Univ Tuscia	Italy	2	31	66	2	33.00	1	0	0
17	BRITO, M	Univ Southampton	England	2	16	61	3	20.33	0	1	1
18	XIAO, XP	Wuhan Univ Technol	Peoples' R China	2	26	61	2	30.50	0	2	0
19	ARON, M	University of Southern California	USA	2	26	57	1	57.00	1	0	0
20	RAWSON, A	Univ Southampton	England	2	9	56	2	28.00	0	1	1

Abbreviations are the same as in Tables 1 and 3. h-Index is calculated from the dataset (local) and as indicated by WoS (global).

To complete the data, this table includes the authors' affiliation to an institution and the country indicated in the WoS database.

Data are ordered first by local h-Index, second by total number of citations, and third by the total number of publications.

Global h-Index is not used in the order because it is not based on the dataset obtained and does not depend on it, although it is considered an additional value that should be considered for our analysis.

A share of 53.33% of the citations in Table 7 is concentrated on the seven authors with more citations, "HERRERA-VIEDMA, E", "XU, ZS", "LIAO, HC", "GARCIA, V", "MARQUES, A. I", "SANCHEZ, J. S", and "MULLINS, M".

None of the authors with the highest local h-Index are from the U.S.A., with China being first in the ranking, with six authors, followed by Spain with four authors and Ireland with three authors, in the first 20. These last three countries have 13 of 20 authors with the highest local h-Index. By looking at the whole dataset, we find the first U.S.A. author in position 19th, "ARON, M", with a local h-Index of 2, as for others in the table, and with 57 citations for two publications and a global h-Index of 26.

Despite that, note that there is an important overlap between some of the authors: MARQUEZ, SANCHEZ, and GARCIA published together all their papers in the dataset, and the same happened with POLAND, MURPHY, and FURXHI (and partially MULLINS and ARVANITIS).

This could be understood as a consequence of being a very specialized research subject, with few people highly cited. Note anyway that, while the publications of the second group ranged from 2019 to 2020, the three publications of the first group are from 2012.

The publications of the other Spanish author, HERRERA-VIEDMA, are more stable in time, with the first in 2018 and the last in 2023, and were made in collaboration with people from China, the country with most of the authors in the list.

4.5. The Most Productive and Influential Institutions

This paper provides another interesting ranking comparing the production and influence of institutions. Note that different authors in the same paper could be linked to different entities.

We can see the results in Table 8, ordered first by the h-Index, representing the influence of each institution, second by the total number of citations (TC), and third by the total number of publications, satisfying the criteria indicated in Section 3.

In this case, Table 5 includes the position of the institution in two well-known rankings, ARWU (Shanghai Academic Ranking of World Universities) and QS (Quacquarelli Symonds World University Rankings). In cases where the institution is not in any of these rankings, we put two hyphens "--" representing this situation.

Finally, the table includes a ratio ($\times 1000$) between the total number of publications that meet the requirements of this paper (TPS) and the total number of publications of each institution, to put it in relative terms (some institutions can produce more papers in risk assessment of artificial intelligence in business, but they can produce more papers in general due to their type or size, for example).

The third institution in terms of influence (h-Index and citations) and production (number of publications included in our research) in the table is "KING ABDULAZIZ UNIVERSITY" from Saudi Arabia, followed by "NATIONAL UNIVERSITY OF SINGAPORE", and "UNIVERSITY OF EDINBURGH" from the United Kingdom.

These top five institutions represent 43.02% of the total citations and 34.34% of the publications in Table 8.

In these data, as well, the youth (and the specialization) of the research subject is evident; the former because of small values of publications and citations compared to more extended topics, and the latter because the papers of these institutions published in the research subject only represent 0.019‰ of the papers these institutions have published in the same period.

Table 8. The 25 most productive and influential institutions on artificial intelligence risk assessment in businesses.

R	Institution	Country	h	TC	TPS	TC/TPS	≥50	≥20	≥10	ARWU	QS	TP	TPS/TP ×1000
1	SICHUAN UNIVERSITY	People's R China	7	330	9	36.67	1	2	2	98	336	160,468	0.06
2	CHINESE ACADEMY OF SCIENCES (1)	People's R China	6	458	8	57.25	2	2	2	--	--	1,112,433	0.01
3	KING ABDULAZIZ UNIVERSITY	Saudi Arabia	4	301	7	43.00	1	2	1	201–300	149	73,484	0.10
4	NATIONAL UNIVERSITY OF SINGAPORE	Singapore	4	200	5	40.00	1	2	1	68	8	218,642	0.02
5	UNIVERSITY OF EDINBURGH	United Kingdom	4	193	5	38.60	1	3	0	40	27	224,252	0.02
6	UNIVERSITY OF LIMERICK	Ireland	4	155	4	38.75	1	2	1	801–900	421	25,434	0.16
7	HOHAI UNIVERSITY	People's R China	4	117	4	29.25	1	2	0	401–500	1001–1200	42,736	0.09
8	UNIVERSITY OF CHINESE ACADEMY OF SCIENCES, CAS (1)	People's R China	4	105	5	21.00	0	2	2	--	--	303,004	0.02
9	UNIVERSITY OF NEW SOUTH WALES SYDNEY	Australia	4	85	6	14.17	0	1	3	77	19	210,937	0.03
10	SOLENT UNIVERSITY	United Kingdom	4	78	4	19.50	0	1	2	--	--	2476	1.62
11	UNIVERSITY OF SOUTHAMPTON	United Kingdom	4	78	4	19.50	0	1	2	151–200	80	152,444	0.03
12	UNIVERSITY OF GRANADA	Spain	3	286	4	71.50	1	2	0	301–400	431	84,885	0.05
13	UNIVERSITAT JAUME I	Spain	3	180	3	60.00	2	1	0	601–700	--	2020	1.49
14	STOCKHOLM UNIVERSITY	Sweden	3	81	3	27.00	1	0	2	101–150	153	86,128	0.03
15	WUHAN UNIVERSITY OF TECHNOLOGY	People's R China	3	78	3	26.00	0	2	1	201–300	801–850	56,434	0.05
16	HENAN POLYTECHNIC UNIVERSITY	People's R China	2	144	2	72.00	1	1	0	901–1000	--	18,054	0.28
17	SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS—CHINA	People's R China	2	122	5	24.40	2	0	0	701–800	--	8196	0.61
18	INSTITUTO POLITECNICO DO PORTO	Portugal	2	116	2	58.00	1	0	1	--	--	12,706	0.16
19	DONGBEI UNIVERSITY OF FINANCE AND ECONOMICS	People's R China	2	96	3	32.00	1	1	0	--	--	3096	0.97
20	TUSCIA UNIVERSITY	Italy	2	66	2	33.00	1	0	0	901–1000	901–950	11,017	0.18
21	UNIVERSITY OF GLASGOW	United Kingdom	2	39	2	19.50	0	1	1	101–150	78	170,885	0.01
22	KAUNAS UNIVERSITY OF TECHNOLOGY	Lithuania	2	38	2	19.00	0	1	1	--	751–760	13,873	0.14
23	RUTGERS UNIVERSITY NEW BRUNSWICK	USA	2	35	3	11.67	0	1	0	101–150	328	245,099	0.01
24	KHALIFA UNIVERSITY OF SCIENCE AND TECHNOLOGY	United Arab Emirates	2	35	2	17.50	0	1	0	601–700	202	18,613	0.11
25	WUHAN INSTITUTE OF TECHNOLOGY	People's R China	2	29	2	14.50	0	1	0	901–1000	--	11,710	0.17

A total of 15 entities from the list of 25 are positioned in all the university rankings included, and 14 entities from the list of 20 are ranked in the top 500 in some of the university rankings included.

