



Article

Ethical Challenges in Primary vs. Secondary Datasets: A Systematic Review of Manipulation and Transparency

Suastika Yulia Riska * , Triyanna Widiyaningtyas , Hakkun Elmunsyah , Siti Sendari 

Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang City, 65145, Indonesia

Abstract—The swift advancements in Artificial Intelligence and Machine Learning have rendered datasets essential; nonetheless, their heightened utilization has engendered intricate ethical dilemmas that are frequently neglected. This study seeks to delineate and highlight ethical concerns associated with the collection of primary data and the reutilization of secondary datasets in computer science research. We employed a Systematic Literature Review (SLR) methodology in accordance with the PRISMA 2020 guidelines, examining 72 publications sourced from five esteemed academic databases (Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Google Scholar) published from 2021 to 2025. The study results indicate that ethical difficulties emerge uniformly in both primary and secondary datasets. Primary datasets primarily face challenges related to privacy threats, anonymization, and Informed Consent, whereas secondary datasets are more susceptible to licensing infringements, dataset repurposing, and insufficient preparation transparency. The three domains that predominantly encountered these challenges were Machine Learning, Computer Vision, and Natural Language Processing. Moreover, practices of data manipulation, including cherry-picking and concealed preparation, were identified as detrimental to scientific integrity. This study's findings underscore the need for enhanced ethical standards for datasets and greater transparency in preparation documentation to ensure the repeatability of data-driven research.

Keywords—*anonymization; data ethics; dataset manipulation; reproducibility; scientific integrity; systematic literature review.*

1. Introduction

Swift advancements in computer science, especially in artificial intelligence, machine learning, digital image processing, and data science, have made datasets a crucial component of the research process (Xu et al., 2021). Nearly all contemporary computational models rely on the quality, completeness, and transparency of datasets to generate accurate outputs. This circumstance has resulted in heightened utilization of datasets, encompassing both directly gathered (primary datasets) and publicly sourced (secondary datasets) data (Gong et al., 2023; Xu et al., 2021). This expansion has also brought up several ethical challenges that are becoming increasingly complex and frequently neglected in research techniques (Böhm et al., 2022; Oniani et al., 2023).

Ethical challenges in primary datasets include data-gathering methods, data validation, protection of subject privacy, and transparent documentation (Cheng et al., 2024). The use of facial photos, voice recordings, location data, medical information, and Internet of Things (IoT) sensors presents a risk of privacy infringements in the absence of informed consent protocols, anonymization, and data protection measures

* Corresponding author.

E-mail Address: suastika.yulia.2505349@students.um.ac.id (S. Y. Riska)

Author E-mail(s): SYR (suastika.yulia.2505349@students.um.ac.id), TW (triyannaw.ft@um.ac.id), HE (hakkun@um.ac.id), SS (siti.sendari.ft@um.ac.id)

Digital Object Identifier 10.32815/jitika.1227

Manuscript submitted 17 November 2025; revised 8 December 2025; accepted 9 December 2025.

ISSN: 2580-8397(O), 0852-730X(P).

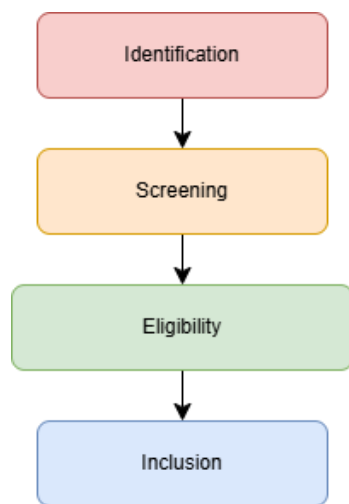


Fig 1. PRISMA Flow Diagram

(Cheng et al., 2024; Kalloniatis et al., 2021; Oniani et al., 2023; Sampaio et al., 2023). Moreover, certain researchers exhibit a propensity to manipulate primary datasets by excluding material deemed detrimental to model performance, doing unrepresentative sampling, or illicitly altering data values (Cheng et al., 2024; Oniani et al., 2023; Pagano et al., 2023a; Shahid et al., 2022). These techniques undermine scientific integrity and exacerbate the reproducibility problem in AI research.

Concurrently, the utilization of secondary databases raises considerable ethical concerns. Numerous datasets sourced from public sources, including Kaggle, UCI, GitHub, and other research platforms, are utilized without consideration for licenses, copyright laws, or the initial intent of the dataset developer (Alam et al., 2023; Alvarado Rojas & Twyman, 2025; Quaranta et al., 2021; Romano et al., 2022). Numerous studies alter secondary datasets without scientific justification, substitute labels without documentation, or employ datasets in contexts divergent from their original intent (Do et al., 2024; Gienapp et al., 2023; Marques et al., 2024). The absence of transparency in the preprocessing phase complicates the replication of study findings, ultimately undermining scientific credibility.

These difficulties demonstrate that dataset ethics is a vital consideration, yet it is hardly addressed in depth within computer science literature (Brown et al., 2024; Fareed et al., 2025; Paullada et al., 2021; ŞAHİN et al., 2025). The majority of studies prioritize model performance, whereas the ethical aspects concerning data are often overlooked (Hanna et al., 2025; Hua et al., 2025; Javed et al., 2023). While previous reviews have addressed general AI ethics or specific domain biases, there is a notable scarcity of research systematically comparing ethical breaches across the lifecycle of 'Primary' versus 'Secondary' datasets. Most existing literature treats dataset ethics as a monolith, failing to distinguish that primary data collection faces distinct challenges (e.g., immediate consent, sensor privacy) compared to reutilization of secondary datasets (e.g., context collapse, where data is used outside its original intended scope). This study bridges this gap by rigorously juxtaposing these two distinct data origins, offering a novel taxonomy of ethical risks that is currently absent in comparative SLR studies (Kaushik et al., 2024; Melhart et al.,

2023).

This study employed a Systematic Literature Review (SLR) methodology to address the subsequent research questions based on these requirements:

- RQ1: What ethical dilemmas emerge in the acquisition of primary data within computer science research?
- RQ2: What ethical concerns emerge in the reutilization of secondary datasets?
- RQ3: In what manner are instances of dataset misuse and manipulation documented in the computer science research literature?
- RQ4: What ethical suggestions are outlined in the literature to enhance the transparency and integrity of dataset-driven research?

