International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 7, July 2025
DOI 10.17148/IMRJR.2025.020706

Data Preparation Strategies for Improving the Performance of Machine Learning Models in Heart Disease Prediction

MERLIN SOFIA S[1], Dr. D. RAVINDRAN[2], DR. G. AROCKIA SAHAYA SHEELA[3]

Research Scholar, Department of Computer Science, St.Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India[1]

Associate Professor, Department of Computer Science, St.Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India[2]

Assistant Professor, Department of Computer Science, St.Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India[3]

Abstract: Heart disease represents a significant health challenge worldwide, requiring efficient classification and prediction approaches for prompt diagnosis and intervention. Reliable prediction models depend on high-quality data, which calls for thorough preprocessing, particularly in areas such as imputation, outlier elimination, and feature selection to improve dataset quality and model effectiveness. Current preprocessing methods, which include managing missing data, identifying outliers, and selecting features, frequently show limitations and biases when dealing with complex datasets. In order to address these issues, this paper introduces the Data Preprocessing algorithm. This algorithm employs ensemble-based approaches to thoroughly tackle these challenges, improving dataset quality and elevating feature significance. The algorithm incorporates data merging techniques to unify various datasets and maintain uniformity. It makes use of the EMERALD algorithm for effective imputation of missing data and the RACE algorithm for successful outlier elimination. Normalization methods like Min-Max scaling are applied to standardize the data, while the DYNAMIC algorithm identifies key features essential for predictive performance.

Keywords: Data preparation, Missing data handling, Outlier elimination, Variable selection, Ensemble methods.

I. INTRODUCTION

Cardiovascular disease is a leading cause of death worldwide, highlighting the urgent requirement for precise classification and prediction techniques to reduce its effects [1]. This range of heart-related issues demands efficient management approaches and early identification to significantly enhance patient outcomes. Machine learning methods have become essential resources for tackling these issues, utilizing datasets to create predictive models that support clinicians in their decision-making activities. Data preprocessing is vital in this scenario as it readies raw data for machine learning applications. This process includes key steps such as cleaning data, normalizing values, addressing missing data, identifying outliers, and selecting features, all intended to improve the quality and dependability of the data for later analysis. Even though they are fundamental, current preprocessing methods frequently have drawbacks that can affect the precision and potency of predictive models. Imputation techniques like regression imputation, mean substitution, and median substitution can introduce biases or inaccuracies, especially in complex datasets with missing values [4]. Similarly, noisy datasets or non-standard data distributions may be difficult for conventional outlier removal methods like the Interquartile Range (IQR) to handle [5]. Synergistic relationships among variables that are essential for predictive accuracy may be overlooked by feature selection methods that rely on statistical metrics or heuristic approaches [6]. The Data Preprocessing algorithm is presented as a solution to these problems. By incorporating sophisticated ensemble techniques created especially for heart disease datasets, this algorithm is intended to fully address these limitations. It seeks to improve data quality through efficient outlier identification and removal using ensemble-based approaches, optimize feature selection by utilizing a variety of criteria to find the most pertinent features for precise prediction models, and improve data integrity and completeness through robust imputation techniques.

The Data Preprocessing algorithm uses a number of cutting-edge methods:
For reliable missing data imputation, use the Ensemble Method for Regression-based Assessment of Lost Data using Weighted Prediction (EMERALD) algorithm. RACE Algorithm used to Clustering Ensemble and robust statistics for precise outlier identification and elimination. For the best feature subset selection, use the Dynamic Algorithm: Diverse Integrated Multi-criteria Ensemble Feature Selection.