In Figure 3, we can see clearly the novelty of this field, with papers cited between the top 22 institutions ranging from 2018 to 2024. The first of the four clusters identified by VOSviewer comprises Chinese Academy of Sciences, Dalian University of Technology, Dongbei University of Finance and Economics, and Universitat Jaume I; the second cluster is formed by Sichuan University, Rutgers University-New Brunswick, and King Abdulaziz University; the third one is composed of the University of Edinburgh, the University of Limerick, and Sejong University; and finally the fourth one holds the National University of Singapore, the Southwestern University of Finance, and Wuhan Technology School.

Again, we see that institutions from China are the most prolific and influential, with 7 Chinese institutions included, accounting for a total of 1479 citations and 41 papers, led by "SICHUAN UNIVERSITY" in first place, followed by "CHINESE ACADEMY OF SCIENCES" in second place and "UNIVERSITY OF CHINESE ACADEMY OF SCIENCES, CAS" in third place. Although the names are similar and they have strong links, we preferred to keep separate the Chinese Academy of Sciences and the University of Academy of Sciences, run by the former [72], just as WoS does. That allows the research to respect the exact affiliation indicated by each author.

At Figure 3, the size of the nodes represents the number of citations, while the connection thickness represents the strength of citation relationships; thicker, stronger, and lighter-color lines represent newer activity, and the darker color represents older activity.

So, we can see that Sichuan University is pivotal in this citation network, with its research widely recognized. Other institutions, like Sejong University at the right, could be considered new research leaders, and the Chinese Academy of Sciences and Sichuan University could be especially suitable for collaboration, as they bridge multiple organizations.

4.6. Country Analysis

The following elaborated ranking in Table 9 allows us to compare productivity and influence between countries and see if it is aligned with the previous rankings.

Table 9. The 15 most productive and influential countries in research on artificial intelligence risk assessment in businesses.

R	Country	h	TC	TPS	Pop	TC/Pop	TPS/Pop	≥100	≥50	≥20
1	CHINA	23	1909	96	1422.58	1.342	0.067	4	6	16
2	USA	10	371	23	343.48	1.080	0.067	0	3	5
3	UNITED KINGDOM	8	217	11	68.68	3.159	0.160	0	0	4
4	AUSTRALIA	5	101	8	26.45	3.818	0.302	0	0	1
5	INDIA	5	85	9	1438.07	0.059	0.006	0	0	2
6	IRELAND	4	155	4	5.20	29.827	0.770	0	1	2
7	MALAYSIA	4	63	4	35.13	1.794	0.114	0	0	1
8	ITALY	4	57	5	59.50	0.958	0.084	0	0	1
9	SPAIN	3	186	7	47.91	3.882	0.146	0	2	1
10	SWEDEN	3	67	3	10.55	6.350	0.284	0	0	1
11	PORTUGAL	3	46	3	10.53	4.368	0.285	0	0	1
12	SAUDI ARABIA	3	24	6	33.26	0.721	0.180	0	0	0
13	TURKEY	2	235	2	87.27	2.693	0.023	1	0	0
14	GREECE	2	90	4	10.24	8.787	0.391	0	1	1
15	NORWAY	2	54	3	5.52	9.783	0.543	0	0	1

The abbreviations are the same as in Tables 1 and 3, except for Pop—population (in millions), TPS/Pop—studies per millions of population, and TC/Pop—citations per millions of population.

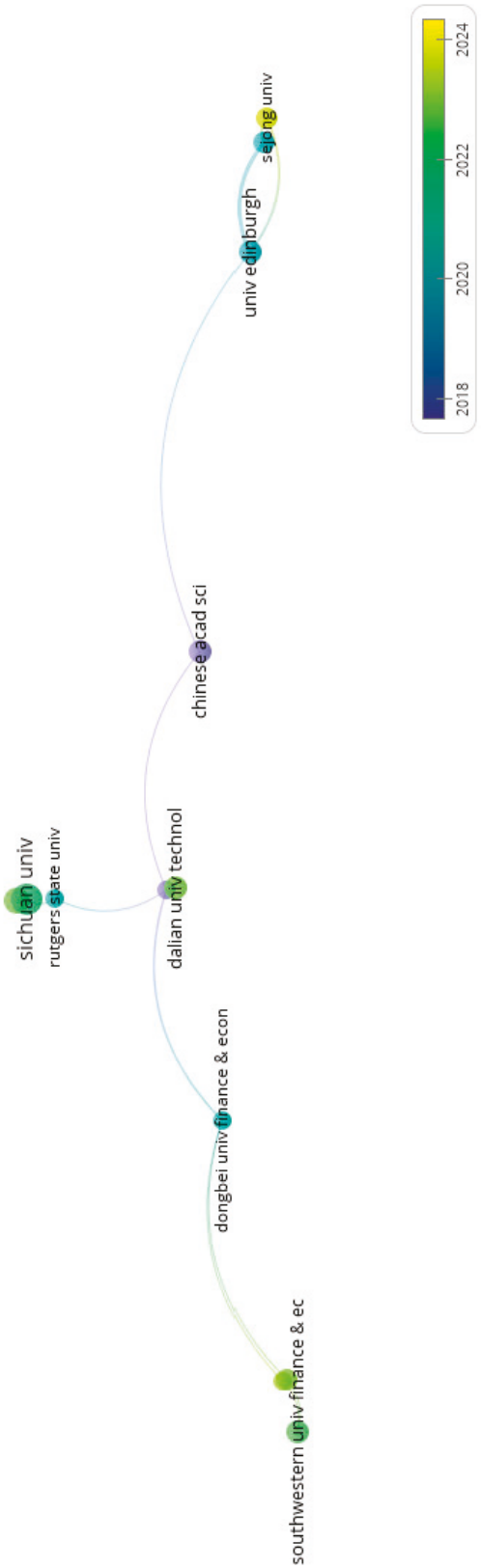


Figure 3. Citations between institutions. Overlay visualization via VOSviewer.

As for the previous ones, it is ordered first by the country's local h-Index, second by the number of citations received, and third by the number of papers published on the research subject, according to the WoS database.

We can see in Table 9 some countries that were identified before, such as China, the U.S.A., the United Kingdom, Australia, Ireland, Italy, Spain, Sweden, Portugal, and Saudi Arabia.

In contrast, there are countries appearing in Table 8 that do not appear in Table 9, such as Singapore, Lithuania, and UAE. This is because other countries have a higher number of active and influential institutions, allowing them to overcome the former in terms of the combination of the h-Index, the number of publications, and the number of citations.

On the other hand, countries appearing in Table 9 for the first time, namely, India, Malaysia, Turkey, Greece, and Norway, have a high number of publications but are highly distributed among different institutions; India, for example, has nine publications from nine different institutions.

These are the reasons why Table 9 has been included in this paper, that is, in order to provide a precise view of the most productive and influential countries.

Countries with larger populations can have some advantages regarding scientific production. To take this into account, the total populations in millions in 2023 according to Our World In Data, a website curated by the University of Oxford [73], have been included, accompanied by the ratios of citations and publications per million inhabitants. Population data for 2024 were not available at the time we wrote this paper.

According to these data, the country with the most publications and citations per inhabitant was Ireland, followed by Norway, Greece, and Australia (all above average).

Additionally, only two countries (China and Turkey) have papers with 100 or more citations, five have papers with 50 or more citations (China, U.S.A., Ireland, Spain, and Greece), and thirteen have papers with 20 or more citations.

