cii Student Papers 2025

Karlsruhe Institute of Technology

aifb

CRITICAL INFORMATION INFRASTRUCTURES RESEARCHGROUP

cii Student Papers - 2025

Research Group Critical Information Infrastructures (cii)
Karlsruhe Institute of Technology
Department of Economics and Management
Institute of Applied Informatics and Formal Description Methods

Web: cii.aifb.kit.edu

Corresponding Editor:
Prof. Dr. Ali Sunyaev
TUS1320 Lehrstuhl für Informationsinfrastrukturen
Bildungscampus 2
74076 Heilbronn, Germany
Phone: +49 7131 26418-121
E-Mail: sunyaev@tum.de

DOI: 10.5445/IR/1000185402

Editorial

Critical information infrastructures (cii) are sociotechnical systems comprising essential software components and information systems with a pivotal impact on individuals, organizations, governments, economies, and society. Bearing this pivotal impact in mind, our research group investigates various research- and practice-driven challenges for cii while looking at the design, development, and evaluation of reliable and secure information systems. The main driver for our research is theorizing on and designing the applications and methods required to create and innovate sociotechnical systems with promising value propositions. With this, we are multifaceted in the use contexts, including the Internet and healthcare industries, as well as industry-specific applications of secure and trustworthy artificial intelligence (AI) models. As we focus on human behavior affecting cii and vice versa, our research enables us to rigorously generate strong theoretical insights while producing research outputs relevant to practical audiences. Our main research contexts are reliable and decentralized information systems within the scope of critical infrastructures, such as blockchain technologies and collaborative decentralized machine learning, digital health systems (e.g., innovative health IT applications), and trustworthy internet-based systems, for example, trustworthy AI and research on the auditing/certification of IT in general.

The last two semesters were both exciting and transformative for our chair. In October 2024, we said farewell to the Karlsruhe Institute of Technology (KIT) after more than six rewarding and fulfilling years and continue our work as part of the Technical University of Munich (TUM), Campus Heilbronn. This transition marks the beginning of a new chapter for our research group, as the Chair of Information Infrastructures at TUM School of Computation and Information Technology (CIT). Yet it is important for us to express our sincere gratitude to the students, colleagues, and community at KIT. The bonds we have formed, the ideas we have developed, and the student journeys we have accompanied are deeply meaningful to us. We cherish the students from KIT that we accompanied throughout the years, whether in our lectures, seminar courses, or as supervisors of their final theses, and are committed to concluding our shared academic journeys with dedication. Thus, we are keen on finishing all started student projects and theses and offer the students the possibility to publish their works in our final student papers miscellany at KIT.

In the past year at KIT, our research group has supervised nearly 150 courseworks and theses of bachelor's and master's students. To us, research is an essential and inseparable part of university education. This is why we follow the research-based teaching and learning paradigm, which allows us to incorporate our research topics directly into students' education and learning experiences. Students benefit from stronger engagement, increased learning performance, and increased skills necessary for life-long learning and the effective dissemination of generated insights and knowledge (Blomster et al., 2014; Christe et al., 2015; Healey, 2005; Nuchwana, 2012; Rueß et al., 2016). We are highly motivated to provide excellent teaching to students, whereby we apply inquiry-based learning methods and actively introduce our research topics to them in various seminars and lectures. As we think that sound research and working in a team go hand in hand, students primarily work in groups during our courses and deal with problems and issues related to sociotechnical challenges in the realm of cii. To ensure high relevance, the course topics generally correspond to what we are currently researching. Students may propose their own research topics or conduct their studies in collaboration with small, medium, or large companies.

Following our cutting-edge information systems research, topics vary from semester to semester. Research topics included but are not limited to disruptive health information systems (Thiebes et al., 2023) and shaping the future of the digital health sector (Sunyaev, Fürstenau, et al., 2024), the secure design of cloud, fog, and edge services (Blume et al., 2023; Brecker et al., 2023), task-congruence in gamified healthcare information systems (Schmidt-Kraepelin et al., 2024), the evaluation of AI explanations for industry experts (Toussaint, Warsinsky, et al., 2024), accountable AI (Du et al., 2024; Nguyen et al., 2024), designing and implementing requirements for distributed ledgers (Leinweber et al., 2023), adoption and trust concerns regarding the use of AI in autonomous vehicles (Renner et al., 2023), the effects of trust in organizational security practices and protective structures on employees' security-related precaution-taking (Greulich et al., 2024), the emergence and consequences of consumer skepticism toward web seals (Lins et al., 2024), and theory development for transparency of information privacy practices (Dehling & Sunyaev, 2023, 2024). Our team supports students throughout the research process, helping them identify and organize problems, apply appropriate research methods consistently, develop and communicate approaches to solutions, and write research papers.

Involving students in daily work and bringing research to students provides many benefits to the students, our research group, the research community, and practice in general. Students engage with present-day practice problems that research is trying to solve. Moreover, they can apply the theoretical principles and knowledge acquired in previous lectures while working on their seminar papers, deepening their understanding. By offering research-based learning courses, students can gain first-hand experience in self-reliant research and scientific writing and benefit from their now enhanced skillset for writing upcoming seminars, bachelor's, and master's theses. Consequently, we believe our students' works are of high value. Nonetheless, in the past, only few students continued their research after attending a seminar or a lecture, and their works often disappeared into drawers despite the disruptive and valuable insights students have come up with. As a research group, we always appreciate the work of students and started publishing the best works in a miscellany dedicated to making them available to a broader audience. Our previous collections (see Sunyaev, Du, et al., 2024; Sunyaev et al., 2021, 2022, 2023) have encouraged students to continue their research in various forms, including volunteer work in their free time, as part of their work as a student assistant in our research group, and even pursuing a PhD. Students, furthermore, regularly interact with organizations during our courses, which can lead to collaborations and even pave the way for upcoming jobs. Moreover, it encourages great students to incorporate their insights into our research and, sometimes, together with these students, we advance and publish exceptional results in conference proceedings or journals (e.g., Bodynek et al., 2023; Drossos et al., 2025; Furmanek et al., 2024; Hasse et al., 2024; Hofmann et al., 2024; Hu et al., 2023; Jeck et al., 2025; Sproll et al., 2025; Zeller et al., 2025).

In the spirit of making students' works available, we continue the idea of publishing the best student works from our courses and are delighted to present this collection for the fifth time in a row. In this work, we bring together the best student works from the summer term of 2024 and the winter term of 2024/25. Contributions in this anthology come from two different courses that provide students with a broad range of topics related to cii:

Emerging Trends in Digital Health:

The seminar *Emerging Trends in Digital Health* aims to provide insights into current topics in the field of information systems with a focus on innovative digital healthcare systems. Students can choose to work on many different topics around the lectures and research topics of the research group, including genomics (Thiebes, Toussaint, et al., 2020), distributed ledger technology (Beyene et al., 2022; Hu et al., 2024), AI (Leiser et al., 2023; Thiebes et al., 2021), digital transformation and employee-driven digital innovation in the healthcare sector (Guse et al., 2022, 2024, 2025), and gamification in healthcare (Schmidt-Kraepelin et al., 2023). An example of our interdisciplinary work in this field is a recent systematic mapping study on explainable AI for omics data. The study investigates current machine learning approaches for biomedical data and applied explainable AI methods by systematically analyzing extant literature. In doing so, the study provides a research discipline-spanning overview and identifies open shortcomings of explainable AI for omics data and suggests several future research directions (Toussaint, Leiser, et al., 2024).

Digital Health:

The course *Digital Health* introduces master's students to digitization in healthcare. Students learn about the theoretical foundations and practical implications of various topics surrounding digitization in healthcare, including health information systems, telematics, big healthcare data, and patient-centered healthcare (e.g., Guse et al., 2022; Pandl et al., 2021; Rädsch et al., 2021; Thiebes, Schlesner, et al., 2020; Warsinsky et al., 2021). After an introductory session on the challenge of digital transformation in healthcare, the following sessions focus on an in-depth exploration of selected topics that represent current challenges in research and practice. Students work in groups of three to four on specific topics and must write a course paper. One recent example of our contributions to advancing knowledge in the field of digital health is a scenario-based factorial survey on the privacy-utility trade-offs in genetic data sharing (Thiebes et al., 2024). Genetic data sharing raises privacy risks not only for individuals but also for their families and friends. While most studies focus on personal privacy-utility trade-offs, our research shows that interdependent trade-offs also shape disclosure decisions. The study finds clear effects of both personal and interdependent considerations, though the influence of social distance remains uncertain.

Student Works in this Miscellany

We selected the student works representing excellent and intriguing studies from these courses. The student works in this book cover a wide range of research problems, including an investigation of privacy policies in direct-to-consumer genetic testing and a framework for their automated analysis, a systematic mapping study of explainable AI systems, and a structured literature review unraveling the role of adaptive gamification in digital health interventions.

- Aldag, Kocker, Lei, and Pietsch investigate the role of adaptive gamification in digital health interventions, such as fitness and calorie tracking apps, by focusing on the personalization of gamified elements based on user traits and context. Through a structured literature review, they analyze how gamification strategies interact with user characteristics like the Hexad User Types and Big Five Personality Traits. Results suggest that personalized gamification can enhance engagement and motivation, supporting health behavior change. However, the study also identifies a lack of long-term evaluations and context-specific adaptations, highlighting directions for future research.
- Gänsauer, Weinreuter, Ben Aoun, and Pietsch perform a systematic mapping study analyzing 780 studies. They categorize 27 evaluation-ready AI systems by key characteristics, including data modality, explainability technique, and evaluation method. Their findings reveal a discrepancy in the usage of feature relevance techniques, which remain underutilized in genomics and pharmacology, and visual explainability methods dominate imaging-based models. They conclude that there is a need for standardized evaluation metrics and cross-domain benchmarks for such systems.
- Faust presents a framework for the automated analysis of privacy policies in the direct-to-consumer genetic testing (DTC-GT) sector using the Longformer model. By categorizing policy content into 22 DTC-GT-specific categories and evaluating extraction quality across multiple dimensions, the student highlights both the potential and limitations of current privacy policies. While the model shows fair to excellent performance in extracting information from well-structured categories (e.g., company sale, data deletion), it struggles with vague or incomplete policies, particularly in areas like research and data storage. The findings underscore ongoing privacy concerns and the need for greater clarity and standardization in DTC-GT privacy disclosures.

We are grateful to the students who revised and improved their work for this publication and to our dedicated research group members who mentored them. Your commitment to academic excellence makes this publication possible.

Farewell

To our students at KIT: We sincerely thank you for your energy, your curiosity, and the trust you placed in us throughout the years. Teaching and learning with you has been one of the most rewarding parts of our journey. As we move on to TUM, we carry with us not only the research and coursework we shared but, above all, the memories of your dedication, creativity, and growth. You will always be an important part of our story, and we hope you continue to pursue your paths with courage and passion.

To our colleagues and friends at KIT, to whom we are grateful for the collaboration in supervising excellent students, in shared projects, and in publishing joint research. The inspiring discussions with you and the friendships that we developed at KIT enriched our work and shaped our thinking. Your support and shared commitment to advancing research and teaching have left a lasting mark on us.

Although our journey now continues at TUM, a part of us will always remain at KIT, in the connections we built, the knowledge we shared, and the memories we carry forward. We are committed to honoring this legacy by continuing to support and publish excellent student work, now at our new academic home in Heilbronn. We are looking forward to building new collaborations, mentoring new generations of students, and continuing to explore the challenges and opportunities of sociotechnical systems in the years to come.

Sincerely,

Ali Sunyaev, Benjamin Sturm, Guangyu Du, Manuel Schmidt-Kraepelin, Niclas Kannengießer, Philipp A. Toussaint, Scott Thiebes, Yannick Heß

Miscellany Team 2025

Supervising Research Associates

References

Beyene, M., Toussaint, P. A., Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2022). A Scoping Review of Distributed Ledger Technology in Genomics: Thematic Analysis and Directions for Future Research. *J Am Med Inform Assoc*, *29*(8), 1433–1444. **https://doi.org/10.1093/jamia/ocac077**

Blomster, J., Venn, S., & Virtanen, V. (2014). Towards Developing a Common Conception of Research-Based Teaching and Learning in an Academic Community. *Higher Education Studies*, *4*(4), 62–75. **https://doi.org/10.5539/hes.v4n4p62**

Blume, M., Lins, S., & Sunyaev, A. (2023). Uncovering Effective Roles and Tasks for Fog Systems. In G. A. Papadopoulos, F. Rademacher, & J. Soldani (Eds.), *Service-Oriented and Cloud Computing*. *ESOCC 2023*. *Lecture Notes in Computer Science* (Vol. 14183, pp. 119–135). Springer. **https://doi.org/10.1007/978-3-031-46235-1_8**

Bodynek, M., Leiser, F., Thiebes, S., & Sunyaev, A. (2023). Applying Random Forests in Federated Learning: A Synthesis of Aggregation Techniques. *Wirtschaftsinformatik 2023 Proceedings*, Article 46. 18th International Conference on Wirtschaftsinformatik (WI2023), Paderborn, Germany. **https://aisel.aisnet.org/wi2023/46**

Brecker, K., Lins, S., Trenz, M., & Sunyaev, A. (2023). Artificial Intelligence as a Service: Trade-Offs Impacting Service Design and Selection. *Proceedings of the 44th International Conference on Information Systems (ICIS)*.

Christe, D., Shah, A., Bhatt, J., Powell, L., & Kontsos, A. (2015). Raising interest in STEM education: A research-based learning framework. *2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services*, 167–169. **https://doi.org/10.1109/ETTLIS.2015.7048192**

Dehling, T., & Sunyaev, A. (2023). A Design Theory for Transparency of Information Privacy Practices. *Information Systems Research*, *Articles in Advance*, 1–22. **https://doi.org/10.1287/isre.2019.0239**

Dehling, T., & Sunyaev, A. (2024). *A design theory for transparency of information privacy practices : [Appendix]*. **https://doi.org/10.5445/IR/1000170914**

Drossos, T., Kirste, D., Kannengießer, N., & Sunyaev, A. (2025). Automated market makers: toward more profitable liquidity provisioning strategies. *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 358–365. **https://doi.org/10.1145/3672608.3707833**

Du, G., Lins, S., Blohm, I., & Sunyaev, A. (2024). *My Fault, Not AI's Fault. Self-Serving Bias Impacts Employees' Attribution of AI Accountability*. 45th International Conference on Information Systems, Bangkok, Thailand.

Furmanek, L., Lins, S., Blume, M., & Sunyaev, A. (2024). Developing a Hybrid Deployment Model for Highly Available Manufacturing Execution Systems. *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 2095–2100. **https://doi.org/10.1109/MIPRO60963.2024.10569530**

Greulich, M., Lins, S., Pienta, D., Thatcher, J. B., & Sunyaev, A. (2024). Exploring Contrasting Effects of Trust in Organizational Security Practices and Protective Structures on Employees' Security-Related Precaution Taking. *Information Systems Research*, *Articles in Advance*, 1–23. **https://doi.org/10.1287/isre.2021.0528**

Guse, R., Thiebes, S., Hennel, P., Rosenkranz, C., & Sunyaev, A. (2022). How Do Employees Perceive Digital Transformation and its Effects? A Theory of the Smart Machine Perspective. *ICIS 2022 Proceedings*. International Conference on Information Systems (ICIS) 2022, Copenhagen, Denmark. **https://aisel.aisnet.org/icis2022/digit_nxt_gen/digit_nxt_gen/6**

Guse, R., Thiebes, S., Winterhoff, P., Alzate, M. V., Stangier, J., & Sunyaev, A. (2024). The development of value proposition in healthcare in the course of digital transformation. *Academy of Management Proceedings*, *2024*(1), 16562. **https://doi.org/10.5465/AMPROC.2024.16562abstract**

Guse, R., Warsinsky, S., Thiebes, S., & Sunyaev, A. (2025). *Employee-driven digital innovation in healthcare – a scoping review*. Hawaii International Conference on System Sciences. **https://doi.org/10.24251/HICSS.2025.424**

Hasse, F., Leiser, F., & Sunyaev, A. (2024). Informed machine learning for cardiomegaly detection in chest X-rays: a comparative study. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. **https://doi.org/10.1109/ISBI56570.2024.10635719**

Healey, M. (2005). Linking research and teaching: exploring disciplinary spaces and the role of inquiry-based learning. In R. Barnett (Ed.), *Reshaping the University: New Relationships between Research, Scholarship and Teaching* (pp. 67–78). McGraw Hill / Open University Press.

Hofmann, P., Brand, A., Späthe, E., Lins, S., & Sunyaev, A. (2024). AI-based Tools in Higher Education: A Comparative Analysis of University Guidelines. *Proceedings of Mensch Und Computer 2024*, 665–673. **https://doi.org/10.1145/3670653.3677513**

Hu, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2024). Mapping Distributed Ledger Technology Characteristics to Use Cases in Healthcare: A Structured Literature Review. *ACM Transactions on Computing for Healthcare*, *5*(3), Article 15. **https://doi.org/10.1145/3653076**

Hu, S., Usta, A., Schmidt-Kraepelin, M., Warsinsky, S., Thiebes, S., & Sunyaev, A. (2023). *Be Mindful of User Preferences: An Explorative Study on Game Design Elements in Mindfulness Applications*. (No. 15). Article 15. Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, USA.

Jeck, J., Leiser, F., Hüsges, A., & Sunyaev, A. (2025). TELL-ME: toward personalized explanations of large language models. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–18. **https://doi.org/10.1145/3706599.3719982**

Leinweber, M., Kannengießer, N., Hartenstein, H., & Sunyaev, A. (2023). Leveraging Distributed Ledger Technology for Decentralized Mobility-as-a-Service Ticket Systems. In H. Proff (Ed.), *Towards the New Normal in Mobility* (pp. 547–567). Springer Fachmedien Wiesbaden. **https://doi.org/10.1007/978-3-658-39438-7_32**

Leiser, F., Rank, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2023). Medical informed machine learning: A scoping review and future research directions. *Artificial Intelligence in Medicine*, *145*, Article 102676. **https://www.sciencedirect.com/science/article/pii/S0933365723001902**

Lins, S., Greulich, M., Löbbers, J., Benlian, A., & Sunyaev, A. (2024). Why So Skeptical? Investigating the Emergence and Consequences of Consumer Skepticism toward Web Seals. *Information & Management*, *61*(2), Article 103920. **https://doi.org/10.1016/j.im.2024.103920**

Nguyen, L. H., Lins, S., Renner, M., & Sunyaev, A. (2024). Unraveling the Nuances of AI Accountability: A Synthesis of Dimensions Across Disciplines. *ECIS 2024 Proceedings*. 32nd European Conference on Information Systems, Paphos, Cyprus.

Nuchwana, L. (2012). How to Link Teaching and Research to Enhance Students' Learning Outcomes: Thai University Experience. *Procedia - Social and Behavioral Sciences*, *69*, 213–219. **https://doi.org/10.1016/j.sbspro.2012.11.401**

Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2021). How Detection Ranges and Usage Stops Impact Digital Contact Tracing Effectiveness for COVID-19. *Sci Rep*, *11*(1), Article 9414. **https://doi.org/10.1038/s41598-021-88768-6**

Rädsch, T., Eckhardt, S., Leiser, F., Pandl, K. D., Thiebes, S., & Sunyaev, A. (2021). What Your Radiologist Might be Missing: Using Machine Learning to Identify Mislabeled Instances of X-ray Images. *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*, 1294–1303. **https://hdl.handle.net/10125/70769**

Renner, M., Lins, S., Söllner, M., Jarvenpaa, S., & Sunyaev, A. (2023). Artificial Intelligence-Driven Convergence and its Moderating Effect on Multi-Source Trust Transfer. *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS)*, 5208–5217. **https://hdl.handle.net/10125/103271**

Rueß, J., Gess, C., & Deicke, W. (2016). Forschendes Lernen und forschungsbezogene Lehre - empirisch gestützte Systematisierung des Forschungsbezugs hochschulischer Lehre. *Zeitschrift Für Hochschulentwicklung*, *11*(2), 23–44. **https://doi.org/10.3217/ZFHE-11-02/02**

Schmidt-Kraepelin, M., Ben Ayed, M., Warsinsky, S., Hu, S., Thiebes, S., & Sunyaev, A. (2024). Leaderboards in Gamified Information Systems for Health Behavior Change: The Role of Positioning, Psychological Needs, and Gamification User Types. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*, 3444–3453. **https://hdl.handle.net/10125/106800**

Schmidt-Kraepelin, M., Thiebes, S., Warsinsky, S. L., Petter, S., & Sunyaev, A. (2023, April 19). Narrative Transportation in Gamified Information Systems: The Role of Narrative-Task Congruence. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 215. 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany. **https://doi.org/10.1145/3544549.3585595**

Sproll, Y., Heinrich, R., Quang Le, L. B., & Kannengießer, N. (2025). SM-SIM: a simulator for analyzing selfish mining attacks in blockchain systems. *2025 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 1–9. **https://doi.org/10.1109/ICBC64466.2025.11114629**

Sunyaev, A., Du, G., Renner, M., Toussaint, P. A., Thiebes, S., Lins, S., & Erb, Y. (Eds.). (2024). *cii student papers - 2024*. Karlsruher Institut für Technologie (KIT). **https://doi.org/10.5445/IR/1000173991**

Sunyaev, A., Fürstenau, D., & Davidson, E. (2024). Reimagining digital health: advances in patient-centeredness, artificial intelligence, and data-driven research. *Business & Information Systems Engineering*, *66*(3), 249–260. **https://doi.org/10.1007/s12599-024-00870-x**

Sunyaev, A., Renner, M., Toussaint, P. A., Thiebes, S., & Lins, S. (Eds.). (2021). *cii Student Papers - 2021*. Karlsruher Institut für Technologie (KIT). **https://doi.org/10.5445/IR/1000138902**

Sunyaev, A., Renner, M., Toussaint, P. A., Thiebes, S., & Lins, S. (Eds.). (2022). *cii Student Papers - 2022*. Karlsruher Institut für Technologie (KIT). **https://doi.org/10.5445/IR/1000150078**

Sunyaev, A., Renner, M., Toussaint, P. A., Thiebes, S., & Lins, S. (Eds.). (2023). *cii Student Papers - 2023* [PDF]. Karlsruher Institut für Technologie (KIT). **https://doi.org/10.5445/IR/1000162178**

Thiebes, S., Gao, F., Briggs, R. O., Schmidt-Kraepelin, M., & Sunyaev, A. (2023). Design Concerns for Multiorganizational, Multistakeholder Collaboration: A Study in the Healthcare Industry. *Journal of Management Information Systems*, *40*(1), 239–270. **https://doi.org/10.1080/07421222.2023.2172771**

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy Artificial Intelligence. *Electronic Markets*, *31*(2), 447–464. **https://doi.org/10.1007/s12525-020-00441-4**

Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2020). Distributed Ledger Technology in Genomics: A Call for Europe. *European Journal of Human Genetics*, *28*, 139–140. **https://doi.org/10.1038/s41431-019-0512-4**

Thiebes, S., Schmidt-Kraepelin, M., Toussaint, P. A., & Lyytinen, K. (2024). Privacy-Utility Trade-Offs in Genetic Data Sharing and the Moderating Role of Social Distance: An Interdependent Privacy Calculus. *ICIS 2024 Proceedings*, Article 8. International Conference on Information Systems (ICIS) 2024, Bangkok, Thailand. **https://aisel.aisnet.org/icis2024/security/security/8/**

Thiebes, S., Toussaint, P. A., Ju, J., Ahn, J. H., Lyytinen, K., & Sunyaev, A. (2020). Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing. *J Med Internet Res*, *22*(1), Article e14890. **https://doi.org/10.2196/14890**

Toussaint, P. A., Leiser, F., Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2024). Explainable artificial intelligence for omics data: a systematic mapping study. *Briefings in Bioinformatics*, *25*(1), Article bbad453. **https://doi.org/10.1093/bib/bbad453**

Toussaint, P. A., Warsinsky, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2024). Designing Gamification Concepts for Expert Explainable Artificial Intelligence Evaluation Tasks: A Problem Space Exploration. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*, 1338–1347. **https://hdl.handle.net/10125/106542**

Warsinsky, S., Schmidt-Kraepelin, M., Thiebes, S., & Sunyaev, A. (2021). Are Gamification Projects Different? An Exploratory Study on Software Project Risks for Gamified Health Behavior Change Support Systems. *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*, 1305–1314. **http://hdl.handle.net/10125/70771**

Zeller, S. C., Kandora, P.-N. K., Kirste, D., Kannengießer, N., Rebennack, S., & Sunyaev, A. (2025). Automated market makers: a stochastic optimization approach for profitable liquidity concentration. *2025 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 1–9. **https://doi.org/10.1109/ICBC64466.2025.11114638**

Table of Contents

Editorial ..... **I**

Ali Sunyaev, Benjamin Sturm, Guangyu Du, Manuel Schmidt-Kraepelin, Niclas Kannengießer, Philipp A. Toussaint, Scott Thiebes, Yannick Heß

How do Adaptive Gamified Health Information Systems Affect Health Behaviour Change? **1**

Kolja Aldag, Lucia Kocker, Yuqi Lei, Felix Pietsch

Explainable Artificial Intelligence for Biomedical Data: A Systematic Mapping Study **24**

Robin Gansäuer, Maria Weinreuter, Hichem Ben Aoun, Felix Pietsch

Privacy Policy Feature Extraction for Direct-to-Consumer Genetic Testing **42**

Luisa Faust

How do Adaptive Gamified Health Information Systems Affect Health Behaviour Change?

Emerging Trends in Digital Health, Summer Term 2024

Kolja Aldag
Master's Student
Karlsruhe Institute of Technology
kolja.aldag@student.kit.edu

Lucia Kocker
Master's Student
Karlsruhe Institute of Technology
lucia.kocker@student.kit.edu

Yuqi Lei
Master's Student
Karlsruhe Institute of Technology
yuqi.lei@student.kit.edu

Felix Pietsch
Master's Student
Karlsruhe Institute of Technology
felix.pietsch@student.kit.edu

Abstract

Background: *Digital health interventions, such as fitness apps and calorie tracking applications, offer promising tools to support health behaviour change. To make the process more enjoyable, gamification is a commonly used feature in digital health interventions.*

Objective: *This paper explores the concept of adaptive gamification and its influence on health behaviour change, focusing on the personalization of gamified elements to user-specific and context-specific factors.*

Methods: *Through a structured literature review, key gamification elements were identified, and their interactions with user characteristics, specifically the Hexad User Types and the Big Five Personality Traits were analysed.*

Results: *The findings highlight that for certain individuals, tailored gamification strategies can significantly influence user engagement and motivation, which are critical for successful health behaviour interventions. However, gaps remain in the long-term assessment of these interventions and the exploration of context-specific adaptations in the context of digital health. The potential for future research to fill these gaps by conducting long-term studies and investigating their effectiveness in various use-specific scenarios is identified.*

Conclusion: *This study provides valuable insights for developers of health apps and digital health interventions, emphasizing the potential of personalized gamification to change health outcomes.*

Keywords: adaptive gamification, digital health interventions, health behavior change, personalization, hexad user types, big five personality traits, health information systems

Introduction

Obesity and higher-than-optimal BMI have caused about 5 million deaths in 2019. As in 2022, about one in eight adults worldwide is overweight (World Health Organization, 2024). Apart from tragic health consequences, obesity also poses a large weight upon the world's economies. Research predicts the global cost of obesity and overweight to reach $3 trillion per year by 2030 and more than $18 trillion by 2060 (Okunogbe et al., 2021). These facts and figures stress the importance of effective and widely available health interventions to address this growing health issue. Due to their availability and popularity, digital health interventions offer promising tools, such as fitness apps or calorie tracking applications.

