Data Duplication Detection and Removal System Using Machine Learning

ANSH BALGOTRA
Department of Information technology, Maharaja Agrasen Institute of Technology, New Delhi, India
anshbalgotra@gmail.com

Abstract— The problem of missing data is a critical issue in various domains, as it can lead to inaccurate analysis and flawed decision-making. Traditional methods for handling missing values have been replaced by machine learning techniques, which offer more efficient solutions. Research in this area has explored various approaches to data imputation, analyzing their strengths and limitations. A systematic literature review of studies from 2016 to 2021 identified key factors influencing the effectiveness of thesemethods, providing valuable insights for researchers and data analysts. In parallel, the rapid expansion of data storage and processing has led to challenges in managing large -scale information, particularly in deduplication. Duplicate data, originating from multiple sources, complicates storage efficiency and retrieval accuracy. Cloud service providers have adopted data deduplication techniques to optimize storage costs and bandwidth usage. However, the conflict between encryption for security and deduplication efficiency presents a challenge. To address this, hybrid chunking methods, such as the Two Threshold Two Divisor (TTTD) and Dynamic Prime Coding (DPC) algorithm, have been proposed. These techniques improve deduplication performance while balancing security requirements. Furthermore, entity resolution plays a crucial role in information integration, aiming to consolidate and organize data from diverse sources. Deduplication, as a key step in this process, enhances data quality by identifying and eliminating redundant records. Research in this domain spans machine learning, data mining, and information retrieval, focusing on both supervised and unsupervised approaches. By analyzing various methodologies, researchers can refine existing techniques to improve accuracy, processing speed, and computational efficiency. Overall, advancements in machine learning, deduplication, and entity resolution contribu te to more effective data management, addressing challenges in missing data imputation, secure deduplication, and large-scale information integration.

Keywords— **Missing Data**, **Data Quality**, **Machine Learning**, **Processing Speed**, **Computational Efficiency**, **Structured Data**, **Unstructured Data**, **Database Management**, **Encryption, Accuracy**, **Performance**

1. Introduction

Data quality is a critical factor in data ana lytics, and missing va lues present a significant challenge in real-world datasets. When data is incomplete, it can introduce bia s, reduce accuracy, and ultimately affect decision-making processes. Missing va lues arise due to various reasons, including survey non-responses, human errors in data entry, or system failures during data collection. Since missing data is inevitable, it must be addressed before data preprocessing to ensure accurate and reliable analysis.

In recent years, machine learning (ML) has emerged as a powerful tool for handling missing data, offering more advanced imputation techniques compared to traditional methods such as deletion or mean substitution. However, despite the growing body of research on ML-ba sed imputation methods, studies rema in fra gmented across different publications. There is a need for a comprehensive review that consolidates findings, assesses improvements, and identifies gaps in existing approaches. This study aims to explore and evaluate ML techniques for missing value imputation through a systematic literature review (SLR). The research will examine recent trends, analyze the effectiveness of ML models, highlight their strengths and limitations, and suggest potentia l directions for future research in this doma in. By providing a structured overview, this study seeks to enhance understanding and contribute to the advancement of ML-based solutions for missing data problems. A total of 684 research articles from various scientific databases were analyzed using search engines, and 94 of them were selected as prima ry studies. Finally, several recommendations were given to guide future researchers in applying machine learning to impute missing values.

1.1 Background

Missing values are a common challenge in real-world datasets, often affecting the quality and accuracy of data ana lytics. Missing va lues can happen for multiple reasons, such as respondents' refusal to answer, manual typing errors, or equipment malfunctions [1]– [3]. The presenceof missing values can introduce bia ses and reduce the relia bility of analytical outcomes. Da ta loss can occur due to various reasons, including respondents' unwillingness to provide answers, human errors during data entry, or technical failures in data collection systems. Moreover, Janssen et al. [4] highlight that having a complete dataset can greatly influence the decision-making process in an organization. For example, low-quality data will lead to inaccurate analysis, which will result in the wrong decisions begin made Since missing data is an unavoidable issue, it must be appropriately addressed before proceeding with data preprocessing to ensure high-quality outcomes.

1.2 Research Problem

Although several studies have explored ML-based imputation methods, the research rema ins scattered across various journals and conferences. There is a need for a comprehensive assessment of ML approaches for handling missing data, including eva luating their effectiveness, identifying their limitations, and understanding improvements made over time. Despite numerous proposed models, there is a la ck of a structured review that consolidates findings and tracks advancements in this area.