2. Methods

This study uses a Systematic Literature Review (SLR) approach to identify, evaluate, and synthesize ethical issues in primary data collection and the reuse of secondary datasets in computer science research. The SLR methodology adopts the PRISMA 2020 standard, which is commonly used in systematic research in the fields of computer engineering and software engineering (Castillo & Grbovic, 2022).

2.1. Research Design

The SLR in this study was conducted through five main interconnected stages, as shown in Fig 1.

Based on Fig. 1, the research process follows the standard PRISMA stages: Identification, Screening, Eligibility, and Inclusion. The Identification stage involved defining the research scope and executing a search strategy across reputable academic databases to identify potential articles relevant to the research questions. The Screening stage entailed reviewing titles and abstracts, followed by an Eligibility assessment where full-text articles were evaluated against specific inclusion and exclusion criteria to ensure methodological relevance. Finally, articles that met all criteria proceeded to the Inclusion stage, where data extraction and synthesis were conducted through descriptive and thematic analyses to derive structured, objective evidence.

2.2. Database Sources and Search Strategies

This study's literature search was conducted across five esteemed academic databases predominantly used in computer science: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and Google Scholar. The literature search approach included a thorough amalgamation of terms to encompass the notions of ethics, dataset misuse, and the breadth of the computer science area. The search queries utilized integrated phrases pertaining to ethical concerns and research fields, employing the Boolean operators AND and OR. The utilized search strings were:

```

("ethical issues" OR "data ethics" OR "dataset misuse" OR "data manipulation" OR "dataset reuse" OR "research misconduct" OR "data collection ethics") AND ("computer science" OR "machine learning" OR "data science").
  
```

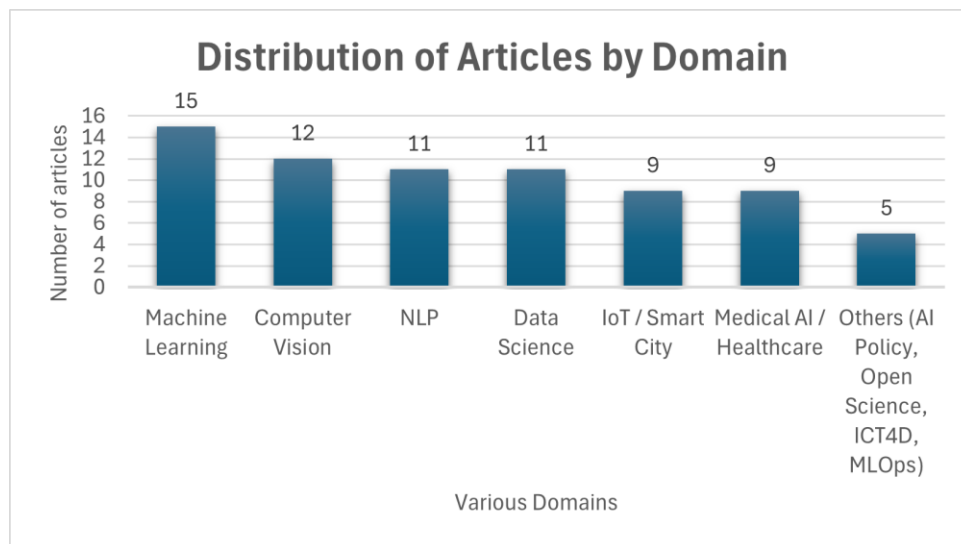


Fig 2. Distribution of article by domain

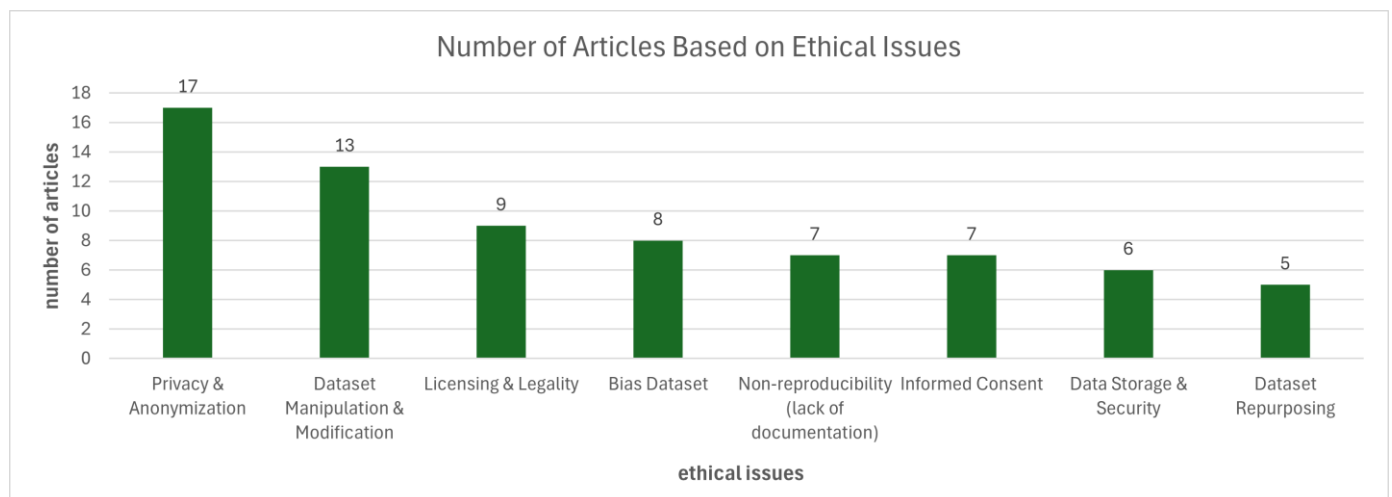


Fig 3. Number of articles based on ethical issues

The use of this string aims to ensure maximum and systematic identification of relevant articles in accordance with the predetermined research questions (Mancin et al., 2024). The search string was applied to the Title, Abstract, and Keywords fields to ensure high relevance. Since syntax varies across databases, the query was adapted accordingly (e.g., using 'Near/5' operators in WoS and specific wildcard handling in IEEE Xplore) while maintaining the semantic equivalence of the boolean logic.

2.3. Inclusion and Exclusion Criteria

Inclusion criteria were strictly applied to ensure the relevance and quality of articles analysed in this systematic review (Klimova et al., 2023). Articles were included if they discussed ethical issues related to the collection of primary datasets or the use of secondary datasets, and if they focused on the domains of computer science, machine learning, AI, data

mining, or data science, and were published between 2021 and 2025. Additional requirements included full-text access, publication in English or Indonesian, and inclusion in reputable scientific journals or conference proceedings, defined as venues indexed in Scopus (Q1–Q4) or the Web of Science (Core Collection) to ensure peer-review rigor and scientific quality. Meanwhile, exclusion criteria were used to weed out articles that did not substantially address the ethical aspects of dataset use, were editorials, opinions, blogs, or non-scientific reports, did not provide analyzable dataset method details, only released new datasets without discussing their ethical implications, or were duplicate publications/preprint versions of the same article.