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 7, July 2025
DOI 10.17148/IMRJR.2025.020706

By putting forth The Data Preprocessing Algorithms as an extensive preprocessing framework designed especially for heart disease datasets, this paper advances the field. It seeks to show better results than current approaches, offering a strong basis for improving the precision and dependability of predictive models in clinical decision support systems and epidemiological research centered on the diagnosis and treatment of heart disease. This paper is structured as follows: In order to identify current constraints and difficulties, Section 2 examines preprocessing techniques used in heart disease research. With techniques like EMERALD for robust imputation, RACE for outlier detection, and DYNAMIC for feature selection, Section 3 describes the Data Preprocessing Algorithm in detail. Experimental results showing The Data Preprocessing algorithm effectiveness in enhancing data quality and predictive accuracy are presented in Section 4, where its performance is compared to that of conventional techniques. Section 5 wraps up by highlighting important discoveries, talking about the implications for further study, and suggesting possible avenues to improve preprocessing in the analysis of data related to heart disease. The Data Preprocessing Algorithms, which seek to maximize data quality and improve predictive accuracy in clinical settings, constitute a substantial advancement in preprocessing methodologies for heart disease datasets overall.

II. RELATED WORKS

The importance of preprocessing heart disease datasets has received a lot of attention lately, mostly because of the difficulties and complexities that come with using Electronic Health Records (EHRs) [11]. EHRs, a key part of Medical Big Data (MBD), offer tremendous potential for precision medicine advancements, according to Sorkhabi et al. [7]. However, because of their intrinsically "dirty" nature, these records are frequently inappropriate for direct mining. In order to overcome this, the authors suggest PEPMED, a methodical pre-processing technique that greatly enhances data utility and accuracy using hybrid approaches designed to address the unique difficulties presented by EHRs.

Zemlianyi and Baibuz[8] investigate different imputation techniques designed for coronary heart disease data in order to address missing data, a prevalent problem in heart disease datasets. They contend that it is excessive and leads to the loss of important information to remove observations with missing values, a common practice in libraries like NumPy and Pandas. Rather, they suggest a number of imputation techniques, such as deep learning-based approaches, mean/median imputation, k-nearest neighbors (kNN), and multiple imputation by chained equations (MICE). According to their analysis, techniques such as fillna_2steps_rg provide time savings and increased classification accuracy, indicating that customized imputation techniques can greatly improve the results of data analysis.

Similar to this, Hu and Du[9] offer a novel framework for managing missing data in EHRs that makes use of constrained support vector machines (cSVM) and the Gaussian Process Latent Variable Model (GPLVM). Their approach produces a reliable tool for forecasting hospital readmissions among patients with heart failure by accurately imputed missing values and incorporating uncertainty into the predictions. When compared to conventional techniques, their results show better performance in terms of both imputation accuracy and predictive capabilities.

Berkelmans et al. [10] compare the efficacy of different imputation methods in clinical settings. According to their findings, as long as crucial predictor variables are present, simpler techniques like median imputation outperform more intricate ones. This implies that simple imputation techniques can be just as successful in real-world scenarios as intricate algorithms, despite being simpler to use and comprehend.

The effect of feature selection on the imputation of missing values in medical datasets is examined by Liu et al. [21]. They show that pattern recognition and predictive accuracy are enhanced when feature selection and imputation are combined. Selecting the right feature selection algorithm is essential, though, with decision tree models working better with higher-dimensional datasets and genetic algorithms and information gain models working better with lower-dimensional ones.

In the field of anomaly detection, Nanehkaran et al. [12] suggest an unsupervised, density-based method that uses DBSCAN clustering to find outliers in data related to heart disease. Their approach improves on conventional clustering techniques that suffer from noise sensitivity and initial point selection by adjusting clustering parameters to achieve high accuracy in diagnosing cardiac abnormalities.

Numerous studies have focused on feature selection, a crucial preprocessing step. Using a floating window with adaptive size, Javeed et al. [13] present a novel feature selection technique that greatly improves the accuracy of heart disease prediction when paired with neural network classifiers. This approach shows its efficacy in improving feature sets for improved model performance, outperforming many other approaches currently in use.