1.2 Research Problem

This study aims to conduct a systematic literature review (SLR) on ML-based missing value imputation techniques. The

primary objectives include:

i. Identifying recent trends in ML applications for handling missing values.

ii. Evaluating ML methods used to address the missing data problem.

iii. Ana lyzing the strengths and limitations of different ML-based imputation techniques.

2. **LITERATURE REVIEW**

De-duplicating search results and studies effectively during the systematic review process is essential. A comprehensive understanding of the data set's characteristics and proper validation of de- duplication outcomes are also critical.

This prevents compromising the quality of the review and the reliability of its results. The approach proposed in this opinion review is compatible with the personal style of the libraria n but can be universally used by any researcher conducting a systematic review for the purpose of getting reproducible and reliable results.

Document the process - document your de-duplication process in the review protocol clearly. This should include the criteria used to decide duplicates, the softwa re/tools used, and the choices made while de- duplicating so that other people can replicate it. Methodological transparency increases the credibility of the review.

Utilize reference management softwa re (e.g., EndNote, Zotero, Mendeley) to organize and arrange search results. They have automatic (default) de-duplication features that assist in recognizing at least the exact duplicates and eliminate obvious duplicates. e.g., in EndNote the default is author, year, and title. They can also assist in recognizing duplicates based on defined criteria e.g., volume, issue, and pa ges which involve choosing a method best suited for that project e.g., the Bramer method for EndNote.

Manual check - manually check the duplicates identified by some of the automated tools. Determine matching criteria - establish specific criteria for matching e.g., titles, authors, publication dates, and other bibliographic details to determine whether records are duplicates. Establish a cut-off for matching so that potentia lly valuable studies are not lost. Manage multiple versions of one article - mark different versions e.g., conference abstracts, and full-text papers. Consider whether to mana ge them as different records init ia lly or duplicates according to their content (typically screening in full text will address this issue).

Colla borate with non-journal sources - expect to de- duplicate other sources like conference proceedings, reports, and theses alongside journal articles. Notice their unique indexing and citation styles.

Oversee updates and duplicate searching of the ongoing systematic review plan. If your review involves multiple rounds of searching, use the de-duplication process to identify studies already included in previous rounds to avoid the falla cy that updating the search pla n is as easy as continuing where you left off.

Resolve discrepancies - in the event of discrepancies or doubt, refer to your review team to m a ke informed decisions regarding the status of potentially duplicate records. Document decisions - document all decisions taken during the de-duplication process, including the reasons for excluding or keeping records. Transparency in decision-making increases the reproducibility of the review.

Keep in mind that although machines can speed up the de-duplication process, m a nua l checking is usually inevitable and still necessary for accurate duplicate identification, particularly when titles and abstracts cannot be used for differentiation. Transparency, completeness, and consistency are essential guidingprinciples in de-duplicating studies in systematic reviews.

3. Methodology

The following terms explain the various types of de- duplication processes: (1) Exact match de-duplication: This method examines precise matches in key fields, such as unique identifiers or customer IDs. If the same information is shown on multiple records, these duplicates are removed; (2) Fuzzy match de- duplication: Fuzzy de-duplication techniques use algorithms to determine the similarity between records, even if they do not have exact matches in key fields, allowing for the recognition of duplicates with slight differences or misspellings; and (3) Rule-based de-duplication: Rule-based de-duplication involves defining specific rules or criteria to identify duplicates. These rules can be based on data patterns, business logic, or specific requirements" [5].

3.1 Existing Techniques for Data Deduplication

The problem of data duplication has been extensively researched, and various techniques have been proposed over the years. The most notable methods include automated and manual searching, rule-based algorithms, and artificia l intelligence-driven approaches. Below is an analysis of different techniques employed by researchers in the past, along with their limitations.

i. **Hybrid Automated and Manual Searching (2013)** A research team from Fourth Military Medical University proposed a combination of automated and manual searching for duplicate identification. The study categorized duplicates into Type-I (duplicates among different databases) and Type-II (duplicate publications in different journals/issues). While the automated method efficiently identified most Type-I duplicates, the manual method wa s more effective for Type-II duplicates. However, the manual approach led to a higher incidence of incorrect items in Type-I duplicates, particula rly from the EMBASE database. This approach, though comprehensive, wa s time-intensive and prone to human error. [6]

ii. **Comparison of De-Duplication Features in Reference ManagementSoftware(2015)** Canadian researchers examined de-duplication tools availa ble in RefWorks, EndNote, and Mendeley. The study measured false positives, false nega tives, and the time taken for de-duplication. [7] The effectiveness of these tools varied based on user expertise and the complexity of the dataset. While the study

provided valuable insights, the manual involvement required for certa in steps made these tools less efficient for large datasets.[8]

iii. Systematic Review Assistant-De-duplication Module (SRA-DM) (2015)
Bond University researchers developed the SRA-DM to improve duplicate detection. Compared to EndNote's de-duplication process, SRA-DM

demonstrated superior sensitivity (84%) and specificity (100%), detecting 42.86% more duplicates. Despite these improvements, SRA-DM struggled with undetected duplicates due to metadata discrepancies, limiting its reliability.