2.4. Screening Process

The screening process was conducted independently by two authors to minimize selection bias. Any disagreements

regarding article inclusion were resolved through discussion and consensus, involving a third author as an adjudicator when necessary. This triangulation ensured the reliability of the selection process. The article selection and screening process was carried out systematically through four main stages of the PRISMA protocol. The first stage, Identification, was the initial search that produced a number of articles from the five main databases. The second stage, Screening, involved screening the titles and abstracts of articles to eliminate publications that were not relevant to the dataset's ethical scope. The third stage, Eligibility, is carried out through a comprehensive full-text evaluation to ensure that each article meets the established inclusion and exclusion criteria. Finally, the included stage is the determination of articles that meet all requirements as the final sample for the Systematic Literature Review (SLR) analysis (O'Dea et al., 2021; Peters et al., 2022; Sarkis-Onofre et al., 2021).

2.5. Data Extraction Techniques

Key information from each article that has passed the selection process is systematically recorded in an extraction table. This process aims to collect core data relevant to the research question (Falcetta et al., 2023; Foo et al., 2021; Schmidt et al., 2025). The extracted data includes the year of publication, database source, research domain, type of dataset discussed (primary or secondary), form of ethical issues raised, case studies or examples of dataset violations, ethical recommendations proposed by the article authors, and evaluation methods or research approaches used. This detailed data extraction forms the basis for the next stage of analysis.

2.6. Data Analysis Techniques

Data analysis was conducted using two main approaches. First, descriptive analysis was used to examine publication patterns, annual trends, topic distribution, and the spread of ethical issues across the various research domains reviewed. Second, thematic analysis was applied to identify and synthesize key themes in depth. These themes were classified into four categories, namely: ethics of primary data collection, ethics of secondary dataset use, reported patterns of dataset manipulation, and ethical recommendations from previous studies. Synthesizing across these two approaches enabled the collection of structured, objective evidence supporting the scientific reproducibility of research results.

3. Result

Findings from the Systematic Literature Review (SLR) process were conducted based on the PRISMA 2020 protocol. Findings include publication distribution, classification of ethical issues in primary and secondary datasets, and patterns of dataset manipulation in the computer science research literature.

3.1. General Literature Statistics

A literature search of five databases (Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Google Scholar) yielded 1,213 articles in the initial identification stage. After screening the titles and abstracts, 412 articles were deemed relevant to dataset

ethics. The full-text evaluation process filtered the articles down to 124, and after the eligibility stage based on inclusion-exclusion criteria, 72 articles were included in the final sample for this SLR. The distribution of articles by domain is shown in Fig. 2.

The distribution of publications shows an upward trend over the past five years, in line with the growing attention to artificial intelligence ethics, data bias, and model transparency. The largest area of research is machine learning (37%), followed by data science (21%), computer vision (18%), NLP (11%), and health informatics (13%). This shows that dataset ethics issues arise frequently in fields that rely on large amounts of data and complex preprocessing (Ahmed et al., 2025; Hassan et al., 2025; Khoei & Singh, 2025; Rani et al., 2025).

3.2. Categories of Ethical Issues in Primary and Secondary Datasets

This subchapter presents in-depth thematic results regarding the categories of ethical issues found in the literature. Based on an analysis of 72 selected articles, a number of dominant ethical issues have been identified and classified. The frequency distribution of the most frequently encountered ethical issues in computer science research, both in primary data collection and in the use of secondary data, is presented visually in Fig 3. The categories are not mutually exclusive; a single paper may be categorized under multiple ethical issues.

Based on Fig 3, thematic analysis of the selected articles identified a number of ethical issues that most frequently appeared in the primary and secondary datasets. Privacy and anonymization issues were the main concerns raised by 17 articles, especially in primary datasets involving the collection of sensitive information (Gadotti et al., 2024; Lison et al., 2021; Neves et al., 2023; Vovk et al., 2023; Zuo et al., 2021a). Many studies were found to neglect privacy protections, such as failing to anonymize facial or location data, thereby risking the identification of individuals without their consent. The second most significant category was Dataset Manipulation and Modification by 13 articles, where issues of data manipulation, such as deleting or altering data without clear methodological justification, can compromise research integrity and lead to biased results, and affect reproducibility (Chen et al., 2024; Qadhi et al., 2024; Sizo et al., 2025; Wang et al., 2025).

Additionally, Licensing and Legality by 9 articles is a concern, highlighting license and copyright violations, especially in datasets with specific licensing restrictions (Hao et al., 2025; Li et al., 2025; Peukert & Windisch, 2025). Dataset Bias by 8 articles is also significant, as unrepresentative data can potentially produce unfair models, particularly in sensitive fields such as medical and financial analysis (Norori et al., 2021; Pagano et al., 2023a; Tasci et al., 2022). Meanwhile, Non-reproducibility (7 articles) and Informed Consent (7 articles) often occur due to a lack of transparency in the preprocessing process and failure to ensure clear consent from human subjects in the collection of personal data (Sciaccotta et al., 2025; Van Vaerenbergh et al., 2025). Two other important issues are Data Storage and Security (6 articles), which highlights the importance of encryption and access control for sensitive data, and Dataset Repurposing (5 articles), which is the use of datasets for purposes other than their original intent, which can lead to invalid findings and violate ethical principles (Alabdulatif et al., 2023; Farhadighalati et al., 2025).

3.3. *Patterns of Dataset Manipulation in Computer Science Literature*

Analysis of the publications revealed three predominant patterns of dataset modification. The initial trend is data cherry-picking, wherein researchers choose particular subsets of the dataset to achieve elevated accuracy without disclosing the selection methodology (Saceleanu et al., 2023; Schetinger et al., 2023). The second pattern encompasses data falsification and relabeling, which denotes the alteration of data labels or attributes without a justifiable methodological rationale, such as the elimination of minority groups to mitigate imbalance and enhance model performance (Altalhan et al., 2025; Jui & Rivas, 2024; Timmons et al., 2023). The third trend is Hidden Preprocessing. This category was identified from articles that conducted reproducibility audits or post-hoc analyses, in which the authors attempted to replicate prior studies but failed due to undocumented steps. It is characterized by the omission of significant preprocessing steps (such as noise removal, normalization, or data augmentation) in certain research, rendering the claimed results ambiguous and challenging to repeat (Domingos et al., 2022; ElSabagh et al., 2025; Valizadeh et al., 2025). These manipulative patterns signify a significant threat to scientific integrity and the reproducibility of research within the domain of computer science.

