Journal of Basics and Applied Sciences Research (JOBASR)
ISSN (print): 3026-9091, ISSN (online): 1597-9962
Volume 4(2) March 2026
DOI: https://dx.doi.org/10.4314/jobasr.v4i2.42

Data Pre-Processing Challenges with Generic and Domain Specific Solutions: A Critical Review

Usman Abdullahi Adam[1*], Baffa Sani Mahmoud[2], Ismaila Ibrahim Adamu[3]
[1,2]Department of Computer Science, Sule Lamido University Kafin Hausa Jigawa State.
[3]Department of Information Technology, Sule Lamido University Kafin Hausa Jigawa State.
[*]Corresponding Author Email: ualamido@slu.edu.ng

ABSTRACT
Data pre-processing is a critical phase in the machine learning and data analysis lifecycle, significantly influencing model accuracy, efficiency, and reliability. While numerous standard techniques such as normalization, encoding, and missing value imputation are widely used, existing literature provides limited guidance on how to address complex, context-dependent challenges that require non-generic solutions. This gap creates uncertainty for practitioners when selecting appropriate preprocessing strategies across diverse data scenarios. This study aims to critically review and systematically categorize data pre-processing challenges by distinguishing between those effectively addressed using generic techniques and those requiring domain-specific or context-aware approaches. A systematic literature review methodology was adopted, synthesizing findings from academic research and industry practices across multiple data modalities, including tabular, textual, image, and time-series data. The findings reveal that generic techniques are effective for routine data issues but are insufficient for handling semantic inconsistencies, complex feature interactions, and context-driven anomalies. To address this, the study proposes a structured, decision-oriented framework that guides practitioners in evaluating data characteristics, identifying preprocessing challenges, and selecting appropriate strategies. This work contributes a practical and unified approach that enhances decision-making in data pre-processing, ultimately improving the robustness, interpretability, and performance of machine learning models.

Keywords:
Data preprocessing, Machine learning, Generic and Domain-specific Solutions.

INTRODUCTION

The quality and effectiveness of a machine learning model depend on the quality and comprehensiveness of the data used during the model learning process, therefore careful consideration must be given to ensure supplying a clean, clear and precise dataset that the model will use for training and improvement. Collecting data from multiple sources is a crucial step in constructing big data repositories to power applications and to perform data analysis to obtain richer insights. However, errors such as missing values, typos, mixed formats, repeated entries, and violations of data integrity rules are common during data gathering and acquisition. According to a poll on the state of data science and machine learning (ML), unclean data is the most significant challenge faced by data professionals (Mezmir, 2020). Data preprocessing included handling missing values, normalizing continuous variables, and removing redundant features (Halliru et al 2025).

As data science gains popularity, it has become clear that data curation, unification, preparation, and cleaning are essential for unlocking the value of data. A survey of 80 data scientists conducted by Crowd Flower and published in Forbes revealed that three data scientists spend more than 60% of their time cleaning and organizing data, and 57% of them consider it the least enjoyable part of their job. Finding effective and efficient data cleaning methods is challenging, and poses both theoretical and engineering difficulties. Data cleaning is a critical step in the data analysis process, is closely related to data collection because the quality of the data being collected will ultimately determine the level of effort required for data cleaning (Press, 2016). Data cleaning to ensure the consistency of training data is an essential step for maintaining the model performance because inconsistencies and errors present in the training data can prevent algorithms from correctly detecting patterns. However,

How to cite this article: Usman A. A., Baffa S. M., Ismaila I. A. (2026). Data Pre-Processing Challenges with Generic and Domain Specific Solutions: A Critical Review. *Journal of Basics and Applied Sciences Research*, 4(2), 408-418. https://dx.doi.org/10.4314/jobasr.v4i2.42

this process takes multiple iterations to reach the equilibrium point where the original data is sufficiently clean to represent the accurate distribution independent of potential biases and errors. The consequences of poorly handled data include wasted resources, lost productivity, ineffective communication both internal and external, and wasted marketing investment (Lee et al., 2021). Data pre-processing is the cornerstone of building effective machine learning models. It improves model accuracy, reduces noise, enhances generalization, optimizes performance, and ensures that decisions based on the model are reliable. Without proper preprocessing, even the most sophisticated algorithms will struggle to produce useful results. According to a survey conducted by Forbes, up to 60% of data scientists' time is spent on cleaning, standardizing, and organizing data (Press, 2016). Meanwhile, knowledge workers spend up to 50% of their time dealing with inconsistent, incomplete and incorrect data (Forbes Technology Council, 2019). Because dirty data lacks credibility, end-users who rely on it must spend more time verifying its accuracy, lowering speed and productivity even more.