A thorough assessment of feature selection methods for heart disease prediction was carried out by Dissanayake and Md. Johar [14]. They discover that the best accuracy and precision are obtained when backward feature selection is used

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 7, July 2025
DOI 10.17148/IMRJR.2025.020706

in conjunction with decision tree classifiers. This emphasizes how crucial it is to choose the right feature subsets for enhanced predictive performance.

Using methods such as principal component analysis, chi-squared testing, and relief, Spencer et al. [15] further investigate the interaction between feature selection and classification techniques. Their results highlight how performance varies based on the combination of machine learning algorithms and feature selection, with Chi-squared feature selection and BayesNet yielding the best outcomes.

In addition to highlighting the significance of early heart disease detection, Kadhim and Radhi [16] offer a model that makes use of machine-learning algorithms that have been optimized. Their method, which consists of processing patient data, training models, and optimizing hyperparameters, achieves high accuracy, especially when using the Random Forest algorithm. This highlights how machine learning can improve clinical decision-making.

Despite these developments, there are still a number of drawbacks to current methods, including biases introduced during imputation, poor handling of complex data distributions in outlier detection, and traditional feature selection methods' inadequate consideration of synergistic feature relationships. By incorporating cutting-edge ensemble-based methods designed especially for heart disease datasets, the suggested Data Preprocessing algorithm seeks to address these drawbacks. The Data Preprocessing Algorithm enables more dependable clinical decision support systems and epidemiological studies centered on the prognosis and management of heart disease by utilizing techniques like EMERALD for robust imputation, RACE for precise outlier identification, and DYNAMIC for optimized feature selection.

III. METHODOLOGY

3.1 Data Preprocessing algorithm

A thorough preprocessing framework called the Data Preprocessing Algorithm was created to combine and improve several datasets related to heart disease. The algorithm guarantees that the data is consistent, clean, and prepared for sophisticated machine-learning applications[22],[23].

The five heart disease data sets—"cleveland.csv," "hungarian.csv," "longbeach.csv," "switzerland.csv," and "statlog.csv"—are loaded and combined at the beginning of the Data Preprocessing algorithm. First, the five data sets will be read into memory and combined into a single data set. It is crucial to make sure that every header entry is consistent and to eliminate any duplicate entries that might arise from data points that overlap between data sets.

The next issue, which arises frequently in medical datasets, is missing data after the datasets have been merged. Therefore, the EMERALD method (Ensemble Method for Regression-based Assessment of Lost Data Using Weighted Prediction) is used by the Data Preprocessing algorithm to fill in the missing data. EMERALD begins by separating the data into two categories: complete (also known as "train data") and missing (also known as "test data"). Three regression models (M5P Tree, Gaussian Processes, and Linear Regression) are trained on a subset of the entire dataset for each feature with missing values. The Mean Squared Error (MSE) is used to evaluate each model's performance, and inverse MSEs are used to compute the weighted prediction. The missing values are filled in using the weighted ensemble prediction, which is more dependable than relying just on one model.

The RACE (Robust Anomaly Detection and Cluster-based Ensemble) algorithm is used in the Data Preprocessing algorithm to detect and remove outliers after handling missing values. RACE determines the enhanced Z-scores after first determining the medians and Median Absolute Deviations (MADs) for every feature. The 95th percentile of these z-scores is then used to identify the anomalies. To remove the outliers, RACE employs an enhanced k-means clustering technique that uses the Manhattan distance to determine the distance between each data point and the centroid of clusters. Data points are identified as anomalies if they appear more distant from the centroids of their respective clusters than a predetermined threshold. RACE eliminates outliers from the dataset by combining statistical and clustering-based approaches.

Normalisation is a preprocessing step in the machine learning pipeline [7] [9]. All feature values are subjected to Min-Max Normalization using the Data Preprocessing algorithm, which ensures that every feature value falls within the range of 0 to 1. As a result, we promise that during the training of the machine learning model, features with larger scales won't overshadow others. DYNAMIC (Different Integrated Multi-criteria Ensemble Feature Selection): We use DYNAMIC to choose the most relevant features from the data in the last stage of preprocessing. To produce a strong feature set, DYNAMIC combines several feature selection methods.