4. Discussion

4.1. *The influence of resource scarcity*

The Influence of Resource Scarcity on Ethical Breaches as A critical driver of the identified ethical lapses is the tension between resource constraints and the pressure to publish ('publish or perish'). As highlighted by recent studies (Jayachandran & Arni, 2023), the high cost and difficulty of obtaining high-quality, legally cleared primary data often push researchers toward 'quick fixes'—such as scraping private data or ignoring licensing terms of secondary datasets (Hassan et al., 2025). In contexts of data scarcity, scientific integrity dictates that researchers should prioritize transparency over inflated performance. Documenting the limitations of small sample sizes is a valid scientific finding; however, the trend suggests a preference for manipulating data to achieve State-of-the-Art results, undermining the field's long-term progress.

4.2. *Comparison of Ethical Issues in Primary and Secondary Datasets*

The distribution analysis indicates that ethical difficulties occur uniformly in both primary and secondary datasets, each comprising 50%, signifying that dataset ethics is a pervasive challenge for research involving direct data collection as well as for studies employing public datasets. Although they occur at comparable frequencies, variations in the nature of the data pose distinct ethical challenges. Primary datasets often present concerns such as inadequate Informed Consent, privacy hazards associated with sensitive data (including facial photos, voice recordings, location data, and sensor information), opaque data validation procedures, sample bias, and security vulnerabilities in data storage (Khatriwada et al., 2024a; Wan et al., 2022; Zuo et al., 2021b). In contrast, secondary datasets are more

susceptible to violations of dataset licenses, repurposing beyond their original context, alterations to dataset labels or structures without explicit disclosures, insufficient pretreatment documentation, and challenges related to reproducibility (Alberto et al., 2023; Augustine et al., 2024; Hao et al., 2025). This balance affirms that both categories of datasets necessitate robust ethical norms; original datasets are particularly susceptible to human privacy concerns, whilst secondary datasets are more prone to copyright violations, undisclosed alterations, and methodological distortions.

4.3. *The Dominance of Ethical Issues in Machine Learning, Computer Vision, and NLP*

The investigation indicates that ethical concerns are mostly focused in the fields of Machine Learning (21%), Computer Vision (17%), Natural Language Processing (15%), and Data Science (15%). The elevated concentration arises from four primary similarities among these three domains: significant reliance on extensive datasets, considerable complexity in preprocessing, substantial risk of bias and data manipulation, and performance requirements that frequently prompt researchers to engage in potentially unethical optimization practices (Fan et al., 2021; Mondal et al., 2025; Pagano et al., 2023b; Werner de Vargas et al., 2023). In the domain of Computer Vision, prevalent infractions include the unauthorized use of human photographs and the absence of data anonymization, exemplified by the neglect to obscure faces or vehicle license plates (Haley, 2025; Laishram et al., 2025). In the realm of NLP, ethical concerns frequently pertain to the reutilization of datasets and the demographic biases inherent in language data that lacks representativeness (Bear Don't Walk IV et al., 2022; Mohamed et al., 2024). Ultimately, in the realm of Machine Learning, ethical concerns predominantly pertain to dataset manipulation and insufficient transparency in the preprocessing phase, which considerably impacts model repeatability (Dhanda et al., 2025; Hanna et al., 2025; Javed et al., 2023).

4.4. *Analysis of the Most Frequently Occurring Ethical Issue Categories in the Dataset*

The dominance of Privacy and Anonymization issues, particularly in Computer Vision, correlates strongly with the increasing ubiquity of biometric surveillance and facial recognition technologies. Meanwhile, the prevalence of 'Hidden Preprocessing' in Machine Learning domains suggests a systemic reproducibility crisis driven by competitive benchmarks. The prevalence of this issue is justifiable, as the majority of datasets utilized in computer science research, especially in Machine Learning and Computer Vision, contain personal or sensitive information, such as facial data, voice recordings, geographic location, or medical records (Khatriwada et al., 2024b; Martinez-Martin et al., 2021). The second most prevalent category is Dataset Manipulation or Modification (25%), encompassing activities such as cherry-picking, relabeling, and concealed filtering. These methods are especially perilous as they might artificially enhance model performance and compromise scientific integrity (Chen et al., 2024; Doskaliuk et al., 2025). Moreover, concern about dataset licensing and legality (17.5%) became prominent, particularly in studies that employ secondary datasets without verifying

license terms, which may limit certain commercial or research applications (Alberto et al., 2023; Martorana et al., 2022). Furthermore, concerns regarding Dataset Bias (15%) and Non-reproducibility (12.5%) were significant, highlighting that the absence of comprehensive preprocessing reports by numerous researchers not only complicates the replication of studies but also renders them susceptible to methodological bias (Norori et al., 2021; Pagano et al., 2023a; Tasci et al., 2022). These ethical considerations demonstrate that the challenges of data-driven research pertain not just to the utilization of data but also to its selection, processing, storage, and transparent reporting.

Furthermore, a qualitative analysis of the reviewed articles reveals a concerning correlation between data transparency and reported performance metrics. Papers categorized under 'Hidden Preprocessing' or 'Cherry-picking' frequently reported State-of-the-Art results (often exceeding 95% accuracy) without providing reproducible baselines. In contrast, studies that explicitly documented data limitations and ethical constraints tended to report more modest performance Figs. This suggests that the lack of dataset ethics—specifically the omission of rigorous cleaning steps—may artificially inflate model performance, creating a false perception of scientific advancement while undermining reproducibility.

4.5. Implications of Findings for Computer Science Research Practices

The results of this Systematic Literature Review (SLR) have significant practical implications for data-driven research methodologies in computer science. The results of dataset manipulation and privacy breaches highlight the need for enhanced ethical standards for datasets at academic institutions to ensure that the resulting models are impartial and reliable. The elevated prevalence of non-reproducibility highlights the necessity for transparency in preprocessing; the scientific community must be incentivized to utilize open-code, open-data logs, and comprehensive pipeline documentation as primary factors for reproducibility. Third, the presence of dataset bias underscores the need to use more equitable and representative datasets through fairness audits and dataset representation assessments to avert model bias against specific populations. Ultimately, concerns over repurposing and licensing infringements in public data necessitate the judicious use of secondary datasets, requiring researchers to verify the license, intended purpose, and ethical implications before use.

