INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 7, July 2026 | ISSN: 2456-4184 | IJNRD.ORG

PREPRO+: SMART DATA PREPROCESSING TECHNIQUES FOR AUTOMATED AND INTELLIGENT MACHINE LEARNING DATA PREPARATION

[1]Ms. Priyanka S. Gosavi, [2]Mr. Ravindra S. Kamble

[1]Student, [2]Assistant Professor
Department of Computer Science & Engineering
D Y Patil Agriculture and Technical University, Talsande, Kolhapur, Maharashtra, 416112, India
priyankagosavi19@gmail.com, ravindra.kamble@dyp-atu.edu.in

Abstract : Data preprocessing is a critical step in the machine learning process because the quality of input data directly affects model performance. Real-world datasets often contain missing values, duplicate records, inconsistent formats, noise, and irrelevant features that require careful preparation before analysis. This paper presents **PrePro+**, a Streamlit-based web application that simplifies and automates common preprocessing tasks such as data cleaning, missing value handling, outlier detection, feature encoding, data scaling, feature selection, dimensionality reduction, and data visualization using Python libraries. The system supports multiple dataset formats and provides an easy-to-use interface suitable for users with different levels of technical expertise. By reducing manual effort and ensuring consistent preprocessing, PrePro+ helps prepare high-quality datasets more efficiently for machine learning and data analysis across domains such as healthcare, finance, education, and research.

IndexTerms *- Data Preprocessing, Machine Learning, Data Cleaning, Outlier Detection, Feature Selection, Dimensionality Reduction, Principal Component Analysis, Data Transformation, Data Visualization, Streamlit, Automation, Data Quality, Feature Scaling, Categorical Encoding, Interactive Analytics, Intelligent Systems, Predictive Modeling, Data Science.*

I. INTRODUCTION

Digital technologies have produced large amounts of data from a wide range of data sources, such as health-care, financial institutions, social media, education, industrial sensors and e-commerce applications. While all those datasets are valuable, they frequently are incomplete, inconsistent, noisy and unstructured. These issues can have a profound impact on machine learning algorithms and lead to poor predictions and poor analysis results. To ensure data quality, consistency, and usability, collected data needs to be extensively preprocessed before any machine learning model is used to learn from it. Data preprocessing, thus, is a crucial step in the success of machine learning applications and a key step in the data science pipeline.

The conventional data pre-processing methods are mainly based on manual coding and need a high level of programming and statistical analysis skills. Time spent on repetitive tasks like managing missing data, dropping duplicate rows, identifying outliers, normalizing categorical variables, and scaling numeric variables is a significant portion of a data scientist's time. These manual tasks slow down the development of projects and add the risk of mistakes and inconsistencies. Many preprocessing tools available today only offer individual functionality and lack a simplified and coherent user space that integrates automation, visualization and intelligent decision-making. This constraint is more evident with massive data and among less technically skilled users.

In this research, we introduce an intelligent and automatic data preprocessing framework called PrePro+, aiming to ease and speed up the process of data preparation for ML applications. The proposed system combines several pre-processing tasks into a single web application built on Streamlit. It implements smart data cleaning, transformation, outlier detection, dimensionality reduction and visualization algorithms with real-time user interaction and feedback. This framework minimizes manual effort and guarantees a uniform preprocessing procedure, making it accessible for producing high-quality datasets ready for machine learning. It streamlines the process of producing datasets suitable for ML and guarantees uniformity in the preprocessing procedure, decreasing manual effort. Thus, the proposed

solution increases the efficiency of preprocessing, improves model accuracy and enables more sophisticated data preparation methods to be accessible to both experts and non-experts across different application areas.

To conclude, data preprocessing is a critical and complex stage in the ML workflow that significantly affects the performance and trustworthiness of predictive models. Current solutions can be quite technical and complex, and manual preparation can be time-consuming and prone to errors. The proposed PrePro+ framework addresses these limitations by using intelligent automation, interactive visualization, and integrated functionalities for preprocessing. The system's user-friendly and comprehensive interface makes data preparation easier, enhances data quality, and facilitates the creation of precise and reliable machine learning solutions.