Gamification in health apps is a commonly used feature to address the loss of intrinsic motivation. It incorporates elements such as points, leaderboards and badges to apply a more playful approach to otherwise mundane tasks (Johnson et al., 2016). Current research has been questioning the frequently used "one size fits all"- approach to gamification (Nacke & Deterding, 2017). Instead, rather adaptive and personalized gamification approaches have been suggested, addressing user- or context specific factors and including them in the gamification design (Tondello, 2019).

Context-specific customization focuses on the application context, specifically, for example, whether it is a digital health intervention in mental health or physical fitness. User-specific adaptation focuses on customization based on the users and features that characterize them. For example, their age (Sandford-james et al., 2022), fitness level (Yang et al., 2023) or personality (Ciocarlan et al., 2018). In research, personality is examined in more detail, with two concepts taking center stage: Hexad User Types and the Big Five Personalities. Due to the presence of these concepts in research, this paper also focuses on these theories.

While the effects of adaptive gamification in the context of digital health have been investigated in several studies, there exists no review that combines the findings of various studies to a comprehensive overview analysing the interactions between adaptive gamification elements and user-specific customization in the realm of digital health. This paper aims to address the research gap by exploring the effects of adaptive gamified information systems on health behaviour change. Therefore, the following subgoals are explored:

(1) Identification of "adaptive" elements in existing gamified information systems in the context of digital health
(2) Analysis of user-specific factors, such as age or personality type, influencing adaptive elements
(3) Analysis context-specific factors (e.g., mindfulness application, fitness) defining the overall frame of the gamified health information system and their combination with adaptive gamified elements
(4) Evaluation of influence of adaptive elements on the outcome of health behaviour change (also in comparison to non-adaptive gamification approaches)

Following the introduction, the theoretical background on gamification is laid, especially looking into its elements. Moreover, approaches to user specific factors, specifically the Hexad User Types and the Big Five Personalities are presented. Afterwards the methodology of the structured literature review is explained. The results on the main characteristics of the literature found, the gamification elements used as well as the user specific factors are presented. Consequently, as a key finding, the identified interaction between the user specific factors (i.e., personality types) and gamification elements is explained. The results part is followed by the discussion, addressing principal findings, potential for future research as well as implications and limitations, followed up by a brief conclusion.

Theoretical Background

Health Behaviour Change in Digital Health

Health behaviour includes personal attributes like beliefs, values, emotions, and habits related to health maintenance, influenced by family, social, cultural, and institutional factors (Gochman, 1982). Health behaviour change refers to the deliberate actions that individuals take to adopt a healthier lifestyle for their overall well-being. In order to explore why individuals engage in certain health behaviours and how to promote healthier ones, Health Behaviour Change Theory (HBCT) provides various explanations, considering psychological, social, and environmental factors. Key theories are discussed below.

Health Belief Model (HBM) posits that individuals are more likely to engage in health-promoting behaviours if they believe they are at risk of a serious health problem and believe that taking specific actions can reduce this risk. Key constructs include perceived susceptibility, severity, benefits, barriers, cues to action, and self-efficacy (Rosenstock, 1966).

Transtheoretical Model (TTM) describes behaviour change as a process that involves progress through five stages: precontemplation, contemplation, preparation, action, and maintenance. Tailoring interventions to an individual's current stage could enhance their effectiveness (Prochaska & DiClemente, 1983).

Self-Determination Theory (SDT) focuses on intrinsic and extrinsic motivation. It defines three basic psychological needs that support intrinsic motivation, being autonomy, competence, and relatedness. When these needs are met, individuals are more likely to be motivated and engaged in behaviour change (Deci & Ryan, 1985).

Persuasive strategies are essential for designing effective digital health interventions. Techniques from frameworks such as the Persuasive Systems Design (PSD) model by Oinas-Kukkonen and Harjumaa are widely used. The PSD model is comprehensive, combining various strategies to guide the design of persuasive gamified systems, thereby promoting positive behaviour change (Oinas-Kukkonen & Harjumaa, 2009). Gamified health behaviour change interventions often are based on HBCTs. Therefore, HBCTs are commonly integrated into current research and health behaviour change interventions can also be found within interventions examined in this review.

Gamification

Gamification is defined as the use of elements of game design in a non-game context, such as health, productivity, finance, education, or any other domain. It should be emphasised that this does not involve whole game technologies or full games, but only design elements in any non-game context in various ways and arrangements (Deterding et al., 2011). When gamification is adapted to the characteristics of the individual user or the context in which gamification takes place, it is called adaptive gamification (Böckle et al., 2017). Personalized, user-centred and adaptive mechanisms are used. There are several game elements that are used in gamification systems. To categorize the gamification elements, the classification of Koivisto et. al will be used. The gamification elements are grouped into the categories "Achievement/Progression" including points, challenges and leaderboards, "Social" including cooperation and competition, "Immersion" including avatars, virtual words, and in-game rewards, "Non-digital elements" including financial rewards or motion tracking as well as "Miscellaneous elements" such as virtual currency and game rounds (Koivisto & Hamari, 2019).

User-Specific Factors

The Big Five Personalities is a framework that comes from psychology and is not specifically designed for gamification as the factors were identified in the 1990s (Costa & McCrae, 1992). The Big Five Personalities measure five different dimensions of personality traits that describe an individual's character. The dimensions are described as follows:

- People with high scores on **neuroticism** are considered as worrying, emotional and vulnerable, whereas people with low scores are emotionally stable, calm, unemotional and hardy.
- **Extraversion** shows the differences between personalities with a high social activity, being talkative, fun loving and active, and introverts that are quiet, sober and passive.
- High scores on **openness to experience** can be seen on personalities that are imaginative, creative and curious, whereas closedness refers to uncreative, uncurious and conservative personalities.
- The personality trait of **agreeableness** differentiates between people that are soft-hearted, trusting and generous, and ruthless, suspicious and stingy people.
- **Conscientiousness** shows the distinction between the will to achieve vs. undirectedness. People with high scores on conscientiousness are considered as hardworking, well-organized, punctual and ambitious. People with low scores are lazy, disorganized, late and aimless (Costa & McCrae, 1992).

The Hexad User Types are a model for distinguishing between different player types. It is derived from SDT, which provides insights into behavioural motivation (Ryan & Deci, 2000). The six types are built on these motivators, four of them are intrinsically, two are extrinsically motivated:

- Relatedness drives **Socialisers**, i.e., their desire to engage in social interactions and form connections.
- The pursuit of independence and self-expression drives **Free Spirits**. They are artistic and curious.
- Achievers are driven by mastery. They aspire to grow intellectually, acquire new abilities, and improve themselves. They desire to overcome challenges.
- **Philanthropists** are driven by a sense of purpose and meaning. They have a selfless desire to help others and improve their lives without expecting anything in return.
- Change serves as an extrinsic motivator for **Disruptors**. They aim to induce change by directly upsetting the system or via other users.
- Extrinsic rewards are what motivates **Players**. They will only take actions required to obtain rewards from a system, without doing anything more (Marczewski, 2015).

Both models are commonly used in adaptive gamification to distinguish between different individuals and tailor the gamification design based on their characteristics.

Methodology

Data Collection

The data collection followed a structured literature review approach, outlined in Figure 1. A "TITLE-ABS-KEY"-search with the following search string was conducted in the databases Scopus and PubMed: *TITLE-ABS-KEY((Gamif*) AND (Adaptive OR Individual OR Personal* OR Tailored) AND (Health OR Wearables OR Fitness OR Sports OR Wellness OR Wellbeing))*.

While PubMed has been chosen because of its background in life sciences and biomedicine, Scopus was used to get coverage of a broader field of research. Having entered the search string into both databases, 1805 results have been obtained. After removal of duplicates and papers not in English, 1400 papers have remained for further analysis within the relevance coding process. Literature that met at least one of the following exclusion criteria, was excluded from the review: "Grey Literature", "no health context", "no mention of gamification" or "no focus on adaptiveness".

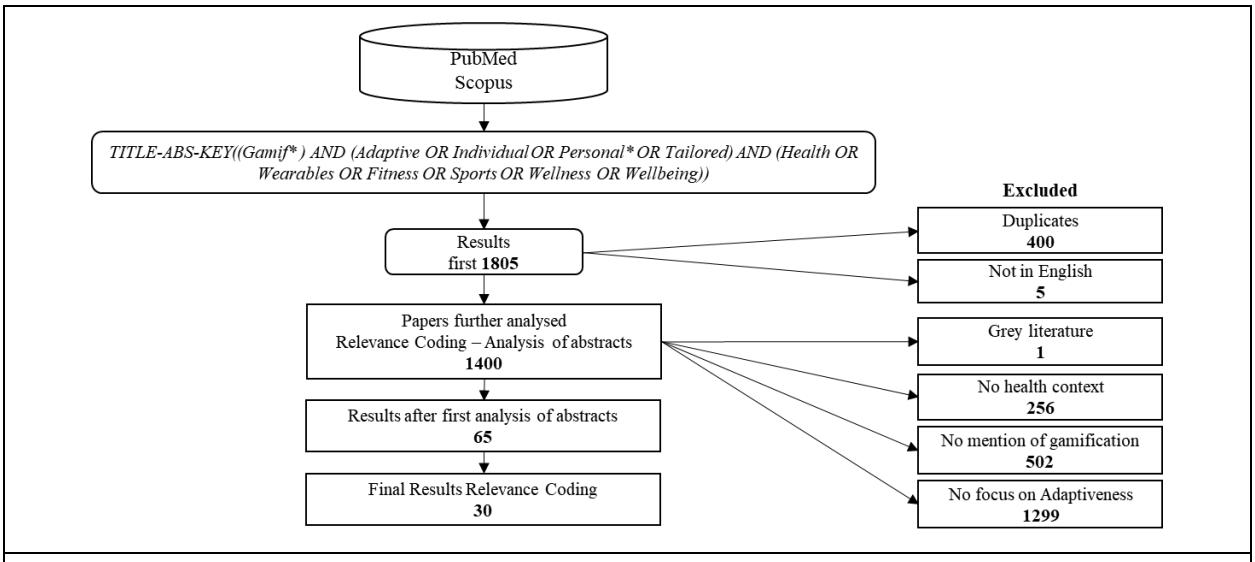

Figure 1. Data collection

The first 20 abstracts were read by four individuals each to gain a common understanding of the exclusion criteria. Afterwards, each abstract has been read and checked for the exclusion criteria by one individual. 65 papers were initially found to be relevant. Out of those 65 papers, each abstract has been checked a

second time, this time by four individuals, resulting in 30 papers remaining. Out of those papers, six papers were later identified as not completely fitting and therefore not used in the analysis.

Data Analysis

The data analysis followed the process outlined by Webster and Watson (2002). All results obtained in the previous step, data collection, have been transferred to a concept coding matrix (see appendix A1 for a shortened version, the detailed table is available by the authors upon request.). The table comprises several topics and elements such as the empirical approach, health context, user specific factors, adaptive elements and the overall influence of adaptive gamification. Through an in depth-examination of each paper by four individuals, the populated fields were discussed and filled with information. This allowed a quick and comprehensive overview of the investigated topics. Especially the connection between the adaptive gamified elements and the user specific factors and also between the context specific factors became more prevalent and will be discussed in the following chapter.

Results

Main Characteristics of Literature Found

An analysis of the bibliographic data of the 30 selected papers published between 2013 and 2024 was performed. As shown in Figure 2, there was a significant increase in the number of papers published after 2015, with six papers related to the topic published in 2023. This reflects the increased academic interest in this area.

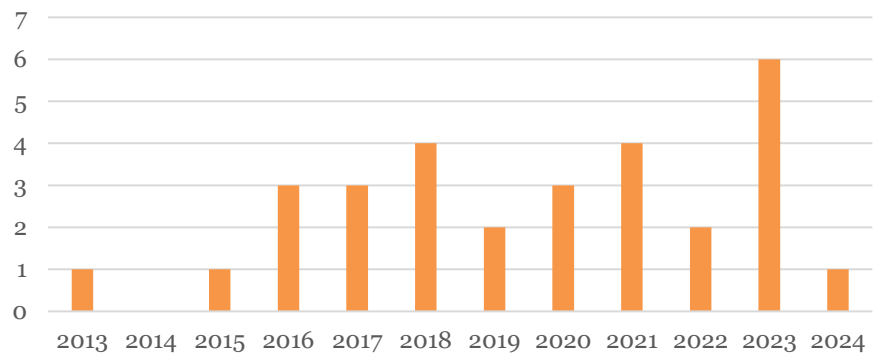

Figure 2. Years of Publication

Additionally, as shown in Figure 3, the most frequently used empirical approach among the selected papers is the mixed-method approach, which usually combines qualitative and quantitative approaches, with a total of 13 papers using this method. The second most frequent approach is quantitative, utilized in 11 papers. Some papers also adopt conceptual methods or conduct literature reviews. Notably, there are no papers that exclusively use qualitative methods.

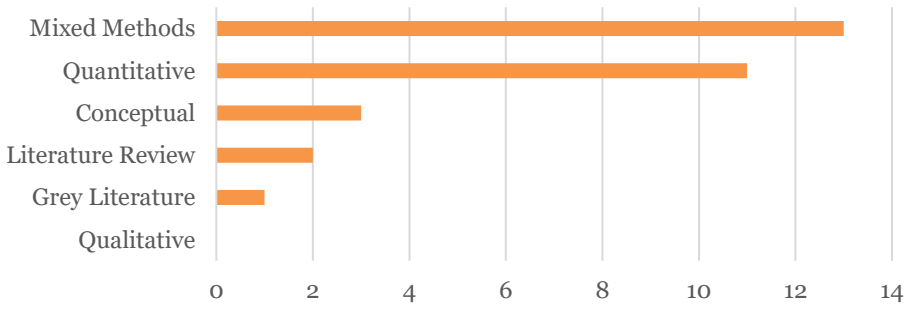

Figure 3. Empirical Approach

In order to address subgoal (3), the targeted health-context in each paper has been analysed. As shown in Figure 4, apart from six studies that did not specify any specific health context, the remaining papers covered six different types of health contexts: physical activity, mental health, cognitive function, healthy eating, health education and smoking cessation. Some studies did focus on multiple health behaviours.

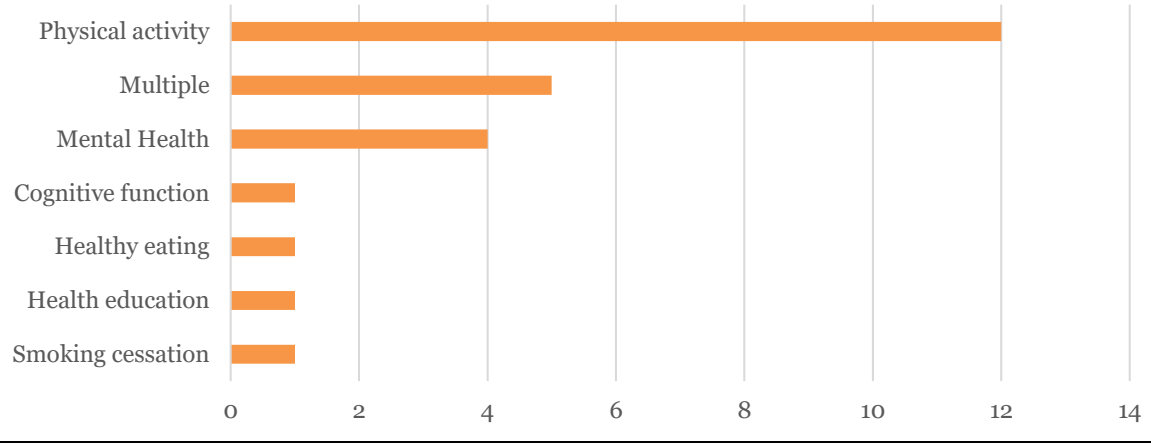

Figure 4. Analysis of Context-Specific Factors

Physical activity emerged as the most researched health context with twelve papers addressing it. A more detailed classification of the twelve papers targeting physical activity has been conducted, as shown in Figure 5. Among them, four papers specifically focused on fitness, making it the most targeted health behaviour, followed closely by multiple targeted behaviours and mental health, which also received more attention.

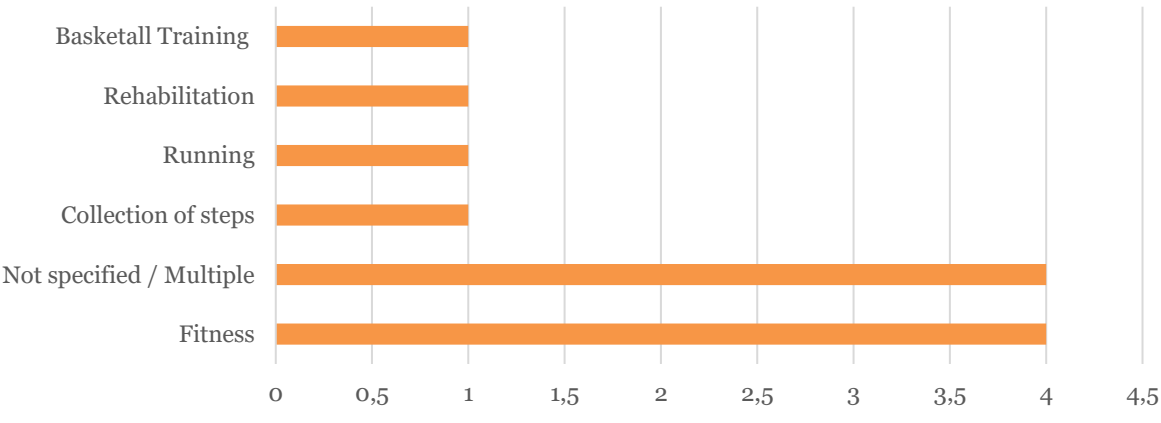

Figure 5. Analysis of Physical Activity

Main Characteristics of Literature Found

To address subgoal (1), the gamification elements utilized in each paper were identified during the literature analysis process. A total of 86 different elements were observed. Table 1 shows the most used elements. The column “Total” shows the total times each element was used. The column “Mentions by Paper” shows in which paper the element has been used. The numbers in the cell correspond to the literature IDs. A breakdown of corresponding ID to paper can be found in appendix A1. The entire overview of all elements can be found in appendix A2.

An element has been added to the overview, once it has been used in an analysis in the research paper, regardless of whether the research examined has found any correlations between the element and the correlating factor (e.g., personality type).

Element	Total	Mentions by paper (ID)	Element	Total	Mentions by paper (ID)
Rewards	15	2, 4, 6, 7, 15, 18, 21, 23, 24, 25, 26, 27, 28, 29, 30	Unlockable Content, Access	7	2, 6, 23, 24, 25, 26, 30
Social Collaboration, Cooperation, Teams	13	2, 6, 7, 15, 18, 20, 21, 24, 25, 26, 28, 29, 30	Avatar, Virtual Character	6	2, 6, 21, 24, 26, 27
Leaderboards	13	4, 9, 14, 15, 19, 23, 24, 25, 26, 27, 28, 29, 30	Custom Goal	5	2, 6, 7, 15, 28
Points	12	2, 6, 9, 18, 20, 23, 24, 25, 26, 27, 29, 30	Achievements, Collectibles	5	4, 21, 24, 25, 26, 30
Not Specified	10	1, 3, 5, 8, 10, 11, 12, 13, 16, 22	Feedback	5	9, 15, 24, 27, 28
Challenge	10	2, 4, 6, 20, 21, 24, 25, 26, 27, 30	Social Network, Friend Invite	5	20, 21, 25, 26, 30
Social Competition	10	2, 6, 15, 18, 19, 21, 25, 26, 29, 30	Learning	4	23, 25, 26, 30
Levels	9	17, 18, 21, 24, 25, 26, 27, 29, 30	Nonlinear Gameplay, (Branching) Choices	4	18, 23, 25, 30
Personalization	8	7, 15, 20, 23, 25, 26, 28, 30	Easter Eggs	4	23, 25, 26, 30
Progress (Bar)	8	9, 18, 21, 23, 24, 26, 27, 30	Certificates	4	23, 25, 26, 30
Badges	7	2, 6, 9, 23, 25, 26, 27	Gifting	4	24, 25, 26, 30
Knowledge Sharing	7	2, 6, 7, 9, 25, 26, 30			

Table 1. Most Common Gamification Elements

Analysis of User-Specific Factors

In the analysis of 30 research papers, 37 different approaches to user-specific adaptive gamification were identified to investigate subgoal (2). This discrepancy arises because some studies investigated multiple factors rather than focusing on one. An overview of the identified factors can be found in Figure 6. An in-depth overview of each paper and its corresponding user specific factor can be found in appendix A1.

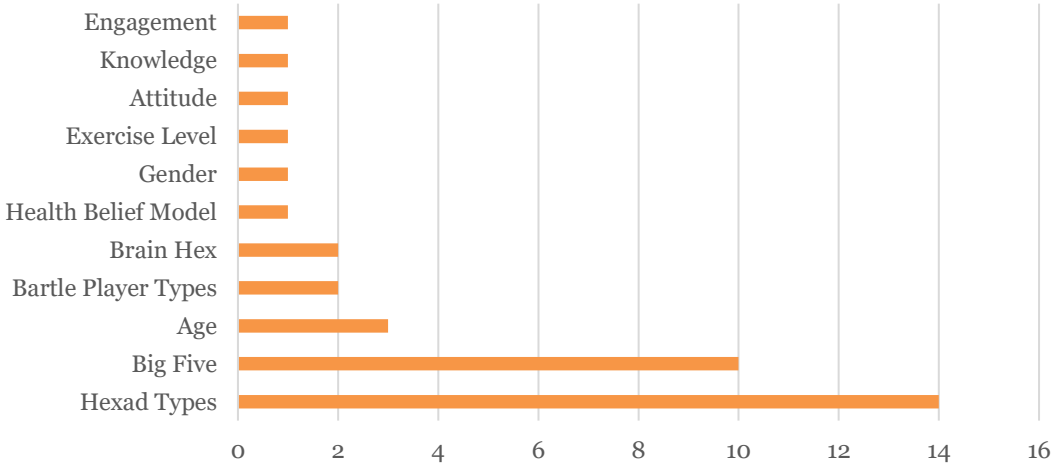

Figure 6. Analysis of User-Specific Factors

The Hexad User Types emerged as the most frequently used framework, with 14 papers using this concept to tailor gamification elements to users. As previously explained in the section on user-specific factors, the Hexad User Types are specifically designed for understanding player personalities, making them a natural fit for gamification research and therefore explaining their prevalence. Following the Hexad User Types, the Big Five Personality Traits were the second most used framework. Ten papers utilized the Big Five model, likely due to its widespread acceptance and application in psychological and behavioural studies (Costa & McCrae, 1992). In three cases, the gamification experience has been adapted to the age group. In two of those cases, the factor age has been combined the Big Five Personalities (ID 9; Mendez et al. (2019)/ ID 14; Jia et al. (2017)). Apart from the Hexad User Types, two further approaches to gamer types, BrainHex and the Bartle Player Types, have also been observed. However, far less frequently than the Hexad User Types. BrainHex and Bartle Player Types have each only been used in two papers (BrainHex: ID 29; Orji et al. (2014) and ID 24; Sienel et al. (2021); Bartle: ID 5; Jozani et al. (2018) and ID 24; Sienel et al. (2021)). Engagement, knowledge, attitude, exercise level and gender have each only been examined in one case. Overall, through the analysis of user specific factors it became clear that the Hexad User Types and the Big Five Personalities are by far the most common concepts in the research examined. Therefore, the following chapter investigating the interactions between the user specific factors and the gamification elements focuses on those concepts.

Interactions Between Personality Types and Adaptive Gamification Elements

Subgoal (4) aimed to evaluate the influence of adaptive elements on the outcome of health behaviour change. To address this topic, the interaction between user-specific factors, i.e., personality types described through the Big Five Personalities or the Hexad User Types and their interactions with adaptive gamification elements are analysed.

Big Five Personalities

First, the results showing the interactions between the gamification elements and the Big Five Personalities are presented. All correlations are shown in Table 2.

The numbers in the cells refer to the literature IDs. The correlating literature to each ID can be found in appendix A1. Negative correlations are marked with a minus, which can have two meanings: On the one hand, the gamification element can be effective for users with low scores on the personality type. On the other hand, the gamification element can demotivate users with high scores on the personality type. It is important to note that just because no correlation is pictured in the table, it does not imply that there is no connection at all. Most papers examined did not investigate all possible combinations of items, therefore emptiness can sometimes mean that there was no investigation rather than there is no connection. In the following, some of the key interactions are explained.

Achievement / Progression

The gamification elements in the achievement/progression domain show many positive correlations, especially for the users with high scores on extraversion and agreeableness. Goals can advise users and encourage them to reach those goals (ID 28; Orji et al. (2017)). Extraversive people find points helpful in supporting habit tracking (ID 27; Jia et al. (2017)). For users with high scores on neuroticism rewards are perceived as enjoyable as they bring concrete recognition for efforts and accomplishments (ID 27; Jia et al. (2017)). Badges can be noticed as helpful and enjoyable too, but also as meaningless as in the smoking cessation domain (ID 27; Jia et al. (2017)).

Social

Cooperation and normative influence show negative correlations for users with high scores on neuroticism and openness. An explanation can be that those users do not like to cooperate with others as they are mostly strangers (ID 7; Ndulue et al. (2022)). Additionally, they feel like stories from other users don't help them as they have the feeling that those don't contain new relevant information for them. Users with high scores on openness state that those stories could be fake news and do not trust social media at all (ID 7; Ndulue et al. (2022)).