To address these challenges and assist researchers in maintaining scientific integrity, we synthesize our findings into a preliminary ethical framework. Table 1 presents a proposed checklist to guide researchers in the ethical handling of both primary and secondary datasets throughout the research lifecycle.

5. Conclusion and Future Research

5.1. Conclusion

This article provides a systematic literature review of ethical concerns associated with original data collection and the reuse of secondary datasets in computer science research. A study found that ethical issues related to datasets occur equally across all data categories. Primary datasets primarily encounter issues of privacy and anonymization, Informed Consent, data

Table 1. Proposed Ethical Checklist for Primary and Secondary Datasets

Phase	Check	Primary Action	Dataset	Secondary Action	Dataset
Planning	☐	Has IRB/Ethical clearance been obtained?		Is the dataset license compatible with the research goal (e.g., Non-Commercial)?	
Collection	☐	Is Informed Consent explicit and documented?		Verify the original purpose vs. new context (Contextual Integrity – ensuring data flow appropriateness).	
Processing	☐	Are anonymization techniques irreversible?		Document all filtering/resizing steps to prevent "Hidden Preprocessing".	
Analysis	☐	Check for sampling bias in demographic representation.		Ensure no "Cherry-picking" of subsets to inflate accuracy.	
Publication	☐	Are raw/anonymized data sharing protocols established?		Cite the original dataset creators properly.	

validation, and storage security, but secondary datasets are more susceptible to licensing infringements, data repurposing, undocumented relabeling, and insufficient preparation transparency. The findings indicate that the research domains most susceptible to ethical concerns are Machine Learning, Computer Vision, and Natural Language Processing, correlating with their significant reliance on extensive and intricate datasets. The principal ethical concerns, categorized, are privacy and anonymization (30%), dataset manipulation or modification (25%), and violations of dataset legality and licensing (17.5%). The findings of this systematic literature review indicate that the ethical challenges associated with datasets are multifaceted, encompassing legal, methodological, technical, and social dimensions. Consequently, there is a necessity for enhanced ethical literacy, the implementation of rigorous preprocessing documentation standards, and improved adherence to licenses and regulations in data-centric research within the domain of computer science.

5.2. Recommendations for Further Research

Based on the findings and conversations, various strategic recommendations and prospects for further research exist. The fundamental recommendation is to establish a more precise Dataset Ethics Framework for computer research, encompassing criteria for original data collection, verification of public dataset licensing, and comprehensive audit protocols for datasets. Furthermore, empirical investigations of dataset manipulation techniques within the academic community are essential to understanding the underlying determinants and consequences of behaviors such as cherry-picking on academic integrity. From a technical standpoint, it is essential to provide tools for automated dataset audits that can identify ethical breaches and concealed data alterations, significantly aiding reviewers and

research ethics committees. An extensive examination of reproducibility in machine learning research is essential, emphasizing efficient documentation practices to enhance transparency in dataset preprocessing and versioning. Ultimately, in line with technological advancement, additional research should investigate Dataset Ethics within the realms of Industry and Smart Society (including IoT and smart city data) and examine Cross-Cultural Bias in Global Datasets to address imbalanced representation in international datasets.

Data availability

All data produced or examined during this study are present in this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

All authors participated in the study design, writing, and manuscript revision. SYR drafted the initial manuscript, TW and HE revised it, and SS supervised the study. All authors have reviewed and approved the final manuscript.