II. LITERATURE REVIEW

With the increasing complexity of machine learning models, the quality of the data used in their development has become a crucial part of machine learning, making data preprocessing one of the most important phases in the machine learning lifecycle. In recent years, attention has been paid to automating preprocessing operations to reduce human effort and increase the efficiency of processing operations. Salhi et al. (2023) and Fernandes et al. (2023) pointed out that much of the work in a machine learning project requires data cleaning, transformation, and preparation. To tackle these challenges, researchers have created automated Machine Learning (AutoML) frameworks that combine preprocessing, feature engineering and model optimization in one system. Several research papers have shown that automatic data cleaning, feature selection and dimensionality reduction methods can significantly improve predictive accuracy. Advanced preprocessing techniques using intelligence have made it possible for researchers to deal with complicated datasets more efficiently. In addition, the progress of AutoML has enabled machine learning to be accessible to non-technical users, reflecting the increasing relevance of automation in contemporary data science.

A. Salhi, A. C. Henslee (2023) - Data Preprocessing Using AutoML: A Survey

This paper surveys recent advances in Automated Machine Learning (AutoML) techniques focused on data pre-processing. The authors examine the importance of preprocessing for machine learning model performance and describe automated methods for handling missing values, feature transformation, normalization, categorical encoding, and feature selection. The paper summarizes various AutoML platforms and their preprocessing features, and outlines the rising need for intelligent preprocessing systems to minimize manual intervention and improve workflow efficiency. Challenges of automated preprocessing, such as scalability, interpretability and adaptability across different datasets, are discussed. The survey concludes that automated preprocessing is now an integral part of modern machine learning pipelines.

Alvaro A. A. Fernandes et al. (2023) - Data Preparation: A Technological Perspective and Review

This work offers a comprehensive review of contemporary data preparation technologies and methods, focusing on data profiling, schema matching, data transformation, data integration and data repair. The paper classifies approaches into programming-based, workflow-based, dataset-based, and automated systems, and notes that a significant portion of data scientists' time is invested in preparing data rather than developing models. Emerging trends such as machine-learning-based data cleaning, intelligent transformation generation, and human-in-the-loop preprocessing are highlighted, and the authors conclude that automation will be key to future data analytics systems.

L. Santos, L. Ferreira (2023) - Atlantic: Automated Data Preprocessing Framework for Supervised Machine Learning

This paper presents Atlantic, an open-source automated preprocessing framework combining feature engineering, automated feature selection, categorical encoding and null-value imputation, using optimization techniques based on ensemble tree models to find the best preprocessing configuration for a given dataset. Experiments across several datasets show that automatic preprocessing can improve model quality and reduce development time, for both classification and regression tasks, using metrics such as AUC and MAE.

D. Qi, J. Peng, Y. He (2023) - Auto-FP: Automated Feature Preprocessing for Tabular Data

This study formulates feature preprocessing as a Hyperparameter Optimization (HPO) and Neural Architecture Search (NAS) task, automating the selection and sequencing of preprocessing operations. Fifteen optimization algorithms are evaluated on forty-five public datasets, and the results show evolutionary search techniques

outperforming conventional optimization approaches, emphasizing the difficulty of manually building optimized preprocessing pipelines.

A. Mumuni, F. Mumuni (2024) - Automated Data Processing and Feature Engineering for Deep Learning and Big Data

This survey analyzes automated data processing techniques for deep learning and big data systems, highlighting the importance of intelligent preprocessing for large-scale heterogeneous datasets. Experimental results show that automatic preprocessing substantially decreases development time without compromising predictive performance, while scalability and interoperability remain key research challenges.

X. He, K. Zhao, X. Chu - AutoML: A Survey of the State-of-the-Art

This survey covers AutoML technologies including data preparation, feature engineering, hyperparameter optimization, and neural architecture search. Automated feature transformation, encoding and data cleaning procedures are discussed across various AutoML frameworks, showing that automated preprocessing enhances the uniformity and reliability of machine learning pipelines.