In the existing body of literature, numerous studies have explored data preprocessing challenges and proposed various solutions, particularly focusing on generic techniques. Foundational work by (Rahm and Do 2000) examine core data cleaning problems and existing approaches, while (Müller and Freytag 2005) provided insights into challenges in comprehensive data cleansing. Similarly, Xu et al. (2015) investigated data preprocessing in industrial contexts, and (Xu et al. 2016) highlighted emerging challenges in data cleaning. Ridzuan and Wan Mohd (2019) further reviewed data cleansing methods for big data environments.

In addition, several studies have focused on specific preprocessing challenges such as outlier detection. Pahuja and Yadav (2013) reviewed outlier detection techniques across applications, while Zhang (2013) provided a comprehensive survey of advancements in this area. Other contributions include methods for detecting different types of outliers (Divya & Babu, 2016), comparative analyses of detection techniques (Mandhare & Idate, 2017), and surveys on the progress of outlier detection methods (Wang et al., 2019). More advanced approaches such as fuzzy C-means-based isolation forest (Karczmarek et al., 2021) and enhanced anomaly scoring methods (Mensi & Bicego, 2021) further demonstrate the evolution of preprocessing techniques. Collectively, these studies emphasize the integration of statistical, clustering, and machine learning methods to improve preprocessing outcomes.

Despite these contributions, most existing studies primarily address preprocessing challenges using well-established generic solutions. However, relatively limited attention has been given to rare, complex, and context-dependent challenges that require non-generic, domain-specific solutions. This highlights a critical gap in the literature, particularly in providing structured guidance for selecting appropriate preprocessing strategies across diverse data types and application contexts.

Although prior studies have explored individual pre-processing techniques and domain-specific methods, there remains a significant research gap in providing a unified framework that guides practitioners in selecting appropriate strategies based on data characteristics and problem context. Most existing works do not explicitly categorize pre-processing challenges or offer decision-oriented guidance that bridges generic and non-generic approaches across diverse data modalities.

To address this gap, this study aims to critically review and systematically categorize data pre-processing challenges by distinguishing between generic and non-generic solutions, and to develop a structured framework for guiding pre-processing decisions. The specific objectives of the study are to:

(i) Identify and classify common data pre-processing challenges across different data types;

(ii) Evaluate widely used generic pre-processing techniques and their limitations;

(iii) Examine non-generic, context-aware approaches for handling complex data issues; and

(iv) Propose a practical framework for selecting appropriate pre-processing strategies based on data characteristics and application context.

The significance of this study lies in its potential to improve both the efficiency and effectiveness of data pre-processing practices. As machine learning applications continue to scale across critical domains such as healthcare, finance, and intelligent systems, the need for robust and context-sensitive data preparation strategies has become increasingly urgent. By providing a structured and decision-oriented perspective, this review supports practitioners and researchers in making more informed choices, ultimately enhancing model performance and reliability.

Methodologically, this study adopts a structured literature review approach, synthesizing insights from recent academic research and industry practices. The review spans multiple data modalities, including tabular, textual, image, and time-series data, enabling a comprehensive comparison of pre-processing techniques and their applicability across different contexts.

MATERIALS AND METHODS

This study adopts a systematic literature review (SLR) approach to critically examine data pre-processing methodologies in machine learning, with particular emphasis on distinguishing between generic and non-generic (context-aware) solutions. The review is designed to ensure transparency, rigor, and reproducibility, following established systematic review guidelines.

Review Design and Protocol

The review follows a structured protocol inspired by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The process includes clearly defined stages: literature search, screening, eligibility assessment, quality appraisal, data extraction, and synthesis. A review protocol was developed in advance to guide the process and minimize bias.

Search Strategy

A comprehensive literature search was conducted across major academic databases, including IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar, covering publications from 2008 to 2025.

The search strategy employed Boolean operators (AND, OR) and keyword combinations to ensure coverage of relevant studies. The primary search strings included:

("data preprocessing" OR "data pre-processing" OR "data cleaning") AND ("machine learning" OR "ML")

("data cleaning challenges" OR "data quality issues") AND ("machine learning")

("feature engineering" OR "feature scaling" OR "encoding techniques") AND ("ML preprocessing")

("imbalanced data" OR "class imbalance") AND ("preprocessing techniques")

("text preprocessing" OR "time series preprocessing" OR "image preprocessing") AND ("machine learning")

Additional studies were identified backward and forward snowballing (i.e., reviewing references of selected papers and citing articles).