		Big Five Personalities				
		NEU	**EXT**	**OPE**	**AGR**	**CON**
Gamification elements	**Achievement / progression**					
	Progress bar / Levels	27	27			27
	Challenge				27	
	Badges	27, -7				
	Points	27	27	27		
	Social					
	Cooperation / Knowledge Sharing	-7	7	-7, -28	7, 28	7
	Normative Influence	-7	7	-7	7	7
	Social Comparison		27, 28	-28	14	
	Social Competition	-7	28	-7, -28	7, 28	7
	Personalization	-7	7, 28	-7, 28	7, 28	7, 28
	Customization		28	-28	28	
	Immersion					
	Rewards	27, -7	7, 28	-7, -28	7, 28	7
	Virtual Character			-27		
	Simulation		28		28	28
	Miscellaneous					
	Punishment		28		28	
	Self-monitoring / feedback		28	-28	28	28

NEU: neuroticism, EXT: extraversion, OPE: openness (to experience), AGR: agreeableness, CON: conscientiousness

Table 2. Correlations Between Big Five Personalities and Gamification Elements

However, users with high scores on extraversion, agreeableness, or conscientiousness find nor-mative influence helpful as it brings hope, and determination, and serves as kind of a peer support (ID 27; Jia et al. (2017)). It can allow feeling less alone and sharing your stories (ID 7; Ndulue et al. (2022)). Cooperation can keep the users accountable and focused as they do not want to disappoint others (ID 7; Ndulue et al. (2022)). Help can be offered to each other and the users highlight the opportunity to make friends with similar problems (ID 28; Orji et al. (2017)).

Users with high scores on openness do not like competing in gamified health apps because they enjoy competition in entertainment games more (ID 7; Ndulue et al. (2022)). In the domain of risky health behaviour they can perceive losing as discouraging and competition itself as stressful and not helpful (ID 28; Orji et al. (2017)). Social comparison is described as invasive and reduces the self-confidence of users (ID 28; Orji et al. (2017)). For users with high scores in neuroticism comparison by leaderboards is observed to be not motivating because they assume that other users cheat (ID 7; Ndulue et al. (2022)).

Introverted people in the health domain have been found to not prefer to use leaderboards as they do not want to share their progress and feel like health-related routines are not appropriate for competition (ID 28; Orji et al. (2017)). Extroverted people, however, are motivated by competition and comparison as it has been found to make the users committed, accomplished, and challenged (ID 28; Orji et al. (2017)). The same has been observed to apply to users with high scores on agreeableness. Beyond, they prefer to start at a leaderboard from the bottom and go up from time to time because it brings greater enjoyment to them (ID 27; Jia et al. (2017)).

While customization is user-controlled, personalization is system-controlled. This distinction of the gamification elements leads to different results in their effectiveness, especially for users with high scores on openness. While customization is found not to be helpful because it is difficult to do (ID 28; Orji et al. (2017)), personalization can lead to less abstract and more useful systems as well as increased confidence of the users (ID 28; Orji et al. (2017)). But, the motivation does not solely come from personalization, it has been found to only works if the users are already motivated (ID 28; Orji et al. (2017)).

If users are not motivated, personalization has not been found to help (ID 7; Ndulue et al. (2022)). Users with high scores on neuroticism have been found to not trust personalization as they are sceptical about privacy. Customization and personalization can be effective for people with high scores on extraversion, agreeableness, and conscientiousness. The users state that personalization on the one hand convinces them and makes them more aware of their behaviour (ID 7; Ndulue et al. (2022)). Customization on the other hand increases system relevance, and the personal touch and makes the users feel connected.

Immersion

For users with high scores on extraversion, agreeableness, and conscientiousness, simulations have shown positive correlations. It is stated that the users can see the short- and long-term consequences of their actions which motivates them in the domain of risky health behaviour (ID 28; Orji et al. (2017)). Feedback is perceived as engaging and leads to positive reinforcement (ID 28; Orji et al. (2017)). Punishment has been found to motivate the first two mentioned personality types as it is challenging, engaging, and introduces consequences to the users (ID 28; Orji et al. (2017)). Finally, rewards have been found to give the users a feeling of accomplishment and determination and motivate them (ID 27; Jia et al. (2017)). All in all, gamification elements have been found to show most positive correlations for users with high scores on agreeableness and extraversion. Negative or no correlations for the elements for people who are open and have high scores on neuroticism, but it has to be mentioned that there are exceptions, for example in the domain of achievement and progression.

Hexad User Types

Second, the results demonstrating the interactions between gamification elements and Hexad User Types are presented. All correlations are detailed in Table 3.

Achievement / Progression

The correlations between Hexad User Types and gamification elements in the “Achievement / Progression” category reveal distinct preferences for adaptive gamification designs. Achievers and Players have been found to have positive correlations with the elements challenges, points, badges, with these correlations appearing frequently across multiple sources (ID 2, ID 20, ID 23, ID 30). This consistency could be explained by a motivation to achieve goals and measure progress through tangible rewards. Disruptors also have been found to have positive correlations with the elements challenges and points, possibly due to their competitive nature and desire for measurable success (ID 30; Krath and Von Korflesch (2021)). Philanthropists have been found to positive correlate with levels and unlockable content (ID 2; Altmeyer et al. (2019)), suggesting a preference for long-term engagement and investment in the game. The same elements have been observed multiple times (ID 2, ID 23, ID 30) among Achievers, which further stresses their significance for players who are motivated by progression and unlocking new content.

Social

In the “Social” category, the element cooperation positively correlates with the user types socializers and players, which can be possibly explained by them enabling partnerships, support, accountability, and reducing feelings of isolation (ID 15; Orji et al. (2018)). Social competition positively correlated with the user types Disruptors and Free Spirits, suggesting that competitive social interactions provide significant motivation (ID 30; Krath and Von Korflesch (2021)). The Transtheoretical Model of Behaviour Change (“TTM”) suggests that people pass through several successive stages of change (“SoC”) when altering behaviour. Interestingly, in a high TTM stage, social competition and cooperation—both driven by the relatedness motive—were perceived as significantly more persuasive. A possible reason for this is the fear of falling behind other users, which can negatively impact motivation. These findings indicate that the SoC

itself is an important factor to consider when designing persuasive, gamified interventions in the context of physical activity (ID 2; Altmeyer et al. (2019)). However, no correlation has been found between competition, free spirits, and philanthropists (ID 15; Orji et al. (2018)). As previously explained, personalization and customization aim to tailor systems but use different approaches: customization is user-controlled, while personalization is system-controlled. Research suggests customization can increase system effectiveness by giving users a sense of autonomy (ID 15; Orji et al. (2018)). The results show customization positively correlates with socializers and disruptors. This can be attributed to customization providing control, choice, and a personal touch, which improves the system's appeal and ease of use. Conversely, personalization is correlated with socializers and free spirits, which can be explained by it enhancing system usefulness, relevance, and trust. However, personalization has a negative correlation with the disruptors and no correlation with achievers, philanthropists, or players, as disruptors prefer influencing the system themselves (ID 15; Orji et al. (2018)). Many positive correlations between socializers and the elements leaderboards and comparison have been found, indicating that recognition and social status are key drivers for this group (ID 30; Krath and Von Korflesch (2021)).

		Hexad User Types					
Gamification elements		**AC**	**DI**	**FS**	**PH**	**PL**	**SO**
	Achievement / progression						
	Challenge	2, 20, 30		2, 30	30	2, 30	4, 30
	Points	2, 23, 30		2, 30	2	2, 20, 23, 30	2, 18
	Badges	2, 23, 30				2, 23, 30, 4	30
	Level / Unlockable Content	30, 23, 2		23, 2		30, 2	
	Achievements	4, 23			4	4, 30	
	Custom Goal	2	-15	2, 18	2		2
	Social						
	Cooperation / Knowledge Sharing	2, 30		2, 30	2, 3, 30	2, 15	2, 20, 30, 15
	Leaderboard / Comparison	4, 30		4	4	15, 4, 30, 23	4, 30, 15
	Social Competition	2, 30	30, 15			2, 30	2, 30, 15, 4
	Personalization		-15	15			15
	Customization		15	20, 23		30	15
	Immersion						
	Rewards	2, 30			2	4, 2, 15, 23, 30	
	Simulation		-15		15		15
	Virtual Character					2	
	Miscellaneous						
	Punishment		-15			15	15
	Self-Monitoring / Feedback		-15				15

AC: achiever, DI: disruptor, FS: free spirit, PH: philanthropist, PL: player, SO: socializer

Table 3. Correlations Between Hexad User Types and Gamification Elements

Immersion

In the "Immersion" category, the a positive correlation between simulations and virtual characters and the types Free Spirits and Socializers has been found (ID 2; Altmeyer et al. (2019)). This suggests that these players appreciate immersive and narrative experiences that allow for deep engagement with the game world. Within the TTM framework, the connection between the virtual character gamification element and the Achiever type has been found to be notably stronger for users in the Low-TTM stage than for those in the High-TTM stage (ID 2; Altmeyer et al. (2019)). Philanthropists have been found to positively correlate with the element simulations (ID 15; Orji et al. (2018)), indicating their preference for realistic and meaningful gamification experiences. Achievers and Players positively correlate with the element rewards (ID 30; Krath and Von Korflesch (2021)), reflecting their goal-oriented nature and desire for recognition. Disruptors show a negative correlation with simulations (ID 15; Orji et al. (2018)) possibly due to a lack of interest in immersive experiences and a greater focus on competitive and dynamic game elements.

Miscellaneous

In the miscellaneous elements category, the negative correlation with punishment and self-monitoring/feedback among Disruptors and Free Spirits (ID 15; Orji et al. (2018)) suggests a rejection of restrictive and controlling elements. These players often prefer greater freedom and autonomy in their gaming experience. Socializers, on the other hand, show a positive correlation with self-monitoring/feedback (ID 15; Orji et al. (2018)), indicating they value feedback for improving social interactions and personal growth. Achievers and Players seem to be more receptive to these elements, as they can help measure and enhance performance. The correlations between Hexad User Types and gamification elements could significantly influence individuals' behaviour by aligning the gamification design with their specific motivations. Achievers and Players seem to prefer challenges and rewards, fostering goal-oriented behaviour. Disruptors show positive correlations with by competition, also responding well to leaderboards. Philanthropists and Socializers seem to value cooperation and customization, promoting community-focused engagement. Free Spirits show positive correlations with immersive and personalized experiences, encouraging creative exploration. Understanding these correlations can help future research create tailored strategies that enhance user engagement and sustain meaningful interactions.

Discussion

Principal Findings

In the exploration of gamification elements across 30 research papers, a total of 86 distinct gamification elements has been identified, providing a broad landscape of elements employed in adaptive gamification research. Rewards, social collaboration, leaderboards, points, challenges, and social competition are the elements most frequently mentioned. These elements typically encourage users through external incentives or interactions centred around the community. Further exploration of the correlations between gamification elements and user personality traits focuses on the most used frameworks: the Hexad User Types and the Big Five Personalities. There is already existing research on adaptive gamification and the impact of different elements on these two most used frameworks. According to the findings, gamification elements are most positively associated with users who have high levels of agreeableness and extraversion under the Big Five framework. Conversely, for individuals high in openness and neuroticism, these elements often show negative or no significant correlations, though it's important to note that the domain of achievement and progression is an exception.

Additionally, the way Hexad User Types relate to gamification elements could significantly influence individual behaviour by tailoring the gamification design to meet specific personal motivations. Specifically, Achievers and Players prefer challenges and rewards. Disruptors show positive correlations in competitive contexts and respond well to leaderboards. Philanthropists and Socializers appear to value cooperation and customization, promoting community-centred engagement. Free Spirits exhibit positive correlations with immersive and personalized experiences, which support their creative exploration.

Future Research

While the current study provides valuable insights into the user-factor specific adaptive gamification and their influence on user engagement and behaviour, there still remain many aspects unexplored, calling for future research. Firstly, looking at the initial research objective, it is noteworthy that the subgoals three and four could not be explored in depth. Subgoal three was to analyse context-specific factors (e.g., mindfulness application, fitness) defining the overall frame of the gamified health information system and their combination with adaptive gamified elements. However, apart from a study by Ndulue et al. (2022) no papers investigated looked specifically at adapting gamification to the healthcare context. Almost all studies explored focused on user-specific adaptation. Therefore, no meaningful conclusions can be drawn about this subgoal within the scope of this paper, calling for further research.

Moreover, the results on subgoal four, the "evaluation of influence of adaptive elements on the outcome of health behaviour change" must be explored deeper. While the results give an indication of positive correlations between the user-specific factors and gamification elements, it is important to note that this does not necessarily imply health behaviour change. Most studies use motivation as a proxy to measure the impact of adaptive gamification on health behaviour. Questions are for example "How motivated are you to do xy". This is not necessarily meaningful about whether people behave differently. Studies that investigate the actual health behaviour are largely missing. No studies that examine adaptive gamification outcomes in comparison to "normal" gamification approaches were found. Finally, no long-term studies examining the impact of gamification on actual behaviour change rather than just motivational shifts could be identified. Future research should focus on long term studies tracking users over extended periods to observe whether gamified interventions lead to sustained behaviour changes. Looking into research gaps and the potential for future research emphasizes that there is still a lot of potential in this new and rapidly evolving field of research.

Implications and Limitations

From a theoretical perspective, this review contributes to the understanding of how adaptive elements in gamification can be tailored to different user personalities and specific contexts to enhance user engagement and motivation. The findings suggest that personalization based on the Big Five Personality Traits and Hexad User Types does lead to higher motivation and could potentially lead to more effective health behaviour interventions.

From a practical perspective, the findings can be used by developers of health apps and digital health interventions to create more individualized and engaging experiences for users. By using adaptive gamification elements that align with the users' personality traits, app designers can potentially increase user retention and the effectiveness of health behaviour change interventions. This can be particularly valuable when designing interventions for diverse user groups with different motivations and preferences and personalities.

Some important literature may have been missed due to the limitations of the Scopus and PubMed databases. Additionally, mostly short-term motivational effects were examined, and long-term data were lacking, which results in a limited understanding of the long-term effects of adaptive gamification. Furthermore, there has been no research on actual behaviour change, only on changes in motivation. Finally, some of the studies included in the review have small sample sizes, which affects the overall quality of the findings.

Conclusion

This study has investigated how adaptive gamification approaches can improve health behaviour change interventions. Overall, the findings highlight the importance of tailoring gamification strategies to individual user characteristics, such as personality traits and user types, as well as the specific context of the health intervention. This approach could have the potential to significantly influence user engagement and motivation, which are crucial for effective health behaviour change.

The analysis of user-specific factors revealed that the Hexad User Types and the Big Five Personality Traits are the most used frameworks in the studies examined. These frameworks help in understanding how different users respond to various gamification elements, allowing for more personalized and effective

interventions. The study also identified a wide range of gamification elements, with rewards, social collaboration, and leaderboards being the most used.

However, the review also identified several gaps in the existing research. There is a need for more studies examining the long-term effects of adaptive gamification on actual health behaviour change, rather than just short-term motivational shifts. Additionally, more research is needed to explore the impact of context-specific factors in combination with adaptive gamification elements in the health context. By addressing these gaps, researchers can further advance the field of adaptive gamification in digital health and develop more effective and personalized health interventions.

References

Aljabali, R. N., & Ahmad, N. (2018). A Review on Adopting Personalized Gamified Experience in the Learning Context. 2018 IEEE Conference on e-Learning, e-Management and e-Services (IC3e),

Altmeyer, M., Lessel, P., Jantwal, S., Muller, L., Daiber, F., & Krüger, A. (2021). Potential and effects of personalizing gameful fitness applications using behavior change intentions and Hexad user types. *User Modeling and User-Adapted Interaction*, *31*(4), 675-712. **https://doi.org/10.1007/s11257-021-09288-6**

Altmeyer, M., Lessel, P., Muller, L., & Krüger, A. (2019). Combining Behavior Change Intentions and User Types to Select Suitable Gamification Elements for Persuasive Fitness Systems. In H. Oinas-Kukkonen, K. T. Win, E. Karapanos, P. Karppinen, & E. Kyza (Eds.), *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems* (Vol. 11433, pp. 337-349). Springer International Publishing. **https://doi.org/10.1007/978-3-030-17287-9_27**

Böckle, M., Novak, J., & Bick, M. (2017). TOWARDS ADAPTIVE GAMIFICATION: A SYNTHESIS OF CURRENT DEVELOPMENTS. 25th European Conference on Information Systems (ECIS),

Brandl, L. C., & Schrader, A. (2023). Clustering on Player Types of Students in Health Science – Trial and Data Analyses. In R. Röhrig, N. Grabe, M. Haag, U. Hübner, U. Sax, C. Oliver Schmidt, M. Sedlmayr, & A. Zapf (Eds.), *Studies in Health Technology and Informatics*. IOS Press. **https://doi.org/10.3233/SHTI230698**

Busch, M., Mattheiss, E., Orji, R., Marczewski, A., Hochleitner, W., Lankes, M., Nacke, L. E., & Tscheligi, M. (2015). Personalization in Serious and Persuasive Games and Gamified Interactions. CHI PLAY '15: The annual symposium on Computer-Human Interaction in Play,

Carlier, S., Coppens, D., De Backere, F., & De Turck, F. (2021). Investigating the Influence of Personalised Gamification on Mobile Survey User Experience. *Sustainability*, *13*(18), 10434. **https://doi.org/10.3390/su131810434**

Ciocarlan, A., Masthoff, J., & Oren, N. (2018). Kindness is Contagious: Study into Exploring Engagement and Adapting Persuasive Games for Wellbeing. UMAP '18: 26th Conference on User Modeling, Adaptation and Personalization,

Costa, P. T., & McCrae, R. R. (1992). The Five-Factor Model of Personality and Its Relevance to Personality Disorders. *Journal of Personality Disorders*, *6*(4), 343-359. **https://doi.org/10.1521/pedi.1992.6.4.343**

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US. **https://doi.org/10.1007/978-1-4899-2271-7**

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". MindTrek '11: Academic MindTrek 2011,

Feng, Z., Lau, N., Zhu, M., Liu, M., Refati, R., Huang, X., & Lee, K.-p. (2023). Behavioural design of gamification elements and exploration of player types in youth basketball training. *Smart Learning Environments*, *10*(1), 56. **https://doi.org/10.1186/s40561-023-00278-2**

Gochman, D. S. (1982). Labels, Systems and Motives: Some Perspectives For Future Research and Programs. *Health Education Quarterly*, *9*(2-3), 167-174. **https://doi.org/10.1177/109019818200900213**

Gosetto, L., Pittavino, M., Falquet, G., & Ehrler, F. (2023). Personalization of Mobile Apps for Health Behavior Change: Protocol for a Cross-sectional Study. *JMIR Research Protocols*, *12*, e38603. **https://doi.org/10.2196/38603**

Jia, Y., Liu, Y., Yu, X., & Voida, S. (2017). Designing Leaderboards for Gamification: Perceived Differences Based on User Ranking, Application Domain, and Personality Traits. CHI '17: CHI Conference on Human Factors in Computing Systems,

Jia, Y., Xu, B., Karanam, Y., & Voida, S. (2016). Personality-targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances. CHI'16: CHI Conference on Human Factors in Computing Systems,

Johnson, D., Deterding, S., Kuhn, K.-A., Staneva, A., Stoyanov, S., & Hides, L. (2016). Gamification for health and wellbeing: A systematic review of the literature. *Internet Interventions*, *6*, 89-106. **https://doi.org/10.1016/j.invent.2016.10.002**

Jozani, M. M., Maasberg, M., & Ayaburi, E. (2018). Slayers vs Slackers: An Examination of Users' Competitive Differences in Gamified IT Platforms Based on Hedonic Motivation System Model. In P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies. Learning and Teaching* (Vol. 10925, pp. 164-172). Springer International Publishing. **https://doi.org/10.1007/978-3-319-91152-6_13**

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, *45*, 191-210. **https://doi.org/10.1016/j.ijinfomgt.2018.10.013**

Krath, J., & Von Korflesch, H. F. O. (2021). Player Types and Game Element Preferences: Investigating the Relationship with the Gamification User Types HEXAD Scale. In X. Fang (Ed.), *HCI in Games: Experience Design and Game Mechanics* (Vol. 12789, pp. 219-238). Springer International Publishing. **https://doi.org/10.1007/978-3-030-77277-2_18**

Lyu, S., & Bidarra, R. (2023). Procedural generation of challenges for personalized gait rehabilitation. FDG 2023: Foundations of Digital Games 2023,

Marczewski, A. (2015). *Even ninja monkeys like to play: gamification, game thinking & motivational design*. Gamified UK.

Martin-Niedecken, A. L., & Götz, U. (2016). Design and Evaluation of a Dynamically Adaptive Fitness Game Environment for Children and Young Adolescents. CHI PLAY '16: The annual symposium on Computer-Human Interaction in Play,

Mendez, J. I., Ponce, P., Meier, A., Peffer, T., Mata, O., & Molina, A. (2019). Framework for promoting social interaction and physical activity in elderly people using gamification and fuzzy logic strategy. 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP),

Moreno-Blanco, D., Sánchez-González, P., Gárate, F. J., Cáceres, C., Solana-Sánchez, J., Tormos-Muñoz, J. M., & Gómez, E. J. (2020). New Approaches for Personalizing Daily Activity Monitoring in mHealth Applications. In J. Henriques, N. Neves, & P. De Carvalho (Eds.), *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019* (Vol. 76, pp. 1181-1186). Springer International Publishing. **https://doi.org/10.1007/978-3-030-31635-8_143**

Nacke, L. E., & Deterding, S. (2017). The maturing of gamification research. *Computers in Human Behavior*, *71*, 450-454. **https://doi.org/10.1016/j.chb.2016.11.062**

Ndulue, C., Oyebode, O., Iyer, R. S., Ganesh, A., Ahmed, S. I., & Orji, R. (2022). Personality-targeted persuasive gamified systems: exploring the impact of application domain on the effectiveness of behaviour change strategies. *User Modeling and User-Adapted Interaction*, *32*(1-2), 165-214. **https://doi.org/10.1007/s11257-022-09319-w**

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive Systems Design: Key Issues, Process Model, and System Features. *Communications of the Association for Information Systems*, *24*. **https://doi.org/10.17705/1CAIS.02428**

Okunogbe, A., Nugent, R., Spencer, G., Ralston, J., & Wilding, J. (2021). Economic impacts of overweight and obesity: current and future estimates for eight countries. *BMJ Global Health*, *6*(10), e006351. **https://doi.org/10.1136/bmjgh-2021-006351**

Orji, R., Nacke, L. E., & Di Marco, C. (2017). Towards Personality-driven Persuasive Health Games and Gamified Systems. CHI '17: CHI Conference on Human Factors in Computing Systems,

Orji, R., Tondello, G. F., & Nacke, L. E. (2018). Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types. CHI '18: CHI Conference on Human Factors in Computing Systems,

Orji, R., Vassileva, J., & Mandryk, R. L. (2014). Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. *User Modeling and User-Adapted Interaction*, *24*(5), 453-498. **https://doi.org/10.1007/s11257-014-9149-8**

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology*, *51*(3), 390-395. **https://doi.org/10.1037/0022-006X.51.3.390**

Ren, L., Yan, J., Zhu, Z., & Du, M. (2024). Personalization Characteristics and Evaluation of Gamified Exercise for Middle-Aged and Older People: A Scoping Review. *Journal of Aging and Physical Activity*, *32*(2), 287-299. **https://doi.org/10.1123/japa.2022-0224**

Rosenstock, I. M. (1966). Why people use health services. *The Milbank Memorial Fund Quarterly*, *44*(3), Suppl:94-127. **http://www.ncbi.nlm.nih.gov/pubmed/5967464**

Sandford-james, A., Parkes, J., & Campbell, J. (2022). What is the practical utility of the MyCognitionPRO platform for monitoring and preventing cognitive decline in a real-world context of people over 50years-old experiencing cognitive ageing? *Alzheimer's & Dementia*, *18*(S2), e064139. **https://doi.org/10.1002/alz.064139**

Sienel, N., Münster, P., & Zimmermann, G. (2021, 2021). Player-Type-based Personalization of Gamification in Fitness Apps. 14th International Conference on Health Informatics,

Tondello, G. (2019). Dynamic Personalization of Gameful Interactive Systems. **http://hdl.handle.net/10012/14807**

Tondello, G. F., Mora, A., & Nacke, L. E. (2017). Elements of Gameful Design Emerging from User Preferences. CHI PLAY '17: The annual symposium on Computer-Human Interaction in Play,

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., & Nacke, L. E. (2016). The Gamification User Types Hexad Scale. CHI PLAY '16: The annual symposium on Computer-Human Interaction in Play,

Wanderley De Oliveira, L., & Teixeira De Carvalho, S. (2020). A Gamification-Based Framework for mHealth Developers in the Context of Self-Care. 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS),

Webster, J., & Watson, R. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii-xxiii. **http://www.jstor.org/stable/4132319**

Wen, T., & Guo, Y. (2023). Combining Game User Types and Health Beliefs to Explore the Persuasiveness of Gamification Strategies for Fitness Systems. In C. Stephanidis, M. Antona, S. Ntoa, & G. Salvendy (Eds.), *HCI International 2023 Posters* (Vol. 1833, pp. 199-209). Springer Nature Switzerland. **https://doi.org/10.1007/978-3-031-35992-7_28**

World Health Organization. (2024). Obesity and overweight. *World Health Organization*. **https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight**

Yang, Y., University of, W., Goh, K. Y., National University of, S., Teo, H. H., National University of, S., Tan, S. S. L., & National University of, S. (2023). Compete with Me? The Impact of Online Gamified Competition on Exercise Behavior. *Journal of the Association for Information Systems*, *24*(3), 912-935. **https://doi.org/10.17705/1jais.00806**

Zhao, Z., Arya, A., Orji, R., & Chan, G. (2020). Effects of a Personalized Fitness Recommender System Using Gamification and Continuous Player Modeling: System Design and Long-Term Validation Study. *JMIR Serious Games*, *8*(4), e19968. **https://doi.org/10.2196/19968**

Appendix

A1 Concept Coding Matrix

ID	Title	Author (year)	Empirical approach *(Conceptual, quantitative, qualitative, mixed methods)*	Context-specific factors *(Health behaviour / Mental health / Informative / Rehabilitation / Disctraction From pain / Professional training)*	Healthcare context *(Specific use context)*	User-specific factors *(Player types, Hexad User Types, Personas, Age, Personality type)*
1	What is the practical utility of the MyCognitionPRO platform for monitoring and preventing cognitive decline in a real-world context of people over 50years-old experiencing cognitive ageing?	Sandford-james et al. (2022)	Mixed methods	Health behaviour	Cognitive function	Age
2	Combining behavior change intentions and user types to select suitable gamification elements for persuasive fitness systems	Altmeyer et al. (2019)	Mixed methods	Health behaviour	Physical activity	Hexad User Types
3	New Approaches for Personalizing Daily Activity Monitoring in mHealth Applications	Moreno-Blanco et al. (2020)	Conceptual	Health behaviour	Mental Health	-
4	Combining Game User Types and Health Beliefs to Explore the Persuasiveness of Gamification Strategies for Fitness Systems	Wen and Guo (2023)	Mixed methods	Health behaviour	Physical activity	Hexad User Types AND HBM
5	Slayers vs Slackers: An Examination of Users' Competitive Differences in Gamified IT Platforms Based on Hedonic Motivation System Model	Jozani et al. (2018)	Conceptual	No health context	-	Player personalities (according to Bartle and Robson)