Another aspect that this paper evaluates is the diversity of nationalities of the teams that collaborate in each paper. In Figure 4 we can see in green the number of papers that Bibliometrix considers an intra-country collaboration, and in red the number of papers that Bibliometrix considers an inter-country collaboration.

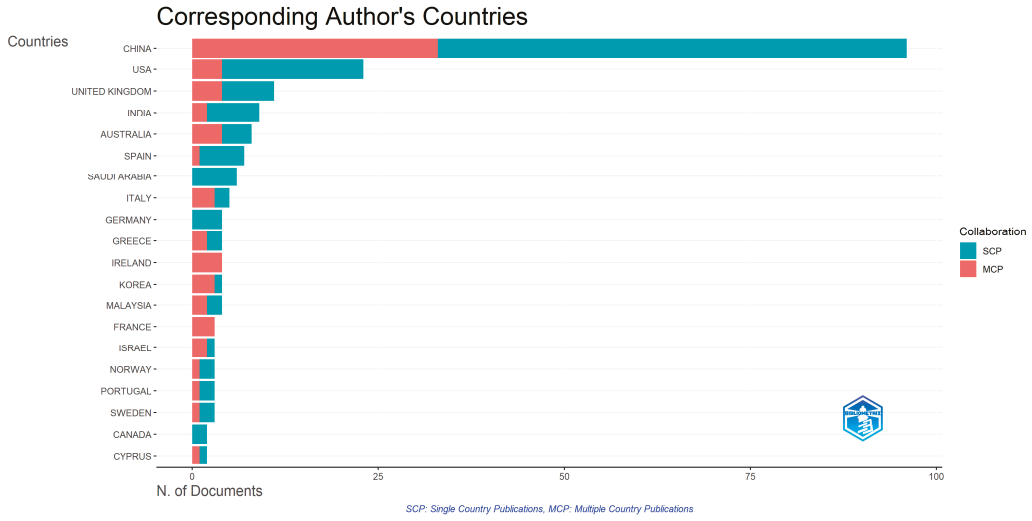


Figure 4. Corresponding author's countries and diversity of nationalities.

The data analyzed show that the total production of French and Irish corresponding authors is in collaboration with other countries, and that when the corresponding author is

from Saudi Arabia, Germany, or Canada, all the production is intra-country. This could contribute to the fact that countries with an important AI industry and influence, like Germany or France, are not in Table 9.

The average percentage of inter-country collaboration in the top 20 is 41.32% and 41.75% if we omit the outliers (0% and 100%).

Please note that Bibliometrix uses the corresponding author's country to assign the publication's country. We should be aware of this because, for instance, articles in which Herrera-Viedma, a Spanish author, has collaborated, are assigned to China, as the corresponding author has this nationality. So, we can have articles that contribute to the production or influence of some authors but are assigned to a different country (in the MCP bar).

Figure 5, generated with the Bibliometrix tool, shows graphically the collaboration links between different countries. The thickness of red lines is related to the production of each country, and the lines between different countries show the strength of collaboration between them.

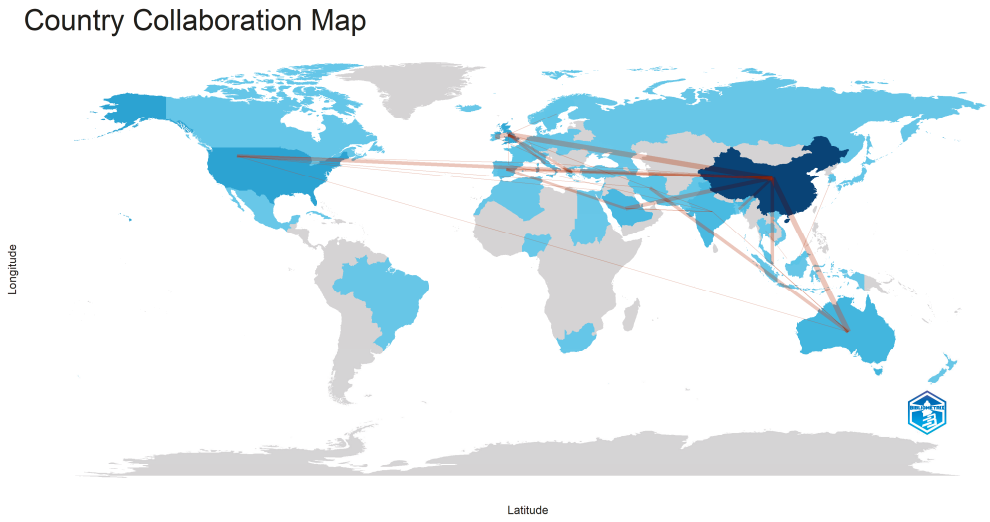


Figure 5. Collaboration between countries.

4.7. Landscapes and Evolution of Artificial Intelligence's Risk Assessment in Business

As we have seen in the previous sections of this paper, the selected subject is very young, scholarly speaking; although this bibliometric analysis has enough data to reach some important conclusions, numbers are very low compared to older subjects.

The landscape in Figure 6, based on the most used words through Keywords Plus, that is, on the frequency that these words appear in the titles of the papers referenced by each publication [74], throws no surprises: "risk-assessment", "model", "classification", "prediction", and "feature-selection" or "performance" constitute the basis of the challenges any AI tool has to deal with, so we can consider the basis is shared with many AI research topics.

If, alternatively, we look at the Bibliometrix treemap using author's keywords in Figure 7, we can see the landscape is slightly different: although we have "machine learning" and "risk assessment" as two of the main keywords, together with "artificial intelligence" and "deep learning", we can see that terms like "credit risk", "credit scoring", and "risk management" now have more importance.

Additionally, if we look in detail at the Bibliometrix list of the 25 most used author's keywords, we can see a new word at the "tail" of the list more related to risk assessment, namely, "cybersecurity".



Figure 6. Bibliometrix 25-word treemap (Keywords Plus).

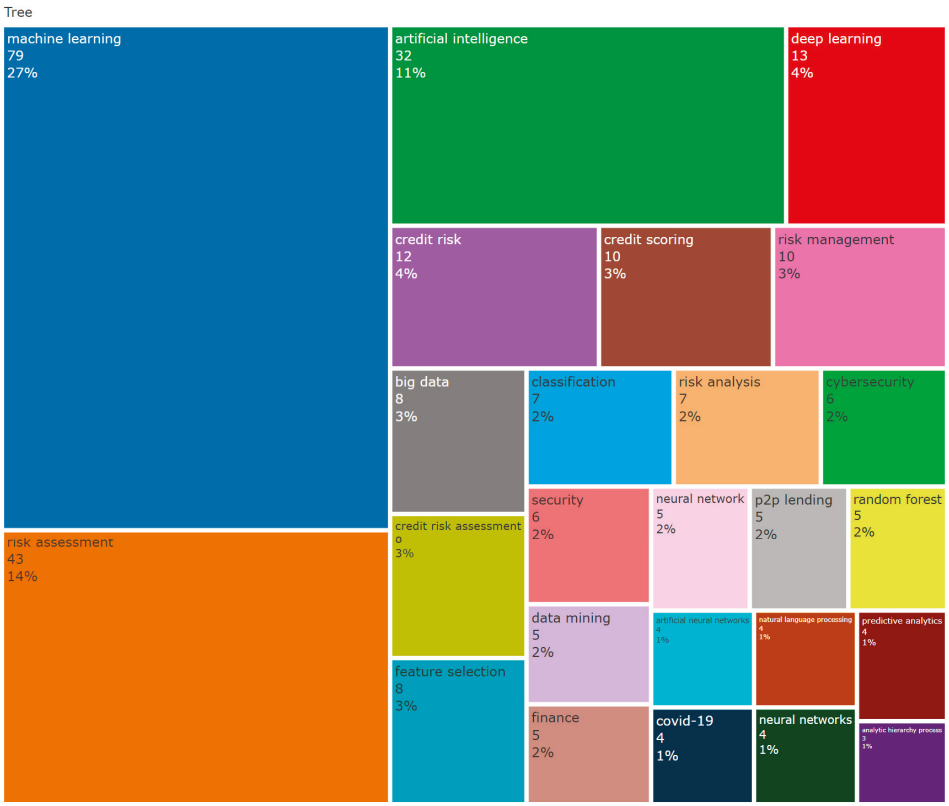


Figure 7. Bibliometrix 25-word treemap (author's keywords).