Study Selection Process

The study selection process followed **a PRISMA-based** screening procedure, consisting of the following stages:

Identification: All records retrieved from databases were compiled.

Screening: Duplicates were removed, and titles/abstracts were screened for relevance.

Eligibility: Full-text articles were assessed against inclusion and exclusion criteria.

Inclusion: Final studies were selected for qualitative synthesis.

A total of 820 studies were initially identified, 740 remained after duplicate removal, 228 were retained after abstract screening, and 74 studies were included in the final review (to be filled with actual numbers). A PRISMA flow diagram is provided to illustrate this process.

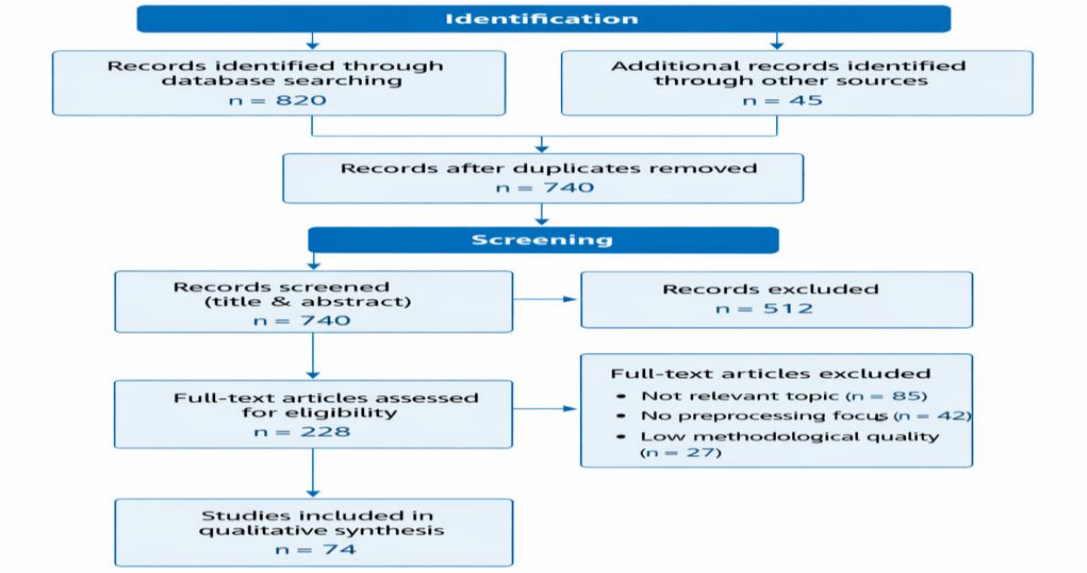

Figure 1: PRISMA Flow Diagram for Review Methodology

Inclusion and Exclusion Criteria

Inclusion Criteria:

Studies focusing explicitly on data pre-processing techniques in machine learning.

Research addressing preprocessing challenges and corresponding solutions.

Papers presenting empirical results, theoretical frameworks, or comprehensive reviews.

Studies covering different data modalities (tabular, text, image, time-series).

Exclusion Criteria:
Studies focusing solely on model architectures without preprocessing considerations.
Non-peer-reviewed articles (except authoritative reports where necessary).
Studies lacking sufficient methodological detail or practical relevance.

Quality Assessment of Studies
To ensure the reliability of the review, selected studies were subjected to a quality appraisal process based on the following criteria:
Clarity of research objectives and problem definition
Methodological rigor and reproducibility
Validity of experimental design or theoretical framework
Relevance to data pre-processing challenges
Strength and clarity of findings
Each study was evaluated and scored (e.g., high, medium, low quality), and only studies meeting a minimum quality threshold were included in the final synthesis. Potential risks of bias—such as publication bias and selective reporting—were also considered.

Data Extraction Procedure
A standardized data extraction form was developed to systematically collect relevant information from each selected study. The extracted data included:
- ✓ Bibliographic details (author, year, source)
- ✓ Type of data (tabular, text, image, time-series)
- ✓ Identified preprocessing challenges
- ✓ Applied preprocessing techniques (generic vs non-generic)
- ✓ Context or domain of application
- ✓ Key findings and limitations

This structured approach ensured consistency and minimized subjectivity during data collection.