References

- Ahmed, M. M., Okesanya, O. J., Oweidat, M., Othman, Z. K., Musa, S. S., & Lucero-Prisno Iii, D. E. (2025). The ethics of data mining in healthcare: Challenges, frameworks, and future directions. *BioData Mining*, 18(1), 47. <https://doi.org/10.1186/s13040-025-00461-w>
- Alabdulatif, A., Thilakarathne, N. N., & Kalinaki, K. (2023). A novel cloud enabled access control model for preserving the security and privacy of medical big data. *Electronics*, 12(12), 2646.
- Alam, G., McChesney, I., Nicholl, P., & Rafferty, J. (2023). Open datasets in human activity recognition research—Issues and challenges: A review. *IEEE Sensors Journal*, 23(22), 26952–26980.
- Alberto, I. R. I., Alberto, N. R. I., Ghosh, A. K., Jain, B., Jayakumar, S., Martinez-Martin, N., McCague, N., Moukheiber, D., Moukheiber, L., & Moukheiber, M. (2023). The impact of commercial health datasets on medical research and health-care algorithms. *The Lancet Digital Health*, 5(5), e288–e294.
- Altalhan, M., Algarni, A., & Alouane, M. T.-H. (2025). Imbalanced data problem in machine learning: A review. *IEEE Access*. <https://ieeexplore.ieee.org/abstract/document/10845793/>
- Alvarado Rojas, A., & Twyman, M. (2025). Cultural data markets: Interpreting the popularity of public datasets. *New Media & Society*, 14614448251359631. <https://doi.org/10.1177/14614448251359631>
- Augustine, S. P., Bailey-Marren, I., Charton, K. T., Kiel, N. G., & Peyton, M. S. (2024). Improper data practices erode the quality of global ecological databases and impede the progress of ecological research. *Global Change Biology*, 30(1), e17116. <https://doi.org/10.1111/gcb.17116>
- Bear Don't Walk IV, O. J., Reyes Nieva, H., Lee, S. S.-J., & Elhadad, N. (2022). A scoping review of ethics considerations in clinical natural language processing. *JAMIA Open*, 5(2), ooac039.
- Böhm, S., Carrington, M., Cornelius, N., De Bruin, B., Greenwood, M., Hassan, L., Jain, T., Karam, C., Kourula, A., Romani, L., Riaz, S., & Shaw, D. (2022). Ethics at the Centre of Global and Local Challenges: Thoughts on the Future of Business Ethics. *Journal of Business Ethics*, 180(3), 835–861. <https://doi.org/10.1007/s10551-022-05239-2>
- Brown, N., Xie, B., Sarder, E., Fiesler, C., & Wiese, E. S. (2024). Teaching Ethics in Computing: A Systematic Literature Review of ACM Computer Science Education Publications. *ACM Transactions on Computing Education*, 24(1), 1–36. <https://doi.org/10.1145/3634685>
- Castillo, S., & Grbovic, P. (2022). The APISSE methodology for systematic literature reviews in engineering. *IEEE Access*, 10, 23700–23707.
- Chen, Z., Chen, C., Yang, G., He, X., Chi, X., Zeng, Z., & Chen, X. (2024). Research integrity in the era of artificial intelligence: Challenges and responses. *Medicine*, 103(27), e38811.
- Cheng, L., Han, J., & Nasirov, J. (2024). Ethical considerations related to personal data collection and reuse: Trust and transparency in language and speech technologies. *International Journal of Legal Discourse*, 9(2), 217–235. <https://doi.org/10.1515/ijld-2024-2010>
- Dhanda, S. S., Panwar, D., Lin, C.-C., Sharma, T. K., Rastogi, D., Bindewari, S., Singh, A., Li, Y.-H., Agarwal, N., & Agarwal, S. (2025). Advancement in public health through machine learning: A narrative review of opportunities and ethical considerations. *Journal of Big Data*, 12(1), 154. <https://doi.org/10.1186/s40537-025-01201-x>
- Do, S., Ollion, É., & Shen, R. (2024). The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. *Sociological Methods & Research*, 53(3), 1167–1200. <https://doi.org/10.1177/00491241221134526>
- Domingos, L. C., Santos, P. E., Skelton, P. S., Brinkworth, R. S., & Sammut, K. (2022). A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. *Sensors*, 22(6), 2181.
- Doskaliuk, B., Zimba, O., Yessirkepov, M., Klishch, I., & Yatsyshyn, R. (2025). Artificial intelligence in peer review: Enhancing efficiency while preserving integrity. *Journal of Korean Medical Science*, 40(7). <https://synapse.koreamed.org/articles/1516090023>
- ElSabagh, A. A., Azab, S. S., & Hefny, H. A. (2025). A comprehensive survey on Arabic text augmentation: Approaches, challenges, and applications. *Neural Computing and Applications*, 37(10), 7015–7048. <https://doi.org/10.1007/s00521-025-11020-z>
- Falcetta, F. S., De Almeida, F. K., Lemos, J. C. S., Goldim, J. R., & Da Costa, C. A. (2023). Automatic documentation of professional health interactions: A systematic review. *Artificial Intelligence in Medicine*, 137, 102487.
- Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, 9, 652801.
- Fareed, M., Fatima, M., Uddin, J., Ahmed, A., & Sattar, M. A. (2025). A Systematic Review of Ethical Considerations of Large Language Models (LLMs) in Healthcare and Medicine. *Frontiers in Digital Health*, 7, 1653631.
- Farhadighalati, N., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., & Barata, J. (2025). A systematic review of access control models: Background, existing research, and challenges. *IEEE Access*. <https://ieeexplore.ieee.org/abstract/document/10850915/>
- Foo, Y. Z., O'Dea, R. E., Koricheva, J., Nakagawa, S., & Lagisz, M. (2021). A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. *Methods in Ecology and Evolution*, 12(9), 1705–1720. <https://doi.org/10.1111/2041-210X.13654>
- Gadotti, A., Rocher, L., Houssiau, F., Crețu, A.-M., & De Montjoye, Y.-A. (2024). Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29), eadn7053. <https://doi.org/10.1126/sciadv.adn7053>
- Gienapp, L., Kircheis, W., Sievers, B., Stein, B., & Potthast, M. (2023). A large dataset of scientific text reuse in open-access publications. *Scientific Data*, 10(1), 58.
- Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162, 107268.
- Haley, P. (2025). The Impact of Biometric Surveillance on Reducing