ID	**Title**	**Author (year)**	**Empirical approach** *(Conceptual, quantitative, qualitative, mixed methods)*	**Context-specific factors** *(Health behaviour / Mental health / Informative / Rehabilitation / Disctraction From pain / Professional training)*	**Healthcare context** *(Specific use context)*	**User-specific factors** *(Player types, Hexad User Types, Personas, Age, Personality type)*
6	Potential and effects of personalizing gameful fitness applications using behavior change intentions and Hexad user types	Altmeyer et al. (2021)	Mixed methods	Health behaviour	Physical activity	Hexad User Types AND TTM
7	Personality-targeted persuasive gamified systems: exploring the impact of application domain on the effectiveness of behaviour change strategies	Ndulue et al. (2022)	Quantitative	Health behaviour	Healthy eating and smoking cessation	Big Five Personalities
8	A gamification-based framework for mhealth developers in the context of self-care	Wanderley De Oliveira and Teixeira De Carvalho (2020)	Mixed methods	Health behaviour	Mental Health	Hexad User Types
9	Framework for promoting social interaction and physical activity in elderly people using gamification and fuzzy logic strategy	Mendez et al. (2019)	Conceptual	Health behaviour	Physical activity	Big Five Personalities AND Levels of Attitute, Knowledge and Engagement AND Age
10	A Review on Adopting Personalized Gamified Experience in the Learning Context	Aljabali and Ahmad (2018)	Literature review	No health context	-	-
11	Personalization Characteristics and Evaluation of Gamified Exercise for Middle-Aged and Older People: A Scoping Review	Ren et al. (2024)	Literature review	Health behaviour	Physical activity	-

ID	Title	Author (year)	Empirical approach *(Conceptual, quantitative, qualitative, mixed methods)*	Context-specific factors *(Health behaviour / Mental health / Informative / Rehabilitation / Disctraction From pain / Professional training)*	Healthcare context *(Specific use context)*	User-specific factors *(Player types, Hexad User Types, Personas, Age, Personality type)*
12	Personalization in serious and persuasive games and gamified interactions	Busch et al. (2015)	Grey Literature	No health context	-	-
13	Design and evaluation of a dynamically adaptive fitness game environment for children and young adolescents	Martin-Niedecken and Götz (2016)	Quantitative	Health behaviour	Physical activity	-
14	Designing leaderboards for gamification: Perceived differences based on user ranking, application domain, and personality traits	Jia et al. (2017)	Quantitative	Health behaviour	Physical activity	Big Five Personalities AND age AND Gender
15	Personalizing persuasive strategies in gameful systems to gamification user types	Orji et al. (2018)	Mixed methods	Health behaviour	Mental health	Hexad User Types
16	Kindness is contagious: Study into exploring engagement and adapting persuasive games for wellbeing	Ciocarlan et al. (2018)	Mixed methods	Health behaviour	Mental health	Big Five Personalities
17	Procedural generation of challenges for personalized gait rehabilitation	Lyu and Bidarra (2023)	Mixed methods	Rehabilitation	Physical activity	-
18	Behavioural design of gamification elements and exploration of player types in youth basketball training	Feng et al. (2023)	Mixed methods	Health behaviour	Physical activity	Hexad User Types

ID	**Title**	**Author (year)**	**Empirical approach** *(Conceptual, quantitative, qualitative, mixed methods)*	**Context-specific factors** *(Health behaviour / Mental health / Informative / Rehabilitation / Disctraction From pain / Professional training)*	**Healthcare context** *(Specific use context)*	**User-specific factors** *(Player types, Hexad User Types, Personas, Age, Personality type)*
19	Compete with Me? The Impact of Online Gamified Competition on Exercise Behavior	Yang et al. (2023)	Quantitative	Health behaviour	Physical activity	Exercise level
20	Effects of a Personalized Fitness Recommender System Using Gamification and Continuous Player Modeling: System Design and Long-Term Validation Study	Zhao et al. (2020)	Mixed methods	Health behaviour	Physical activity	Hexad Types
21	Personalization of Mobile Apps for Health Behavior Change: Protocol for a Cross-sectional Study	Gosetto et al. (2023)	Mixed methods	Health behaviour	-	Hexad User Types AND Big Five Personalities AND Theory of planned behaviour
22	Clustering on Player Types of Students in Health Science - Trial and Data Analyses	Brandl and Schrader (2023)	Quantitative	Health education	Health education	Hexad User Types
23	Investigating the influence of personalised gamification on mobile survey user experience	Carlier et al. (2021)	Quantitative	No direct focus on health	Surveys in Healthcare	Hexad User Types
24	Player-type-based personalization of gamification in fitness apps	Sienel et al. (2021)	Mixed methods	Health behaviour	Physical activity	Hexad User Types AND bartle player types AND Big Five Personalities AND Brain Hex
25	The Gamification User Types Hexad Scale	Tondello et al. (2016)	Quantitative	No health context	Not specified	Hexad User Types AND Big Five Personalities

ID	Title	Author (year)	Empirical approach *(Conceptual, quantitative, qualitative, mixed methods)*	Context-specific factors *(Health behaviour / Mental health / Informative / Rehabilitation / Disctraction From pain / Professional training)*	Healthcare context *(Specific use context)*	User-specific factors *(Player types, Hexad User Types, Personas, Age, Personality type)*
26	Elements of Gameful Design Emerging from User Preferences	Tondello et al. (2017)	Mixed Methods	No health context	Not specified	Hexad User Types AND Big Five Personalities
27	Personality,targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances	Jia et al. (2016)	Quantitative	General health habits, e.g., running, eating fruits, drinking water	Not specified	Big Five Personalities
28	Towards Personality-driven Persuasive Health Games and Gamified Systems	Orji et al. (2017)	Quantitative	General health bevahiour	eg. Drinking	Big Five Personalities
29	Modeling the efficacy of persuasive strategies for different gamer types in serious games for health	Orji et al. (2014)	Mixed Methods	Health Behaviour	Weight loss & calorie tracking	BrainHex
30	Player Types and Game Element Preferences: Investigating the Relationship with the Gamification User Types HEXAD Scale	Krath and Von Korflesch (2021)	Quantitative	No health context	Not specified	Hexad User Types

Table A1. Concept Coding Matrix (Short Version)

Note: The detailed concept coding matrix is available by the authors upon request.

A2 Overview of Gamification Elements by Paper

Element	Total	ID (Paper mentioned)	Element	Total	ID (Paper mentioned)
Rewards	15	2, 4, 6, 7, 15, 18, 21, 23, 24, 25, 26, 27, 28, 29, 30	Social Status	2	26, 30
Social Collaboration / Cooperation / Teams	13	2, 6, 7, 15, 18, 20, 21, 24, 25, 26, 28, 29, 30	Tips	1	30
Leaderboards	13	4, 9, 14, 15, 19, 23, 24, 25, 26, 27, 28, 29, 30	Self-Monitoring	1	15
Points	12	2, 6, 9, 18, 20, 23, 24, 25, 26, 27, 29, 30	Objective	1	18
Not Specified	10	1, 3, 5, 8, 10, 11, 12, 13, 16, 22	Renovation	1	18
Challenge	10	2, 4, 6, 20, 21, 24, 25, 26, 27, 30	Puzzle	1	18
Social Competition	10	2, 6, 15, 18, 19, 21, 25, 26, 29, 30	Novelty	1	18
Levels	9	17, 18, 21, 24, 25, 26, 27, 29, 30	Sensation	1	18
Personalization	8	7, 15, 20, 23, 25, 26, 28, 30	Acknowledgement	1	18
Progress Bar	8	9, 18, 21, 23, 24, 26, 27, 30	Stats	1	18
Badges	7	2, 6, 9, 23, 25, 26, 27	Chance	1	18
Knowledge Sharing	7	2, 6, 7, 9, 25, 26, 30	Time	1	18
Unlockable Content / Access	7	2, 6, 23, 24, 25, 26, 30	Rarity	1	18
Avatar / Virtual Character	6	2, 6, 21, 24, 26, 27	Economy	1	18
Achievements / Collectibles	6	4, 21, 24, 25, 26, 30	Reputation	1	18
Custom Goal	5	2, 6, 7, 15, 28	Rivalry	1	19
Feedback	5	9, 15, 24, 27, 28	Awareness Cues	1	19
Social Network / Friend Invite	5	20, 21, 25, 26, 30	Discussion board	1	24
Learning	4	23, 25, 26, 30	Assessment	1	24
Nonlinear Gameplay / (Branching) Choices	4	18, 23, 25, 30	Records	1	24
Easter Eggs	4	23, 25, 26, 30	Schedule	1	24
Certificates	4	23, 25, 26, 30	Number Limit	1	24
Gifting	4	24, 25, 26, 30	Performance Graph	1	24
Cheating	3	2, 6, 20	Permadeath	1	24
Social / Peer Pressure	3	4, 18, 26	Time Limit	1	24
Storytelling / Narrative	3	18, 26, 30	Crowning	1	24
Lottery / Chance	3	23, 25, 30	Topic	1	24
Trading / Virtual Economy	3	25, 26, 30	Difficulty Selection	1	24

Element	Total	ID (Paper mentioned)	Element	Total	ID (Paper mentioned)
Social Discovery	3	25, 26, 30	Prize Pacing	1	24
Administrative Roles	3	25, 26, 30	Torture Break	1	24
Exploratory Tasks	3	25, 26, 30	Scarlet Letter	1	26
Creativity Tools	3	25, 26, 30	Glowing Choice	1	26
Quests	3	25, 26, 30	Beginner's Luck	1	26
Innovation Platforms	3	25, 26, 30	Signposting	1	26
Voting Mechanisms	3	25, 26, 30	Anchor	1	26
Development Tools	3	25, 26, 30	Juxtaposition	1	26
Meaning and Purpose	2	23, 26	Power-Ups or Boosters	1	26
Punishment	2	15, 28	Humanity hero	1	26
Simulation	2	15, 28	Free Lunch	1	26
Social Status	2	26, 30	Mystery Box	1	26
Bragging / Pride	2	24, 29	Theme	1	26
Anonymity	2	25, 30	Boss Battles	1	26
Anarchic Gameplay	2	25, 30	Meaningful Choices	1	26

Table A2. Gamification Elements by Paper

Explainable Artificial Intelligence for Biomedical Data: A Systematic Mapping Study

Digital Health, Winter Term 2024/25

Robin Gansäuer
Master's Student
Karlsruhe Institute of Technology
robin.gansaeuer@student.kit.edu

Maria Weinreuter
Master's Student
Karlsruhe Institute of Technology
maria.weinreuter@student.kit.edu

Hichem Ben Aoun
Master's Student
Karlsruhe Institute of Technology
hichem.aoun@student.kit.edu

Felix Pietsch
Master's Student
Karlsruhe Institute of Technology
felix.pietsch@student.kit.edu

Abstract

Background: *Artificial intelligence (AI) in biomedicine must be explainable to ensure trust, accountability, and safety. However, most explainable AI (XAI) approaches are evaluated primarily on technical grounds, with limited systematic assessment and reliable metrics, hindering translation into practice.*

Objective: *This study aims to identify and categorize XAI systems for biomedical data that fulfill criteria for evaluation readiness, defined by output demonstration, input data availability, and practical applicability.*

Methods: *A systematic mapping study was conducted using a Scopus search from 2020 to 2025, retrieving 1,178 journal and conference papers. Restricting to 2023 to 2025 yielded 780. Screening identified 27 systems meeting evaluation readiness criteria, classified by AI evaluation, AI method, agnosticism, data type, input samples, medical field, medical use case, scope, stage, XAI evaluation, XAI method, and XAI technique.*

Results: *Imaging-oriented XAI systems dominate, with 63% (n = 17) using radiology or histopathology pipelines and visual heatmaps such as Grad-CAM and saliency maps. Tabular (n = 4) and graph-based (n = 3) approaches are less common. Feature relevance methods, especially SHAP (n = 6), prevail. Grad-CAM is frequent in imaging (n = 4), while graph-based techniques are emerging (n = 4). Evaluation strategies emphasize functionality-grounded approaches, 52% (n = 14), with fewer application-grounded (n = 11) and rare human-grounded (n = 2) assessments. Most XAI systems provide code and, in some cases, web applications, enabling reproducibility, e.g., MICA and CIS2MS.*

Conclusion: *Evaluation-ready XAI systems in biomedicine are scarce, imaging-dominated, and specialized. Local explanations and architecture modification methods are underused, and cross-domain benchmarking is essential. The lack of standardized XAI metrics limits comparability. This mapping offers a starting point for validation. Its Scopus-only search, single-round coding, and assumed evaluation readiness criteria motivate broader search, multistage review, and human-grounded research.*

Keywords: explainable AI, biomedical data, evaluation readiness, XAI evaluation

Introduction

In a widely cited study, a pneumonia-prediction model erroneously inferred that asthma patients had a lower mortality risk. This misclassification was not due to any biological factor but rather to the fact that these patients were more frequently admitted directly to intensive care, where they received more immediate and aggressive treatment. Such spurious correlations, if embedded in opaque systems, could mislead decision-makers and result in harmful or even fatal recommendations (Caruana et al., 2015).

This example reflects a broader imperative within biomedicine: Artificial intelligence (AI) systems must not only be accurate but also explainable. In high-stakes contexts, such as clinical diagnostics, treatment planning, or triage, the opacity of many AI systems raises pressing concerns regarding trust, accountability, and safety (Barredo Arrieta et al., 2020; Babic et al., 2021). Healthcare professionals must understand AI model reasoning to make sound decisions and avoid embedding systemic biases (Antoniadi et al., 2021). Regulatory pressures reinforce this need. For instance, the European Union's General Data Protection Regulation requires that meaningful information be provided about the logic of algorithmic decisions (Kim et al., 2024). Similarly, ethical frameworks and policy initiatives worldwide increasingly call for transparent and auditable AI in medicine (World Health Organization, 2023; European Union, 2024).

Over the past five years, explainable AI (XAI) has emerged as a response to these demands, offering methods to interpret and communicate the internal logic of complex models. Techniques such as saliency maps, feature attribution, concept-based explanations, and surrogate models have been proposed and applied across diverse biomedical domains, including radiology, genomics, and electronic health records (Barredo Arrieta et al., 2020; Antoniadi et al., 2021). However, a growing body of research warns that many of these XAI systems remain disconnected from real-world clinical utility (Ghassemi et al., 2021; Chen et al., 2022; Donoso-Guzmán et al., 2025). Most are evaluated on technical grounds, without empirical validation involving clinicians, patients, or decision-makers. As a result, it remains unclear which XAI solutions improve understanding, trust, or decision-making under real constraints (Kim et al., 2024).

Recent systematic reviews confirm that few biomedical XAI systems undergo human-centered evaluation (Chen et al., 2022; Donoso-Guzmán et al., 2025). In medical imaging, for example, explainable models are rarely tested with user studies or in situ validation, and when evaluated, metrics are often vague or inconsistently applied (Chen et al., 2022). As a result, advances in XAI frequently fail to translate into clinical practice. Many AI models also stall at the prototype stage. In one review of over 250 systems for neonatal and pediatric intensive care, none had reached clinical deployment (Schouten et al., 2024). These findings underscore the need to distinguish between conceptual innovation and *evaluation readiness*, the degree to which a system is validated, transparent, and documented for experimental or clinical testing.

To address this gap, this study conducts a systematic mapping study (SMS) of XAI systems for biomedical data with the goal of identifying which systems fulfill criteria for evaluation readiness. Specifically, the study asks:

Which XAI systems for biomedical data fulfill criteria for evaluation readiness?

Evaluation readiness is defined as compliance with the three criteria: Output demonstration, input data availability, and practical applicability. This construct provides a structured lens for distinguishing between exploratory XAI proposals and systems sufficiently mature to progress from algorithmic development to applied evaluation. For research, this implies a shift from predominantly technical assessments toward comprehensive validation strategies that consider usability, human interaction, and contextual robustness. For practice, evaluation readiness offers a means to triage among the proliferation of XAI systems, identifying those that are suitable for integration into clinical workflows or formal studies. By bridging the gap between innovation and implementation, evaluation readiness enables both communities to collectively advance translational progress in biomedical AI.

Following the Introduction, the Background section provides an overview of XAI, its core principles, biomedical data, and associated evaluation challenges, while also introducing evaluation readiness with its three dimensions as a proposed framework for assessing system maturity. The Systematic Mapping Study section outlines the methodology for identifying and categorizing evaluation-ready systems. The Results section presents the findings, reporting 27 systems and summarizing their characteristics. The Discussion section covers the principal findings, implications for practice and research, and limitations and future work. Finally, the Conclusion distills overarching insights and highlights avenues for subsequent research.

Background on Explainable AI and Biomedical Data

Explainable AI

Explainable AI seeks to enhance the transparency and interpretability of AI models, ensuring that their decision-making processes are comprehensible to human users (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020; Doshi-Velez & Kim, 2017). While no universally accepted definition exists (Barredo Arrieta et al., 2020), XAI broadly aims to bridge the gap between model complexity and human understanding, fostering trust, accountability, and safety in high-stakes domains (Ribeiro et al., 2016; Holzinger et al., 2019).

At its core, XAI introduces explanations that clarify the relationship between input features and model predictions, fostering both interpretability, which ensures that explanations are clear and comprehensible to human users, and fidelity, which describes how well an explanation reflects the actual reasoning process of a model (Doshi-Velez & Kim, 2017). These explanations can be generated ante hoc, where models are inherently interpretable, such as decision trees and linear regression, or post hoc, where interpretability techniques, such as Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME), provide insights into otherwise opaque black-box models (Adadi & Berrada, 2018; Ribeiro et al., 2016).

Despite significant advancements, fundamental challenges persist. XAI must balance interpretability with predictive performance, define standardized evaluation metrics, and adapt explanations to different user groups and domains (Barredo Arrieta et al., 2020).

Biomedical Data

Biomedical data, encompassing genomic sequences, electronic health records, and medical imaging, form the backbone of contemporary healthcare research, driving significant progress in diagnostics and personalized medicine (Shortliffe & Cimino, 2014). However, these data sources pose notable challenges for XAI due to their complexity, diversity, and sensitivity.

Genomic data illustrate this potential clearly, as advancements in sequencing technologies have uncovered genetic variants crucial to understanding disease mechanisms, particularly in cancer and rare genetic disorders (Miotto et al., 2018). Yet, the inherent sparsity and high dimensionality of genomic data require robust computational approaches for reliable interpretation (Miotto et al., 2018).

Electronic health records integrate structured clinical data with unstructured clinical notes, providing valuable longitudinal insights into patient health (Shortliffe & Cimino, 2014). Despite their predictive potential, data inconsistencies and missing values often hinder their clinical usability (Weiskopf & Weng, 2013; Hripcsak & Albers, 2013; Lewis et al., 2023).

Medical imaging is fundamental to disease diagnosis and monitoring. Recent advances in machine learning, especially convolutional neural networks (CNNs) and vision transformers, have significantly improved the extraction of clinically relevant imaging features (Miotto et al., 2018). Nevertheless, effectively integrating imaging data with genomic and clinical information remains a substantial technical hurdle.

Evaluation of Explainable AI in Biomedical Data

Evaluating XAI in biomedical data presents distinct challenges due to the complexity and heterogeneity of these datasets. Effective frameworks must go beyond generic technical metrics to ensure clinical and research relevance, incorporating domain-specific criteria such as biological plausibility, diagnostic accuracy, and interpretability consistency (Doshi-Velez & Kim, 2017).

A widely used taxonomy distinguishes between three levels of evaluation for XAI: functionality-grounded, human-grounded, and application-grounded assessments (Zhou et al., 2021). Functionality-grounded evaluation relies on formal computational tests without involving human participants and is primarily used to benchmark algorithmic behavior. Human-grounded evaluation incorporates simplified tasks with representative users or domain experts to examine explanation clarity, usability, or intuitiveness. Application-grounded evaluation takes place in real-world settings, embedding XAI systems into authentic decision-making workflows such as clinical environments and assessing their impact on trust, performance,

and outcomes (Zhou et al., 2021). This structured perspective enables a nuanced assessment of XAI and highlights the varying degrees to which explainability methods approximate real-world applicability, thereby underscoring the importance of aligning evaluation strategies with clinical practice.

For genomic data, explanations should align with biological pathways and molecular interactions, supporting biomarker discovery and disease understanding (Adadi & Berrada, 2018). Similarly, in electronic health records, temporal and contextual fidelity is essential for uncovering meaningful patient trajectories and mitigating data inconsistencies (Bernabé et al., 2023). In medical imaging, evaluations must validate the relevance of highlighted regions against expert annotations while assessing improvements in diagnostic performance. Multimodal approaches integrating genomic, clinical, and imaging data require interpretability across data types (Miotto et al., 2018).

Beyond technical performance, ethical considerations such as fairness, accountability, and data privacy necessitate robust evaluation frameworks. While general metrics provide a foundation, biomedical XAI demands assessment methods that reflect applicability and patient impact (Barredo Arrieta et al., 2020; Holzinger et al., 2019).

Extensive research has explored the integration of XAI into biomedical data analysis. In omics, the absence of standardized evaluation workflows remains a central obstacle to clinical translation (Toussaint et al., 2024). In medical decision support, model-agnostic techniques continue to dominate, yet stronger collaboration between AI researchers and healthcare professionals is needed to ensure practical adoption (Prentzas et al., 2023). In medical imaging, recent reviews highlight persistent challenges in aligning XAI explanations with clinical decision-making (Chen et al., 2022).

Despite these advances, few studies focus on evaluation-ready XAI systems, leaving this area understudied and limiting their broader adoption in experimental validation settings (Holzinger et al., 2019).

Concept of Evaluation Readiness

XAI systems in biomedicine continue to multiply, yet rigorous empirical validation in real clinical settings or with domain experts remains scarce (Jung et al., 2023; Prentzas et al., 2023). Consequently, much of the literature remains at the proof-of-concept level, leaving essential aspects such as usability, reproducibility, and applicability unexplored (Donoso-Guzmán et al., 2025). To bridge this gap, research traditions, namely evaluability assessment, organizational readiness, and technology maturity models, are drawn upon, all of which emphasize readiness as a requisite for meaningful evaluation (Leviton et al., 2010; Damschroder et al., 2009; Lavin et al., 2022).

In program evaluation, the tradition of evaluability assessment highlights that interventions should only be evaluated once their objectives are clearly articulated, resources are secured, and appropriate data are available for analysis (Leviton et al., 2010). Conducting evaluation in the absence of such preconditions risks generating results that are inconclusive or misleading. In biomedical XAI, this translates into the need for clarity about what the system is designed to achieve, transparency about the resources required to run it, and access to data that can meaningfully test its claims.

In implementation science, the concept of organizational readiness underscores that the success of evaluation and adoption depends not only on technical design but also on the broader system context (Weiner, 2009). Commitment from stakeholders, alignment with workflows, and adequate capacity are all required to ensure that innovations can be tested and used effectively (Damschroder et al., 2009). Applying this logic to XAI highlights that evaluation cannot occur in isolation from end-users. If explanations are unintelligible to clinicians or if workflows cannot accommodate them, empirical evaluation will likely fail regardless of technical accuracy.

Finally, technology maturity models formalize the principle that innovations progress through stages of development, and that only after reaching a sufficient level of maturity does field evaluation make sense (Mankins, 2009). For example, the Technology Readiness Levels framework specifies that prototypes should be demonstrated in relevant environments before being subjected to large-scale testing (Lavin et al., 2022). For biomedical XAI, this underscores the expectation that systems should first demonstrate stable performance and plausible explanatory outputs before clinical trials or user studies are attempted.

Together, these strands of literature converge on a common insight. Evaluation requires preconditions of clarity, feasibility, and maturity. This construct is operationalized through three dimensions:

Dimension 1: Output Demonstration. The system must have shown that its explanatory outputs (e.g., feature attributions) function as intended. This reflects the technology maturity perspective that prototypes must demonstrate reliable behavior before field evaluation is warranted (Lavin et al., 2022).

Dimension 2: Input Data Availability. Relevant, high-quality biomedical datasets must be accessible to support independent testing and reproducibility. This criterion follows from evaluability assessment's emphasis on the indispensability of data for credible evaluation (Leviton et al., 2010).

Dimension 3: Practical Applicability. The system must be usable and feasible in its intended context, including workflow integration and user interpretability. This dimension draws on implementation science, which identifies feasibility and acceptance as critical preconditions for evaluating and embedding innovations (Damschroder et al., 2009).

These three dimensions are not claimed to capture all aspects of evaluation readiness. Rather, they serve as a pragmatic framework applied in the screening phase of the SMS to identify XAI systems sufficiently mature for empirical evaluation. Together, the dimensions provide a lens for distinguishing exploratory XAI proposals from systems ready to progress from algorithmic development to applied validation.

Systematic Mapping Study

A systematic mapping study was conducted to survey evaluation-ready XAI systems and assess their alignment with various medical fields. The procedure followed the methodological framework proposed by Petersen et al. (2008) and was adapted to fit the specific needs of this study (Toussaint et al., 2024).

The resulting five-step process employed is depicted in Figure 1. The process begins with the formulation of the research question in Step 1. In Steps 2 and 3, a systematic search and rigorous screening of relevant studies were conducted. In Step 4, key facets for analysis were identified, and in the final Step 5, data extraction was performed to map XAI systems to their respective medical domains.

This approach directly addresses the challenges related to XAI in healthcare applications as discussed in the Background subsection Evaluation of Explainable AI in Biomedical Data, offering structured insights into the current landscape of evaluation-ready systems.

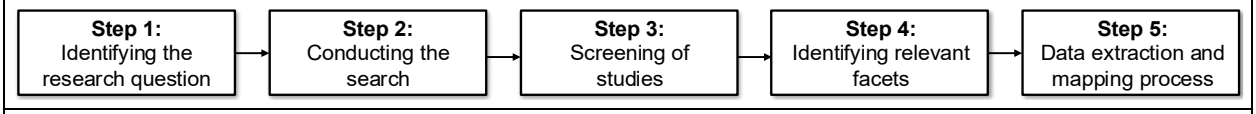

Figure 1. Procedure of the Systematic Mapping Study, Based on Petersen et al. (2008) and Following Toussaint et al. (2024)

Conducting the Search and Screening of Studies

Steps 2 and 3 of the systematic mapping process encompass conducting the search and screening the studies. The search and screening process is outlined in Figure 2.

To find relevant studies, a systematic search using the Scopus database was initially conducted. The search query was developed with a predefined search string, which was tested and refined to improve relevance. The final search string included additional filters, restricting the date range to January 2020 through January 2025, selecting only journal articles and conference papers, and limiting the language to English.

Conducting the search resulted in a set of 1,178 studies. To prioritize the most recent advancements, the set was further refined to include only studies published between January 2023 and January 2025, resulting in a subset of 780 initial studies to undergo screening.

After conducting the search, the studies were screened in three phases:

Phase 1: Title and Abstract Screening. In Phase 1, the 780 initial studies were evenly distributed among the four authors. Each author reviewed the titles and abstracts of their assigned studies based on four criteria, as illustrated in Figure 2. These criteria were initially tested on 20 studies per researcher. Afterward, the criteria were slightly refined to ensure a common understanding. Studies that did not meet any of these criteria were excluded, resulting in 120 studies being selected for further screening.