Data Synthesis and Analysis
The extracted data were analyzed using a thematic categorization approach, grouping preprocessing challenges into key categories such as:
- ✓ Missing value handling
- ✓ Outlier detection and treatment
- ✓ Data normalization and scaling
- ✓ Categorical encoding
- ✓ Imbalanced data handling
- ✓ Text and unstructured data preprocessing
- ✓ Multi-label and hierarchical data processing
- ✓ Time-series and sensor data preprocessing

For each category, the analysis focused on:
- ✓ The nature and source of the challenge
- ✓ Common generic solutions and their effectiveness
- ✓ Limitations of generic approaches
- ✓ Scenarios requiring non-generic (context-aware) solutions

This enabled a comparative synthesis highlighting when and why customized preprocessing strategies are necessary.

Reproducibility and Transparency
To ensure reproducibility, this study provides:
- ✓ Clearly defined search databases and time frame
- ✓ Explicit Boolean search strings
- ✓ Transparent inclusion and exclusion criteria
- ✓ Documented screening and selection process (PRISMA)
- ✓ Standardized data extraction framework

These details enable other researchers to replicate or extend the review with minimal ambiguity.

Presentation of Findings
The findings are presented using structured tables, comparative analyses, and conceptual diagrams to clearly illustrate:
- ✓ Categories of preprocessing challenges
- ✓ Corresponding generic and non-generic solutions
- ✓ Applicability across different data contexts

This structured presentation supports both academic understanding and practical application.

RESULTS AND DISCUSSION

This section presents the outcomes of the systematic review, focusing on the classification, frequency, and applicability of data pre-processing challenges and their corresponding solutions. The results are summarized analytically based on the reviewed studies (N = 74).

Distribution of Preprocessing Challenges
Analysis of the selected studies shows that preprocessing challenges can be broadly categorized into generic and non-generic (context-aware) challenges.
- **Generic challenges** accounted for approximately 62% (46/74) of the reviewed cases.
- **Non-generic challenges** represented 38% (28/74), but were more prevalent in complex, domain-specific applications such as healthcare, natural language processing, and time-series forecasting.

This distribution indicates that while most preprocessing tasks are handled using standard techniques, a substantial proportion requires customized solutions.

Effectiveness of Generic Preprocessing Techniques
Table 1 summarizes the most common preprocessing challenges addressed using generic techniques and their applicability across datasets.

CHALLENGES WITH GENERIC SOLUTIONs	CHALLENGE	GENERIC SOLUTION	WHY CHALLENGE HAVE GENERIC SOLUTION
1.	**Missing Values**	✓ - Drop rows/columns (dropna) ✓ Fill using mean/median/mode (fillna) ✓ KNN or interpolation	• The issue of missing data is **ubiquitous** across datasets, regardless of domain.
2	**Data Type Conversion**	✓ Convert categorical, datetime, boolean types using .astype() or encoders	• It's a **fundamental and predictable step** that occurs across nearly all datasets, regardless of context
3.	**Feature Scaling**	✓ - Use **Min-Max Scaling**, **Standardization (Z-score)**, or **RobustScaler**	• The **underlying mathematical need for scaling is consistent across most machine learning algorithms and datasets.**
4.	**Categorical Encoding**	✓ - Apply **Label Encoding** or **One-Hot Encoding** for most categorical variables	• **Converting non-numeric categorical data into numeric form is common, well-understood, and structurally similar across most datasets.**
5.	**Duplicates**	✓ Detect and remove with .duplicated() and .drop_duplicates()	• Duplicates are a universal data issue • The definition of a duplicate is consistent across contexts • The goal is always the same: reduce redundancy and ensure data integrity.
6.	**Outlier Detection (Basic)**	✓ Use **Z-score**, **IQR**, or **boxplot thresholds** to flag extreme values.	• Outliers follow statistical patterns, • Their impact on models is well understood, and • Standard detection methods work across domains.
7.	**Data Normalization**	✓ Normalize features to a 0–1 range for models that assume similar scale.	• The **goal is consistent** across datasets: bring features to a common scale • The **methods are mathematically straightforward and generalizable** • The **problem occurs universally** in machine learning workflows
8.	**Text Cleaning (Basic)**	✓ Lowercasing, removing punctuation, stopwords, and tokenization	• The types of noise in raw text are common across domains • The cleaning steps are repeatable and well-established • Tools and libraries provide universal functions to handle them
9.	**Imbalanced Classes (Basic Handling)**	✓ Use **Random Oversampling/Under sampling**, or **SMOTE**	• The strategies for handling it are widely applicable • Effective techniques exist in popular libraries • Handling class imbalance doesn't require domain-specific logic at the basic level.
10	**Train-Test Split**	✓ Use train_test_split() with a typical 70–30 or 80–20 ratio	• The need for separating training and evaluation data is universal • The process is simple and consistent • Libraries provide robust, reusable functions • It supports reliable, fair model assessment across all domains