If we look in Figure 8 at these author's keywords in a chronological way, we can easily see the recent emergence of interest in risk assessment in AI, noting that the peak year was

2023, with 43 occurrences, and is still echoing today. Of course, there are more general terms, such as "machine learning" with a larger size of the bubble (more occurrences), and others, such as "neural network", more present over time, but if we focus on the risk topics, the risk assessment is the most important subject of research.

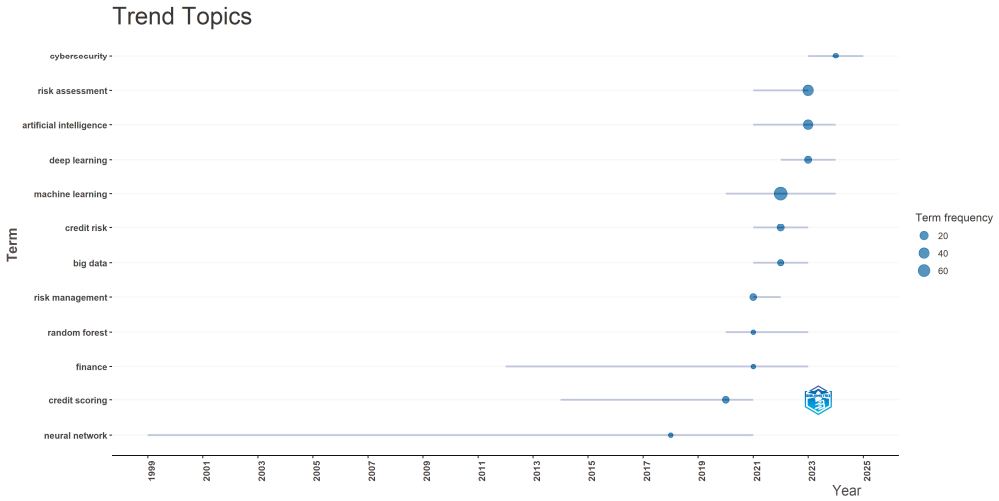


Figure 8. Trends in the co-occurrence of author's keywords over time.

4.8. Current Emerging Issues in Artificial Intelligence's Risk Assessment in Business

This paper uses SciMat to elaborate on the strategic map and identify emerging issues in artificial intelligence risk assessment in business.

The periods are annual between 2018 and 2024, including all the publications returned by the search strategy explained in Section 3. A unique period was allocated comprising years between 1997 and 2017, both inclusive, given that it was an interval with very few publications, i.e., only 13 in total (0.62 publications per year on average, considerably less than in any year from 2018 until 2023).

Taking this into account, Figures 9–11. were elaborated, using the SciMat tool with the following parameters [33]: We chose words as the unit of analysis, including author's words, source words, and added words; we also used a frequency reduction for all the periods of 2; a co-occurrence matrix; network reduction of 2; normalization by an equivalence index; a simple center algorithm as the clustering algorithm, with network size between 3 and 12; a union mapper as document mapper; h-Index and sum of citations as quality measures; an evolution map based on an inclusion index; and Jaccard's index for an overlapping map.

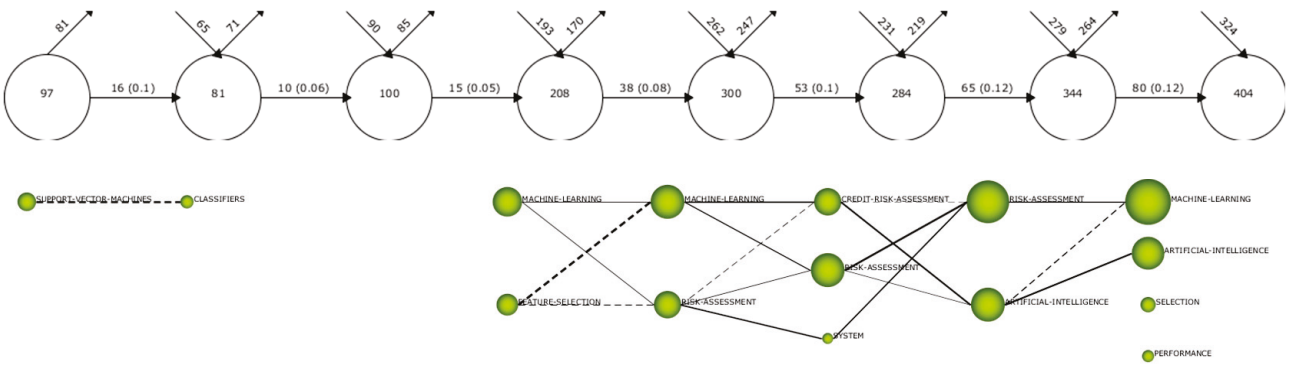


Figure 9. SciMat overlapping map and document count evolution map.

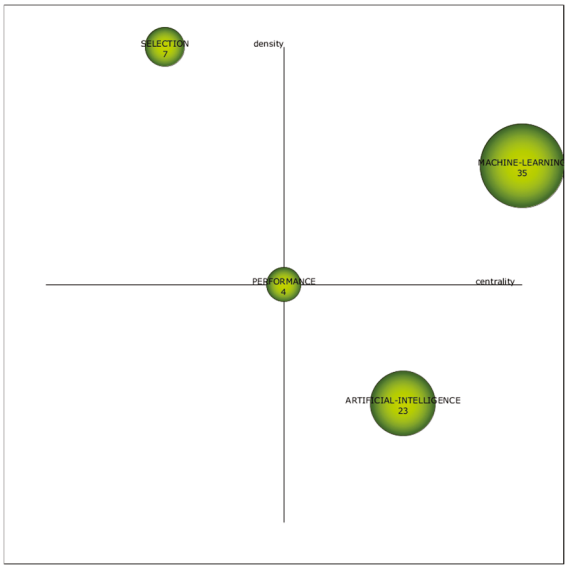


Figure 10. SciMat document count strategic map.