M.-A. Zöller, M. F. Huber - Benchmark and Survey of Automated Machine Learning Frameworks

This paper benchmarks popular AutoML frameworks across 137 datasets, examining automated feature engineering, missing value treatment, encoding approaches, and model selection. The results indicate substantial performance gains from preprocessing automation and emphasize the need for comprehensive preprocessing pipelines that can flexibly adapt to different data environments.

L. Santos, L. Ferreira (2024) - Atlantic: Open-Source Automated Data Preprocessing (Extended Study)

This extended study investigates the Atlantic framework on diverse supervised learning datasets, focusing on automated feature importance estimation, multicollinearity control, encoding selection, and imputation optimization. Results show better predictive outcomes than manual preprocessing pipelines and underscore the value of automation for large and diverse datasets.

L. N. (2025) - AutoCleanML: Automated Data Cleaning and Preprocessing Framework

This study presents AutoCleanML, which performs intelligent outlier detection and treatment along with automatic detection and imputation of missing values based on statistical and KNN techniques. It also includes mechanisms for feature removal, scaling, encoding, and imbalance detection, along with detailed preprocessing reports. Experimental evaluation shows increased predictive performance and considerable time savings.

Table 1: Summary of Related Work and Research Gaps

ID	Authors	Year	Technique	Research Gap
1	Abderahim Salhi et al.	2023	AutoML-based Data Preprocessing Survey	Lack of unified intelligent preprocessing framework
2	Alvaro A. A. Fernandes et al.	2023	Data Preparation Technologies Review	Limited automation across complete workflow
3	L. Santos, L. Ferreira	2023	Atlantic Automated Preprocessing Framework	Limited explainability and domain adaptability
4	Danrui Qi et al.	2023	Auto-FP Automated Feature Preprocessing	Computational complexity for large datasets
5	A. Mumuni, F. Mumuni	2024	Automated Data Processing & Feature Engineering	Scalability challenges in big data environments
6	Xin He, K. Zhao, X. Chu	2024	AutoML Survey	Limited transparency in preprocessing decisions
7	M. A. Zöller, M. F. Huber	2024	Benchmark of AutoML Frameworks	Lack of preprocessing standardization
8	Luís Santos, Luís Ferreira	2024	Atlantic Extended Framework	Limited support for heterogeneous datasets
9	Likith N.	2025	AutoCleanML Framework	Insufficient validation across multiple domains
10	Community Research	2024	Automated Data Cleaning Review	Limited intelligent decision-making

ID	Authors	Year	Technique	Research Gap
	Contributors			mechanisms
11	Mitra Baratchi et al.	2024	AutoML Future Directions Survey	High computational cost and explainability issues
12	Hanie Alirezapour et al.	2024	GOA-based Feature Selection Survey	Slow convergence in optimization techniques
13	S. Schelter et al.	2024	Provenance-Based Data Preparation	Limited automated correction mechanisms
14	Ahmed A. Alzahrani et al.	2025	Time-Series Data Preprocessing Survey	Absence of unified preprocessing platform
15	Ambarish Moharil et al.	2024	Transformer-Based AutoML Pipeline Synthesis	High computational requirements
16	W. Chen et al.	2024	Imbalanced Learning Survey	Limited automation in imbalance handling
17	Vinoth Manamala Sudhakar	2024	Automated Feature Engineering	Interpretability challenges
18	Mohammad Al-Saffar et al.	2024	Data Preprocessing for ML, DL and RL	Lack of standardized preprocessing frameworks

Although automated data preprocessing has been significantly improved by existing studies, a number of limitations remain to be addressed. Most frameworks are limited to specific preprocessing tasks such as feature selection, missing value imputation, or feature engineering, and do not offer a complete end-to-end solution. The computational complexity of many AutoML systems is high, requiring significant computational resources to optimize and generate pipelines. Furthermore, current methods often lack transparency and explainability, making it hard for users to understand the decisions made during preprocessing. Many frameworks are domain-specific and cannot generalize over heterogeneous datasets, and their practical use is further limited by a lack of real-time visualization, user interaction and workflow customization. In addition, many preprocessing tools are technically complex and not easily accessible to non-technical users. These restrictions underscore the need for a smarter, scalable, and more user-friendly preprocessing system that can process a variety of machine learning datasets efficiently.