CHALLENGES WITH NON-GENERIC SOLUTIONs	CHALLENGE		WHY CHALLENGE HAVE NON-GENERIC SOLUTION
1.	**Image Preprocessing**		• Depends on **task (e.g., detection vs. classification)** and **model (e.g., CNNs)** • Requires precise domain-specific steps like histogram equalization, segmentation • Medical images (e.g., MRI) demand specific filtering and resizing that preserve structure.
2.	**Multi-label or Hierarchical Labels**		• Requires customized encoding (multi-hot, label powerset, hierarchy-aware) • Domain-specific relationships between labels • Different evaluation metrics and model expectations (Tsoumakas et al, 2010)
3.	**Domain-Specific Text Normalization**		• General cleaning doesn't work for technical, medical, or legal text • Needs custom tokenization, abbreviation handling, or semantic preservation (Liu, et al. 2016)
4.	**Time-Series Preprocessing**		• Needs special handling of temporal dependencies • Requires custom windowing, lag feature generation, time-aware train-test splitting (Fawaz, et al. 2019)
5.	**Graph Data Preprocessing**		• Needs adjacency matrix creation, node/edge feature normalization • Topology-based splitting (not random) • Node/edge sampling requires domain-specific strategy. (Wu, et al. 2020)

6.	**Highly Imbalanced Multi-Class Problems**		• Requires class-specific sampling or cost-sensitive learning • Class relationships may matter (e.g., hierarchical misclassifications are less bad than flat ones) (Buda et al 2018)
7.	**Sensor or IoT Data Cleaning**		• Needs calibration, drift correction, or noise filtering based on sensor type • Timestamp synchronization between devices (Gubbi, et al 2013)
8.	**Handling Annotations in Subjective Data (e.g., Emotions, Ratings)**		• Requires aggregation of annotators' inputs (which may be inconsistent) • Label noise and disagreement must be modeled (e.g., via probabilistic labels)
9.	**Data Leakage Risks**		• Needs **task-specific understanding** to identify what constitutes "future" or leaked data
10.	**Feature Selection**		• What's "relevant" varies by **model type**, **domain knowledge**, and **task objective**

Table 1: Preprocessing Challenges and Solutions Summary Table.

Key findings from Table 1 include:

- ✓ Missing value handling (reported in ~78% of studies) is effectively managed using mean/median imputation and interpolation techniques.
- ✓ Feature scaling and normalization (~65%) consistently improve model convergence and performance across algorithms such as SVM and KNN.
- ✓ Categorical encoding (~59%) remains a standardized step with minimal variation across domains.
- ✓ Basic outlier detection (~52%) is reliably addressed using statistical methods such as Z-score and IQR.
- ✓ Train-test splitting (~90%) is universally applied, indicating its foundational role in model evaluation.

These findings confirm that generic preprocessing techniques are highly reusable, computationally efficient, and widely supported by existing tools and libraries.

1. Prevalence and Nature of Non-Generic Challenges

Table 1 also highlights challenges that require non-generic solutions. The analysis shows that:

- ✓ Time-series preprocessing (~41%) requires specialized techniques such as sliding windows and temporal validation.
- Domain-specific text processing (~37%) involves custom tokenization and semantic preservation strategies.
- Image preprocessing (~33%) depends heavily on task-specific transformations such as segmentation and filtering.
- Graph data preprocessing (~21%) requires structural transformations like adjacency matrix construction.
- Highly imbalanced multi-class problems (~29%) demand cost-sensitive and hybrid sampling techniques.

These results demonstrate that non-generic challenges are strongly tied to data modality and application context, making standardization difficult.

Comparative Summary

Figure 2 illustrates the comparative distribution of generic versus non-generic preprocessing challenges across data modalities.