- Violent Crime: Strategies for Apprehending Criminals While Protecting the Innocent. *Sensors*, 25(10), 3160.
- Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3), 100686.
- Hao, J., Nalley, A., Yeung, A. W. K., Tanaka, R., Ai, Q. Y. H., Lam, W. Y. H., Shan, Z., Leung, Y. Y., AlHadidi, A., & Bornstein, M. M. (2025). Characteristics, licensing, and ethical considerations of openly accessible oral-maxillofacial imaging datasets: A systematic review. *Npj Digital Medicine*, 8(1), 412.
- Hassan, H., Manan, M., Sehar, S., Shahzadi, A., & Fatima, M. (2025). Automation in Web Data Extraction: Opportunities, Challenges, and Ethical Considerations. *International Journal of Computing and Data Science*, 1(2), 37–46.
- Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry*, 24(3), 383–394. <https://doi.org/10.1002/wps.21352>
- Javed, H., Muqet, H. A., Javed, T., Rehman, A. U., & Sadiq, R. (2023). Ethical frameworks for machine learning in sensitive healthcare applications. *IEEE Access*, 12, 16233–16254.
- Jayachandran, J., & Arni, V. (2023). Traversing the Ethical Landscape of Data Scraping for AI. Available at SSRN 4666354. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4666354
- Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, 15(8), 3095–3125. <https://doi.org/10.1007/s13042-023-02083-2>
- Kalloniatis, C., Lambrinoudakis, C., Musahl, M., Kanatas, A., & Gritzalis, S. (2021). Incorporating privacy by design in body sensor networks for medical applications: A privacy and data protection framework. *Computer Science and Information Systems*, 18(1), 323–347.
- Kaushik, P., Sharma, K., Mahawar, M. K., Wasim, J., Dey, G., & Nibiya, S. A. (2024). Ethical Considerations in Data Mining and Analytics. *2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, 1516–1521. <https://ieeexplore.ieee.org/abstract/document/10895012/>
- Khatiwa, P., Yang, B., Lin, J.-C., & Blobel, B. (2024a). Patient-generated health data (PGHD): Understanding, requirements, challenges, and existing techniques for data security and privacy. *Journal of Personalized Medicine*, 14(3), 282.
- Khatiwa, P., Yang, B., Lin, J.-C., & Blobel, B. (2024b). Patient-generated health data (PGHD): Understanding, requirements, challenges, and existing techniques for data security and privacy. *Journal of Personalized Medicine*, 14(3), 282.
- Khoei, T. T., & Singh, A. (2025). Data reduction in big data: A survey of methods, challenges and future directions. *International Journal of Data Science and Analytics*, 20(3), 1643–1682. <https://doi.org/10.1007/s41060-024-00603-z>
- Klimova, B., Pikhart, M., Benites, A. D., Lehr, C., & Sanchez-Stockhammer, C. (2023). Neural machine translation in foreign language teaching and learning: A systematic review. *Education and Information Technologies*, 28(1), 663–682. <https://doi.org/10.1007/s10639-022-11194-2>
- Laishram, L., Shaheryar, M., Lee, J. T., & Jung, S. K. (2025). Toward a Privacy-Preserving Face Recognition System: A Survey of Leakages and Solutions. *ACM Computing Surveys*, 57(6), 1–38. <https://doi.org/10.1145/3673224>
- Li, B., Liu, C., Fan, L., Chen, S., Zhang, Z., & Liu, Z. (2025). Open Source, Hidden Costs: A Systematic Literature Review on OSS License Management. *IEEE Transactions on Software Engineering*. <https://ieeexplore.ieee.org/abstract/document/11072183/>
- Lison, P., Pilán, I., Sanchez, D., Batet, M., & Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4188–4203. <https://aclanthology.org/2021.acl-long.323/>
- Mancini, S., Sguanci, M., Andreoli, D., Soekeland, F., Anastasi, G., Piredda, M., & De Marinis, M. G. (2024). Systematic review of clinical practice guidelines and systematic reviews: A method for conducting comprehensive analysis. *MethodsX*, 12, 102532.
- Marques, N., de Melo Soares, C. D., de Melo Casali, D., Guimarães, E. C., Fava, F. G., da Silva Abreu, J. M., Moras, L. M., da Silva, L. G., Matias, R., & de Assis, R. L. (2024). Retrieving biodiversity data from multiple sources: Making secondary data standardised and accessible. *Biodiversity Data Journal*, 12, e133775.
- Martinez-Martin, N., Luo, Z., Kaushal, A., Adeli, E., Haque, A., Kelly, S. S., Wieten, S., Cho, M. K., Magnus, D., & Fei-Fei, L. (2021). Ethical issues in using ambient intelligence in health-care settings. *The Lancet Digital Health*, 3(2), e115–e123.
- Martorana, M., Kuhn, T., Siebes, R., & van Ossenbruggen, J. (2022). Aligning restricted access data with FAIR: A systematic review. *PeerJ Computer Science*, 8, e1038.
- Melhart, D., Togelius, J., Mikkelsen, B., Holmgård, C., & Yannakakis, G. N. (2023). The ethics of AI in games. *IEEE Transactions on Affective Computing*, 15(1), 79–92.
- Mohamed, Y. A., Mohamed, A. H., Kannan, A., Bashir, M., Adiel, M. A., & Elsadig, M. A. (2024). Navigating the ethical terrain of ai-generated text tools: A review. *IEEE Access*. <https://ieeexplore.ieee.org/abstract/document/10813359/>
- Mondal, S., Maity, R., & Nag, A. (2025). A Comprehensive Survey of Heart Disease Prediction Approaches: Methods, Applications, Performance Analysis, Datasets, Research Challenges, and Future Scopes. *Archives of Computational Methods in Engineering*. <https://doi.org/10.1007/s11831-025-10438-x>
- Neves, F., Souza, R., Sousa, J., Bonfim, M., & Garcia, V. (2023). Data privacy in the Internet of Things based on anonymization: A review. *Journal of Computer Security*, 31(3), 261–291. <https://doi.org/10.3233/JCS-210089>
- Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, 2(10). [https://www.cell.com/patterns/fulltext/S2666-3899\(21\)00202-6?dgcid=raven_jbs_etoc_email](https://www.cell.com/patterns/fulltext/S2666-3899(21)00202-6?dgcid=raven_jbs_etoc_email)
- O'Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T. H., Gurevitch, J., Page, M. J., Stewart, G., Moher, D., & Nakagawa, S. (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: A PRISMA extension. *Biological Reviews*, 96(5), 1695–1722. <https://doi.org/10.1111/brev.12721>
- Oniani, D., Hilsman, J., Peng, Y., Poropatich, R. K., Pamplin, J. C., Legault, G. L., & Wang, Y. (2023). Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digital Medicine*, 6(1), 225.
- Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., & Oliveira, E. L. (2023a). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 15.
- Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., & Oliveira, E. L. (2023b). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 15.
- Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11). [https://www.cell.com/patterns/fulltext/S2666-3899\(21\)00184-7](https://www.cell.com/patterns/fulltext/S2666-3899(21)00184-7)
- Peters, M. D., Godfrey, C., McInerney, P., Khalil, H., Larsen, P., Marnie, C., Pollock, D., Tricco, A. C., & Munn, Z. (2022). Best practice guidance and reporting items for the development of scoping review protocols. *JBIM Evidence Synthesis*, 20(4), 953–968.
- Peukert, C., & Windisch, M. (2025). The economics of copyright in the