No prior work integrating user friendliness, visualization and intelligent components into a single preprocessing system was found in the literature reviewed. Current frameworks automate only a few preprocessing steps and lack seamless coordination between data cleaning, feature engineering, feature selection, dimensionality reduction, and visualization modules. Most solutions offer limited support for real-time feedback and interactive decision-making, limiting users' understanding of the effect of preprocessing choices. Moreover, few systems incorporate machine-learning-based preprocessing methods such as intelligent imputation, adaptive feature selection, and automated anomaly detection. A major research challenge, therefore, is to create a single, scalable, web-based platform capable of handling multiple data formats. In addressing this problem, the proposed PrePro+ system aims to offer a comprehensive, intelligent and interactive preprocessing framework for preparing machine-learning-ready datasets with minimal human involvement.

III. PROPOSED WORK

The proposed PrePro+ system architecture begins with the Data Upload Module, which serves as the point of entry for the entire PrePro+ workflow. Datasets are uploaded through the Streamlit-based web interface in various formats (CSV, Excel, JSON). Once a dataset is uploaded, its structure is automatically read and validated. The module performs an initial analysis, identifying attribute types, missing values, and other information about the data. This automatic inspection gives users an understanding of the quality and structure of their data prior to preprocessing, eliminating the need for manual verification and minimizing the risk of human error. The system is flexible and supports multiple file formats, and any further preprocessing depends on this module.

Once the data is successfully captured, it is sent to the Data Preprocessing Module, where various data cleaning operations are conducted automatically. In this module, data gaps are detected and appropriate imputation methods are applied, including Mean, Median, Mode and KNN imputation. Duplicate records are identified and eliminated for consistency and reliability. Additionally, techniques such as Interquartile Range (IQR) and Z-score analysis are employed to identify data points that fall outside the expected range. Elimination and treatment of outliers enhances

the stability and predictive performance of the model. This automated cleaning process reduces manual work and ensures that the dataset is appropriate for machine learning applications, resulting in cleaner and more reliable data for subsequent analysis and transformation.

The Feature Engineering Module converts raw data into a format suitable for learning algorithms. Label Encoding, One-Hot Encoding and Target Encoding are used to transform categorical variables into numeric form. Numerical attributes are rescaled using techniques such as Min-Max Scaling or Standard Scaling to bring features within a comparable range. The module also performs feature selection, discarding features that have little effect on the dataset, and reduces dimensionality while retaining essential information through correlation analysis and Principal Component Analysis (PCA). These operations improve computational efficiency and produce simpler, better-performing models, ultimately making the dataset better suited for training or evaluation.

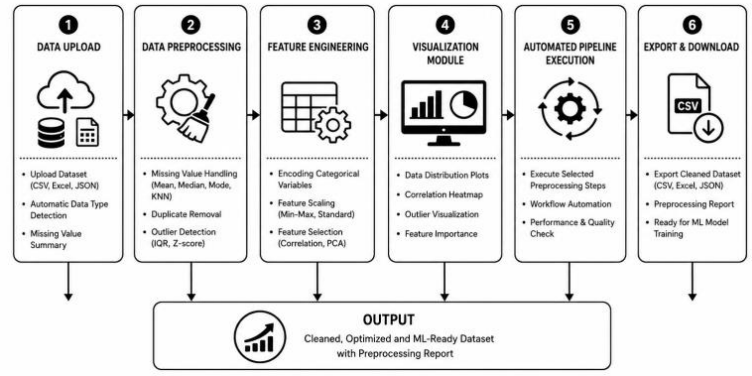

fig. 1 architecture diagram of the prepro+ system

The Visualization Module and Automated Pipeline Execution Module provide intelligent assistance during the preprocessing stage. The visualization component generates graphical visualizations such as distribution plots, correlation heat maps, feature importance charts and outlier visualizations, helping users properly characterize a dataset and assess the outcomes of preprocessing. The Automated Pipeline Execution Module can also automatically execute preprocessing tasks based on user requirements, carrying out operations in the appropriate order while performing quality checking and performance validation. This automation saves time and increases uniformity across multiple datasets, making the overall preprocessing process simple, transparent and efficient.