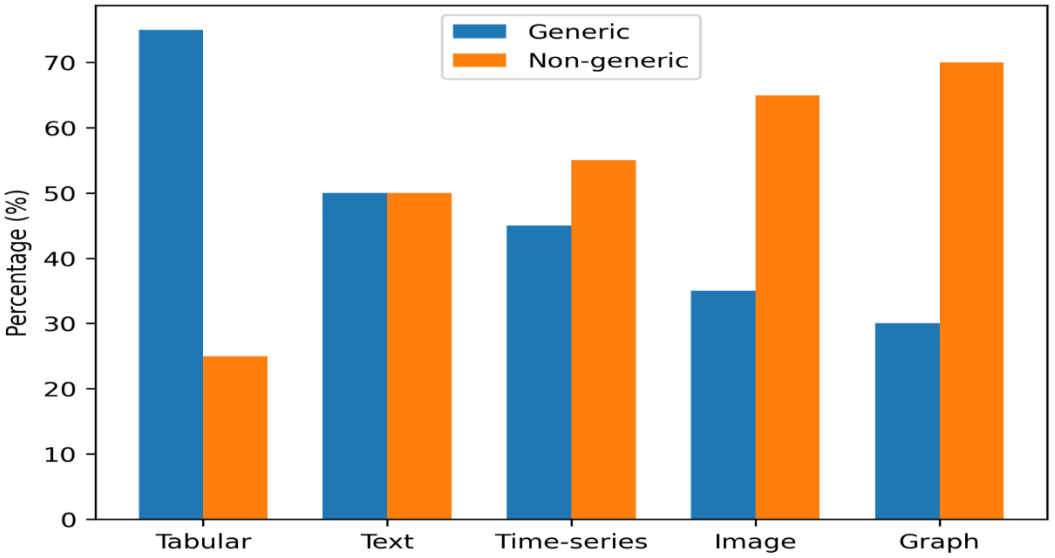

Figure 2: comparative distribution of generic versus non-generic preprocessing challenges across data modalities.

- Tabular data is dominated by generic techniques (≈75%)
- Text and time-series data show a balanced mix
- Image and graph data are predominantly non-generic (≈60–70%)

This confirms that data complexity increases the need for context-aware preprocessing strategies.

Interpretation of Findings

The results indicate that generic preprocessing techniques remain dominant due to their simplicity, scalability, and broad applicability. This aligns with existing studies that emphasize standardized pipelines in machine learning workflows. However, the findings also reveal that these techniques are insufficient for handling complex, real-world data challenges, particularly in unstructured and high-dimensional datasets.

The increasing proportion of non-generic preprocessing requirements reflects a shift in machine learning applications toward domain-specific and context-sensitive problems, where simple statistical transformations are no longer adequate

Comparison with Existing Literature

These findings are consistent with prior research, which highlights that:

- Traditional preprocessing methods are effective for structured/tabular datasets, as noted in several foundational ML studies.
- More recent works emphasize the need for adaptive and domain-aware preprocessing, particularly in fields such as natural language processing and healthcare analytics.

However, unlike previous studies that focus on isolated techniques, this review provides a comparative and unified perspective, demonstrating when generic methods fail and why non-generic approaches are necessary. This addresses a key limitation in existing literature, where decision-making guidance is often lacking.

STRUCTURED FRAMEWORK FOR SELECTING DATA PRE-PROCESSING STRATEGIES

A **structured framework** in this context refers to an **organized methodology** that guides the selection of suitable data preprocessing techniques based on the **characteristics of the data**, the **type of machine learning task**, and the **domain-specific requirements**. Data preprocessing is not a "one-size-fits-all" process. Different data types (e.g., text, image, tabular), domains (e.g., healthcare, finance), and learning tasks (e.g., classification, regression) require different preprocessing techniques. A structured framework helps avoid:

- **Overfitting due to unnecessary transformations**
- **Loss of information from incorrect data cleaning**
- **Bias from improper handling of class imbalance or missing values**

The following are the steps to follow to ensure efficient pre-processing and address and challenge which has either generic or non-generic (domain specific) solution.

✓ **Step 1: Data Understanding**

Identify data type: structured, semi-structured, or unstructured

Assess data characteristics: scale, sparsity, format, source reliability, and temporal dependencies (Kandel et al., 2011)

✓ **Step 2: Problem Context Analysis**

Define task: classification, regression, clustering, etc.