Phase 2: Full Paper Screening. In Phase 2, the full text of the remaining 120 studies was reviewed by all four authors. Phase 2 was crucial at this point, since the titles and abstracts alone did not always clearly indicate whether the proposed XAI systems were genuinely evaluation-ready. In this phase, the concept of evaluation readiness, grounded in evaluability assessment (Leviton et al., 2010), was applied to determine whether systems met the necessary preconditions for meaningful evaluation. A study was included in the final set only if all authors agreed that the XAI system met three criteria for evaluation readiness, as shown in Figure 2. As a result, the number of studies was reduced to 24.

Phase 3: Full Review Screening. In Phase 3, each author screened one review paper on XAI methods in healthcare, selected from 2022 and 2023. This helped address the strict filtering criteria, which initially included only publications from 2023 to 2025, ensuring coverage of the most promising XAI systems emerging around this pivotal period. The review papers included either studies excluded in Phase 1, mostly due to missing XAI system maturity or missing evaluation, or additional studies identified unsystematically. The XAI systems mentioned in these review papers were evaluated against all criteria defined in Phases 1 and 2, and three additional XAI systems meeting the criteria were added to the final selection.

Ultimately, 27 studies reporting evaluation-ready XAI systems were selected.

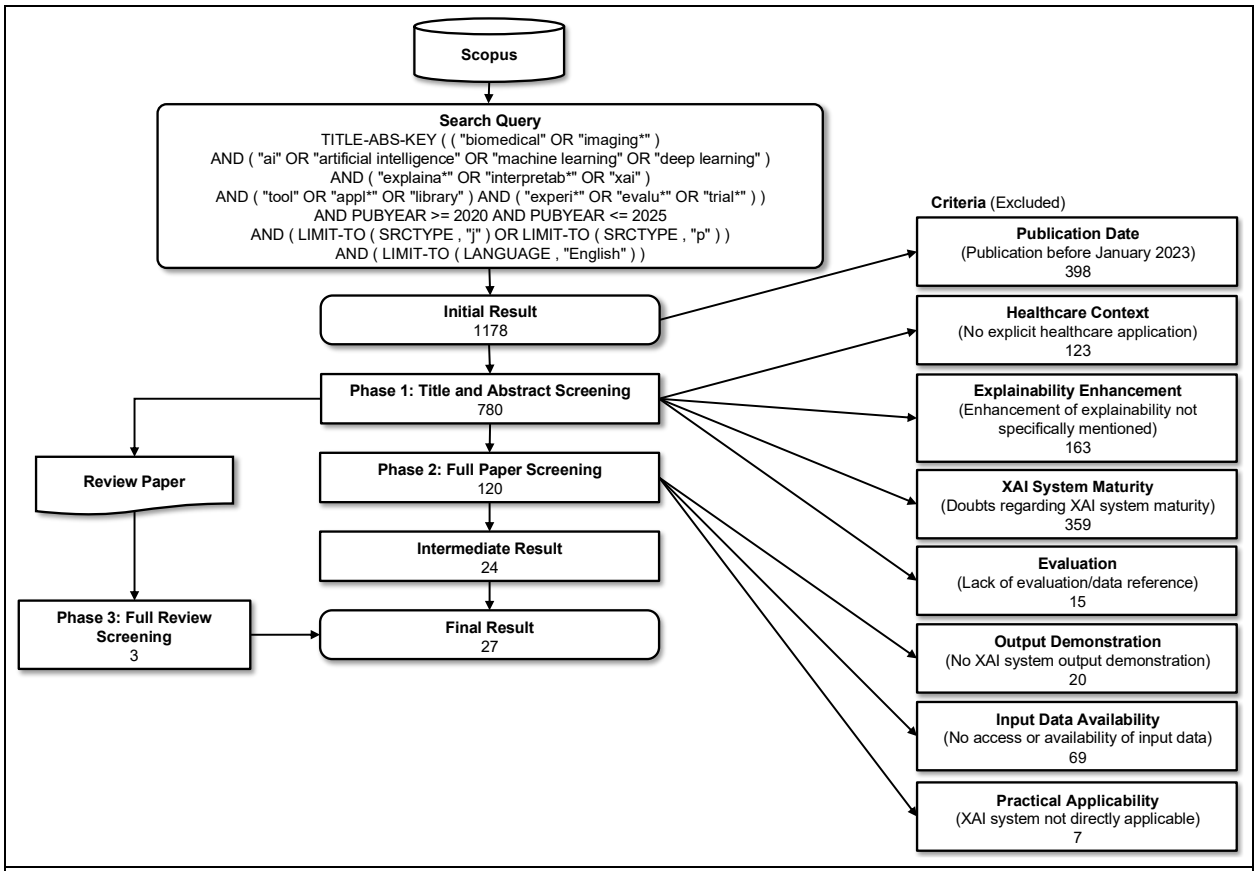

Figure 2. Search and Screening Process

Identifying Relevant Facets

The next step involved the identification of relevant facets to develop a classification scheme. Related review papers (Toussaint et al., 2024; Nazir et al., 2023; Markus et al., 2021; Gupta & Seeja, 2024; Karim et al., 2023; Muhammad & Bendechache, 2024) were analyzed to identify general facets used to structure overviews of XAI in healthcare. Additionally, specific facets extending beyond those related to XAI methods or healthcare applications were brainstormed individually, based on information given in the final set of 27 studies. These specific facets focus on characteristics of evaluation-ready XAI systems, such as online references or the evaluation methods applied in the studies.

Following this approach, each researcher independently compiled a list of relevant facets. These lists were then collectively discussed, refined, and consolidated into a final selection. Table 1 provides an overview of the selected facets, along with an explanation for each.

Facet	**Explanation**	**References**
AI Evaluation	Describes metrics applied to assess the performance of the AI model.	Toussaint et al., 2024
AI Method	Describes the AI methods for which the XAI system is designed to provide explanations.	Toussaint et al., 2024; Nazir et al., 2023
Agnosticism	Describes whether the XAI system generalizes across models or is model-specific.	Gupta & Seeja, 2024; Muhammad & Bendechache, 2024; Karim et al., 2023; Nazir et al., 2023
Data Type	Describes how biomedical information is stored and processed in the XAI system.	Karim et al., 2023
Input Samples	Describes the number of data points used to evaluate the AI or XAI model.	N/A (authors' proposed facet)
Medical Field	Describes the medical specialty where the XAI system is applied or intended for use.	Toussaint et al., 2024; Nazir et al., 2023
Medical Use Case	Describes the purpose for which the XAI system was developed.	Markus et al., 2021
Scope	Describes whether explanations focus on global model behavior or single predictions.	Gupta & Seeja, 2024; Muhammad & Bendechache, 2024; Karim et al., 2023; Nazir et al., 2023
Stage	Describes whether explanations are integrated into the model or added post-prediction.	Gupta & Seeja, 2024; Nazir et al., 2023; Markus et al., 2021
XAI Evaluation	Describes evaluation methods applied to assess the performance of the XAI model.	Zhou et al., 2021
XAI Method	Describes the approach used to provide explanations to human users.	Gupta & Seeja, 2024; Karim et al., 2023; Nazir et al., 2023; Barredo Arrieta et al., 2020
XAI Technique	Describes specific techniques used in the XAI method to provide explainability.	Nazir et al., 2023

Table 1. Developed Classification Scheme

Data Extraction and Mapping Process

In the final step, information from the 27 studies was extracted, coded according to predefined facet categories, and mapped to the corresponding facets to ensure comparability.

Results

List of XAI Systems

The initial search identified 1,178 studies, from which 27 were ultimately selected based on their presentation of evaluation-ready XAI systems for biomedical data. A full list of these systems, including authors, years, descriptions, and GitHub repositories, is provided in the Appendix (Table A1).

Each XAI system is documented alongside its output samples as presented in the respective studies. Notably, studies processing imaging data, such as Melanoma Classification (Gamage et al., 2024), present a diverse range of intuitive visual outputs through heatmaps. Melanoma Classification generates heatmaps aimed at enhancing clinicians' confidence in melanoma diagnosis. The study showcases outputs from different AI models and input instances and provides snapshots of the final application interface.

In addition, the input data used to generate XAI outputs are available for each system. These datasets may either be publicly accessible datasets referenced in the respective studies or datasets provided directly by the authors. XAI systems such as MICA (Bie et al., 2024), which are trained and tested on large publicly available datasets, are particularly notable in this regard. MICA, designed for skin lesion diagnostics, is built on three publicly available datasets: Derm7pt, PH^2, and Skin-Con. The study explicitly details the number of data points extracted from each dataset and specifies the proportions used to split the dataset.

Finally, each XAI system is linked to its source code, enabling direct experimentation and reproducibility. While all studies provide GitHub repositories to facilitate the application of their XAI systems, some additionally offer web-based applications, further enhancing accessibility and user interaction across different user groups. Among these, CIS2MS (Rasouli et al., 2024) distinguishes itself through a particularly user-friendly web interface. Developed as an explainable machine learning approach for predicting the conversion from clinically isolated syndrome (CIS) to multiple sclerosis (MS), its web-based application allows users to interactively test the system's predictions with various input data.

Characteristics of XAI Systems

The previously introduced facets are analyzed to characterize XAI systems in biomedical research, specifically with regard to data types, explainability techniques, evaluation approaches, medical use cases, and additional characteristics such as explainability stage, domain agnosticism, and scope.

Data Types. Examining the distribution of data types (Figure 3A), a significant proportion of XAI systems in biomedical research relies on imaging data, with imaging-oriented XAI systems comprising 63% (n = 17) of the examined systems. This includes applications in radiology and histopathology, where methods such as gradient-weighted class activation mapping (Grad-CAM) and saliency maps are frequently employed to generate spatially interpretable heatmaps. Notable examples include glioma segmentation (Zeineldin et al., 2024), otitis media assessment (Chen et al., 2024), and autism spectrum disorder prediction (Garcia & Kelly, 2024), all of which utilize CNNs for feature extraction and visualization. By contrast, systems utilizing tabular data (n = 4) primarily focus on genomics and structured clinical datasets, as seen in drug-drug interaction models (Wang et al., 2024) and multimodal Alzheimer's disease subtyping (Yang et al., 2024). Graph-based approaches (n = 3) are gaining traction in niche areas such as psychiatry (Zheng et al., 2025) and pathology (Jaume et al., 2021), uncovering structural relationships and enhancing insights.

XAI Techniques. Regarding explainability techniques (Figure 3B), feature relevance methods such as SHAP (n = 6) dominate the landscape, particularly in oncology (Shetab Boushehri et al., 2023), genomics (Yang et al., 2024), and pharmacology (Wang et al., 2024), where they elucidate the contributions of molecular markers and multimodal features. Grad-CAM (n = 4) remains a cornerstone for imaging-based applications, enabling visual explanations for models used in dermatology (Gamage et al., 2024), radiology (Chen et al., 2024), and neurology (Garcia & Kelly, 2024). Graph-based explainability techniques, including graph neural networks (GNNs) and local graph neural networks (n = 4), are increasingly applied in psychiatric and neurological disorder prediction (Zheng et al., 2025).

XAI Evaluations. Evaluation strategies for XAI models (Figure 3C) exhibit a strong emphasis on functionality-grounded approaches, which account for 52% (n = 14) of the studies. These evaluations prioritize the quantitative assessment of model explainability, particularly in domains requiring algorithmic

benchmarking such as genomics (Yang et al., 2024) and drug interaction modeling (Wang et al., 2024). Application-grounded evaluations (n = 11) are common in imaging and clinical decision-support applications, where model explanations are validated using expert annotations and real-world performance, as observed in radiology (Zeineldin et al., 2024) and histopathology (Jaume et al., 2021). Human-grounded evaluations (n = 2) are less frequently implemented but play a crucial role in assessing the interpretability and usability of XAI outputs, making them more applicable in psychiatric and neurological applications (Zheng et al., 2025).

Medical Use Cases. The primary medical use case of XAI systems in biomedical research and clinical practice (Figure 3D) is the verification and enhancement of diagnostic processes, accounting for 44% (n = 12) of the examined XAI systems. These include imaging-based applications such as glioma segmentation (Zeineldin et al., 2024) and otitis media classification (Chen et al., 2024), where visual explanations enhance clinical confidence in AI-assisted diagnoses. Knowledge discovery (n = 9) constitutes another major application, particularly in genomics (Yang et al., 2024) and immunology (Shetab Boushehri et al., 2023), where XAI methods facilitate the identification of disease markers and therapeutic targets. Lastly, management-focused XAI systems (n = 6) support personalized medicine and treatment planning, exemplified by psychiatric diagnosis tools (Zheng et al., 2025) and MS conversion prediction models (Rasouli et al., 2024).

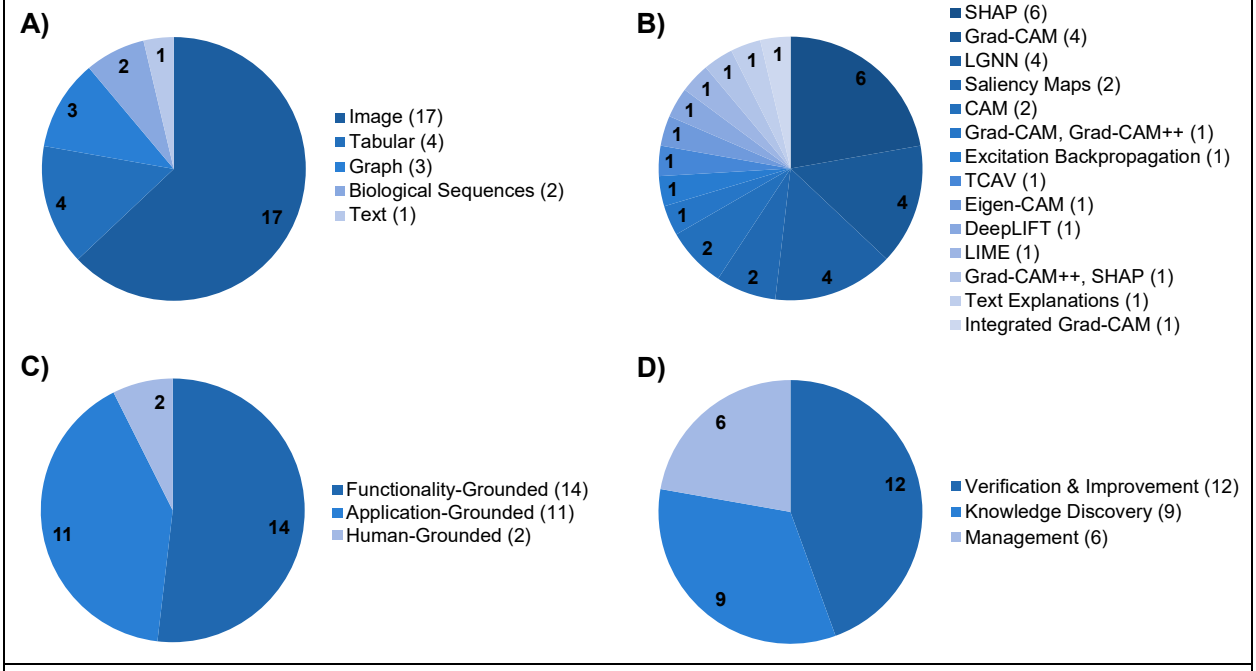

Figure 3. Data Types (A), XAI Techniques (B), XAI Evaluations (C), and Medical Use Cases (D)

Additional Characteristics. Beyond data types, explainability techniques, evaluation methods, and medical applications, additional characteristics further define the landscape of XAI systems in biomedicine. Post hoc explainability (n = 14) slightly exceeds ante hoc approaches (n = 13). Most systems are domain-specific (n = 21), tailored for specialized applications in neurology (Garcia & Kelly, 2024) and oncology (Shetab Boushehri et al., 2023), while a smaller subset is domain-agnostic (n = 6). Finally, the distribution of global (n = 14) vs. local (n = 13) explainability underscores the balance between providing holistic model transparency and focusing on instance-level interpretability within XAI.

Explainability Methods in Medical Fields

Explainability methods are mapped to medical fields to examine which approaches have been considered applicable for XAI in medicine, as illustrated in Figure 4. This chart displays the number of studies associated with each specific combination of categories.

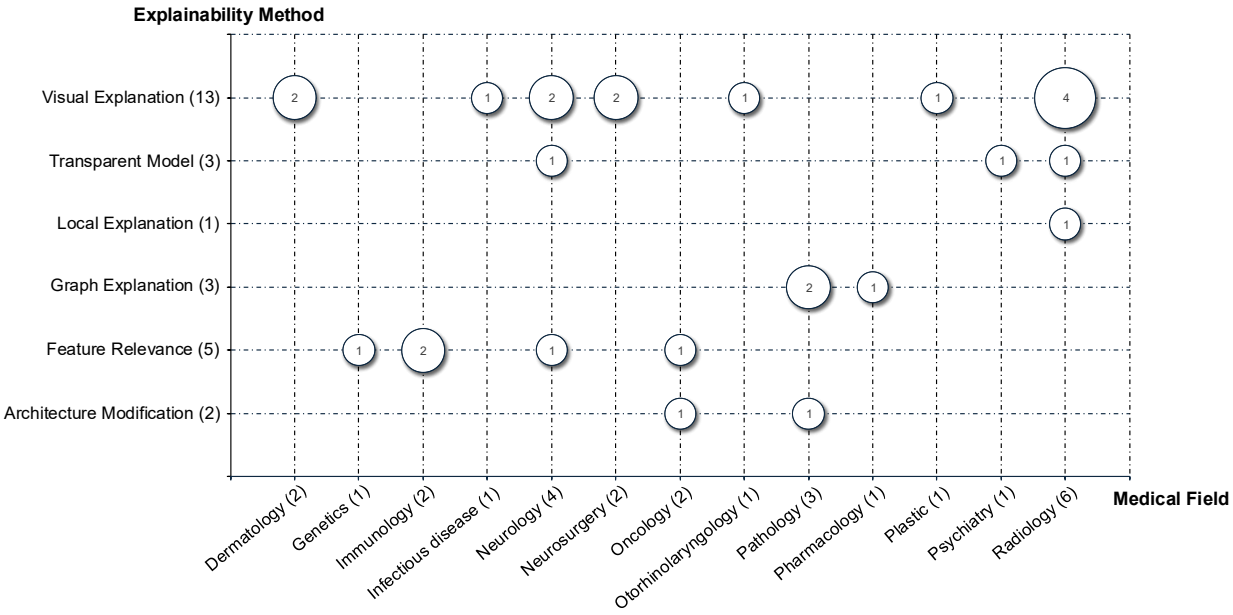

Figure 4. Bubble Chart of Explainability Methods and Medical Fields

Prevalence of Visual Explanation Methods. The results indicate that visual explanation methods, such as Grad-CAM and saliency maps, are the most widely utilized approaches, reflecting their ability to produce spatially interpretable outputs. These methods are particularly prevalent in imaging-intensive fields. In radiology, XAI systems for glioma segmentation (Zeineldin et al., 2024) and otitis media evaluation (Chen et al., 2024) employ visual heatmaps to highlight diagnostically relevant regions, aiding clinical decision-making. Similarly, in neurology, XAI systems predicting autism spectrum disorder (Garcia & Kelly, 2024) and Alzheimer's disease subtypes (Yang et al., 2024) leverage these techniques to elucidate critical brain regions, enhancing interpretability and alignment with clinical reasoning. Dermatological applications, such as Melanoma Classification (Gamage et al., 2024), utilize these methods to identify visual features, such as texture and color.

Feature Relevance Methods Across Genomic and Multimodal Data. Feature relevance methods, including SHAP and LIME, play a pivotal role in domains that require a granular understanding of individual features' contributions to model outputs. In oncology, for instance, XAI systems for cancer detection (Shetab Boushehri et al., 2023) employ SHAP to highlight genetic markers and imaging phenotypes critical to diagnostic and prognostic predictions. Similarly, in genetics and immunology, feature relevance methods are instrumental in XAI systems for drug-drug interaction prediction (Wang et al., 2024) and genomic analysis (Yang et al., 2024), where insights into biomarkers and molecular interactions are essential for understanding disease mechanisms and optimizing therapeutic strategies and outcomes.

Emergence of Transparent Model Methods. Transparent models, which integrate explainability into their fundamental architecture, have gained traction in fields that require a deep understanding of data relationships. In psychiatry, for example, BrainIB (Zheng et al., 2025) employs graph-based approaches to identify subgraph biomarkers, offering clinically consistent and interpretable insights into brain network connectivity for psychiatric disorder diagnosis. In pathology, transparent latent variable models are used in tissue analysis XAI systems (Jaume et al., 2021), enabling a nuanced exploration of tissue structure-function relationships and advancing precision in cancer detection and classification efforts.

Graph Explanation Methods in Specialized Domains. Graph-based explanations, while less prevalent overall, demonstrate significant potential in domains requiring the analysis of structured data. In pharmacology, drug-drug interaction prediction XAI systems (Wang et al., 2024) employ GNNs to model and interpret interactions between drugs. Similarly, in pathology, GNN-based tissue mapping XAI systems (Jaume et al., 2021) facilitate the discovery of relationships within histopathological datasets, offering insights into tissue-level processes.

Discussion

Principal Findings

This study systematically mapped evaluation-ready XAI systems for biomedical data, providing an overview of their availability, characteristics, and the distribution of explainability methods across medical fields. The findings highlight three critical insights: (1) the scarcity of XAI systems that meet evaluation readiness criteria, (2) the fragmented and highly specialized nature of existing systems, and (3) the uneven adoption of explainability methods across medical fields, revealing notable gaps.

Scarcity of Evaluation-Ready XAI Systems. The first key finding is that evaluation-ready XAI systems remain scarce, despite the increasing volume of research on XAI in biomedicine. From an initial pool of 780 studies, only 27 were identified as meeting evaluation readiness criteria and could be assessed in real-world settings. The majority of XAI research continues to focus on methodological advancements, algorithmic innovations, or conceptual discussions, but relatively few studies provide practically usable, fully documented, and reproducible systems. One of the major barriers to evaluation readiness is limited accessibility of input data. Many studies employ proprietary or sensitive biomedical datasets, which are often unavailable or accessible only upon request, limiting external validation efforts. Furthermore, a significant proportion of proposed systems lack publicly available implementations, such as GitHub repositories or interactive web applications, making independent replication and clinical experimentation difficult. Even among systems that theoretically fulfill evaluation readiness conditions, many lack practical accessibility, further restricting their applicability in experimental and clinical environments.

Fragmentation and Limited Generalizability. The second major finding concerns the high degree of specialization and fragmentation among existing XAI systems. Most identified models are tailored to narrow, well-defined biomedical applications, with limited potential for generalization beyond their original use case. Approximately half of the identified XAI systems rely on ante hoc approaches, meaning they are inherently interpretable but often lack the flexibility required for broader adoption across different medical contexts. Furthermore, the study revealed that human-grounded evaluations are exceedingly rare, with only 2 of the 27 systems involving direct validation by healthcare professionals. This raises concerns about the practical utility and reliability of current XAI implementations, as models that lack human expert assessment may fail to address the real interpretability needs of clinical practitioners. The predominance of functionality-grounded evaluation methods over human- and application-grounded approaches further reinforces this issue, as it suggests that many XAI systems are still being assessed primarily from a technical perspective, rather than in real-world healthcare scenarios.

Uneven Distribution of Explainability Methods. The third key insight reveals that XAI methods are unevenly distributed across medical domains, with some explanation methods remaining underutilized. The results show a clear preference for visual explanation methods, such as Grad-CAM and saliency maps, which are dominant in radiology and neurology, two fields where imaging-based AI models are widely adopted. In contrast, feature relevance methods, including SHAP and LIME, are more commonly applied in oncology and immunology, where AI models focus on identifying biomarkers or genetic variations. However, the study also identifies a notable lack of local explanations and architecture modification methods, despite their relevance for personalized medicine and patient-specific AI interpretations. Additionally, graph-based explainability methods, which have shown potential for modeling complex biological relationships, remain underrepresented. These approaches are more commonly used in fields such as psychiatry and pharmacology but have yet to see widespread adoption across biomedical AI applications. Another striking observation is that, although radiology accounts for the highest share of XAI adoption, the remaining XAI systems are relatively evenly distributed across other medical fields, such as oncology, neurology, pathology, and pharmacology. This pattern suggests that XAI development is driven by domain-specific requirements rather than by efforts to create broadly applicable, cross-disciplinary solutions.

Implications for Practice and Research

From a practice perspective, healthcare professionals can now engage with a curated selection of evaluation-ready XAI systems without requiring bespoke datasets or complex configurations. This accessibility lowers technical barriers, allowing clinicians, radiologists, and bioinformaticians to assess the

relevance of XAI within their specific use cases. However, XAI adoption remains fragmented across medical fields, limiting interdisciplinary applications. While Grad-CAM dominates in radiology, it is rarely applied in genomics or pharmacology. To address this, cross-domain benchmarking is essential. Establishing shared evaluation protocols would improve comparability between explainability techniques and foster trust in their deployment across medical disciplines (Toussaint et al., 2024).

From a research perspective, this study provides a structured foundation for evaluating XAI systems, setting the stage for more systematic assessments. Unlike prior work that primarily focuses on developing new explainability methods, the study shifts toward structured evaluation, categorizing systems based on their readiness for evaluation using predefined facets. This classification enables future research to build on existing XAI systems rather than designing new methods in isolation. Additionally, the findings highlight the need for more differentiation in XAI evaluation between image-based methods, where most approaches currently concentrate, and other biomedical modalities, which remain underexplored. Finally, while a set of evaluation-ready XAI systems is identified, the findings reveal a significant gap in standardized XAI evaluation metrics. Establishing consensus on these metrics will be essential for ensuring reliable and comparable assessments across domains (Doshi-Velez & Kim, 2017).

Limitations and Future Work

This study has three key limitations. First, the literature search was restricted to Scopus, potentially excluding relevant studies from other databases such as PubMed, IEEE Xplore, and arXiv. While review papers helped mitigate this, earlier foundational works may have been overlooked. Second, the categorization of XAI systems was conducted in a single-round coding process based on predefined categories, which, despite consensus validation, introduces a degree of subjectivity in mapping and classification. A multistage expert review could improve classification robustness. Third, this study assumes that evaluation readiness is determined by the three criteria of output demonstration, input data availability, and practical applicability. However, since the identified XAI systems were neither deployed nor tested, this assumption remains uncertain. Nevertheless, defining these criteria aimed to maximize the likelihood of evaluation readiness.

Future work should expand the literature search to include additional databases, preprints, and gray literature for a more comprehensive understanding of evaluation-ready XAI systems. The classification process should incorporate multistage expert validation to refine evaluation criteria and improve reproducibility. In addition, future studies should build on this study's systematically compiled list of 27 XAI systems by testing their actual evaluation readiness through direct application. This might include examining expert interaction, validating clinical usability, and identifying adoption barriers. Integrating human-grounded evaluation, where medical professionals test and refine XAI explanations, would further enhance their applicability.