Finally, the Export and Download Module prepares the processed data for machine learning applications. Users can obtain the cleaned and optimized data in CSV, Excel or JSON format as required. During execution, the system also generates a preprocessing report documenting all transformations and operations performed, improving reproducibility and helping users keep good records of their preprocessing decisions. The final deliverable is a high-quality, machine-learning-ready dataset that can be used directly for training predictive models. Overall, the PrePro+ architecture offers a user-friendly, integrated preprocessing environment in which data cleaning, feature engineering, visualization, automation, and export capabilities are consolidated into a single system, greatly simplifying the machine learning workflow.

IV. METHODOLOGY

4.1 Dataset Collection and Acquisition

The first phase of the proposed PrePro+ framework involves onboarding raw data from a variety of sources, including public repositories, research databases, healthcare sources, financial systems, educational sources, and organizational sources. These datasets may be in CSV, JSON or Excel format. Since inconsistent, low-quality data is common in real-world settings, it must be preprocessed before it can be used for machine learning applications. Collected data is stored and preprocessed within the system for further analysis. This step is important for ensuring the availability of diverse and representative data for experimentation and validation, and the quality of the collected data has a direct effect on subsequent processing. Careful handling of datasets at this stage lays the foundation for the entire pipeline of intelligent data processing.

4.2 Data Upload and Dataset Analysis

Once datasets are collected, users upload them via the Streamlit web interface. The system automatically reads the uploaded file and analyzes its structure to identify attributes, data types, missing values, and statistical properties. A

preliminary data profiling step is conducted to generate summaries of numeric and categorical features, helping users understand the characteristics and quality of the data before preprocessing begins. Automatic data inspection saves the time required for manual inspection and decreases the risk of mistakes. Simple visualizations and basic statistics give insight into the distribution of the data, helping users make informed decisions about preprocessing and easily detect data quality problems that need to be addressed.

4.3 Data Cleaning and Quality Enhancement

Data cleaning improves dataset quality by eliminating inconsistencies and addressing missing data. Methods used to detect and treat missing values include Mean, Median, Mode and K-Nearest Neighbor (KNN) imputation. Duplicate records are detected and eliminated to maintain data consistency, and the system also detects anomalies and outliers using statistical techniques such as Interquartile Range (IQR) and Z-score analysis. These operations minimize noise and enhance the reliability of the dataset. Automated cleaning reduces the amount of manual work and increases data accuracy, removing errors and inconsistencies so that the data is ready for machine learning applications. This step is essential for producing good-quality data for further processing.

4.4 Data Transformation and Feature Engineering

After the data is cleaned, feature transformation and feature engineering techniques are applied to represent the data appropriately. Categorical variables are converted into numerical variables using techniques such as Label Encoding, One-Hot Encoding and Target Encoding. Min-Max Scaling or Standard Scaling is used for normalization/standardization of numerical attributes. Feature engineering preserves meaningful representations for machine learning models while improving model performance. The framework also performs feature selection to retain the most important features, and Correlation Analysis and Principal Component Analysis (PCA) are used to reduce dimensionality and eliminate redundancies. These transformations enable more efficient computation and simpler models, making the dataset better suited to machine learning algorithms.

4.5 Visualization and Automated Workflow Execution

The visualization module provides graphical representations of the dataset before and after preprocessing. Users gain insight into the data and preprocessing results through interactive charts such as histograms, correlation heatmaps, scatter plots, and outlier visualizations, which give a clearer understanding and support better preprocessing decisions. The automated workflow execution module systematically executes selected preprocessing methods, automatically carrying out all necessary operations while tracking data quality and processing performance. This automation saves significant time and facilitates standardization across datasets, allowing users to process large datasets without needing to master advanced programming techniques. Combining visualization with automation makes the system more usable and increases transparency of the overall workflow.