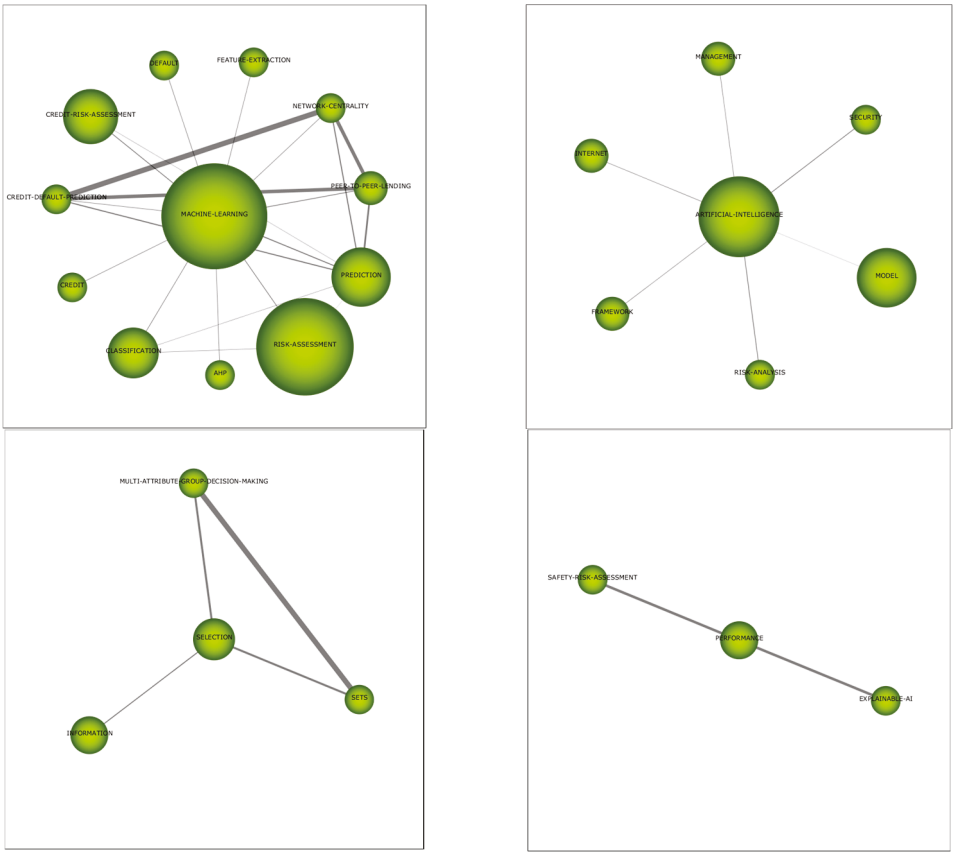


Figure 11. SciMat cluster networks.

We can see in the thematic overlapping map that for each period selected, the number of words entering and exiting is relevant, with similar numbers of words entering and exiting around the research subject; this shows a very dynamic research subject, although a total of 53, 65, and 80 keywords were maintained for periods 2022, 2023, and 2024, respectively. By looking at this progression, we can appreciate a potential progressive stabilization of the research subject.

We can see by looking at the evolution map that in the period between 2021 and 2024, the term "risk assessment" was used frequently in most of the dataset papers, although in 2024 the term does not appear unless looking at the cluster networks of artificial intelligence, machine learning, and performance.

Figure 10 shows the strategic map for the last period analyzed (2024). As Cobo et al. explain [33], this map is divided into quadrants regarding Callon's centrality and density factors. The upper-left quadrant shows the niche of very developed themes (such as selection), the upper-right quadrant shows the motor themes (such as machine learning in this case), and at the lower-right quadrant, corresponding to basic or transversal themes, we can find artificial intelligence, which does not show unexpected results.

Despite this, if we delve into detail in these terms' cluster networks in Figure 11, we can see some interesting findings around the importance of risk assessment and risk analysis, safety, and explainability of artificial intelligence systems.

5. Discussion

This paper identifies the top 25 most productive and influential journals for risk assessment of AI in businesses. The top 13 journals in the list (52%) concentrate 78.95% of the citations, with the top 4 having 51.59% of the citations; so, a certain concentration can be identified, although it is not considered relevant.

Applying the Herfindhal–Hirschman concentration index to evaluate the concentration of the market in specific journals [75] shows that it is not a highly concentrated market (0.12 in citations, 0.06 in papers published).

This paper has also identified an important increase in the number of papers published since 2018 on the research subject of risk assessment of artificial intelligence in business. Note that 2018 is the first year with 10 publications, compared to 9 in the 10 previous years and 13 in total in the 20 previous years.

This increase in the topic "risk assessment in artificial intelligence on business" is consistent with the increase in publications about "artificial intelligence", as we have seen in the results.

This paper also confirms that this subject is a very novel one, with a relevant number of data to enable a bibliometric analysis. However, the number of citations and publications is still far from that of other more mature topics, and is also significantly lower in relation to the total number of publications of the institutions where corresponding authors are affiliated.

The novelty of this research subject signals it as an emerging theme with great potential development opportunities.

This is also evident when we look at the publishing dates of the three most influential articles: 2005, 2015, and 2018, while the term artificial intelligence was coined in 1955. We could say that only in recent years, once the significant progress of artificial intelligence has been recognized, has society started analyzing its risks.

As could be expected in a novel research subject, the local h-Index (calculated within the given dataset) of the 20 most productive and influential authors is very low, ranging in the top 20 between 4 and 2. One interesting finding is that only in the cases of China and Spain can we find very influential authors (by looking at the global h-Index) contributing to the risk assessment of artificial intelligence in business.

We have also identified the most productive and influential institutions, with the ones from China, Spain, and the United Kingdom having the highest number of citations.

When we look at the ranking of the top 15 countries, data are not exactly the same because other countries are more influential and have, in some cases, more publications. However, we see China and the United Kingdom as the leaders, with the U.S.A. also included.

We should note that, with the European Union AI regulation becoming effective on 1 August 2024 [8], other regions are leading the research on AI risk assessment in business instead, although in most cases, these non-EU countries' researchers are collaborating with EU researchers on an inter-country basis.

This is aligned with Bradford's law, claiming that "there are a few very productive periodicals, a larger number of more moderate producers, and a still larger number of constantly diminishing productivity" [76].

Another positive finding in this paper is that explainable AI and cyber security are both part of the emerging concepts related to the research on risk assessment of AI in business, so transparency and security of the AI systems are identified as a clear trend by the results obtained.

According to our research results, a scientist could probably consider China, Spain, and the United Kingdom as the best places to look for research institutions to work with. However, relying solely on quantitative data can be misleading, as AI risk assessment challenges are global, and advances are intertwined across countries and institutions. As the data show, it is not just about which country is leading, but about fostering collaboration and openness to diverse perspectives. While China, Spain, and the UK are leading because of their research throughput, other countries like the US, France, and Germany also contribute significantly to high-quality AI research, particularly due to their strong AI industries. However, the fact that the corresponding author is not of these countries or that the collaboration between these and other countries is rare is penalizing these countries in the overall ranking.

An analogous situation happens when looking solely at journals from MDPI and Elsevier as the most influential for an author to submit his or her work. It is crucial for science to prioritize those journals with strong peer review standards, rigorous editorial policies, and a proven track record of publishing impactful, high-quality work. Therefore, any journal from other publishers meeting these criteria could be a perfectly suitable candidate for submitting research.

At this point, we found several gaps in the research about risk assessment in organizations. For instance, a search for explainable AI in our dataset shows only 11 papers, all between 2021 (1) and 2024 (4), mostly centered on AI in financial (6), medicine (2), and insurance (2) sectors. In a search for standardized AI risk assessment tools, we found NIST AI risk management framework and MIT AI risk database references (2), but none indexed at the WoSCC dataset. Only one of the papers (from 2021) in our dataset [77] deals with corporate digital responsibility (that is, managing opportunities while addressing the related risks of digital transformation, a key to AI adoption), and was cited 38 times according to our dataset. When we searched for "agents", given that agentic AI seems one of the recent fields to be explored, only one result was found. Furthermore, when looking for challenges related to intellectual property, consumer protection, democracy, or multimodal AI risks, zero results were found. In the specific case of explainable AI, we can find works included in the analyzed dataset such as Bharodia and Chen [78], Chen et al. [79], Cho and Shin [80], Pnevmatikakis et al. [81], and Yang et al. [82].