Conclusion

This study provides the first systematic mapping of biomedical XAI systems through the lens of evaluation readiness, to the best of the available evidence. While the field is rapidly expanding, most systems remain at the proof-of-concept stage: 63% of systems are imaging-oriented, using radiology or histopathology pipelines and visual heatmaps, such as Grad-CAM and saliency maps. Tabular and graph-based approaches are less common, and feature relevance methods, particularly SHAP, prevail, with graph-based techniques emerging and Grad-CAM frequently used in imaging. Most systems provide code and, in some cases, web applications (e.g., MICA and CIS2MS), enabling reproducibility. Evaluation strategies emphasize functionality-grounded approaches (52%), with fewer application-grounded and rare human-grounded assessments, underscoring the gap between algorithmic innovation and evaluability. The contribution is twofold. First, it consolidates a fragmented landscape. Second, it operationalizes evaluation readiness to guide progression from algorithmic development to empirical validation. Although bounded by a Scopus-only search, single-round coding, and the selected readiness criteria, the study lays the groundwork for systematic evaluation and highlights readiness for clinical impact. By identifying which systems are evaluation-ready, it provides a roadmap to prioritize empirical testing where clinical translation is most plausible. Future work should refine readiness criteria and extend them across domains to ensure that explainability research remains aligned with clinical practice.

References

Abdel-Alim, T., Tio, P., Kurniawan, M., Mathijssen, I., Dirven, C., Niessen, W., Roshchupkin, G., & van Veelen, M.-L. (2023). Reliability and agreement of automated head measurements from 3-dimensional photogrammetry in young children. *The Journal of Craniofacial Surgery, 34*(6), 1629–1634. **https://doi.org/10.1097/SCS.0000000000009448**

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. **https://doi.org/10.1109/ACCESS.2018.2870052**

Anand, A., Kadian, T., Shetty, M. K., & Gupta, A. (2022). Explainable AI decision model for ECG data of cardiac disorders. *Biomedical Signal Processing and Control, 75*, Article 103584. **https://doi.org/10.1016/j.bspc.2022.103584**

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences, 11*(11), Article 5088. **https://doi.org/10.3390/app11115088**

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science, 373*(6552), 284–286. **https://doi.org/10.1126/science.abg1834**

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. **https://doi.org/10.1016/j.inffus.2019.12.012**

Bernabé, C. H., Queralt-Rosinach, N., Silva Souza, V. E., Bonino da Silva Santos, L. O., Mons, B., Jacobsen, A., & Roos, M. (2023). The use of foundational ontologies in biomedical research. *Journal of Biomedical Semantics, 14*, Article 21. **https://doi.org/10.1186/s13326-023-00300-z**

Bie, Y., Luo, L., & Chen, H. (2024). MICA: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(2), 837–845. **https://doi.org/10.1609/aaai.v38i2.27842**

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for health care: Predicting pneumonia risk and hospital 30-day readmission. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, & G. Williams (Eds.), *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). Association for Computing Machinery. **https://doi.org/10.1145/2783258.2788613**

Chen, B., Li, Y., Sun, Y., Sun, H., Wang, Y., Lyu, J., Guo, J., Bao, S., Cheng, Y., Niu, X., Yang, L., Xu, J., Yang, J., Huang, Y., Chi, F., Liang, B., & Ren, D. (2024). A 3D and explainable artificial intelligence model for evaluation of chronic otitis media based on temporal bone computed tomography: Model development, validation, and clinical application. *Journal of Medical Internet Research, 26*, Article e51706. **https://doi.org/10.2196/51706**

Chen, H., Gomez, C., Huang, C.-M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *npj Digital Medicine, 5*, Article 156. **https://doi.org/10.1038/s41746-022-00699-2**

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4*, Article 50. **https://doi.org/10.1186/1748-5908-4-50**

Dickinson, Q., & Meyer, J. G. (2022). Positional SHAP (PoSHAP) for interpretation of machine learning models trained from biological sequences. *PLOS Computational Biology, 18*(1), Article e1009736. **https://doi.org/10.1371/journal.pcbi.1009736**

Ding, X., Chen, X., Sullivan, E. E., Shay, T. F., & Gradinaru, V. (2024). Fast, accurate ranking of engineered proteins by target-binding propensity using structure modeling. *Molecular Therapy, 32*(6), 1687–1700. **https://doi.org/10.1016/j.ymthe.2024.04.003**

Donoso-Guzmán, I., Kacafírková, K. S., Szymanski, M., Jacobs, A., Parra, D., & Verbert, K. (2025). A systematic review of user-centred evaluation of explainable AI in healthcare. *arXiv*. **https://doi.org/10.48550/arXiv.2506.13904**

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. **https://doi.org/10.48550/arXiv.1702.08608**

Elbouknify, I., Bouhoute, A., Fardousse, K., Berrada, I., & Badri, A. (2023). CT-xCOV: A CT-scan based explainable framework for COVID-19 diagnosis. In *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)* (pp. 1–8). IEEE. **https://doi.org/10.1109/WINCOM59760.2023.10322985**

European Union. (2024). 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). *Official Journal of the European Union, L,* 12 July 2024, 1–144. **https://data.europa.eu/eli/reg/2024/1689/oj**

Gamage, L., Isuranga, U., Meedeniya, D., De Silva, S., & Yogarajah, P. (2024). Melanoma skin cancer identification with explainability utilizing mask guided technique. *Electronics, 13*(4), Article 680. **https://doi.org/10.3390/electronics13040680**

Gao, S., Zhou, H., Gao, Y., & Zhuang, X. (2023). BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability. *Medical Image Analysis, 89*, Article 102889. **https://doi.org/10.1016/j.media.2023.102889**

Garcia, M., & Kelly, C. (2024). 3D CNN for neuropsychiatry: Predicting autism with interpretable deep learning applied to minimally preprocessed structural MRI data. *PLOS ONE, 19*(10), Article e0276832. **https://doi.org/10.1371/journal.pone.0276832**

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. **https://doi.org/10.1016/S2589-7500(21)00208-9**

Gupta, J., & Seeja, K. R. (2024). A comparative study and systematic analysis of XAI models and their applications in healthcare. *Archives of Computational Methods in Engineering, 31*(7), 3977–4002. **https://doi.org/10.1007/s11831-024-10103-9**

Halinkovic, M., Fabian, O., Felsoova, A., Kveton, M., & Benesova, W. (2024). Intrinsically explainable deep learning architecture for semantic segmentation of histological structures in heart tissue. *Computers in Biology and Medicine, 177*, Article 108624. **https://doi.org/10.1016/j.compbiomed.2024.108624**

He, Q., Summerfield, N., Dong, M., & Glide-Hurst, C. (2024). Modality-agnostic learning for medical image segmentation using multi-modality self-distillation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1–5). IEEE. **https://doi.org/10.1109/ISBI56570.2024.10635881**

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), Article e1312. **https://doi.org/10.1002/widm.1312**

Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association, 20*(1), 117–121. **https://doi.org/10.1136/amiajnl-2012-001145**

Jaume, G., Pati, P., Anklin, V., Foncubierta, A., & Gabrani, M. (2021). HistoCartography: A toolkit for graph analytics in digital pathology. In M. Atzori, N. Burlutskiy, F. Ciompi, Z. Li, F. Minhas, H. Müller, T. Peng, N. Rajpoot, B. Torben-Nielsen, J. van der Laak, M. Veta, Y. Yuan, & I. Zlobec (Eds.), *Proceedings of the MICCAI Workshop on Computational Pathology* (Proceedings of Machine Learning Research, Vol. 156, pp. 117–128). PMLR. **https://proceedings.mlr.press/v156/jaume21a.html**

Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon, 9*(5), Article e16110. **https://doi.org/10.1016/j.heliyon.2023.e16110**

Karim, M. R., Islam, T., Shajalal, M., Beyan, O., Lange, C., Cochez, M., Rebholz-Schuhmann, D., & Decker, S. (2023). Explainable AI for bioinformatics: Methods, tools and applications. *Briefings in Bioinformatics, 24*(5), Article bbad236. **https://doi.org/10.1093/bib/bbad236**

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence, 7*, Article 1456486. **https://doi.org/10.3389/frai.2024.1456486**

Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., Lange, D., Baydin, A. G., Sharma, A., Gibson, A., Zheng, S., Xing, E. P., Mattmann, C., Parr, J., & Gal, Y. (2022). Technology

readiness levels for machine learning systems. *Nature Communications, 13*, Article 6039. **https://doi.org/10.1038/s41467-022-33128-9**

Leviton, L. C., Khan, L. K., Rog, D., Dawkins, N., & Cotton, D. (2010). Evaluability assessment to improve public health policies, programs, and practices. *Annual Review of Public Health, 31*, 213–233. **https://doi.org/10.1146/annurev.publhealth.012809.103625**

Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R. O., & Gupta, A. (2023). Electronic health record data quality assessment and tools: A systematic review. *Journal of the American Medical Informatics Association, 30*(10), 1730–1740. **https://doi.org/10.1093/jamia/ocad120**

Liu, F., Wang, H., Liang, S.-N., Jin, Z., Wei, S., Li, X., & Alzheimer's Disease Neuroimaging Initiative. (2023). MPS-FFA: A multiplane and multiscale feature fusion attention network for Alzheimer's disease prediction with structural MRI. *Computers in Biology and Medicine, 157*, Article 106790. **https://doi.org/10.1016/j.compbiomed.2023.106790**

Mankins, J. C. (2009). Technology readiness assessments: A retrospective. *Acta Astronautica, 65*(9–10), 1216–1223. **https://doi.org/10.1016/j.actaastro.2009.03.058**

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, Article 103655. **https://doi.org/10.1016/j.jbi.2020.103655**

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics, 19*(6), 1236–1246. **https://doi.org/10.1093/bib/bbx044**

Muhammad, D., & Bendechache, M. (2024). Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. *Computational and Structural Biotechnology Journal, 24*, 542–560. **https://doi.org/10.1016/j.csbj.2024.08.005**

Naveed, S., Stevens, G., & Robin-Kern, D. (2024). An overview of the empirical evaluation of explainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. *Applied Sciences, 14*(23), Article 11288. **https://doi.org/10.3390/app142311288**

Nazir, S., Dickson, D. M., & Akram, M. U. (2023). Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine, 156*, Article 106668. **https://doi.org/10.1016/j.compbiomed.2023.106668**

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. In G. Visaggio, M. T. Baldassarre, S. G. Linkman, & M. Turner (Eds.), *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE 2008)* (pp. 68–77). BCS Learning & Development. **https://doi.org/10.14236/ewic/EASE2008.8**

Prentzas, N., Kakas, A. C., & Pattichis, C. S. (2023). Explainable AI applications in the medical domain: A systematic review. *arXiv*. **https://doi.org/10.48550/arXiv.2308.05411**

Rasouli, S., Dakkali, M. S., Azarbad, R., Ghazvini, A., Asani, M., Mirzaasgari, Z., & Arish, M. (2024). Predicting the conversion from clinically isolated syndrome to multiple sclerosis: An explainable machine learning approach. *Multiple Sclerosis and Related Disorders, 86*, Article 105614. **https://doi.org/10.1016/j.msard.2024.105614**

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 1135–1144). Association for Computing Machinery. **https://doi.org/10.1145/2939672.2939778**

Schouten, J. S., Kalden, M. A. C. M., van Twist, E., Reiss, I. K. M., Gommers, D. A. M. P. J., van Genderen, M. E., & Taal, H. R. (2024). From bytes to bedside: A systematic review on the use and readiness of artificial intelligence in the neonatal and pediatric intensive care unit. *Intensive Care Medicine, 50*(11), 1767–1777. **https://doi.org/10.1007/s00134-024-07629-8**

Shetab Boushehri, S., Essig, K., Chlis, N.-K., Herter, S., Bacac, M., Theis, F. J., Glasmacher, E., Marr, C., & Schmich, F. (2023). Explainable machine learning for profiling the immunological synapse and functional characterization of therapeutic antibodies. *Nature Communications, 14*, Article 7888. **https://doi.org/10.1038/s41467-023-43429-2**

Shortliffe, E. H., & Cimino, J. J. (Eds.). (2014). *Biomedical informatics: Computer applications in health care and biomedicine* (4th ed.). Springer. **https://doi.org/10.1007/978-1-4471-4474-8**

Tasnim, J., & Hasan, M. K. (2023). CAM-QUS guided self-tuning modular CNNs with multi-loss functions for fully automated breast lesion classification in ultrasound images. *Physics in Medicine & Biology, 69*(1), Article 015018. **https://doi.org/10.1088/1361-6560/ad1319**

Tchetchenian, A., Zekelman, L., Chen, Y., Rushmore, J., Zhang, F., Yeterian, E. H., Makris, N., Rathi, Y., Meijering, E., Song, Y., & O'Donnell, L. J. (2024). Deep multimodal saliency parcellation of cerebellar pathways: Linking microstructure and individual function through explainable multitask learning. *Human Brain Mapping, 45*(12), Article e70008. **https://doi.org/10.1002/hbm.70008**

Toussaint, P. A., Leiser, F., Thiebes, S., Schlesner, M., Brors, B., & Sunyaev, A. (2024). Explainable artificial intelligence for omics data: A systematic mapping study. *Briefings in Bioinformatics, 25*(1), Article bbad453. **https://doi.org/10.1093/bib/bbad453**

Wang, Y., Yang, Z., & Yao, Q. (2024). Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. *Communications Medicine, 4*, Article 59. **https://doi.org/10.1038/s43856-024-00486-y**

Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science, 4*, Article 67. **https://doi.org/10.1186/1748-5908-4-67**

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association, 20*(1), 144–151. **https://doi.org/10.1136/amiajnl-2011-000681**

World Health Organization. (2023, May 16). *WHO calls for safe and ethical AI for health*. **https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health**

Xie, K., Hou, Y., & Zhou, X. (2024). Deep centroid: A general deep cascade classifier for biomedical omics data classification. *Bioinformatics, 40*(2), Article btae039. **https://doi.org/10.1093/bioinformatics/btae039**

Yadav, P. K., Burks, T., Vaddi, S., Qin, J., Kim, M., Ritenour, M. A., & Vasefi, F. (2024). Detection of *E. coli* concentration levels using CSI-D+ handheld with UV-C fluorescence imaging and deep learning on leaf surfaces. In M. S. Kim & B.-K. Cho (Eds.), *Sensing for Agriculture and Food Quality and Safety XVI* (Proceedings of SPIE, Vol. 13060, Paper 1306006). SPIE. **https://doi.org/10.1117/12.3014017**

Yang, Z., Wen, J., Abdulkadir, A., Cui, Y., Erus, G., Mamourian, E., Melhem, R., Srinivasan, D., Govindarajan, S. T., Chen, J., Habes, M., Masters, C. L., Maruff, P., Fripp, J., Ferrucci, L., Albert, M. S., Johnson, S. C., Morris, J. C., LaMontagne, P., … Davatzikos, C. (2024). Gene-SGAN: discovering disease subtypes with imaging and genetic signatures via multi-view weakly-supervised deep clustering. *Nature Communications, 15*, Article 354. **https://doi.org/10.1038/s41467-023-44271-2**

Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., Burgert, O., & Mathis-Ullrich, F. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Scientific Reports, 14*, Article 3713. **https://doi.org/10.1038/s41598-024-54186-7**

Zhao, Q., Zhang, Y., Zhu, M., Gu, S., Gao, Y., Yang, X., & Zhao, L. (2024). DUE: Dynamic uncertainty-aware explanation supervision via 3D imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6335–6343). Association for Computing Machinery. **https://doi.org/10.1145/3637528.3671641**

Zheng, K., Yu, S., Li, B., Jenssen, R., & Chen, B. (2025). BrainIB: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems, 36*(7), 13066–13079. **https://doi.org/10.1109/TNNLS.2024.3449419**

Zhong, Z., Li, J., Sollee, J., Collins, S., Bai, H., Zhang, P., Healey, T., Atalay, M., Gao, X., & Jiao, Z. (2025). Multi-modality regional alignment network for Covid X-ray survival prediction and report generation. *IEEE Journal of Biomedical and Health Informatics, 29*(5), 3293–3303. **https://doi.org/10.1109/JBHI.2024.3417849**

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics, 10*(5), Article 593. **https://doi.org/10.3390/electronics10050593**

Zou, L., Goh, H. L., Liew, C. J. Y., Quah, J. L., Gu, G. T., Chew, J. J., Kumar, M. P., Ang, C. G. L., & Ta, A. W. A. (2023). Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections. *IEEE Transactions on Artificial Intelligence, 4*(2), 242–254. **https://doi.org/10.1109/TAI.2022.3153754**

Appendix

A1 Evaluation-Ready XAI Systems

XAI System	Authors	Year	Description and GitHub Repository
APPRAISE	Ding et al.	2024	APPRAISE enhances explainability in protein engineering via physics-informed pairwise modeling to rank binding propensities. **GitHub:** https://github.com/GradinaruLab/APPRAISE
Autism 3D CNN	Garcia & Kelly	2024	Autism 3D CNN enhances autism diagnosis from brain sMRI using a 3D deep learning model, integrating guided Grad-CAM for interpretable feature identification. **GitHub:** https://github.com/garciaml/Autism-3D-CNN-brain-sMRI
BayeSeg	Gao et al.	2023	BayeSeg enhances medical image segmentation via an interpretable Bayesian framework, decomposing shape and appearance features for better generalizability. **GitHub:** https://zmiclab.github.io/projects.html
BrainIB	Zheng et al.	2025	BrainIB enhances psychiatric diagnosis using an information bottleneck-based GNN, analyzing fMRI data to identify explainable and clinically relevant biomarkers. **GitHub:** https://github.com/SJYuCNEL/brain-and-Information-Bottleneck
CAM-QUS	Tasnim & Hasan	2023	CAM-QUS enhances breast ultrasound diagnosis with modular CNNs and novel loss functions, integrating XAI to highlight clinical lesions. **GitHub:** https://github.com/jarin090/CAM-QUS-guided-modular-CNNs-for-BUS-classification
CIS2MS	Rasouli et al.	2024	CIS2MS predicts CIS-to-MS conversion using XGBoost and SHAP, improving transparency and aiding personalized treatment with interpretable feature importance. **GitHub:** https://github.com/rasoulisaeid/CIS2MS
CraniumPy	Abdel-Alim et al.	2023	CraniumPy quantifies cranial shape using AI, ensuring privacy, avoiding age bias, and providing an explainable FP score linked to clinical severity. **GitHub:** https://github.com/T-AbdelAlim/CraniumPy
CT-xCOV	Elbouknify et al.	2023	CT-xCOV enables explainable COVID-19 diagnosis from CT scans by integrating lung segmentation, deep learning classification, and XAI. **GitHub:** https://github.com/ismailelbouknify/CT-xCOV
DeepCentroid	Xie et al.	2024	DeepCentroid employs ensemble learning with a cascade structure, enhancing interpretability by identifying biologically significant features. **GitHub:** https://github.com/xiexiexiekuan/DeepCentroid
DeepMSP	Tchetchenian et al.	2024	DeepMSP clusters cerebellar pathways using multimodal saliency, integrating structural and functional data with explainable deep networks for brain analysis. **GitHub:** https://github.com/SlicerDMRI/DeepMSP
DUE	Zhao et al.	2024	DUE enhances 3D medical imaging with uncertainty-aware explanation supervision, improving deep learning predictability by addressing spatial and data variances. **GitHub:** https://github.com/AlexQilong/DUE
Ensemble XAI	Zou et al.	2023	Ensemble XAI predicts mortality risk in pneumonia and COVID-19 using SHAP and Grad-CAM++, providing deep learning explainability. **GitHub:** https://github.com/IHIS-HealthInsights/Interpretation-Methods-Voting-dashboard
Gene-SGAN	Yang et al.	2024	Gene-SGAN uncovers disease subtypes by linking phenotypic and genetic data via multi-view, semi-supervised clustering with explainable latent variables. **GitHub:** https://github.com/zhijian-yang/GeneSGAN

XAI System	Authors	Year	Description and GitHub Repository
Histo Cartography	Jaume et al.	2021	HistoCartography simplifies entity-graph analysis for histopathology, integrating preprocessing, machine learning, and XAI tools for tissue structure analysis. **GitHub:** https://github.com/BiomedSciAI/histocartography
IEDL	Halinkovic et al.	2024	IEDL enables explainable heart tissue segmentation using nuclei features, multiscale inputs, and attention mechanisms to emulate histopathologists' decisions. **GitHub:** https://github.com/mathali/IEDL-segmentation-of-heart-tissue
KnowDDI	Wang et al.	2024	KnowDDI predicts drug interactions using GNNs and biomedical knowledge graphs, enriching drug representations with interpretable subgraphs for robustness. **GitHub:** https://github.com/LARS-research/KnowDDI
MAG-MS	He et al.	2024	MAG-MS enhances medical image segmentation with modality-agnostic self-distillation, improving accuracy, robustness, and adaptability. **GitHub:** https://github.com/kisonho/magnet
Melanoma Classification	Gamage et al.	2024	Melanoma Classification combines CNNs and vision transformers with mask-guided segmentation, enhancing diagnosis via Grad-CAM. **GitHub:** https://github.com/LU-Bio-Vision/Melanoma-identification--FYP
MICA	Bie et al.	2024	MICA enhances skin lesion diagnosis by semantically aligning medical images with clinical concepts at multiple levels, improving explainability and accuracy. **GitHub:** https://github.com/Tommy-Bie/MICA
MPS-FFA	Liu et al.	2023	MPS-FFA improves sMRI-based Alzheimer's diagnosis with multiplane and multiscale attention, integrating feature fusion and similarity discrimination. **GitHub:** https://github.com/LiuFei-AHU/MPSFFA
MRANet	Zhong et al.	2025	MRANet generates explainable radiology reports and predicts survival by leveraging region-specific visual grounding and a survival attention mechanism. **GitHub:** https://github.com/zzs95/MRANet
PoSHAP	Dickinson & Meyer	2022	PoSHAP adapts SHAP for positional interpretation of LSTM models on biological sequences, revealing peptide properties and amino acid dependencies. **GitHub:** https://github.com/jessegmeyerlab/positional-SHAP
ScifAI	Shetab Boushehri et al.	2023	ScifAI enables explainable ML and predictive analysis for imaging flow cytometry, improving antibody screening and functional characterization. **GitHub:** https://github.com/marrlab/scifAI-notebooks
ST-CNN-GAP-5	Anand et al.	2022	ST-CNN-GAP-5 detects cardiac disorders from ECG data using a deep neural network, enhancing explainability via SHAP to highlight clinically relevant features. **GitHub:** https://github.com/tusharkadian/BSPC
3D Otitis Media	Chen et al.	2024	3D Otitis Media uses a 3D CNN-based XAI system to evaluate chronic otitis media from CT scans, providing interpretable heatmaps. **GitHub:** https://github.com/huntlylee/3D-Otitis-Media
TransXAI	Zeineldin et al.	2024	TransXAI enables accurate glioma segmentation in brain MRIs using a hybrid Transformer-CNN framework, providing surgeon-understandable heatmaps for XAI. **GitHub:** https://github.com/razeineldin/TransXAI
YOLOv8	Yadav et al.	2024	YOLOv8 enhances rapid, noninvasive E. coli detection on citrus peels using the CSI-D+ system, integrating Eigen-CAM for explainable bacterial classification. **GitHub:** https://github.com/rigvedrs/YOLO-V8-CAM

Table A1. Full List of Evaluation-Ready XAI Systems

Privacy Policy Feature Extraction for Direct-to-Consumer Genetic Testing

Digital Health, Winter Term 24/25

Luisa Faust
Master's Student
Karlsruhe Institute of Technology
ufwim@student.kit.edu

Abstract

Background: *Direct-to-consumer genetic testing (DTC-GT) has gained popularity in recent years due to social media marketing. However, while DTC-GT services collect highly sensitive genetic and biological data, their privacy policies are often lengthy, vague and difficult to understand for consumers, leading to privacy concerns. Existing research deals with the analysis of privacy policies in general but there is no domain-specific, automated tool to extract and assess privacy policy content in DTC-GT. This paper aims to close that gap, presenting a framework for the automated analysis of privacy policies of DTC-GT companies using the Longformer model.*

Objective: *The study focuses on the extraction of content of DTC-GT privacy policies by dividing them into 22 categories adapted to DTC-GT. Furthermore, an evaluation framework across multiple dimensions is developed to assess the extraction of critical privacy related features.*

Methods: *The model training consists of three steps, including the already existing pretraining of the Longformer model, an unsupervised domain adaption and a supervised fine-tuning. The unsupervised adaption uses 16,062 health-related privacy policies to provide exposure to domain-specific legal terminology. 29 DTC-GT policies were manually labeled of which 19 were used for the supervised fine-tuning.*

Results: *The model was evaluated on the 10 remaining manually labeled DTC-GT privacy policies, extracting content into 22 categories (220 possible scores). Since 58 pairs lacked content, 162 scores were analyzed. A custom evaluation assessed completeness, semantic similarity, precision, and information density. Based on weighted scores, results were labeled EXCELLENT (≥ 0.80), FAIR (0.60–0.79), POOR (0.40–0.59), or INADEQUATE (< 0.40). Most extractions (n = 59) were FAIR, 7 were EXCELLENT and 39 POOR. Best results appeared in categories related to company sale (mean = 0.79) and data deletion (mean = 0.72). Weakest in research and storage (means ~0.42). Top companies were 24Genetics (0.76), GenePlanet (0.67), and MyHeritage (0.67). A negative correlation (r = –0.73) between policy length and information density suggests that lengthy, vague policies reduce extraction quality.*

Conclusion: *The results demonstrate that while the framework effectively extracts policy content, the variability and vagueness within policies pose challenges for achieving consistent semantic and syntactic quality. In addition, by showing that essential content cannot be extracted due to its absence across various categories, significant privacy concerns are underlined.*

Keywords: privacy policy, direct-to-consumer genetic testing, natural language processing, longformer, data protection, automated text analysis, digital health

Introduction

In the age of big data, more people are concerned of their privacy, especially in areas where unique personal data is involved. One area in which highly sensitive data is dealt with is DTC-GT. The National Human Genome Research Institute defines DTC-GT as "genetic tests sold directly to consumers to provide information about their genetic information (generally ancestry, some health traits and health risks) from a saliva sample" (National Human Genome Research Institute, 2023). Being promoted on social media platforms like Instagram and TikTok, DTC-GT's market size grew over the last years, reaching USD 2.43 billion in 2024, with a compound annual growth rate (CAGR) of 24.43% (BioSpace Editorial Team, 2024).