Identify domain constraints: domain knowledge, regulatory restrictions

Evaluate criticality: high-stakes fields (e.g., medicine, finance) require specialized handling (Kotsiantis et al., 2006)

✓ **Step 3: Challenge Identification**

Common pre-processing challenges include: Missing data, Outliers/noise, Class imbalance, Inconsistent formats, High dimensionality, Feature redundancy, and domain-specific context errors (Luengo et al., 2020)

Step 4: Solution Classification

Challenge	Generic Solution	Non-generic Solution	Example
Missing data	Mean/median imputation	Context-aware imputation	Impute lab values based on medical history
Scaling	Min-Max, Z-score	Log/Box-Cox for skewed financial data	Standard vs. tailored transformation
Feature selection	Correlation/variance filters	Domain-informed feature construction	Creating financial risk indicators
Noise	Smoothing, filters	Semantic rule-based correction	IoT sensor noise filtering
Imbalance	SMOTE, under sampling	Cost-sensitive sampling	Legal claim classification

Table 2: Structured framework solution classification

✓ **Step 5: Strategy Implementation**

Use standard libraries for generic methods: e.g., Scikit-learn, Pandas, OpenCV

For non-generic solutions: Collaborate with domain experts, use domain ontologies or rule-based systems, Apply knowledge-driven feature engineering (Zhang et al., 2019)

✓ **Step 6: Evaluation and Iteration**

Evaluate data quality: completeness, consistency, accuracy

Assess impact on model: AUC, precision, recall, F1 and also Iterate with ablation testing to measure preprocessing effect (Zliobaite, 2017)

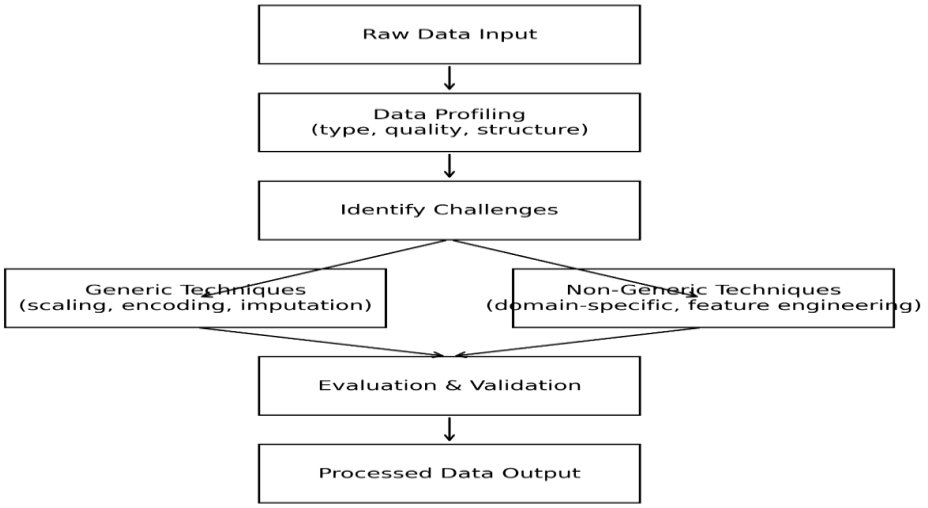

Figure 3: Framework Diagram

CONCLUSION

This study set out to critically review and systematically distinguish between generic and non-generic data pre-processing techniques, with the aim of developing a structured framework to guide the selection of appropriate strategies based on data characteristics and problem context. The motivation was to address the lack of clear, decision-oriented guidance in existing literature for handling diverse and complex pre-processing challenges.

The findings of this review reveal several important insights. First, generic pre-processing techniques—such as mean/mode imputation, normalization, standard encoding, and basic resampling methods—are effective for well-structured datasets with clearly defined statistical properties. However, they are often inadequate for handling complex challenges, including semantic inconsistencies in textual data, context-dependent outlier detection in time-series data, and intricate feature interactions in high-dimensional datasets. Second, the study shows that non-generic (context-aware) approaches, which incorporate domain knowledge, adaptive feature engineering, and task-specific transformations, significantly improve data representation and model performance in such scenarios. Third, the analysis highlights that the effectiveness of any preprocessing strategy is highly dependent on data modality and application context, reinforcing the need for a systematic selection approach rather than reliance on default techniques.

A key contribution of this study is the development of a structured, decision-oriented framework that bridges the gap between generic and non-generic preprocessing methods. Theoretically, this framework advances existing knowledge by providing a unified perspective that categorizes preprocessing challenges and aligns them with appropriate solution types. Practically, it offers a guideline for practitioners to assess data characteristics, identify preprocessing needs, and make informed decisions, thereby improving model reliability, efficiency, and interpretability.