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 7, July 2025
DOI 10.17148/IMRJR.2025.020706

Classifier ranking, target correlation, and classifier subset results are used to evaluate features. Only the most important and predictive features are retained as a result. It maximizes model clarity with findings, minimizes overfitting, and shrinks dimensionality. After all preprocessing steps are finished, the Data Preprocessing Algorithm receives a refined dataset with the chosen features. This new dataset can be used for further machine learning processes.

Data preprocessing fixes issues with data quality by removing noise, outliers, and missing values, enabling reliable and accurate heart disease prediction models [17] [20].

IV. CONCLUSION

Key issues in getting heart disease datasets ready for predictive modeling are successfully addressed by the suggested data preprocessing algorithm. The reliability of data sources is strengthened by removing redundancy and guaranteeing consistency through the integration of multiple datasets. By reducing bias and improving data completeness, the EMERALD algorithm offers reliable imputation of missing values.The RACE algorithm employs a hybrid statistical and clustering approach to precisely identify and remove outliers.By standardizing feature ranges through Min-Max scaling, normalization guarantees balanced contributions throughout model training. By reducing dimensionality while maintaining predictive power, the DYNAMIC feature selection method finds the most informative variables. This ensemble-based framework performs better than traditional preprocessing methods, which are frequently biased and ineffective. In clinical applications, the improved dataset increases predictive accuracy, decreases overfitting, and improves model interpretability. A methodical approach like this offers a strong basis for accurate machine learning-based heart disease diagnosis and prognosis. All things considered, the Data Preprocessing Algorithm is a major development in the processing of medical data that promotes improved clinical judgment and research results.

REFERENCES

[1]. Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. Mobile Information Systems, 2022(1), 1410169.https://doi.org/10.1155/2022/1410169
[2]. Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health and Technology, 10(5), 1137-1144. https://doi.org/10.1007/s12553-020-00438-1
[3]. Benhar, H., Idri, A., & Fernández-Alemán, J. L. (2020). Data preprocessing for heart disease classification: A systematic literature review. Computer Methods and Programs in Biomedicine, 195, 105635.https://doi.org/10.1016/j.cmpb.2020.105635
[4]. Hu, Z., & Du, D. (2020). A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction. PloS one, 15(9), e0237724.https://doi.org/10.1371/journal.pone.0237724
[5]. Ardeti, V. A., Kolluru, V. R., Varghese, G. T., & Patjoshi, R. K. (2022). An Outlier Detection and Feature Ranking based Ensemble Learning for ECG Analysis. Int. J. Adv. Comput. Sci. Appl, 13(6).https://thesai.org/Publications/ViewPaper?Volume=13&Issue=6&Code=IJACSA&SerialNo=86
[6]. Ay, Ş., Ekinci, E., & Garip, Z. (2023). A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. The Journal of Supercomputing, 79(11), 11797-11826.https://doi.org/10.1007/s11227-023-05132-3
[7]. A.AngelPreethi, and Dr.S.Britto Ramesh Kumar, "Dom_Classi: An Enhanced Weighting Mechanism for Domain Specific Words using Frequency based Probability", International Journal of Applied Engineering Research, Vol.14, Issue 1, 2019, pp 140-148
[8]. Zemlianyi, O., & Baibuz, O. (2024). Methods for imputing missing data on coronary heart disease. System technologies, 2(151), 33-49 https://doi.org/10.34185/1562-9945-2-151-2024-04
[9]. S. Parimala, R. Kumutha, F. Iram, M. V. Sunena Rose, N. R., and A. Angelpreethi, "Improved Moth Flame Optimization with Deep Convolutional Neural Network for Colorectal Cancer Classification using Biomedical Images," in *Proc. [Conference Name] *, Jul. 2024, pp. 1–6, https://doi.org/10.1109/icait61638.2024.10690631
[10]. Berkelmans, G. F., Read, S. H., Gudbjörnsdottir, S., Wild, S. H., Franzen, S., Van Der Graaf, Y., ... & Dorresteijn, J. A. (2022). Population median imputation was non-inferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice. Journal of Clinical Epidemiology, 145, 70-80.https://doi.org/10.1016/j.jclinepi.2022.01.011
[11]. A.AngelPreethi, and Dr.S.Britto Ramesh Kumar, "NIC_LBA: Negations and Intensifier Classification of microblog data using Lexicon Based Approach "Journal of Emerging Technologies and Innovative Research (JETIR), Volume 6, Issue 6 , PP 753-759, June 2019.
[12]. Nanehkaran, Y. A., Licai, Z., Chen, J., Jamel, A. A., Shengnan, Z., Navaei, Y. D., & Aghbolagh, M. A. (2022). Anomaly Detection in Heart Disease Using a Density-Based Unsupervised Approach. Wireless Communications and Mobile Computing, 2022(1), 6913043.https://doi.org/10.1155/2022/6913043