However, this fast-growing industry has brought up urgent concerns regarding the transparency and compliance of DTC-GT company's privacy policies. Despite the importance for consumers to understand how companies use their data, numerous DTC-GT privacy policies are verbose and complex, reaching reading levels considerably above the comprehension capacity of most customers, impeding their ability to make educated decisions on how they want the company to handle their data. The problem of consumers not knowing how companies handle their data is especially concerning when customers confuse permission agreements, leading to DTC-GT corporations being allowed to share their genetic data to other parties (Hendricks-Sturrup & Lu, 2019). Furthermore, discrepancies and lack of specificity in privacy policies regarding compliance with regulations like the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) increase issues in guaranteeing transparency, user rights, and safe data practices (Sadri, 2024; Zaeem & Barber, 2020). These issues necessitate a possibility to analyze verbose and long legal texts in a way that makes it easier for consumers to read about the most important issues of their privacy. Although much research on privacy policy analysis exists, the handling of sensitive genetic data requires a domain-specific approach.

This paper aims to address these challenges by developing a Natural Language Processing (NLP)-based Longformer model that extracts important content of the policies by separating them into 22 categories tailored to DTC-GT policies. Providing category-wise information extractions creates the opportunity for customers to look for information that they find interesting in a structured and comprehensive way. The resulted extractions are evaluated by a tailored framework, which includes the comparison to its manually labeled data concerning semantic similarity and completeness while adding precision and information density to not only ensure accurate extraction but also readability and clarity for users. The results of this evaluation highlight potential privacy concerns given the model is only capable to retrieve information that is written within the policy. When the model is not able to extract certain information within a category, the reason is not only the incapability of the model but can also be a result of vague formulations and missing transparencies within the policies, ultimately leading to ratings being lower. The tailored evaluation framework provides insights into the quality of the model extractions as well as the policies themselves.

The remaining part of the paper is structured as follows:

Section *Related Work* provides some background on automated privacy policy analysis and reviews it's related literature; Section *Methodology* describes the methodology, including the model architecture, data preprocessing, and the training framework; Section *Results* elaborates on the results on a category and company level; Finally, Section *Discussion* outlines the findings by highlighting their implications and limitations.

Related Work

Because of the complex and lengthy nature of privacy policies, consumers often do not understand the content nor read the policy at all. For this reason, automating the analysis of privacy policies using NLP provides a solution to the problem of time-consuming and difficult to read policies, adding great value for consumers. The models for the automation have developed into a multidisciplinary field, incorporating classical, empirical, statistical, and advanced techniques (Goldberg & Hirst, 2017; Indurkhya & Damerau, 2010).

Several areas of related research provide relevant background for the analysis in this paper. The systematic research review conducted by Javed and Sajid (2024) offers a broad overview of research papers dealing with various areas of privacy policies. Alongside data handling practices, they also provide input of privacy policies across various sectors such as healthcare. The research mapping by Del Alamo et al. (2022) also

deals with privacy policies but focuses more on the technical methods in NLP, which were used to handle transparency issues of these legal documents. The models discussed within the research mapping implement both supervised and unsupervised learning as well as neural networks.

Classical Approaches

Traditional NLP models are based on rule-based systems and linguistic concepts. The research of d'Aquin et al. (2017) utilized a classical approach by creating a semantic framework called PrivOnto. It analyzed 23,000 data practices from 115 distinct privacy regulations using SPARQL queries and ontology-based annotations. Similarly, Bhatia et al. (2016) analyzed the objectives of privacy policies by using a semi-automatic method that combines crowdworker input and automated components.

Supervised and Unsupervised Learning

The traditional methods offer a useful basis for NLP. However, given the massive volumes of data that are produced daily, the manual labor required for text data analysis has its limitations. Supervised learning provides a solution for these limitations by directly learning the patterns from the data. This increases the scalability and flexibility of models like the Support Vector Machine (SVM), which was utilized by Wilson et al. (2018) to use complex textual patterns for categorizing privacy regulations. Guntamukkala et al. (2015) applied Random Forest (RF) and Decision Tree (DT) algorithms for the assessment of the completeness of privacy policies. They showed that filtering key words and features can increase the accuracy of the model. Costante et al. (2012) further analyzed supervised methods, comparing different algorithms for text classification tasks, including Naïve Bayes, Ridge Regression, and k-Nearest Neighbor. Supervised learning models always require an annotated dataset for its training, resulting in limitations for large and unstructured datasets where labeling the data is either unavailable or costly to produce.

Unsupervised learning is capable of identifying patterns in data on its own, thereby providing a useful addition to supervised models. One example is the use of Latent Dirichlet Allocation (LDA) for topic modeling in privacy rules implemented by Massey et al. (2013). Furthermore, Ramanath et al. (2014) used hidden Markov models (HMM) to automatically sort sections of over 1000 privacy statements and summarized them into topics such as data processing, sharing or deletion to improve the understanding of privacy statements. Unsupervised learning employs unlabeled text to find patterns and generate general-purpose features, whereas supervised learning uses labeled data to develop more particular and focused models, which often improves accuracy. This is why, Talukdar and Biswas (2024) suggested combining both supervised and unsupervised learning to provide a model which can deal with large data while still being tailored to a particular task or domain.

Neural Networks and Transformer Models

Recent developments in neural network models have opened up new possibilities for privacy policy analysis, building on the NLP techniques previously discussed. Neural networks are able to process and model complex, high-dimensional data like text. Both supervised and unsupervised environments can be utilized with neural networks, which allow them to automatically acquire multi-level, hierarchical features that capture both task-specific nuances and generic patterns.

The Convolutional Neural Network (CNN) is one of the neural network models that researchers like Harkous et al. (2018) implemented to automatically divide privacy regulations into semantically coherent sections and classified them according to data practices. Similarly, Lebanoff et al. (Lebanoff & Liu, 2018) applied an LSTM model, which belongs to the recurrent neural network (RNN) family, to classify terms according to their surrounding context, ultimately identifying patterns of vagueness in the language used in privacy policies. Although CNNs and RNNs have made notable progress in handling structured and sequential data, they still struggle with long range dependencies and parallelizing computations.

These issues motivated the invention of the Transformer architecture. Introduced by Vaswani et al. (2017), Transformers use self-attention mechanisms to effectively model both local and global dependencies in texts. These mechanisms enable high performance on tasks such as machine translation, summarization and question answering (Tay et al., 2022). Although transformer models have the ability to recognize complex relationships, the quadratic computational complexity of Transformers poses a significant

challenge when it comes to the processing of long sequences. This issue is especially relevant in fields like law and healthcare, where documents are very lengthy and complex. Beltagy et al. (2020) proposed the Longformer model, a Transformer version designed specifically for managing long text input. It reduces complexity to linear time by using a sliding window attention method. Building on the foundation of Longformer, Xiao et al. (2021) implemented Lawformer, which is a specialized adaptation for Chinese legal documents. This shows that transformer-based models can be used for domain specific areas such as legal documents.

Domain-Specific Gaps

Despite major advances in automated privacy policy analysis, there has been limited study explicitly on the privacy policies of genetic testing providers. Given the intrinsically sensitive nature of DNA information, which necessitates strict transparency and compliance mechanisms to secure its use, this gap is especially concerning. Although studies like Onstwedder et al. (2024) and Hendricks-Sturrup et al. (2019) highlighted the critical need for secure data protocols and increased transparency in DTC-GT, there is yet no research that employs NLP to analyze these domain specific policies, to the best of my knowledge.

This paper uses a neural network architecture based on the Longformer model (Beltagy et al., 2020) to analyse privacy policies of DTC-GT, responding to the gap of scientific work in this specific area.

Methodology

Legal documents dealing with sensitive genetic data like DTC-GT privacy policies are often verbose and lengthy, frequently exceeding the size of traditional Transformer models, such as Google's BERT model, which has a token limit of 512 tokens (Devlin et al., 2019). Furthermore, the exposure to sensitive genetic data leads to a complexity of content that requires a deeper understanding of the local and global context. With the functionality to manage 4,096 tokens and the ability to implement local sliding window attention and global attention masks, the Longformer model is a suitable option to examine long and complex DTC-GT policies.

This section describes the methodology including the procedure of unsupervised training on general health-related privacy policies to capture legal representations, followed by supervised fine-tuning on labeled DTC-GT privacy policies to classify domain specific key topics. Furthermore, the general technical framework is introduced, encompassing an in-depth examination of the model's architecture, attention mechanisms, and training methods.

Model Architecture

This study employs the pre-trained Allenai/Longformer-Base-4096 model for the extraction of content from DTC-GT privacy policies (Beltagy et al., 2020). The model is based on the Longformer architecture and supports three main attention mechanisms.

The first mechanism is the sliding window attention, which creates attention links between each token within a fixed window size, enabling the model to focus on local patterns and relationships in the input data. The default size of tokens is 512 and achieved optimal performance in the original Longformer research (Beltagy et al., 2020). In addition, the default value aligns with the computational patterns of RoBERTa, a BERT pretraining replication study that assesses the influence of hyperparameters and training data volume (Beltagy et al., 2020; Devlin et al., 2019). To balance between computational resources and effective performance in document-level tasks, the default token size of 512 is set for this model.

The second mechanism involves global attention, implemented using the *[CLS]* token, which is set before each sequence and the *[SEP]* token, which signals the end of each sequence. A sequence is established within the model based on the training stage of the procedure. During the unsupervised phase, each privacy policy is tokenized into a single sequence, starting with a *[CLS]* token at the beginning and concluding with a *[SEP]* token at the end of each policy. This guarantees that the model perceives the entire policy as a unified document and assimilates its overall language and structure.

During the supervised fine-tuning phase, each labeled policy is segmented into separate sections, with a *[CLS]* token positioned at the beginning and a *[SEP]* token at the end of each labeled section. This enables

the model to focus on specific parts within the policy and learn how to classify its sections one by one. The distinction between training phases ensures that the model benefits from the global context of the whole policy as well as the detailed, section-specific information.

The third mechanism is the dilated sliding window configuration, which creates attention links by skipping intermediate tokens in the sliding window, allowing the model to capture longer-range dependencies. The dilation rate is set to the default value of one. Because this means that each token attends to its immediate neighboring tokens, it does not differentiate to the general sliding window attention mechanism when set to the default value. Because the most related information in the privacy policies appear in consecutive sentences and the global attention using *[CLS]* and *[SEP]* already provides long-range dependencies, the default value is sufficient for the analysis of DTC-GT privacy policies.

Figure 1 illustrates an example sentence from the privacy policy of 23andMe to further explain the attention mechanisms employed (23andMe, 2024; Xiao et al., 2021). The size of the sliding window as well as the dilated sliding window attention is 1. The tokens, *[CLS]* and *[SEP]*, are selected to perform the global attention.

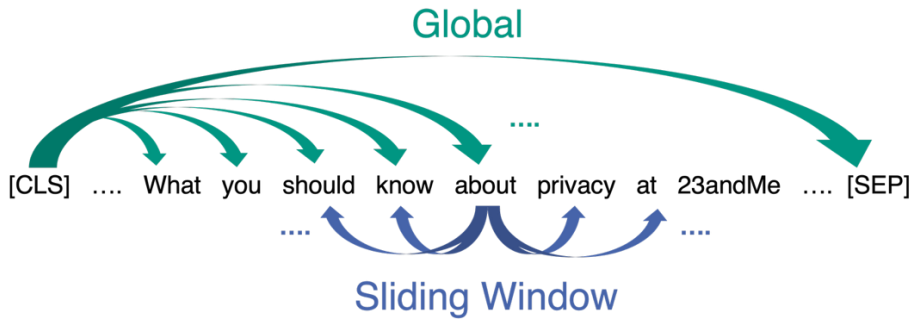

Figure 1. Illustration of the Combination of the Attention Mechanism Used by Longformer

Text Processing Pipeline

To enable the model to process data effectively, it is essential to implement efficient preprocessing of all text that will be included in the model training. For this reason, a processing pipeline was implemented.

The initial step of the pipeline involves the loading of the required data. In the context of unsupervised training, a dataset comprising 1,071,488 English-language privacy policy snapshots sourced from 130,620 unique websites spanning the years 2009 to 2019, established by Amos et al. (2021), was utilized. From this dataset, 16,062 health-related policy snapshots from 6,006 distinct websites were selected specifically for the model to generate an exposure to the usage of language in policies that are related to the health sector. For the model, it was important to not just include one policy per company, but to process several time-series snapshots of the same policy, providing insights into temporal variations. It enabled the learning of policy patterns and structures over a series of time. During the supervised fine-tuning phase, 19 out of the 29 labeled privacy policies from DTC-GT companies were categorized into 22 key topics. These topics, outlined in Table 1, were specifically chosen to analyze the domain specific content in DTC-GT privacy policies. The remaining 10 policies were utilized for the model evaluation process. Prior to using the loaded data within the model, it is essential to perform data cleaning to ensure that the model can effectively process the information. Cleaning the training data involved normalizing line breaks, removing special characters while maintaining sentence structure, normalizing whitespace (converting multiple spaces to a single space), and preserving sentence boundaries to ensure proper segmentation. To further segment the document, the policies were separated into sentences using regular expressions, which ensured that the integrity of each sentence was preserved.

Category	**ID**
Personal Information which is collected and why	1
Specific risks related to disclosure of results	2
Procedure for storage of biological samples	3
Procedure for storage of genetic data	4
That biological samples will be or can be requested to be destroyed	5
That genetic data will be or can be requested to be destroyed	6
Possibility to delete account and the information that was saved	7
Specific length of time for storage of biological samples	8
Specific length of time for storage of genetic data	9
Procedure for handling genetic data should the company be sold or go bankrupt	10
Procedure for handling genetic samples should the company be sold or go bankrupt	11
Arrangements to ensure confidentiality of genetic data and security of biological samples	12
Policy about third parties that may be granted access to genetic data or samples	13
Disclosure of data to law enforcement	14
Plans to use genetic data for health-related research	15
Specific purpose of health-related research	16
Plans to use genetic data for other or unspecified research	17
Statement that no research will be conducted using genetic data	18
Additional consent for health-related research	19
Policy for withdrawing from research	20
Information on whether the health-related research may lead to commercialization or patents	21
Customer's rights (or lack thereof) to commercial benefits from health-related research	22

Table 1. Category Descriptions

In addition, because of Longformer's 4,096 token limit, policies were divided into segments with a maximum of 4,000 tokens, which allowed some additional space for the special tokens *[CLS]* and *[SEP]*. To ensure that no seperation happens in between a sentence, which might lead to confusion of the model, the token counting was conducted using the Longformer tokenizer with sentence-aware chunking. While the LongformerTokenizer was used to convert the policies into tokens, the special tokens *[CLS]* and *[SEP]*

were incorporated as outlined in Section Model Architecture (Beltagy et al., 2020). For a standardized length of all token sequences, the data was padded to the maximum length by appending the *[PAD]* token. When including these padding tokens, it is important to create an attention mask, highlighting *[CLS]* and *[SEP]* to ensure the model does not confuse padding with other tokens.

Training Process

The training of the model consists of a pretraining, which is already provided by Beltagy et al. (2020), an unsupervised training on health related privacy policies and supervised training on labeled DTC-GT privacy policy data.

Pretraining (General Knowledge Foundation)

The Longformer model "allenai/longformer-base-4096" uses a Masked Language Modeling (MLM) for its pretraining, creating a strong foundation for general language tasks by masking random tokens from data, including a diverse corpus of text data like Books Corpus, English Wikipedia, and subsets of the Realnews and Stories dataset (Beltagy et al., 2020).

Unsupervised Legal Domain Adaptation

For the analysis of DTC-GT privacy policies, an exposure to domain-specific legal terminology is needed, which is why unsupervised training was conducted on 16,062 health-related privacy policy snapshots from Amos et al. (2021). The unsupervised domain adaptation was performed on an NVIDIA A100 GPU (40 GB) with the configurations shown in Table 2.

Configuration	**Value**
Batch size	6
Epochs	15
Learning rate	$2 \times 10^{\{-5\}}$
Validation split	10%
Gradient accumulation	4 Steps
Mixed Precision Training	Enabled using PyTorch AMP
Early stopping	Patience of 3 epochs with minimum delta of 0.0001

Table 2. Training Configurations for Unsupervised Domain Adaptation

Because a lower learning rate helps to retain the pretrained knowledge while adapting the model to the new legal specific content without supervision, the learning rate of $2 \times 10^{\{-5\}}$ was chosen for the unsupervised training, ensuring stable and gradual updates and preventing overfitting. The batch size, defined as the number of samples the model processes before updating its internal parameters during training, was set to be maximized without overloading the memory to ensure stable training. The epoch, which is defined as the complete pass through the entire training dataset, was first set to a smaller number of 15, again to not constrain memory.

As shown in Figure 2, after the second pass through the dataset, the training loss stayed constant, leading to an early stopping after 3 epochs. This shows that no further expansion of the epoch size is needed for the unsupervised training process.

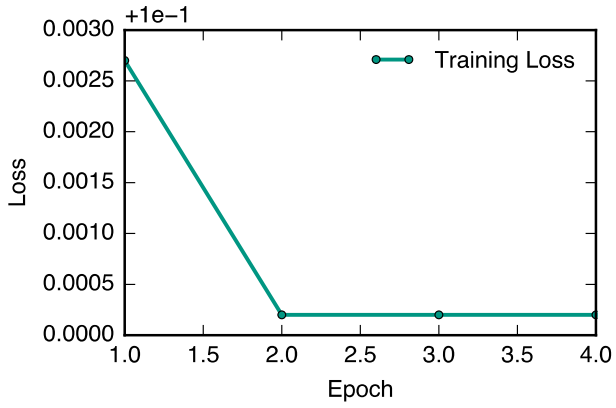

Figure 2. Training Loss During the Unsupervised Pretraining Phase

Supervised Fine-Tuning (Task-Specific Training)

To address DTC-GT-specific privacy language, the model was trained using 19 out of 29 manually labeled DTC-GT privacy policies. These policies were annotated with key topics relevant to genetic privacy and data handling (see Table 1). This approach ensured the model's exposure to genetic-specific language and context, enabling it to learn from sections annotated by categories. The supervised fine-tuning was conducted on an NVIDIA A100 GPU (40 GB) like the unsupervised training but with slightly different configurations shown in Table 3. The learning rate of $3 \times 10^{\{-5\}}$ was selected for supervised fine-tuning to allow faster adaptation to the task-specific labeled data. Since the labeled data provides explicit supervision, this higher learning rate accelerates convergence without risking model instability.

Configuration	**Value**
Batch size	6
Epochs	15
Learning rate	$3 \times 10^{\{-5\}}$
Optimizer	AdamW
Validation split	10%
Mixed Precision Training	Enabled using PyTorch AMP
Early stopping	Patience of 3 epochs with minimum delta of 0.0001

Table 3. Training Configurations for Supervised Domain Adaptation

In the epochs illustrated in Figure 3, the validation error exhibits only a slight decrease following epoch 8. To prevent overfitting, no additional epoch expansion was conducted. Combining unsupervised domain adaptation with supervised fine-tuning helps the model to understand the general legal language while including specific nuances of genetic data privacy, tailoring it for DTC-GT privacy policy analysis. The AdamW optimizer was selected based on its demonstrated success in transformer-based architectures, as highlighted in Beltagy et al. (2020). It is an improved version of the *Adam Algorithm* by Loshchilov and Hutter (2019), which has been implemented in TensorFlow and PyTorch by the community (PyTorch, 2017). Early stopping ensured training efficiency by halting the process if no significant improvement was observed in validation loss. Additionally, mixed precision training was enabled using PyTorch AMP to

optimize computational efficiency and reduce memory consumption (PyTorch, 2024). The training's validation split was 10% for the monitoring of the performance.

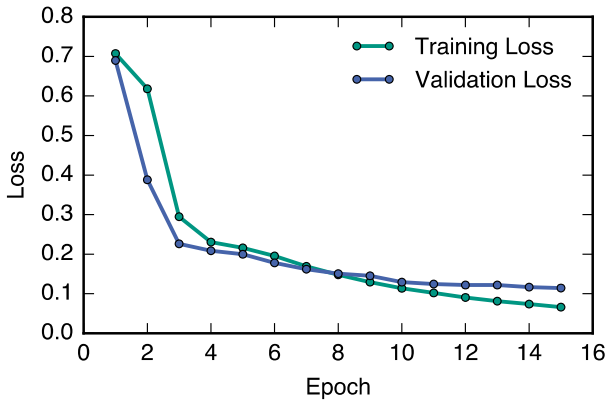

Figure 3. Training and Validation Loss During the Supervised Fine-Tuning Phase

Training Results

During unsupervised training, the average training loss decreased during the first epoch from 0.1027 to 0.1002, demonstrating the model's adaption to the privacy policy language and structure. Afterwards, the training loss stays constant, leading to an early stopping, showing that the model was able to learn fast, leaving it with the small error of 0.1002 (see Figure 2).

For supervised fine-tuning, the train and validation losses steadily decreased, reflecting the model's capability to learn the nuances of genetic privacy topics. The model starts with a training loss of 0.7074 and validation loss of 0.6897 in epoch 1 and stabilizes at 0.0660 training loss and 0.1145 validation loss in epoch 15 (see Figure 3).

Model Evaluation

The model evaluation framework is built to analyze the extracted content of each category and company, including a multi-metric approach consisting of four primary metrics, each weighted dynamically based on text characteristics to capture the variations in length and detail of each policy topic extraction.

Completeness Metric

The first metric, the *Completeness Metric*, is designed to evaluate how much of the labeled (ground-truth) text is captured in the extracted text. It ensures that the model does not miss important information. For the evaluation of the completeness, three components are taken into account.

The first component, weighted at 40% of the overall *Completeness Metric*, is *Sequence Matching*, which assesses the alignment of the extracted text with the labeled text based on sequential similarity. The "SequenceMatcher" class from Python's "difflib" module is used, which calculates a similarity ratio between the extracted and labeled texts using the Ratcliff-Obershelp algorithm (Black, 2021; Python Software Foundation, 2024). It highlights the importance of maintaining the order of information while also considering other aspects like covering the main content (*Word Overlap*) and keeping the text length reasonable (*Length Ratio*).

The second component, the *Word Overlap Analysis*, assesses the ratio of unique words in the labeled text that also appear in the extracted text. Unique words are defined as non-redundant terms from each text after converting both texts to lowercase to provide comparability. This method ensures that key terms in the labeled text are also in the extracted text, regardless of their order. Word overlap holds equal significance to sequence alignment, because it guarantees conceptual consistency while reducing dependency on structural similarity. This is reflected in its weighting of 40%.

The third component is the *Length Ratio*, which ensures that the extracted text is not disproportionately longer or shorter than the labeled text, by penalizing over-extraction when the extracted text is >20% longer than the labeled text. While proportionality is important, it is less critical than the accuracy of structure (*Sequence Matching*) and content (*Word Overlap*). The lower weight of 20% reflects that minor variations in length are tolerable if key information is captured.

The whole *Completeness Metric* is weighted 40%, leaving room for other metrics to account for meaning (semantic) and proportionality (length ratio).

Semantic Similarity Score

The *Semantic Similarity Score* is weighted 40% in the evaluation framework, prioritizing the meaning of extracts without overshadowing the weight of the *Completeness* measure. The score consists of the creation of vector representations called embeddings, the calculation of the similarity between extracted and labeled data and finally the normalization of the scores.

The first step is the *Embedding*, which uses the "all-MiniLM-L6-v2" model from the Sentence Transformers library, a pretrained transformer model specifically designed to create dense vector representations of text "all-MiniLM-L6-v2". Each text is transformed into a numerical vector, which is called an embedding, capturing the semantic meaning of the analyzed text.

For the second step the *Cosine Similarity* is calculated by using the "pytorch_cos_sim"-method from the "sentence-transformers"-library (Sentence Transformers, 2020) which measures the angular similarity between two embeddings, and enables the efficient calculation of semantic similarities. "pytorch_cos_sim" was developed and optimized by the Ubiquitous Knowledge Processing Lab in Darmstadt specifically to use transformer-embeddings, supporting the NVIDIA A100 GPU used in this paper's model (Aarsen, 2024).

The resulting scores are finally normalized to a range between zero and one for consistency and ease of interpretation, where zero indicates no semantic overlap and one indicates a perfect semantic alignment.

Precision Score

The *Precision Evaluation* includes scores for relevance, conciseness and applicability. Altogether, it is weighted 10% in the evaluation framework as it ensures the relevance and conciseness of the extracted text, without allowing it to overshadow more critical metrics like *Completeness* or *Semantic Similarity*.

A lot of privacy policies include irrelevant expressions like "click here", "see above", "contact us", "www", or "http". Starting with a *Base Scores* of 1.0, penalties of -0.1 for each use of such an expressions are added, resulting in the final *Base Score* used for the following calculation. This ensures that the extracted text does not include irrelevant content.

Because privacy policy extraction must present key clauses without irrelevant filler words that could mislead or confuse users, the *Density Factor* is used as a multiplier. This factor is calculated as the ratio of unique words to the total word count in the extract. The precision score is then measured as:

$$\text{Precision Score} = 0.7 + (0.3 * \text{Density Factor})$$

The multiplier is designed to prioritize the importance of text density while maintaining a minimum baseline score. Even if the density is low, the score will not drop below 0.7, ensuring that every text has a basic level of value. The additional 0.3 in the formula is used to reward texts that are more efficient, meaning those with higher density. This proportional reward encourages text to be concise and avoid unnecessary repetition, while still allowing less dense texts to retain a baseline score.

Since key information is to be extracted from the source text during categorization, a penalty for excessive verbosity is added with length normalization. It penalizes proportionally to the excess length of the text and starts with extracts that are >20% longer than the labeled text. This proportional penalty ensures that minor deviations in length do not affect the score excessively but penalizes verbose extracts.

Information Density

If a customer of DTC-GT would like to be informed about the handling of privacy data, it is crucial to avoid unnecessary repetition and to provide meaningful extractions which are concise and can easily be read. For this reason, *Information Density* focuses on *Word Retention* and *Word Efficiency*. Overall, the metric is weighted at 10%, the same as *Precision*, to ensure concise and relevant extractions. It has less weight than *Completeness* and *Semantic Similarity*, as these metrics are more critical for capturing the full meaning and accuracy of key clauses in a privacy policy.

The *Word Retention Score* focuses on the usage of meaningful language and is calculated as:

$$\text{WRS} = \frac{\text{Number of Common Words}}{\text{Total Number of Words in Labeled Text}}$$

Word Efficiency emphasizes the avoidance of repetition and is calculated as:

$$\text{WES} = \frac{\text{Number of Unique Words in Extracted Text}}{\text{Total Number of Words in Extracted Text}}$$

The *Word Retention Score* weighting 60% and the *Word Efficiency Score* weighting 40% creates the balance between the primary need for retaining critical content (*Word Retention*) with the secondary goal of conciseness (*Word Efficiency*).

For a better understanding, Figure 4 provides an overview of the evaluation framework including all relevant metrics.