Iv.Bramer-Method(2016)
Researchers from Erasmus MC-Erasmus University introduced the Bramer method, which leveraged pagination data for duplicate detection. While this method enhanced accuracy, it relied on detailed formatting adjustments, making it complex and time- consuming to implement.[9]

v. Amsterdam Efficient De-Duplication (AED) Method (2019)
The AED method proposed a systematic approach involving accession number tracking and manual assessment. [10] Though the method was claimed to be 100% reliable, it required significant manual intervention for large datasets, which reduced its scalability.

vi. Evaluation of Electronic Methods for De-Duplication (2021)
Canadian researchers compared de-duplication tools such as Ovid, Covidence, and Rayyan. Ovid and Covidence exhibited high specificity, while Rayyan showed high sensitivity. The study, however, did not quantify false positives and false negatives, making it difficult to assess the overall accuracy.[11]

Vii. AI-Based Deduplication - Deduklick (2022) The Swiss research team developed Deduklick, an AI- powered de-duplication algorithm utilizing natural language processing (NLP) and expert-created rules.[12] The system effectively reduced processing time while maintaining high accuracy. Despite its advanta ges, limitations included potential bia ses in decision-making and a lack of real-world implementation data beyond eight datasets.[13]

viii. Automated Systematic Search Deduplicator (ASy SD) (2023)
Developed at the University of Edinburgh, ASySD demonstrated >95% accuracy in duplicate removal across five biomedical datasets, outperforming
previous tools.[14] However, it required signif icant computing resources and struggled with incomplete citation data, making it less practical for non-technical users.



Figure 1: The bar graph showcases the comparative performance of various de-duplication techniques over time, emphasizing improvements in accuracy, specificity, and automation.Proposed Best Technique: AI- Powered Deduplication with NLP and Similarity Scoring

This technique integrates **Artificial Intelligence (AI), NLP, and machine learning models** to automate duplicate detection, ensuring high accuracy, efficiency, and minimal human intervention.

3.1.1 Key Features of the Best Technique

3.1.1.1 Natural Language Processing (NLP)

i. Uses NLP models to analyze textual metadata such as titles, abstracts, and author names.

ii. Identifies minor variations in text, ensuring that duplicate records with slight differences are detected.

iii. Helps in handling different citation styles and format inconsistencies.

3.2.1.2. Metadata Normalization

i. Standardizes bibliographic fields (e.g., author names, journal titles, issue numbers) to reduce discrepancies.

ii. Ensures uniformity in how different databases store information.

iii. Helps in improving the accuracy of duplicate detection by eliminating inconsistencies.

3.2.1.3. Similarity Scoring Algorithm

i. Assigns a similarity score to each record based on metadata fields like **title, author, journal, DOI, year, volume, issue, and page numbers**.

ii. Uses **fuzzy matching techniques** to detect duplicates even when minor changes exist.

iii. Threshold-based classification to determine whether two records are duplicates or unique.

3.2.1.4. Supervised Learning Model

i. Trains machine learning models on large datasets of duplicates and non-duplicates.

ii. Uses **classification algorithms** (such as Support Vector Machines (SVM), Random Forest, or Neural Networks) to predict duplicates with high precision.

iii. Adapts and improves based on feedback from real-world datasets.

3.2.1.4. Automation and Scalability

i. Reduces the time required for manual searching by fully automating the deduplication process.

ii. Capable of processing large datasets efficiently.

iii. Minimizes human intervention while maintaining high accuracy.

3.1.2 Why This is the Best Approach

i. High Accuracy: Uses AI and NLP to detect duplicates more effectively than traditional rule-based or manual methods.

ii. Automation: Reduces the need for manual verification, saving time and effort.

iii. Scalability: Works efficiently on large datasets, making it suitable for real-world applications.

iv. Adaptability: Can be trained and improved over time with new datasets.