The fact that risk assessment is found inside 2024 SCIMat clusters and not in the evolution map nor in the strategic map could be a symptom of this research subject being strongly associated with artificial intelligence and machine learning, as is the case of "explainable-AI" as well.

We can also identify that the performance of artificial intelligence systems is, in 2024, a focus word and that explainability, safety, and risk assessment are strongly linked to performance research as well (note that these could be antagonistic targets in many cases).

Additionally, we found the relative novelty of this subject, showing only in recent years, once the significant progress of artificial intelligence had been recognized, that society started to analyze its risks in a scientific way. Policy-related facts mentioned in the introduction, such as the emergence of the European AI regulation or the NIST AI Risk Management Framework, could probably be other factors facilitating this surge of studies in the last few years. An additional factor related to the increasing focus on AI has been the appearance of LLM-based conversational generative AI, which has enabled easier access to AI systems for the general public, just as the IBM PC did with computers.

6. Conclusions

The findings of this paper determine which are the most productive and influential references in the research subject of risk assessment of artificial intelligence in business. We have identified the following:

- The evolution of production on the research subject through the years;
- The most cited papers;
- The most productive and most influential journals, authors, institutions, and countries;
- The existing collaboration links between countries;
- The landscape and evolution of the selected research subject;
- The emerging concepts that specialized scholars are now dealing with.

Based on the results and the discussion above, we recommend the following actionable insights for scientists, businesses, and policymakers:

1. Focus on explainable AI: Researchers should prioritize developing and implementing explainable AI (XAI) techniques. This will enhance transparency and trust in AI systems, making it easier for stakeholders to understand and manage AI-driven decisions.
2. Enhance cybersecurity measures: Given the increasing integration of AI in business operations, it is crucial to incorporate robust cybersecurity measures. This includes developing AI systems resilient to cyber threats and maintaining data privacy and security.
3. Interdisciplinary collaboration: Encourage collaboration between AI researchers and experts in cognitive biases, behavioral economics, and organizational psychology. This interdisciplinary approach can provide deeper insights into AI's human and organizational impacts, leading to more comprehensive risk assessments.
4. Policy development: Policymakers should consider the implications of AI risk assessment in their regulatory frameworks. This includes creating guidelines that promote the ethical use of AI and addressing potential risks associated with AI deployment in various sectors. Initiatives like the EU AI Regulation could be a reference for them.
5. Continuous monitoring and adaptation: Businesses should establish mechanisms for continuous monitoring of AI systems to identify and mitigate risks promptly. This involves regular audits, updates to AI models, and adapting to new threats and challenges as they arise. Databases such as that at https://airisk.mit.edu (accessed on 10 January 2025) could be helpful for them.

This research offers interesting findings related to the emerging technologies of human–computer Interaction because although AI is revolutionizing the dynamics of human–computer interaction, this advancement presents a dual-edged sword: as the scientific

community initiates rigorous studies on AI risk assessment within organizational contexts, it is imperative to consider critical issues such as data privacy, consumer protection, intellectual property, AI corporate digital responsibility, ethics, multimodal AI risks, bias, and explainable AI. Addressing these factors is essential to ensure the successful integration and interaction of AI systems with human operators.

These challenges, together with the promising line of research in how the social and governmental initiatives, such as regulations and guidelines could have accelerated the scientific progress of AI risk assessment, are signaled as future research directions for our community.

Regarding limitations, it is necessary to state that this study is limited to the WoS Core Collection database. Although this repository is well recognized in all areas, we could expect differences in results and conclusions if another database is used, as discussed in Section 3.

Another issue we have identified is that the results include both the risk assessment of AI in businesses and AI-based risk assessment in businesses. Although that does not reduce the importance of this paper, future research separating both concepts should be interesting for specialized research.

Other interesting future research would be to evaluate the content and significance of key papers or authors as a qualitative analysis or to combine metrics with societal or political impact.

Author Contributions: Conceptualization, J.C.M.-T. and J.V.O.-G.; methodology, J.C.M.-T. and R.O.-B.; writing—original draft preparation, J.C.M.-T.; writing—review and editing, J.V.O.-G., R.O.-B. and S.E.-R. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by INCIBE's Chair funded by the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to legal reasons.

Acknowledgments: This work has been partially supported by INCIBE's Chair funded by the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* **2021**, *54*, 1–36. [CrossRef]
2. Siau, K.; Wang, W. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *J. Database Manag.* **2020**, *31*, 74–87. [CrossRef]
3. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
4. Gerke, S.; Minssen, T.; Cohen, G. Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare. In *Artificial Intelligence in Healthcare*; Academic Press: Cambridge, MA, USA, 2020; pp. 295–336.
5. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]

6. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]
7. *ISO 31000: 2018 (En)*; International Standards Organization International Standards Organization. Risk Management—Guidelines; ISO: Geneva, Switzerland, 2018.
8. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Text with EEA Relevance*; Official Journal of the European Union, L 2024/1689: Maastricht, The Netherlands, 12 July 2024; European Parliament and Council of the European Union.
9. Elham, T. *NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
10. Noyons, E.C.M.; Moed, H.F.; Van Raan, A.F.J. Integrating Research Performance Analysis and Science Mapping. *Scientometrics* **1999**, *46*, 591–604. [CrossRef]
11. Gil-Gomez, H.; Oltra-Badenes, R.; Guerola-Navarro, V.; Zegarra Saldaña, P. Crowdfunding: A Bibliometric Analysis. *Int. Entrep. Manag. J.* **2023**, *19*, 27–45. [CrossRef]
12. Ninkov, A.; Frank, J.R.; Maggio, L.A. Bibliometrics: Methods for Studying Academic Publishing. *Perspect. Med. Educ.* **2022**, *11*, 173–176. [CrossRef]
13. Fahimifar, S.; Janavi, E.; Fadaei, F. Awakening the Beauty: A Journey through Dormant Gems in Strategic Management Literature. *Qual. Quant.* **2024**, *58*, 3331–3362. [CrossRef]
14. Merigo, J.M.; Yang, J.-B. A Bibliometric Analysis of Operations Research and Management Science. *Omega-Int. J. Manag. Sci.* **2017**, *73*, 37–48. [CrossRef]
15. Machado, C.G.; Winroth, M.P.; Ribeiro Da Silva, E.H.D. Sustainable Manufacturing in Industry 4.0: An Emerging Research Agenda. *Int. J. Prod. Res.* **2020**, *58*, 1462–1484. [CrossRef]
16. Ferreira, J.J.M.; Fernandes, C.I.; Kraus, S. Entrepreneurship Research: Mapping Intellectual Structures and Research Trends. *Rev. Manag. Sci.* **2019**, *13*, 181–205. [CrossRef]
17. Passas, I. Bibliometric Analysis: The Main Steps. *Encyclopedia* **2024**, *4*, 65. [CrossRef]
18. Liang, T.-P.; Liu, Y.-H. Research Landscape of Business Intelligence and Big Data Analytics: A Bibliometrics Study. *Expert Syst. Appl.* **2018**, *111*, 2–10. [CrossRef]
19. Batistič, S.; Van Der Laken, P. History, Evolution and Future of Big Data and Analytics: A Bibliometric Analysis of Its Relationship to Performance in Organizations. *Br. J. Manag.* **2019**, *30*, 229–251. [CrossRef]
20. Bawack, R.E.; Wamba, S.F.; Carillo, K.D.A.; Akter, S. Artificial Intelligence in E-Commerce: A Bibliometric Study and Literature Review. *Electron. Mark.* **2022**, *32*, 297–338. [CrossRef] [PubMed]
21. Müßigmann, B.; von der Gracht, H.; Hartmann, E. Blockchain Technology in Logistics and Supply Chain Management—A Bibliometric Literature Review from 2016 to January 2020. *IEEE Trans. Eng. Manag.* **2020**, *67*, 988–1007. [CrossRef]
22. Shi, L.; Mai, Y.; Wu, Y.J. Digital Transformation: A Bibliometric Analysis. *J. Organ. End User Comput.* **2022**, *34*, 1–20. [CrossRef]
23. Fosso Wamba, S.; Bawack, R.E.; Guthrie, C.; Queiroz, M.M.; Carillo, K.D.A. Are We Preparing for a Good AI Society? A Bibliometric Review and Research Agenda. *Technol. Forecast. Soc. Change* **2021**, *164*, 120482. [CrossRef]
24. Jalali, M.S.; Razak, S.; Gordon, W.; Perakslis, E.; Madnick, S. Health Care and Cybersecurity: Bibliometric Analysis of the Literature. *J. Med. Internet Res.* **2019**, *21*, e12644. [CrossRef]
25. Zhang, Y.; Wu, M.; Tian, G.Y.; Zhang, G.; Lu, J. Ethics and Privacy of Artificial Intelligence: Understandings from Bibliometrics. *Knowl. -Based Syst.* **2021**, *222*, 106994. [CrossRef]
26. Albahri, O.S.; AlAmoodi, A.H. Cybersecurity and Artificial Intelligence Applications: A Bibliometric Analysis Based on Scopus Database. *MJCS* **2023**, *2023*, 158–169. [CrossRef]
27. Novelli, C.; Casolari, F.; Rotolo, A.; Taddeo, M.; Floridi, L. Taking AI Risks Seriously: A New Assessment Model for the AI Act. *AI Soc.* **2024**, *39*, 2493–2497. [CrossRef]
28. Novelli, C.; Casolari, F.; Rotolo, A.; Taddeo, M.; Floridi, L. AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act. *DISO* **2024**, *3*, 13. [CrossRef]
29. Slattery, P.; Saeri, A.K.; Grundy, E.A.C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; Thompson, N. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence. *AGI* **2024**, *1*, 1–26. [CrossRef]
30. Bar-Ilan, J. Which H-Index?—A Comparison of WoS, Scopus and Google Scholar. *Scientometrics* **2008**, *74*, 257–271. [CrossRef]
31. Jacso, P. As We May Search—Comparison of Major Features of the Web of Science, Scopus, and Google Scholar Citation-Based and Citation-Enhanced Databases. *Curr. Sci.* **2005**, *89*, 1537–1547.