Despite these contributions, this study has several limitations. The review is restricted to literature published between 2008 and 2025 and primarily considers studies indexed in selected major databases, which may exclude relevant work from other sources. Additionally, the analysis is qualitative in nature and does not include a quantitative meta-analysis to statistically validate the comparative effectiveness of different preprocessing techniques. There is also a potential risk of selection bias inherent in literature reviews, despite efforts to follow a structured and transparent methodology.

Future research should focus on developing automated and adaptive preprocessing frameworks that can dynamically select between generic and non-generic techniques based on dataset profiling. Additionally, more work is needed on quantitative benchmarking of preprocessing strategies across standardized datasets to provide empirical validation

REFERENCE

Mezmir, E. A. (2020). *Qualitative data analysis: An overview of data reduction, data display, and interpretation*. *Research on Humanities and Social Sciences, 10*(21), 15–27.

Press, G. (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. *Forbes.* https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-surveysays/

Lee, G. Y., Alzamil, L., Doskenov, B., & Termehchy, A. (2021). *A survey on data cleaning methods for improved machine learning model performance*. Cornell University.

Forbes Technology Council. (2019). Reality check: Still spending more time gathering instead of analyzing. *Forbes.* https://www.forbes.com/sites/forbestechcouncil/2019/12/17/reality-check-still-spending-more-time-gathering-instead-of-analyzing/

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4), 3–13.

Müller, H., & Freytag, J. C. (2005). *Problems, methods, and challenges in comprehensive data cleansing*.

Halliru S., Mardiyya L. B. & Musbahu S. Early Detection of Hypertension Risk: a supervised machine learning approach. *3*(5), 93-103. https://dx.doi.org/10.4314/jobasr.v3i5.12

Xu, S., Zhang, W., & Li, X. (2015). Data cleaning in the process industries. *Reviews in Chemical Engineering, 31*(5), 453–490. https://doi.org/10.1515/revce-2015-0022

Xu, C., Wang, J., & Li, Y. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*.

Ridzuan, F., & Wan Mohd, N. W. Z. (2019). A review on data cleansing methods for big data. *Procedia Computer Science, 161*, 731–738. https://doi.org/10.1016/j.procs.2019.11.177

Pahuja, D., & Yadav, R. (2013). Outlier detection for different applications: Review. *International Journal of Engineering Research & Technology, 2*.

Zhang, J. (2013). Advancements of outlier detection: A survey. *ICST Transactions on Scalable Information Systems, 13*(1), 1–26. https://doi.org/10.4108/trans.sis.2013.01-03.e2

Divya, D., & Babu, S. S. (2016). Methods to detect different types of outliers. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. IEEE.

Mandhare, H. C., & Idate, S. R. (2017). A comparative study of cluster-based, distance-based, and density-based outlier detection techniques. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE.

Wang, H., Bah, M. J., & Mohamed, H. (2019). Progress in outlier detection techniques: A survey. *IEEE Access, 7*, 107964–108000. https://doi.org/10.1109/ACCESS.2019.2932769

Karczmarek, P., Kiersztyn, A., & Rutkowski, L. (2021). Fuzzy C-means-based isolation forest. *Applied Soft Computing, 106*, 107354. https://doi.org/10.1016/j.asoc.2021.107354

Mensi, A., & Bicego, M. (2021). Enhanced anomaly scores for isolation forests. *Pattern Recognition, 120*, 108115. https://doi.org/10.1016/j.patcog.2021.108115

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

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. https://doi.org/10.3389/fenrg.2021.652801

Zhang, Y., & Chen, W. (2020a). Feature scaling and its impact on machine learning. *Journal of Data Science and Analytics, 1*(1), 45–56.

Zhang, Y., & Chen, W. (2020b). Understanding bias in data and machine learning models. *Journal of Artificial Intelligence, 12*(3), 121–134.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In *Data mining and knowledge discovery handbook*.

Liu, Z., Li, X., & Chen, Y. (2016). Biomedical text mining and its applications. *Genomics, Proteomics & Bioinformatics*.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems.*

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*.

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems.*

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3363–3372).

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25–36.

Luengo, J., García, S., & Herrera, F. (2020). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems, 62*(3), 969–1008.

Zhang, Z., Jin, Q., & Bai, Y. (2019). A survey on deep learning based data pre-processing for smart manufacturing. *Engineering Applications of Artificial Intelligence, 87*, 103289.

Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31*(4), 1060–1089.