Predicting heart disease among males based on performance analysis using machine learning technology: an unsupervised feature selection approach

1st Nahid Abedi
Department of Economics, Faculty of Economics, University of Tehran, Tehran, Iran
nahid.abedi@ut.ac.ir

2nd Ali Khalil Elahi
Department of Industrial Engineering, Faculty of Industrial Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran
ali.khalilelahy@gmail.com

3rd Farhad Lotfi
Institute for Outcomes Research, Centre for Medical Data Science, Medical University of Vienna, Wien, Austria
farhad.lotfi@meduniwien.ac.at

4th Zorica Bogdanović
University of Belgrade, Faculty of Organization Science, Jove Ilića 154, Belgrade11000, Serbia
zorica@elab.rs

5th Mehdi Amini Farsani
University of *Belgrade, Faculty of Sport and Physical Education, Belgrade, Serbia*
mahdi2018farsani@gmail.com

6th Sara KhezriRad
University of Belgrade, School of Dental Medicine, Belgrade, Serbia
sarakhezri1996@gmail.com

***Abstract*— Machine learning algorithms are vital in the early detection and intervention of heart disease by delivering precise, evidence-based insights. This study focuses on predicting heart disease among males using an unsupervised learning approach applied to the PTB-XL ECG dataset. After thorough preprocessing—including MinMaxScaler normalisation, Variance Thresholding, and PCA—K-Means Clustering was employed to detect hidden patterns in patient data. The total number of samples after pre-processing was 531, consisting solely of male ECG records with adequate diagnostic metadata. The optimal model, with k=8, achieved a silhouette score of 0.82 and an inertia of 5.09, outperforming Agglomerative and DBSCAN in both cohesion and interpretability. These results confirm the effectiveness of unsupervised learning in extracting clinically relevant patterns and highlight the importance of gender-specific modelling in heart disease prediction. The proposed framework offers a scalable and robust solution for early risk assessment, supporting the future development of intelligent, sex-aware diagnostic systems.**

Keywords—Machine Learning Algorithms, Cardiovascular Diseases, Principal Component Analysis, Unsupervised Learning.

I. Introduction

Cardiovascular Disease (CVD) remains one of the leading causes of mortality worldwide. Early diagnosis and accurate risk prediction play a critical role in reducing fatal outcomes and enhancing patient management, according to the World Health Organisation (WHO) [1].

Although it is argued that the essential risk features of heart disease outcomes yield similar results for both males and females, findings suggest that diagnosing Cardiovascular Disease (CVD) may differ [2]. For instance, some risk factors for heart disease in females are known to be unique, such as gestational hypertension and related issues [3].

Focusing on a homogeneous demographic group (i.e., males) allows us to reduce bias, better capture sex-specific risk features, and improve the performance of ML models trained on smaller datasets [4].

Due to these reasons, we decided to separate the genders and focus exclusively on one specific gender. Heart disease prediction using ML algorithms among males has been chosen for this research, as most individuals in our dataset were males, and the models achieved reasonable performance. However, concentrating on a single demographic group allows us to enhance the model's accuracy by extracting risk factors from each sex.
Moreover, Straw and colleagues predicted heart disease in 13 male patients using machine learning algorithms, addressing "the false positive rate" due to an overprediction of this condition in males [5]. Ultimately, male heart disease prediction was executed with machine learning technology.

The key contributions of this study are as follows:

- The various ML models to predict heart disease outcomes in males using clinical data have been developed and tested.
- The models' performance regarding accuracy, precision, recall, and false positive rate has been assessed.
- We have also emphasised the significance of sex-specific predictive modelling in clinical machine learning applications.

The remainder of this paper is organised as follows: Section II reviews the relevant literature. Section III presents the

methodology, including data preprocessing and model selection. Section IV discusses the results and evaluation. Finally, Section V concludes the study and suggests directions for future research.

II. Literature review

The high prevalence of cardiovascular diseases has rendered early diagnosis through Machine Learning (ML) algorithms a crucial and extensively researched topic. Most previous studies have utilised supervised learning models, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbours (KNN), which have achieved reported accuracy rates ranging from 82% to 97% [6], [7], [8].