32. Web of Science Core Collection. Available online: https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/web-of-science-core-collection/ (accessed on 9 August 2024).
33. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. An Approach for Detecting, Quantifying, and Visualizing the Evolution of a Research Field: A Practical Application to the Fuzzy Sets Theory Field. *J. Informetr.* **2011**, *5*, 146–166. [CrossRef]
34. Hirsch, J.E. An Index to Quantify an Individual's Scientific Research Output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572. [CrossRef] [PubMed]
35. Hua, P.; Wan, J.; Wu, J. A Perfect Hirsch-Type Index? Experiences Using the Tapered h-Index (hT). Available online: http://eprints.rclis.org/14251/ (accessed on 9 August 2024).
36. Chai, J.-C.; Hua, P.-H.; Rousseau, R.; Wan, J.-K. The Adapted Pure H-Index. In Proceedings of the WIS 2008, Berlin Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting, Berlin, Germany, 8 July 2008.
37. Egghe, L. Mathematical Theory of the H- and g-Index in Case of Fractional Counting of Authorship. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 1608–1616. [CrossRef]
38. Guns, R.; Rousseau, R. Real and Rational Variants of the H-Index and the g-Index. *J. Inf.* **2009**, *3*, 64–71. [CrossRef]
39. Bai, X.; Duan, J.; Li, B.; Fu, S.; Yin, W.; Yang, Z.; Qu, Z. Global Quantitative Analysis and Visualization of Big Data and Medical Devices Based on Bibliometrics. *Expert Syst. Appl.* **2024**, *254*, 124398. [CrossRef]
40. Aria, M.; Cuccurullo, C. *Bibliometrix*: An R-Tool for Comprehensive Science Mapping Analysis. *J. Inf.* **2017**, *11*, 959–975. [CrossRef]
41. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Available online: https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf (accessed on 21 August 2024).
42. Wang, Y.; Wang, S.; Lai, K.K. A New Fuzzy Support Vector Machine to Evaluate Credit Risk. *IEEE Trans. Fuzzy Syst.* **2005**, *13*, 820–831. [CrossRef]
43. Malekipirbazari, M.; Aksakalli, V. Risk Assessment in Social Lending via Random Forests. *Expert Syst. Appl.* **2015**, *42*, 4621–4631. [CrossRef]
44. Liao, H.; Xu, Z.; Herrera-Viedma, E.; Herrera, F. Hesitant Fuzzy Linguistic Term Set and Its Application in Decision Making: A State-of-the-Art Survey. *Int. J. Fuzzy Syst.* **2018**, *20*, 2084–2110. [CrossRef]
45. Sun, J.; Zhang, J.; Li, Q.; Yi, X.; Liang, Y.; Zheng, Y. Predicting Citywide Crowd Flows in Irregular Regions Using Multi-View Graph Convolutional Networks. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 2348–2359. [CrossRef]
46. Chen, N.; Ribeiro, B.; Chen, A. Financial Credit Risk Assessment: A Recent Review. *Artif. Intell. Rev.* **2016**, *45*, 1–23. [CrossRef]
47. Quah, T.-S.; Srinivasan, B. Improving Returns on Stock Investment through Neural Network Selection. *Expert Syst. Appl.* **1999**, *17*, 295–301. [CrossRef]
48. Marqués, A.I.; García, V.; Sánchez, J.S. Exploring the Behaviour of Base Classifiers in Credit Scoring Ensembles. *Expert Syst. Appl.* **2012**, *39*, 10244–10250. [CrossRef]
49. Feng, X.; Xiao, Z.; Zhong, B.; Qiu, J.; Dong, Y. Dynamic Ensemble Classification for Credit Scoring Using Soft Probability. *Appl. Soft Comput.* **2018**, *65*, 139–151. [CrossRef]
50. Marqués, A.I.; García, V.; Sánchez, J.S. Two-Level Classifier Ensembles for Credit Risk Assessment. *Expert Syst. Appl.* **2012**, *39*, 10916–10922. [CrossRef]
51. Huang, X.; Liu, X.; Ren, Y. Enterprise Credit Risk Evaluation Based on Neural Network Algorithm. *Cogn. Syst. Res.* **2018**, *52*, 317–324. [CrossRef]
52. Furxhi, I.; Murphy, F.; Mullins, M.; Arvanitis, A.; Poland, C.A. Practices and Trends of Machine Learning Application in Nanotoxicology. *Nanomaterials* **2020**, *10*, 116. [CrossRef] [PubMed]
53. Luo, J.; Yan, X.; Tian, Y. Unsupervised Quadratic Surface Support Vector Machine with Application to Credit Risk Assessment. *Eur. J. Oper. Res.* **2020**, *280*, 1008–1017. [CrossRef]
54. Xia, Y.; Zhao, J.; He, L.; Li, Y.; Niu, M. A Novel Tree-Based Dynamic Heterogeneous Ensemble Method for Credit Scoring. *Expert Syst. Appl.* **2020**, *159*, 113615. [CrossRef]
55. Afzal, F.; Yunfei, S.; Nazir, M.; Bhatti, S.M. A Review of Artificial Intelligence Based Risk Assessment Methods for Capturing Complexity-Risk Interdependencies: Cost Overrun in Construction Projects. *IJMPB* **2021**, *14*, 300–328. [CrossRef]
56. Adnan, R.M.; Petroselli, A.; Heddam, S.; Santos, C.A.G.; Kisi, O. Short Term Rainfall-Runoff Modelling Using Several Machine Learning Methods and a Conceptual Event-Based Model. *Stoch. Environ. Res. Risk Assess.* **2021**, *35*, 597–616. [CrossRef]
57. Varghese, B.; Chen, F.; Hwang, D.; Palmer, S.L.; De Castro Abreu, A.L.; Ukimura, O.; Aron, M.; Aron, M.; Gill, I.; Duddalwar, V.; et al. Objective Risk Stratification of Prostate Cancer Using Machine Learning and Radiomics Applied to Multiparametric Magnetic Resonance Images. *Sci. Rep.* **2019**, *9*, 1570. [CrossRef]
58. Zhao, M.; Wei, G.; Wei, C.; Wu, J. Improved TODIM Method for Intuitionistic Fuzzy MAGDM Based on Cumulative Prospect Theory and Its Application on Stock Investment Selection. *Int. J. Mach. Learn. Cyber.* **2021**, *12*, 891–901. [CrossRef]