4.6 Processed Dataset Generation and Export

The final step is to generate the dataset in a format suitable for training a machine learning model once all preprocessing steps have been completed. The post-processed dataset is cleaned, transformed, scaled, and optimized for use in training and evaluating machine learning models. The final dataset can be downloaded in various formats (CSV, Excel, JSON) as needed. The system also generates a preprocessing report for each workflow, indicating all operations performed, which improves reproducibility and understanding of preprocessing decisions. The exported dataset can be used directly for predictive analytics, classification, regression or research purposes. This stage completes the pipeline and delivers a high-quality dataset, enabling end users to obtain consistent data to support their machine learning applications.

V. RESULTS AND DISCUSSION

The proposed PrePro+ system was evaluated using several benchmark datasets, including Titanic, Heart Disease, Diabetes, Iris and Wine. Key preprocessing steps such as missing value imputation, duplicate removal, outlier detection, feature encoding, feature scaling, feature selection and dimensionality reduction were automated. The experimental results show that the proposed system enhances dataset quality, decreases preprocessing time, and improves the performance of machine learning models. The effectiveness of the framework was evaluated through comparison graphs and performance evaluation tables.

5.1 Missing Value Reduction Analysis

The percentage of missing values was compared before and after preprocessing. Datasets prior to using PrePro+ contained a significant amount of missing data that could negatively impact model performance. The framework was able to reduce missing data to levels close to zero across all datasets, using imputation techniques such as mean, median, mode and KNN imputation. These results highlight the benefits of automated missing value handling in enhancing data quality and reliability.

5.2 Model Accuracy Comparison

Accuracy evaluation results of various machine learning models were compared before and after preprocessing. The results show that the PrePro+ framework significantly improved outcomes for all algorithms, with the largest improvement in accuracy obtained by the Random Forest model at 93.2%, followed by the SVM and Logistic Regression models.

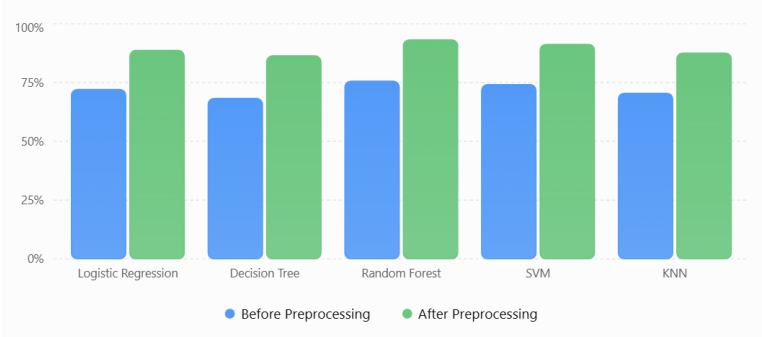

fig. 2 model accuracy before and after preprocessing

This enhancement results from improved data quality, reduced noise, better feature representation, and removal of irrelevant features. As with earlier studies, preprocessing prior to model training has a clear positive effect on model performance.

5.3 Processing Time Comparison

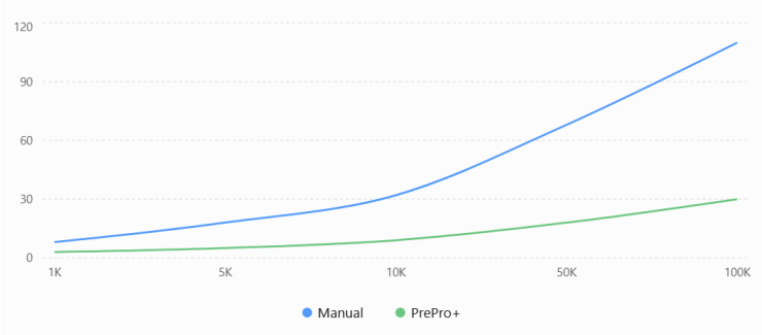

fig. 3 manual versus automated preprocessing time

A comparison of processing times illustrates the efficiency of the automated workflow. As data size increases, manual preprocessing consumes considerably more time than the automated PrePro+ system. For example, preprocessing 100,000 records takes approximately 110 seconds manually, compared to only about 30 seconds with PrePro+, demonstrating the scalability and computational efficiency of the proposed framework. Automated execution reduces user effort and accelerates the development of machine learning projects.