4. Discussion

4.1 Key Improvement Over Previous Studies

4.1.1. Higher Accuracy:

i. Unlike rule-based methods (SRA-DM, Bramer, AED), AI models adapt and improve with training, reducing errors over time.

ii. NLP-based simila rity detection ensures that duplicate records with minor variations are still identified.

4.1.2. Full Automation:

i. Manual methods (Bramer, AED, hybrid searching) required user intervention, which was time-consuming.

ii. AI-powered deduplication eliminates the need for human involvement while ensuring precision.

4.1.3. Scalability:

i. Methods like AED and Bramer struggle with large datasets.

ii. AI-powered techniques process **millions of records** efficiently.

4.1.4. Domain Adaptability:

i. ASySD was designed for biomedical databases.

ii. The proposed method is **flexible across industries**, from research articles to business data.

4.1.5. Reduced False Positives & Negatives:

i. Traditional software like EndNote, Mendeley, andOvid struggled with false duplicate detection.

ii. Machine learning models **refine classification over time** to improve accuracy.

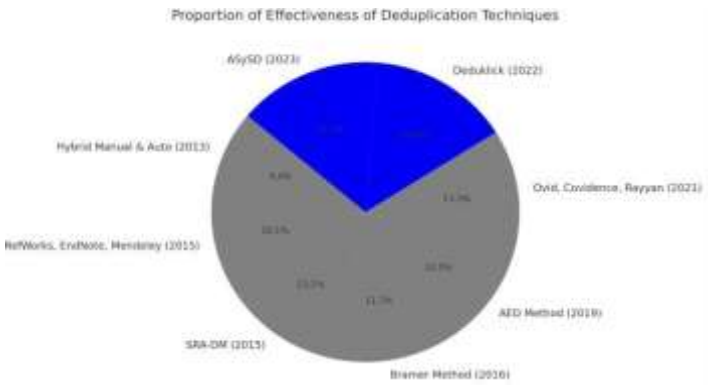


Figure 2: The pie chart illustrates the effectiveness of different de- duplication techniques in terms of accuracy and reliability.

4.2 Limitations and Suggestions For Future Research

4.2.1 Limitations

i. Computational Complexity

The AI-powered deduplication approach requires significant computational power for training and execution, making it resource-intensive for large- scale datasets. High-performance computing environments or cloud-based solutions may be necessary for practical implementation (**Deduklick: AI-Based Algorithm, 2022**).

ii. Data Dependency

The accuracy of the deduplication system relies heavily on metadata completeness and consistency. Problems like duplicate record detection are not well supported by the system and user needs other approaches to deal with duplicate record detection problem.[15] In cases where bibliographic information is missing or formatted inconsistently (e.g., different citation styles or missing author names), the AI model may struggle to correctly identify duplicates (**ASySD: Automated Systematic Search Deduplicator, 2023**).

iii. Domain-Specific Optimization

While the proposed approach performs well on biomedical datasets, its effectiveness in other domains such as legal, financial, and social sciences has not been fully explored. Similar

challenges were noted in earlier studies where de- duplication effectiveness varied across different database structures (**Ovid, Covidence, Rayyan, 2021**).

iv. False Positives & False Negatives

Although AI improves accuracy, some duplicates may still be incorrectly classified due to variations in metadata formatting. The **Bramer method (2016)** highlighted that identifiers such as DOIs and PMIDs are not always reliable, leading to misclassification of records.

v. Real-World Integration Challenges

The implementation of AI-powered deduplication in widely used tools (e.g., EndNote, Zotero, Mendeley) requires API integration and seamless compatibility. The **Amsterdam Efficient De- duplication (AED) Method (2019)** noted that a multi-step manual verification process is still required to ensure accuracy in large datasets.

4.2.2 Suggestions for Future Research

i. Enhancing AI Model Generalization

Future research should explore **transfer learning** to adapt the model to various domains, reducing the need for extensive retraining. Studies like **Deduklick (2022)** indicate that AI models trained on one dataset may not always generalize well to other research fields.

ii. Incorporating Contextual NLP Understanding

Advanced **Natural Language Processing (NLP)** techniques (e.g., **BERT, GPT-based models**) can improve semantic matching between records instead of relying solely on metadata comparisons. This would enhance accuracy and reduce false negatives, similar to the AI-driven approach proposed in **Deduklick (2022).**

iii. Hybrid Model Optimization

A combination of **rule-based approaches** and **deep learning models** could enhance performance, balancing interpretability with accuracy. The **SRA-DM (2015)** showed that hybrid models often outperform fully automated or manual approaches in systematic reviews.

iv. Integration with Blockchain for Data Integrity

Implementing **blockchain-based record tracking** could prevent duplication issues at the source by ensuring transparent and tamper-proof bibliographic records. This approach could mitigate challenges identified in studies like **Bramer Method (2016)** , which struggled with unique identifier inconsistencies.

v. User-Friendly Interface & Cloud-Based Implementation

Developing a **web-based tool** or **cloud-based API** with real-time deduplication capabilities would improve accessibility and usability for researchers. The **ASySD (2023)** method highlighted challenges with software availability, requiring local R Studio execution, which is a barrier for non-technical users.