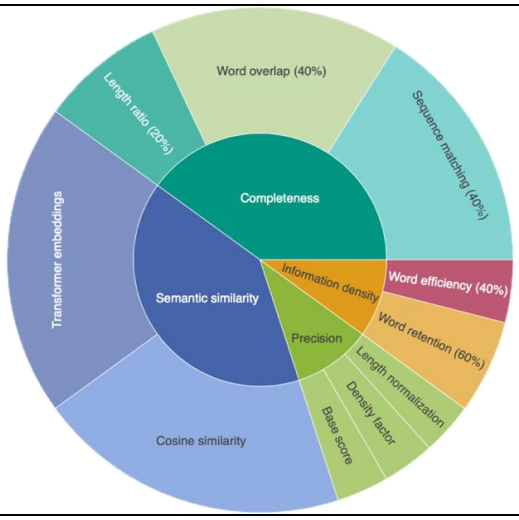

Figure 4. Evaluation Framework for Privacy Policy Extraction

Adding to the whole evaluation metrics is the implementation of dynamic weight adjustment based on text characteristics. For short texts (<100 characters), the *Semantic Similarity* weight is increased by up to 20% while the *Completeness* weight is decreased by up to 10%, calculated as:

$$w_{\{\text{semantic sim}\}} = w_{\{\text{semantic sim}\}} \times \left(1 + 0.2 \times (1 - \text{length factor})\right)$$

$$w_{\{\text{completeness}\}} = w_{\{\text{completeness}\}} \times \left(1 - 0.1 \times (1 - \text{length factor})\right)$$

For long texts (>500 characters), *Completeness* weight is increased by up to 20% while Semantic Similarity weight is decreased by up to 10%, calculated as:

$$w_{\{\text{completeness}\}} = w_{\{\text{completeness}\}} \times (1 + 0.2 \times \text{length factor})$$

$$w_{\{\text{semantic sim}\}} = w_{\{\text{semantic sim}\}} \times (1 - 0.1 \times \text{length factor})$$

Adding dynamic weights ensures the prioritization of the meaning for shorter texts and content coverage for longer texts. This is especially useful regarding the extraction of different content depending on the category as policies provide different amounts of content depending on the subject of the category.

The final scores are converted to qualitative ratings shown in Table 4.

Rating Category	Score Range
EXCELLENT	≥ 0.90
VERY GOOD	≥ 0.80
GOOD	≥ 0.70
FAIR	≥ 0.60
POOR	≥ 0.40
INADEQUATE	< 0.40

Table 4. Qualitative Ratings Based on Score Ranges

Results

After the training, the model is tested on 10 privacy policies from 10 different companies by extracting data and classifying the content in 22 key categories. For each of these classifications, the evaluation framework described in Section Model Evaluation scored the extractions concerning *Completeness*, *Semantic Similarity*, *Precision* and *Information Density*, providing 220 final weighted mean performances, from which 58 are 0 and not included in the overall calculations as the data is missing in the associated policy, leaving 162 scores included in the evaluation analysis shown in Figure 5. The categorization was derived from the scores presented in Table 4.

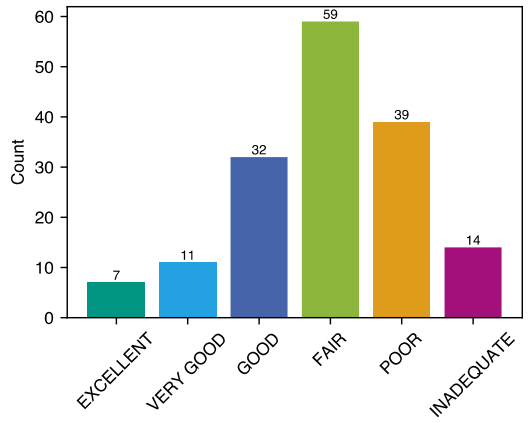

Figure 5. Privacy Policy Rating Distribution

With the majority of 59 policy categories scoring in the FAIR range of 0.60-0.70, the model shows a moderate level of alignment between the extracted content and the labeled content. Still, only 7 policy categories attained the EXCELLENT rating, indicating that a limited number of policies demonstrate superior performance in *Completeness*, *Semantic Similarity*, *Precision*, and *Information Density*.

Even though only 14 category policy extractions where inadequate, many fall below expectations, with a considerable number of 39 in the POOR range. The long tail in the lower performance categories (POOR and INADEQUATE) highlight areas where the model has difficulties to extract the right content. These difficulties are discussed in more detail by looking at performance at both category and company level.

Category-Wise Performance

The performance of the different categories is evaluated by averaging the category-scores over each of the 10 privacy policies, calculating standard deviations (SD) to provide information concerning the variations of the scores.

The results are illustrated in Figure 6 (category numbers are detailed in Table 1, categories labeled as 11, 18, 21, and 22 have no available content in either the policies or extractions).

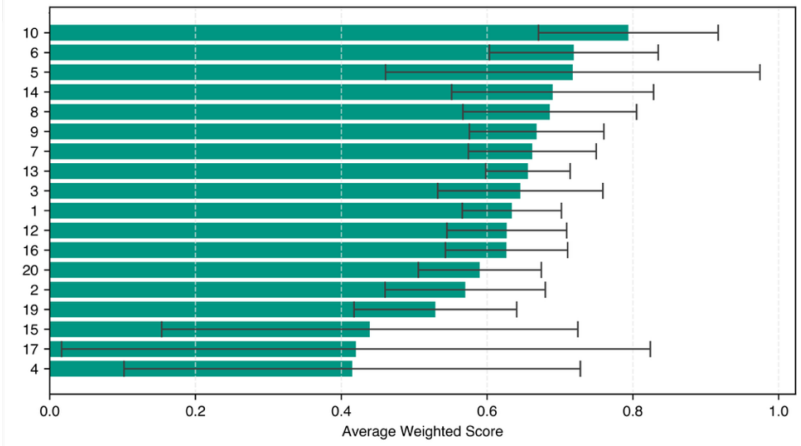

Figure 6. Category Performance with Standard Deviation

The highest performing category is the "procedure for handling genetic data during company sale/bankruptcy" with a mean score of 0.79 and a standard deviation of 0.12. In this category, the privacy policy of the company iGene achieved a high rating of 0.99, while the privacy policy of FamilyTree received a rating of 0.89, indicating that privacy policies in this category are well documented. This is likely due to regulatory requirements such as the GDPR in the EU, which requires a strict handling of sensitive data in transitions, or the HIPAA in the US, which mandates protections for health-related genetic data. The content of policies concerning the handling of genetic data during a company sale or bankrupcy case frequently specify that data may be retained or transferred to the new buyer of the company under similar privacy terms and in compliance with legal, regulatory, or contractual obligations. For instance, GenePlanet references Article 17(3) of the GDPR, which permits data retention in specific cases like legal claims or compliance (Geneplanet, 2023) and MyHeritage similarly notes that personal data may be transferred during an acquisition or restructuring, with safeguards to ensure consistency with existing privacy policies (My Heritage, 2024).

The categories that exhibit the second highest performance are the ones associated with the destruction of biological and genetic data. The category "that biological samples will be or can be requested to be destroyed" has a mean score of 0.72 and a standard deviation of 0.26 while the category "that genetic data will be or can be requested to be destroyed" also has a mean score of 0.72, but with a lower standard deviation of 0.12. The high performance is likely due to legal mandates such as Art. 17 GDPR – Right to erasure ("right to be forgotten"), the HIPAA Security Rule, which mandates the save deletion, if data is not used anymore, or the UK Data Protection Act, which requires clear and secure data destruction procedures for sensitive information.

The categories with the lowest performance apply to health-related research plans and the protocols for the storage of genetic data. The category "plans to use genetic data for other or unspecified research" has a mean of 0.42 (SD: 0.40). The category "plans to use genetic data for health-related research" scores 0.44 (SD: 0.28). The category "additional consent for health-related research" scores 0.53 (SD: 0.11). The category "procedure for storage of genetic data" also scores low with a mean of 0.42 (SD: 0.31). These poor ratings draw attention to shortcomings in regulatory compliance and transparency, which have an immediate effect on the model's performance. Genetic data research plans often lack sufficient specificity, raising concerns about compliance with regulatory frameworks like GDPR Art. 5, which emphasizes fairness, transparency, and limitations on data storage and use (Mittelstadt & Floridi, 2016). Since the model

depends on exact extractions using precise terminology, it has trouble extracting important information when the input regulations themselves are ambiguous or lacking. For instance, vague data storage procedures result in fragmented and inconsistent information, impacting compliance with privacy rules. Knoppers (2014) emphasizes the need for precise protocols for the responsible exchange of genetic and health data. The focus is on responsibility, transparency and the protection of privacy. The problem is worsened by regionally inconsistent regulations and unclear standards that lead to different treatment of genetic data (Knoppers, 2014). This inconsistency affects the quality and accuracy of the analyses.

Company-Wise Performance

The model performed best on the privacy policies of 24Genetics, Geneplanet and MyHeritage with 0.76, 0.67 and 0.67 mean performance (see Figure 7).

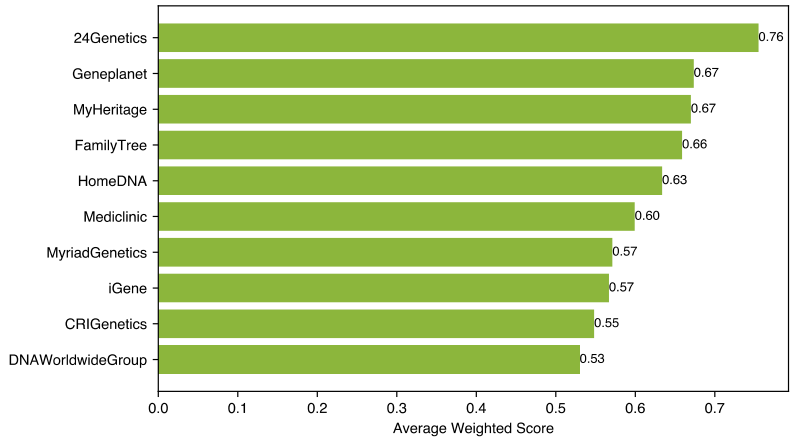

Figure 7. Company Average Performance

Figure 8 shows in more detail how the model performed per company policy through the different categories. The heatmap shows company performance across all categories from Table 1. Each cell displays the weighted score; darker red indicates higher performance. White cells mark missing content with zero scores. 24Genetics, for instance, performs well over all categories, from the lowest of 0.49 in category "additional consent for health related research" to 0.97 in "disclosure of data to law enforcement".

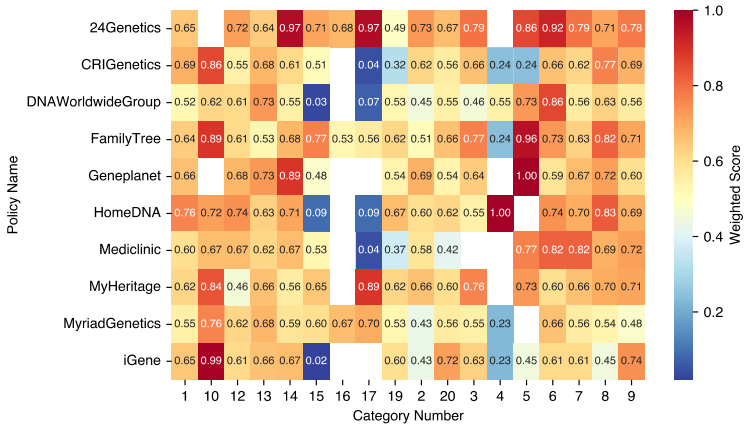

Figure 8. Company Performance Across Categories

Furthermore, as seen by the correlation matrix in Table 5, lengthier policies tend to perform worse, evidenced by a correlation of -0.24 between original length and mean performance. This is particularly influenced by information density, which has a correlation of -0.73 with original length. It suggests that

lengthy and verbose sentences are more challenging for the model to extract, as extended policies frequently contain duplicated or less pertinent information. Furthermore, the model assesses policy extractions with greater information density more favorably. Consequently, section extractions that are more verbose receive lower scores.

	Mean	**Std**	**Len.**	**Words**	**Sent.**	**Compl.**	**Sem.**	**Prec.**	**Dens.**
Mean	1.00	-0.58	-0.24	-0.24	-0.26	0.97	0.97	0.29	0.71
Std	-0.58	1.00	-0.45	-0.45	-0.38	-0.48	-0.66	-0.47	-0.06
Length	-0.24	-0.45	1.00	1.00	0.94	-0.28	-0.18	0.39	-0.73
Words	-0.24	-0.45	1.00	1.00	0.95	-0.28	-0.19	0.40	-0.73
Sent.	-0.26	-0.38	0.94	0.95	1.00	-0.33	-0.21	0.61	-0.82
Compl.	0.97	-0.48	-0.28	-0.28	-0.33	1.00	0.89	0.18	0.75
Sem.	0.97	-0.66	-0.18	-0.19	-0.21	0.89	1.00	0.27	0.66
Prec.	0.29	-0.47	0.39	0.40	0.61	0.18	0.27	1.00	-0.38
Dens.	0.71	-0.06	-0.73	-0.73	-0.82	0.75	0.66	-0.38	1.00

Table 5. Correlation Matrix of Metrics

Mean - Overall weighted performance score; Std - Standard deviation of scores; Words - Number of words in the policy; Sent. - Number of sentences in the policy; Compl. - Completeness metric score; Sem. - Semantic similarity score; Prec. - Precision score; Dens. - Information density score.

The evaluation highlights strengths in categories related to regulatory compliance, such as data handling during corporate transitions and data destruction procedures, while exposing shortcomings in areas like health-related research plans and genetic data storage protocols. Company-wise, shorter and denser policies tend to perform better. The findings emphasize the necessity for clear, concise, and well-documented policies to ensure compliance and enable the model to efficiently extract all category-related information.

Discussion

This study provided an approach to address the issue regarding the lack of transparency in DTC-GT policies, which contain how companies handle their customers' highly sensitive data. The approach includes an automatic extraction model with a corresponding evaluation framework that evaluates various DTC-GT-specific categories. Because of the global and local attention mechanism and the ability to process longer text data, the Longformer model was employed to produce content extraction in 22 different domain-specific categories (Beltagy et al., 2020). The assessment of the extracts provided insights regarding transparency as well as compliance to consumer privacy expectations and regulatory standards.

For this paper, the content was only extracted, leaving room to further simplify and structure the extracted data by formulating a summary of the extractions. The assessment could be readily implemented for category summaries and might be further enhanced using readability metrics to see if the summarization and simplification of data enhance the content, facilitating consumer comprehension of their data privacy settings. This would provide even more concise language than the extraction, as redundant material could be summarized. Building on this, there is potential to also provide a more detailed scoring method which could be displayed to consumers, helping them decide if they want to agree to the privacy policy they are dealing with.

While analyzing the extractions of the categories, the findings revealed both potentials and limitations of the proposed approach. While the framework effectively extracts and evaluates key content, it also underscores significant issues related to the specificity of categories, the quality of privacy policies, and the limitations of available labeled data. The categories in Table 1 show that they are quite similar in some areas. For instance, there are overlapping concepts concerning “biological” and “genetic” data, such as the categories “procedure for storage of biological samples” and “procedure for storage of genetic data” or “specific length of time for storage of biological samples” and “specific length of time for storage of genetic

data". The model has difficulties in distinguishing between categories due to overlaps in the privacy policy, especially when data is classified as both genetic and biological. The presence of just 19 labeled policies, which do not consistently encompass information for all 22 categories, increases the model's challenge in learning to differentiate. Adding more labeled data for extended supervised training would help the model to create more accurate extracts. The data does not only need to include DTC-GT privacy policies but could also include additional text that covers biological and genetic data such as health-related privacy policies like online pharmacies or online healthcare providers. With further data, the model would be capable of comprehending the nuanced distinctions among complex health-related topics. In addition, a larger volume of data would improve the assessment of all categories, as the existing 10 labeled test datasets lack information on categories 18 ("Statement that no research will be conducted using genetic data"), 21 ("Information on whether the health-related research may lead to commercialization or patents"), and 22 ("Customer's rights (or lack thereof) to commercial benefits from health-related research"). Furthermore, the model could add extra attention on "opting out options", which are possibilities to refuse or withdraw permission. This could solve issues like the one in category 20, "Policy for withdrawing from research", where the model is not able to differentiate between opting out of data being used for research or marketing. The ambiguity of the model for certain concepts is not only influenced by the limited number of labeled policies.

Furthermore, the issue of discrepancies in policy substance and structure arises due to the absence of legislative rules. Hendricks-Sturrup et al. (2019) for instance explain that certain DTC-GT companies share users' genetic data to third parties, including academic institutions and pharmaceutical firms, often relying on broad clauses that permit such data sharing. Consumers typically consent to these terms, yet often lack comprehensive awareness due to insufficient transparency and the resulting challenges in fully grasping the implications (Hendricks-Sturrup & Lu, 2019). The results of this study support this observation, as illustrated in Figure 9.

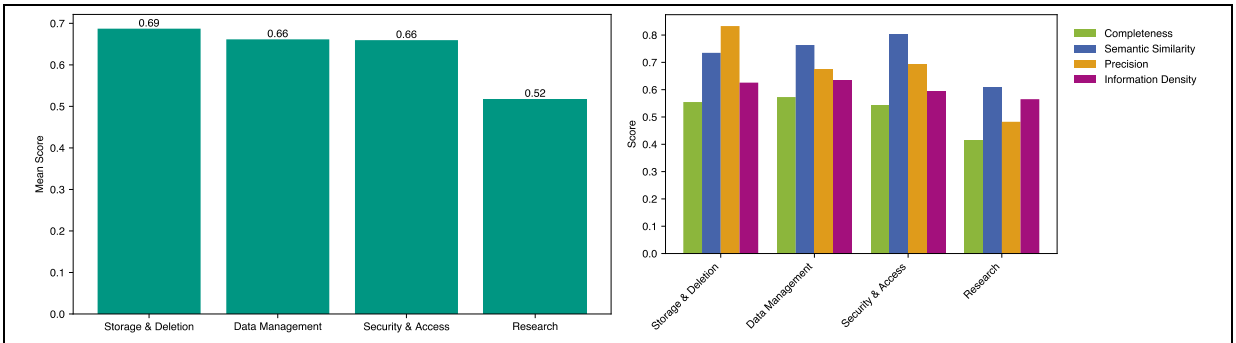

Figure 9. Performance by Category Groups

The overall performance of *research-related* topics is 0.52, significantly inferior to the *security and access* categories at 0.66, *data management* topics also at 0.66, and *storage and deletion* topics at 0.69 (see Table 6 for details concerning the categories included in the various groups). In addition, 3 out of 4 of the above already mentioned categories that are missing in the 10 policies on which the model was tested are *research related*. This indicates that if the policies are ambiguous in their information communication, the model will continually face difficulties in retrieving relevant information.

Besides the improvement of policy content, the model itself can be further enhanced by parameter tuning. The standard settings used for AdamW optimizer, sliding windows, and dilated windows can be fine-tuned for the specific DTC-GT privacy policies. In addition, the global attention mechanism can be explored by expanding beyond *[CLS]*, *[SEP]* and padding tokens by leveraging additional tokens listed in the "Hugging Face" documentation (HuggingFace, 2025).

In conclusion, this study highlights the potential of using NLP-based approaches like the Longformer model to analyze and evaluate DTC-GT privacy policies. However, looking at the extracted data, there are issues in the specificity of categories due to limited labeled data and poor transparency in some policies. Addressing these issues through expanded datasets, parameter optimization, and broader exploration of global attention mechanisms can enhance the model's accuracy and adaptability.

Group	Categories
Storage & Deletion	Procedure for storage of biological samples; Procedure for storage of genetic data; That biological samples will be or can be requested to be destroyed; That genetic data will be or can be requested to be destroyed; Specific length of time for storage of biological samples; Specific length of time for storage of genetic data
Data Management	Personal information which is collected and why; Specific risks related to disclosure of results; Possibility to delete account and the information that was saved; Procedure for handling genetic data should the company be sold or go bankrupt; Procedure for handling genetic samples should the company be sold or go bankrupt
Security & Access	Arrangements to ensure confidentiality of genetic data and security of biological samples; Policy about third parties that may be granted access to genetic data or samples; Disclosure of data to law enforcement
Research	Plans to use genetic data for health-related research; Specific purpose of health-related research; Plans to use genetic data for other or unspecified research; Statement that no research will be conducted using genetic data; Additional consent for health-related research; Policy for withdrawing from research; Information on whether the health-related research may lead to commercialization or patents; Customer's rights (or lack thereof) to commercial benefits from health-related research

Table 6. Category Groups and Their Components

Ultimately, more standardized and stricter guidelines dealing with the disclosure of companies regarding their handling of sensitive biological and genetic data would help to enable better comparability and accuracy in the presentation of data protection. An increase in transparency is crucial not only for the automated evaluation but also for consumers being able to make informed decisions.

References

23andMe. (2024). Consent Document. **https://www.23andme.com/about/consent/**

Aarsen, T. (2024). Multi-GPU support for mine_hard_negatives (#2967). UKP Lab. **https://github.com/UKPLab/sentence-transformers/blob/master/sentence_transformers/util.py**

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., & Mayer, J. (2021). Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. Proceedings of The Web Conference 2021, 22. **https://doi.org/10.1145/3442381.3450048**

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. ArXiv: 2406.01096.

Bhatia, J., Breaux, T., & Schaub, F. (2016). Mining Privacy Goals from Privacy Policies Using Hybridized Task Recomposition. ACM Transactions on Software Engineering and Methodology, 25, 1–24. **https://doi.org/10.1145/2907942**

BioSpace Editorial Team. (2024). Direct-to-Consumer Genetic Testing Market Rising Rapidly at CAGR 24.43%. **https://www.biospace.com/direct-to-consumer-genetic-testing-market-rising-rapidly-at-cagr-24-43-percent**

Black, P. E. (2021, January 8). Ratcliff/Obershelp Pattern Recognition. Dictionary of Algorithms and Data Structures; National Institute of Standards and Technology (NIST). **https://www.nist.gov/dads/HTML/ratcliffObershelp.html**

Costante, E., Sun, Y., Petković, M., & den Hartog, J. (2012). A machine learning solution to assess privacy policy completeness: (Short paper). Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society, 91–96. **https://doi.org/10.1145/2381966.2381979**

d'Aquin, M., Kirrane, S., Villata, S., Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T. B., Russell, N. C., Story, P., Reidenberg, J., Sadeh, N., d'Aquin, M., Kirrane, S., & Villata, S. (2017). PrivOnto: A semantic framework for the analysis of privacy policies. Semantic Web, 9, 1–19. **https://doi.org/10.3233/SW-170283**

Del Alamo, J., Guaman, D., García, B., & Díez Medialdea, A. (2022). A systematic mapping study on automated analysis of privacy policies. Computing, 104. **https://doi.org/10.1007/s00607-022-01076-3**

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. **https://doi.org/10.18653/v1/N19-1423**

Geneplanet. (2023). Privacy Policy. **https://geneplanet.com/eu/privacy-policy**

Goldberg, Y., & Hirst, G. (2017). Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers.

Guntamukkala, N., Dara, R., & Grewal, G. (2015). A Machine-Learning Based Approach for Measuring the Completeness of Online Privacy Policies. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 289–294. **https://doi.org/10.1109/ICMLA.2015.143**

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. 27th USENIX Security Symposium (USENIX Security 18), 531–548.

Hendricks-Sturrup, R., & Lu, C. (2019). Direct-to-Consumer Genetic Testing Data Privacy: Key Concerns and Recommendations Based on Consumer Perspectives. Journal of Personalized Medicine, 9, 25. **https://doi.org/10.3390/jpm9020025**

HuggingFace. (2025). Longformer Documentation. **https://huggingface.co/docs/transformers/model_doc/longformer**

Indurkhya, N., & Damerau, F. J. (2010). Handbook of Natural Language Processing (2nd ed.). Chapman & Hall/CRC.

Javed, Y., & Sajid, A. (2024). A Systematic Review of Privacy Policy Literature. ACM Computing Surveys. **https://doi.org/10.1145/3698393**

Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. The HUGO Journal, 8, 3. **https://doi.org/10.1186/s11568-014-0003-1**

Lebanoff, L., & Liu, F. (2018). Automatic Detection of Vague Words and Sentences in Privacy Policies. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3508–3517). Association for Computational Linguistics. **https://doi.org/10.18653/v1/D18-1387**

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization.

Massey, A. K., Eisenstein, J., Antón, A. I., & Swire, P. P. (2013). Automated text mining for requirements analysis of policy documents. 2013 21st IEEE International Requirements Engineering Conference (RE), 4–13. **https://doi.org/10.1109/RE.2013.6636700**

Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Science and Engineering Ethics, 22(2), 303–341. **https://doi.org/10.1007/s11948-015-9652-2**

My Heritage. (2024). Privacy Policy. https://www.myheritage.com/privacy-policy

National Human Genome Research Institute. (2023). Direct-to-Consumer Genetic Testing FAQ: For Healthcare Professionals. National Institutes of Health. **https://www.genome.gov/For-Health-Professionals/Provider-Genomics-Education-Resources/Healthcare-Provider-Direct-to-Consumer-Genetic-Testing-FAQ**

Onstwedder, S., Jansen, M., Cornel, M., & Rigter, T. (2024). Policy Guidance for Direct-to-Consumer Genetic Testing Services: Framework Development Study. Journal of Medical Internet Research, 26, e47389. **https://doi.org/10.2196/47389**

Python Software Foundation. (2024). Difflib—Helpers for computing deltas: Python Standard Library Documentation. **https://docs.python.org/3/library/difflib.html**

PyTorch. (2017). Torch.optim.AdamW. **https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html**

PyTorch. (2024). Automatic Mixed Precision (AMP) in PyTorch. **https://pytorch.org/docs/stable/amp.html**

Ramanath, R., Liu, F., Sadeh, N., & Smith, N. (2014). Unsupervised Alignment of Privacy Policies using Hidden Markov Models. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2, 605–610. **https://doi.org/10.3115/v1/P14-2099**

Sadri, M. (2024). HIPAA: A Demand to Modernize Health Legislation. The Undergraduate Law Review at UC San Diego, 2(1). **https://doi.org/10.5070/LR3.21252**

Sentence Transformers, T. (2020). All-MiniLM-L6-v2. **https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2**

Talukdar, W., & Biswas, A. (2024). Synergizing Unsupervised and Supervised Learning: A Hybrid Approach for Accurate Natural Language Task Modeling. ArXiv.

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient Transformers: A Survey. ACM Computing Surveys, 55(6). **https://doi.org/10.1145/3530811**

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010.

Wilson, S., Sadeh, N., Smith, N., Schaub, F., Liu, F., Sathyendra, K., Smullen, D., Zimmeck, S., Ramanath, R., Story, P., & Liu, F. (2018). Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations. ACM Transactions on the Web, 13, 1–29. **https://doi.org/10.1145/3230665**

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for Chinese legal long documents. AI Open, 2, 79–84. **https://doi.org/10.1016/j.aiopen.2021.06.003**

Zaeem, R. N., & Barber, K. S. (2020). The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. ACM Transactions on Management Information Systems, 12(1), 1–20. **https://doi.org/10.1145/3389685**