59. McCoy, I.L.; McCoy, D.T.; Wood, R.; Regayre, L.; Watson-Parris, D.; Grosvenor, D.P.; Mulcahy, J.P.; Hu, Y.; Bender, F.A.-M.; Field, P.R.; et al. The Hemispheric Contrast in Cloud Microphysical Properties Constrains Aerosol Forcing. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 18998–19006. [CrossRef] [PubMed]
60. Lappas, P.Z.; Yannacopoulos, A.N. A Machine Learning Approach Combining Expert Knowledge with Genetic Algorithms in Feature Selection for Credit Risk Assessment. *Appl. Soft Comput.* **2021**, *107*, 107391. [CrossRef]
61. Lee, I. Cybersecurity: Risk Management Framework and Investment Cost Analysis. *Bus. Horiz.* **2021**, *64*, 659–671. [CrossRef]
62. Song, X.; Hu, Z.; Du, J.; Sheng, Z. Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud: Evidence from China. *J. Forecast.* **2014**, *33*, 611–626. [CrossRef]
63. Sarkar, S.; Maiti, J. Machine Learning in Occupational Accident Analysis: A Review Using Science Mapping Approach with Citation Network Analysis. *Saf. Sci.* **2020**, *131*, 104900. [CrossRef]
64. Furxhi, I.; Murphy, F.; Mullins, M.; Arvanitis, A.; Poland, C.A. Nanotoxicology Data for in Silico Tools: A Literature Review. *Nanotoxicology* **2020**, *14*, 612–637. [CrossRef] [PubMed]
65. Li, J.-P.; Mirza, N.; Rahat, B.; Xiong, D. Machine Learning and Credit Ratings Prediction in the Age of Fourth Industrial Revolution. *Technol. Forecast. Soc. Change* **2020**, *161*, 120309. [CrossRef]
66. Clarke, R. Principles and Business Processes for Responsible AI. *Comput. Law Secur. Rev.* **2019**, *35*, 410–422. [CrossRef]
67. Rawson, A.; Brito, M.; Sabeur, Z.; Tran-Thanh, L. A Machine Learning Approach for Monitoring Ship Safety in Extreme Weather Events. *Saf. Sci.* **2021**, *141*, 105336. [CrossRef]
68. Adland, R.; Jia, H.; Lode, T.; Skontorp, J. The Value of Meteorological Data in Marine Risk Assessment. *Reliab. Eng. Syst. Saf.* **2021**, *209*, 107480. [CrossRef]
69. Golbayani, P.; Florescu, I.; Chatterjee, R. A Comparative Study of Forecasting Corporate Credit Ratings Using Neural Networks, Support Vector Machines, and Decision Trees. *N. Am. J. Econ. Financ.* **2020**, *54*, 101251. [CrossRef]
70. Ramasamy, L.K.; Khan, F.; Shah, M.; Prasad, B.V.V.S.; Iwendi, C.; Biamba, C. Secure Smart Wearable Computing through Artificial Intelligence-Enabled Internet of Things and Cyber-Physical Systems for Health Monitoring. *Sensors* **2022**, *22*, 1076. [CrossRef]
71. Rahman, M.; Chen, N.; Islam, M.M.; Mahmud, G.I.; Pourghasemi, H.R.; Alam, M.; Rahim, M.A.; Baig, M.A.; Bhattacharjee, A.; Dewan, A. Development of Flood Hazard Map and Emergency Relief Operation System Using Hydrodynamic Modeling and Machine Learning Algorithm. *J. Clean. Prod.* **2021**, *311*, 127594. [CrossRef]
72. CHEN, N. Renamed Chinese Post-Graduate University Holds First Opening Ceremony—Chinese Academy of Sciences. Available online: https://english.cas.cn/newsroom/archive/news_archive/nu2012/201502/t20150215_140273.shtml (accessed on 13 August 2024).
73. Population & Demography Data Explorer. Available online: https://ourworldindata.org/explorers/population-and-demography (accessed on 14 August 2024).
74. KeyWords Plus Generation, Creation, and Changes. Available online: https://support.clarivate.com/ScientificandAcademicResearch/s/article/KeyWords-Plus-generation-creation-and-changes?language=en_US (accessed on 14 August 2024).
75. Rhoades, S.A. The Herfindahl-Hirschman Index. *Fed. Res. Bull.* **1993**, *79*, 188.
76. Bradford, S.C. *Documentation*; C. Lockwood: Monterey, CA, USA, 1953.
77. Bharodia, N.; Chen, W. What Can We Learn from What a Machine Has Learned? Interpreting Credit Risk Machine Learning Models. *JRMV* **2021**, *15*, 1–22. [CrossRef]
78. Chen, D.; Ye, J.; Ye, W. Interpretable Selective Learning in Credit Risk. *Res. Int. Bus. Financ.* **2023**, *65*, 101940. [CrossRef]
79. Cho, S.H.; Shin, K. Feature-Weighted Counterfactual-Based Explanation for Bankruptcy Prediction. *Expert Syst. Appl.* **2023**, *216*, 119390. [CrossRef]
80. Pnevmatikakis, A.; Kanavos, S.; Matikas, G.; Kostopoulou, K.; Cesario, A.; Kyriazakos, S. Risk Assessment for Personalized Health Insurance Based on Real-World Data. *Risks* **2021**, *9*, 46. [CrossRef]
81. Yang, K.; Yuan, H.; Lau, R.Y.K. PsyCredit: An Interpretable Deep Learning-Based Credit Assessment Approach Facilitated by Psychometric Natural Language Processing. *Expert Syst. Appl.* **2022**, *198*, 116847. [CrossRef]
82. Elliott, K.; Price, R.; Shaw, P.; Spiliotopoulos, T.; Ng, M.; Coopamootoo, K.; van Moorsel, A. Towards an Equitable Digital Society: Artificial Intelligence (AI) and Corporate Digital Responsibility (CDR). *Society* **2021**, *58*, 179–188. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Applied Sciences Editorial Office
E-mail: applsci@mdpi.com
www.mdpi.com/journal/applsci



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editor. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editor and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-9268-6