- digital age. *Journal of Economic Surveys*, 39(3), 877–903. <https://doi.org/10.1111/joes.12632>
- Qadhi, S. M., Alduais, A., Chaaban, Y., & Khraisheh, M. (2024). Generative AI, research ethics, and higher education research: Insights from a scientometric analysis. *Information*, 15(6), 325.
- Quaranta, L., Calefato, F., & Lanubile, F. (2021). KGTorrent: A dataset of python jupyter notebooks from kaggle. *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, 550–554. <https://ieeexplore.ieee.org/abstract/document/9463068/>
- Rani, S., Kumar, R., Panda, B. S., Kumar, R., Muften, N. F., Abass, M. A., & Lozanović, J. (2025). Machine learning-powered smart healthcare systems in the era of big data: Applications, diagnostic insights, challenges, and ethical implications. *Diagnostics*, 15(15), 1914.
- Romano, J. D., Le, T. T., La Cava, W., Gregg, J. T., Goldberg, D. J., Chakraborty, P., Ray, N. L., Himmelstein, D., Fu, W., & Moore, J. H. (2022). PMLB v1. 0: An open-source dataset collection for benchmarking machine learning methods. *Bioinformatics*, 38(3), 878–880.
- Saceleanu, V. M., Toader, C., Ples, H., Covache-Busuioc, R.-A., Costin, H. P., Bratu, B.-G., Dumitrascu, D.-I., Bordeianu, A., Corlatescu, A. D., & Ciurea, A. V. (2023). Integrative approaches in acute ischemic stroke: From symptom recognition to future innovations. *Biomedicine*, 11(10), 2617.
- ŞAHİN, E., Arslan, N. N., & Özdemir, D. (2025). Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications*, 37(2), 859–965. <https://doi.org/10.1007/s00521-024-10437-2>
- Sampaio, S., Sousa, P. R., Martins, C., Ferreira, A., Antunes, L., & Cruz-Correia, R. (2023). Collecting, processing and secondary using personal and (pseudo) anonymized data in smart cities. *Applied Sciences*, 13(6), 3830.
- Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., & Lockwood, C. (2021). How to properly use the PRISMA Statement. *Systematic Reviews*, 10(1), 117, [s13643-021-01671-z](https://doi.org/10.1186/s13643-021-01671-z). <https://doi.org/10.1186/s13643-021-01671-z>
- Schetingner, V., Di Bartolomeo, S., El-Assady, M., McNutt, A., Miller, M., Passos, J. P. A., & Adams, J. L. (2023). Doom or Deliciousness: Challenges and Opportunities for Visualization in the Age of Generative Models. *Computer Graphics Forum*, 42(3), 423–435. <https://doi.org/10.1111/cgf.14841>
- Schmidt, L., Mutlu, A. N. F., Elmore, R., Olorisade, B. K., Thomas, J., & Higgins, J. P. (2025). Data extraction methods for systematic review (semi) automation: Update of a living systematic review. *F1000Research*, 10, 401.
- Sciacotta, R., Barone, P., Murdaca, G., Fazio, M., Stagno, F., Gangemi, S., Genovese, S., & Allegra, A. (2025). Decoding Immunodeficiencies with Artificial Intelligence: A New Era of Precision Medicine. *Biomedicine*, 13(8), 1836.
- Shahid, J., Ahmad, R., Kiani, A. K., Ahmad, T., Saeed, S., & Almuhaideb, A. M. (2022). Data protection and privacy of the internet of healthcare things (IoHTs). *Applied Sciences*, 12(4), 1927.
- Sizo, A., Lino, A., Rocha, Á., & Reis, L. P. (2025). Defining quality in peer review reports: A scoping review. *Knowledge and Information Systems*, 67(8), 6413–6460. <https://doi.org/10.1007/s10115-025-02435-0>
- Tasci, E., Zhuge, Y., Camphausen, K., & Krauze, A. V. (2022). Bias and class imbalance in oncologic data—Towards inclusive and transferrable AI in large scale oncology data sets. *Cancers*, 14(12), 2897.
- Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. *Perspectives on Psychological Science*, 18(5), 1062–1096. <https://doi.org/10.1177/17456916221134490>
- Valizadeh, G., Elahi, R., Hasankhani, Z., Rad, H. S., & Shalbaf, A. (2025). Deep Learning Approaches for Early Prediction of Conversion from MCI to AD using MRI and Clinical Data: A Systematic Review. *Archives of Computational Methods in Engineering*, 32(2), 1229–1298. <https://doi.org/10.1007/s11831-024-10176-6>
- Van Vaerenbergh, Y., Hazée, S., & Zwienerberg, T. J. (2025). Open Science: A Review of Its Effectiveness and Implications for Service Research. *Journal of Service Research*, 10946705251338461. <https://doi.org/10.1177/10946705251338461>
- Vovk, O., Pihó, G., & Ross, P. (2023). Methods and tools for healthcare data anonymization: A literature review. *International Journal of General Systems*, 52(3), 326–342. <https://doi.org/10.1080/03081079.2023.2173749>
- Wan, Z., Hazel, J. W., Clayton, E. W., Vorobeychik, Y., Kantarcioglu, M., & Malin, B. A. (2022). Sociotechnical safeguards for genomic data privacy. *Nature Reviews Genetics*, 23(7), 429–445.
- Wang, Y., Cheungpasitporn, W., Ali, H., Qing, J., Thongprayoon, C., Kaewput, W., Soliman, K. M., Huang, Z., Yang, M., & Zhang, Z. (2025). A practical guide for nephrologist peer reviewers: Evaluating artificial intelligence and machine learning research in nephrology. *Renal Failure*, 47(1), 2513002. <https://doi.org/10.1080/0886022X.2025.2513002>
- Werner de Vargas, V., Schneider Aranda, J. A., Dos Santos Costa, R., Da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. *Knowledge and Information Systems*, 65(1), 31–57. <https://doi.org/10.1007/s10115-022-01772-8>
- Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., & Qiu, C.-W. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4). [https://www.cell.com/article/S2666-6758\(21\)00104-1/fulltext](https://www.cell.com/article/S2666-6758(21)00104-1/fulltext)
- Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C., & Al Moubayed, N. (2021a). Data anonymization for pervasive health care: Systematic literature mapping study. *JMIR Medical Informatics*, 9(10), e29871.
- Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C., & Al Moubayed, N. (2021b). Data anonymization for pervasive health care: Systematic literature mapping study. *JMIR Medical Informatics*, 9(10), e29871.



Suastika Yulia Riska received the Bachelor's degree in Information Technology Education from Universitas Negeri Malang, Indonesia, in 2013, and the Master's degree in Information Technology from Institut Teknologi Sepuluh Nopember Surabaya, Indonesia, in 2015. She is currently pursuing a Doctoral degree in Electrical Engineering and Informatics at Universitas Negeri Malang, Indonesia. Since 2015, she has been a Lecturer with Institut Teknologi dan Bisnis Asia Malang. Her research interests include machine learning, data mining, expert systems, and computer vision. She has contributed to several publications in these fields.



Triyanna Widiyaningtyas received the Bachelor's degree from STT Telkom Bandung, Indonesia, in 1998, the Master of Engineering degree from Universitas Gadjah Mada, Indonesia, in 2002, and the Doctoral degree from Universitas Gadjah Mada in 2022. She obtained the Professional Engineer certification from Universitas Brawijaya in 2021. She is currently a Lecturer in the Department of Electrical Engineering at the State University of Malang. Her research interests include knowledge engineering and data science.



Hakkun Elmunsyah received the Bachelor of Engineering degree from Universitas Brawijaya, Indonesia, in 1994, the Master of Engineering degree from Universitas Gadjah Mada, Indonesia, in 2002, and the Doctoral degree from Universitas Negeri Yogyakarta, Indonesia, in 2012. He is currently a Professor with the Department of Electrical Engineering and Informatics, Faculty of Engineering, State University of Malang (UM). His research interests focus on educational technology and information engineering.



Siti Sendari received the Bachelor of Engineering degree from Universitas Brawijaya, Indonesia, in 1996, the Master of Engineering degree from Universitas Gadjah Mada, Indonesia, in 2005, and the Doctor of Engineering degree from Waseda University, Japan, in 2013. She is currently a Professor of Educational Technology and Informatics Engineering. Her research interests include the development of educational technology in informatics.