International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 2, Issue 7, July 2025
DOI 10.17148/IMRJR.2025.020706

[13]. Javeed, A., Rizvi, S. S., Zhou, S., Riaz, R., Khan, S. U., & Kwon, S. J. (2020). Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. Mobile Information Systems, 2020(1), 8843115.https://doi.org/10.1155/2020/8843115

[14]. Dissanayake, K., & Md Johar, M. G. (2021). Comparative study on heart disease prediction using feature selection techniques on classification algorithms. Applied Computational Intelligence and Soft Computing, 2021(1), 5581806.https://doi.org/10.1155/2021/5581806

[15]. Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. Digital health, 6, 2055207620914777.https://doi.org/10.1177/2055207620914777

[16]. Kadhim, M. A., & Radhi, A. M. (2023). Heart disease classification using optimized Machine learning algorithms. Iraqi Journal For Computer Science and Mathematics, 4(2), 31-42.https://doi.org/10.52866/ijcsm.2023.02.02.004

[17]. A.AngelPreethi, S.B.R. Kumar, "Visualizing Big Data Mining: Issues, Challenges and Opportunities" International Journal of Control Theory and Applications, Volume 9, Issue 27, Pages: 455-460, 2016.

[18]. A.AngelPreethi, and Dr.S.Britto Ramesh Kumar, "A Dictionary based Approach for improving the accuracy of opinion mining on big data", International Journal of Research and Analytical Reviews (IJRAR), Vol. 5, Issue 4, Oct-Dec 2018, pp- i836-844.

[19]. Sorkhabi, L. B., Gharehchopogh, F. S., & Shahamfar, J. (2020). A systematic approach for pre-processing electronic health records for mining: A case study of heart disease. International Journal of Data Mining and Bioinformatics, 24(2), 97 120.https://www.researchgate.net/publication/345747207_A_systematic_approach_for_pre-processing_electronic_health_records_for_mining_case_study_of_heart_disease

[20]. Angelpreethi, " Fuzzy Based Sentiment Classification Using Fuzzy Linguistic Hedges for Decision Making", Mapana Journal of Sciences,Vol. 22, Special Issue 2, pp 63-79, 2023 DoI:|https://doi.org/10.12723/mjs.sp2.4

[21]. Liu, C. H., Tsai, C. F., Sue, K. L., & Huang, M. W. (2020). The feature selection effect on missing value imputation of medical datasets. applied sciences, 10(7), 2344.https://doi.org/10.3390/app10072344

[22]. A.AngelPreethi, and Dr.S.Britto Ramesh Kumar, "A methodological framework for opinion mining", International Journal of Computer sciences and Engineering, Vol. 6, special Issue 2, 2018, pp- 6-9.

[23]. A. Angelpreethi and S. B. R. Kumar, "An Enhanced Architecture for Feature Based Opinion Mining from Product Reviews," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, 2017, pp. 89-92.