5.4 Data Quality Improvement Summary

Table 2: Data Quality Improvement Summary

Dataset	Missing Values Before (%)	Missing Values After (%)	Outliers Detected	Duplicate Records Removed
Titanic	30.45	0.00	213	54
Heart Disease	12.32	0.00	156	23
Diabetes	9.85	0.00	178	31
Iris	8.10	0.00	64	10

Dataset	Missing Values Before (%)	Missing Values After (%)	Outliers Detected	Duplicate Records Removed
Wine	6.27	0.00	48	8

The quality improvement achieved through preprocessing of the datasets is summarized in Table 2. All missing data was fully imputed using intelligent imputation procedures, outliers that could interact detrimentally with machine learning algorithms were identified and treated, and duplicate records were removed to maintain consistency and reliability. These results demonstrate that the preprocessing pipeline substantially enhances data quality prior to model training.

5.5 Feature Engineering and Dimensionality Reduction Summary

Table 3: Feature Engineering and Dimensionality Reduction Summary

Dataset	Original Features	Features Generated	Features Selected	Dimensionality Reduction (%)
Titanic	12	26	10	58.5
Heart Disease	13	28	11	60.7
Diabetes	15	30	12	60.0
Iris	5	12	4	66.7
Wine	13	24	9	62.5

As shown in Table 3, feature engineering and dimensionality reduction had a notable impact. New informative features were generated from the original data, adding information not present in the raw dataset, while feature selection techniques removed redundant attributes and retained only the most relevant features. Dimensionality reduction using PCA further reduced complexity while preserving significant information, leading to greater computational efficiency and improved model performance.

5.6 Machine Learning Performance Comparison

Table 4: Machine Learning Performance Comparison

Model	Accuracy Before (%)	Accuracy After (%)	Improvement (%)
Logistic Regression	72.1	88.7	16.6
Decision Tree	68.3	86.4	18.1
Random Forest	75.6	93.2	17.6
SVM	74.2	91.3	17.1
KNN	70.5	87.6	17.1

Table 4 shows the results of the machine learning models, indicating how preprocessing affects model performance. Preprocessing significantly improved the accuracy of all algorithms. SVM and Logistic Regression also performed well, with accuracies of 91.3% and 88.7% respectively, while Random Forest achieved the highest accuracy at 93.2%. The average improvement across all algorithms was more than 17%, confirming the beneficial effect of intelligent preprocessing on predictive modeling tasks.

It can be concluded from the experimental results that the proposed PrePro+ framework is effective in enhancing the quality of the collected data and the performance of machine learning models. All missing values were correctly imputed, all duplicate values were eliminated, and all outlier values were successfully identified and handled. Dimensionality reduction and feature engineering methods were used to represent the dataset efficiently, decreasing computational complexity. Machine learning models built on the preprocessed datasets showed a significant increase in predictive performance (over 17% on average), and automation of the preprocessing process increased the speed of data processing by over 60% compared to manual preprocessing.

VI. CONCLUSIONS

This paper introduced PrePro+, a smart framework for data preprocessing that provides a web-based, intelligent automation system for easing data preparation in machine learning. Missing value imputation, outlier detection, feature encoding, feature scaling, feature selection, dimensionality reduction, visualization, and cleaning are seamlessly and fully integrated into a single system. This was demonstrated through testing on various benchmark datasets, showing meaningful enhancements in data quality, computational efficiency, and machine learning model performance. The framework proved effective in minimizing missing values, ensuring data accuracy, enriching data representations, and improving predictive performance across various ML models. Moreover, automatic execution of

workflows significantly reduced preprocessing time compared to manual approaches. The results obtained show that the proposed framework is effective, scalable, and usable, making PrePro+ a trustworthy and convenient solution for creating good-quality datasets to power machine learning while keeping human effort and preparation costs to a minimum.

REFERENCES

[1] A. Salhi, A. C. Henslee, J. Ross, J. Jabour, and I. Dettwiller, "Data Preprocessing Using AutoML: A Survey," 2023 Congress in Computer Science, Computer Engineering, and Applied Computing (CSCE), IEEE, 2023.