5. Conclusion

In this paper we studied various approaches to secure deduplication. Additionally, it presented some information around data deduplication approaches, data compression approaches, the idea of compressive sensing, attribute-based encryption (ABE) approach for security and deduplication, image compression and encryption methods, deduplication for encrypted data and some performance comparisons of other approaches and the like. This gives some understanding for secure storage of images to public cloud with cryptography and compression techniques. This also gives some understanding toward the use of compression in the face of continuing increasing size and **effects of cyber-attacks both today and soon.** There is also the continuous need for stronger security and access and less capacity to deal with duplicate data in the cloud to improve performance.[16] These secure techniques assist in helping develop secure deduplication. Given the potential we are heading for the discussion of quantum computers and the potential loss of traditional security altogether this secure in some means deduplication of images and for data storage is required nowadays. So, this paper is very helpful for thinking about using compression sensing to achieve simultaneous data compression and security while improving secure deduplication in public cloud.[17] Moving forward we have intentions to improve the current state of the art in this line of work moving into the future.

6. References

[1] R. Deb and A. W.-C. Liew, ''Missing value imputation for the analysis of incomplete traffic accident data,' Inf. Sci., vol. 339, pp. 274–289, 2016,

[2] C.-F. Tsai and F.-Y. Chang, ''Combining instance selection for better missing value imputation,' J. Syst. Softw., vol. 122, pp. 63–71, Dec. 2016,

[3] K. Dhindsa, M. Bhandari, and R. R. Sonnadara, ''What's holding up the big data revolution in healthcare?'' BMJ, vol. 363, pp. 1–2, Dec. 2018,

[4] M. Janssen, H. van der Voort, and A. Wahyudi, ''Factors influencing big data decision-making quality,' J. Bus. Res., vol. 70, pp. 338–345, Jan. 2017,

[5] Dremio. Deduplication. [cited 20 September 2023]. https://www.dremio.com/wiki/deduplication/

[6] Qi X, Yang M, Ren W, Jia J, Wang J, Han G, Fan D. Find duplicates among the PubMed, EMBASE, and Cochrane Library Databases in systematic review. PLoS One. 2013 [PMC free article] [PubMed] [Google Scholar]

[7] Kwon Y, Lemieux M, McTavish J, Wathen N. Identifying and removing duplicate records from systematic review searches. J Med Libr Assoc. 2015;103:184–188.[PMC free article] [PubMed] [Google Scholar]

[8] Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant- Deduplication Module. Syst Rev. 2015;4:6. [PMC free article] [PubMed] [Google Scholar]

[9] Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. 2016;104:240–243. [PMC free article] [PubMed] [Google Scholar]

[10] Otten R, de Vries R, Schoonmade L. Amsterdam Efficient Deduplication (AED) method. Zenodo. 2019 [Google Scholar]

[11] McKeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev. 2021;10:38. [PMC free article] [PubMed] [Google Scholar]

[12] Borissov N, Haas Q, Minder B, Kopp-Heim D, von Gernler M, Janka H, Teodoro D, Amini P. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. Syst Rev. 2022;11:172. [PMC free article] [PubMed] [Google Scholar]

[13] PRISMA. Welcome to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) website! [cited 20 September 2023]. http://prisma-statement.org/Default.aspx .

[14] Hair K, Bahor Z, Macleod M, Liao J, Sena ES. The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. BMC Biol. 2023;21:189.[PMC free article] [PubMed] [Google Scholar]

[15] Li, Lin. (2012) "Data Quality and Data Cleaning in Database Applications." [doctoral dissertation], School of Computing, Edinburgh Napier University.

[16] Ananthakrishna, R., Chaudhuri, S., Ganti, V.: Eliminating Fuzzy Duplicates in Data Warehouses. In: Proc. of Int. Conf. on Very Large Databases, Hong Kong, China, pp. 586–597 (2002)

[17] Manconi, A. *et al*. Removing duplicate reads using graphics processing units. *BMC Bioinformatics* **17** (2016).