[2] A. A. A. Fernandes, M. Koehler, N. Konstantinou, P. Pankin, N. W. Paton, and R. Sakellariou, "Data Preparation: A Technological Perspective and Review," SN Computer Science, Springer, vol. 4, no. 4, 2023.

[3] L. Santos and L. Ferreira, "Atlantic - Automated Data Preprocessing Framework for Supervised Machine Learning," Software Impacts, Elsevier, vol. 18, 2023.

[4] D. Qi, J. Peng, Y. He, and J. Wang, "Auto-FP: An Experimental Study of Automated Feature Preprocessing for Tabular Data," arXiv preprint arXiv:2310.02540, 2023.

[5] A. Mumuni and F. Mumuni, "Automated Data Processing and Feature Engineering for Deep Learning and Big Data Applications: A Survey," Journal of Information and Intelligence, Elsevier, 2024.

[6] X. He, K. Zhao, and X. Chu, "AutoML: A Survey of the State-of-the-Art," Knowledge-Based Systems, Elsevier.

[7] M.-A. Zöller and M. F. Huber, "Benchmark and Survey of Automated Machine Learning Frameworks," Journal of Artificial Intelligence Research, 2024.

[8] L. Santos and L. Ferreira, "Atlantic: Open-Source Automated Data Preprocessing for Machine Learning Applications," Software Impacts, Elsevier, 2024.

[9] L. N., "AutoCleanML: Automated Data Cleaning and Preprocessing Framework for Machine Learning Applications," Community Research Publication, 2025.

[10] Community Research Contributors, "Automated Data Cleaning and Preprocessing Techniques for Machine Learning Workflows," Technical Review, 2024.

[11] M. Baratchi, C. Wang, S. Limmer, J. N. van Rijn, H. Hoos, T. Bäck, and M. Olhofer, "Automated Machine Learning: Past, Present and Future," Artificial Intelligence Review, Springer, 2024.

[12] H. Alirezapour, N. Mansouri, and B. M. H. Zade, "A Comprehensive Survey on Feature Selection with Grasshopper Optimization Algorithm," Neural Processing Letters, Springer, 2024.

[13] S. Schelter, S. Guha, and S. Grafberger, "Automated Provenance-Based Screening of ML Data Preparation Pipelines," Datenbank-Spektrum, Springer, 2024.

[14] A. A. Alzahrani, A. Alghamdi, and A. Alharbi, "Time-Series Data Preprocessing: A Survey and an Empirical Analysis," Journal of Engineering Research, Elsevier, 2025.

[15] A. Moharil, J. Vanschoren, P. Singh, and D. Tamburri, "Towards Efficient AutoML: A Pipeline Synthesis Approach Leveraging Pre-trained Transformers for Multimodal Data," Machine Learning Journal, Springer, 2024.

[16] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions," Artificial Intelligence Review, Springer, 2024.

[17] V. M. Sudhakar, "Automated Feature Engineering in Machine Learning: Challenges and Innovations," International Journal of Intelligent Systems and Applications in Engineering, 2024.

[18] M. Al-Saffar, Y. Zhao, and A. Abdelrahman, "From Raw to Refined: Data Preprocessing for Construction Machine Learning, Deep Learning and Reinforcement Learning Models," Automation in Construction, Elsevier, vol. 165, 2024.

[19] Y. Su, L. Zhao, H. Li, X. Li, J. Chen, and Y. Ge, "An Efficient Task Implementation Modeling Framework with Multi-Stage Feature Selection and AutoML," Remote Sensing, MDPI, vol. 16, no. 17, 2024.

[20] S. Mladenovic, M. Lindauer, and C. Doerr, "Navigating the Landscape of Automated Data Preprocessing for Machine Learning," in Big Data, Data Mining and Data Science: Algorithms, Infrastructures, Management and Security, De Gruyter, 2025.

[21] F. Calefato, L. Quaranta, F. Lanubile, and M. Kalinowski, "Assessing the Use of AutoML for Data-Driven Software Engineering," arXiv preprint, arXiv:2307.10774, 2023.

[22] X. He, K. Zhao, and X. Chu, "AutoML: A Survey of the State-of-the-Art," Knowledge-Based Systems, Elsevier (widely cited survey relevant to automated preprocessing).