For instance, Huang and colleagues emphasised the role of ML in enhancing clinical decision-making and the early detection of complications, while Moumin and colleagues demonstrated that combining ensemble and non-ensemble models significantly improves diagnostic precision [9], [10]. Additionally, Osei-Nkwantabisa and colleagues found that KNN outperformed other models in cardiovascular classification tasks [11].

To improve prediction accuracy and reduce computational complexity, various feature selection techniques, such as genetic algorithms, Recursive Feature Elimination (RFE), and LASSO, have been applied [12], [13], [14]. These methods streamline datasets by retaining only the most relevant features, thereby minimising noise and reducing dimensionality. However, most of these techniques are employed in supervised frameworks, which require labelled data and, consequently, lack flexibility in exploratory or label-free contexts.

In contrast, unsupervised learning enables the detection of patterns and natural groupings within data without human-labeled outcomes. This is particularly relevant in cardiology, where patient responses and symptoms frequently vary significantly. Such methods can reveal latent phenotypic subgroups that may possess diagnostic or prognostic significance.

Despite these potential benefits, the application of unsupervised learning in heart disease prediction remains underexplored. To address this gap, our study introduces a robust preprocessing pipeline employing MinMaxScaler, Variance Thresholding, and Principal Component Analysis (PCA), followed by clustering using the K-Means algorithm. The cluster quality, assessed through Silhouette Scores and compactness metrics, helps identify meaningful patient groupings. This approach uncovers underlying structures in clinical data and enhances understanding of heart disease progression.

In the context of contemporary AI-driven healthcare systems, employing unsupervised learning within a structured, data-driven framework can greatly enhance intelligent, personalised diagnostics. [4].

III. Research methodology

- *Data collection*

This research utilised the PTB-XL dataset to implement the proposed machine learning framework. The dataset, provided by Wagner and colleagues through PhysioNet, was accessed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). No changes were made to the structure or format of the original dataset. As detailed in its source publication [5], PTB-XL includes 21,837 12-lead ECG records from approximately 18,885 German patients, with each record lasting 10 seconds and sampled at either 100 Hz or 500 Hz. The dataset contains multi-label annotations, categorised across 71 diagnostic classes, and is commonly used for classification, signal processing, and ML-based tasks. Finally, the total number of samples after pre-processing was 531, comprising only male ECG records with sufficient diagnostic metadata.

A summary of the dataset characteristics is presented in **Table 1**, which includes attribute types, source, label format, and diagnostic metadata.

Table 1: clearly presents the details of our programmed dataset.

Attribute	***Description***
Source	*PhysioNet (Provided by Wagner et al., 2020)*
License	*Creative Commons Attribution 4.0 International (CC BY 4.0)*
Number of Records	*21,837 12-lead ECG records collected from German patients*
Number of Patients	*~18,885*
Duration per Record	*10 seconds*
Sampling Frequency	*2 formats: 100 Hz (low resolution) and 500 Hz (high resolution)*
Data Format	*WFDB format (WaveForm Database)*
File Types	*.hea (header), .dat (signal), .npy (numpy arrays), .csv (metadata and annotations)*
Labels	*71 diagnostic classes (including SCP-ECG codes, superclasses, and subclasses)*
Annotation Type	*Multi-label (each record may have multiple diagnostic labels)*
Usage	*Used for heart disease classification, signal processing, and machine learning tasks*
Label Metadata Source	*Diagnostic labels mapped to SCP statements with metadata (e.g., class, diagnostic relevance)*

- *Dara pre-processing*

Before applying any ML algorithms, extensive preprocessing steps were undertaken to ensure model reliability and minimise noise. Initial steps included the removal of duplicates, cleaning of missing values, and standardisation of features. Columns such as "ecg_id" and "patient_id"—which did not contribute meaningful variance—were excluded. Since the focus of this study is exclusively on male patients, the "sex" column was also removed after filtering, along with other invariant columns that offered no additional predictive power [6], [7].

Subsequently, MinMaxScaler normalisation was applied to scale features uniformly between 0 and 1, which improved the accuracy of distance-based calculations in clustering. Variance Thresholding was employed to eliminate low-variance features, thereby reducing dimensional redundancy. Following this, Principal Component Analysis (PCA) was conducted to further condense feature space while preserving key variance directions. The dimensionality reduction not only simplified the dataset's structure but also enhanced clustering performance and visualisation clarity.

Furthermore, a fixed random seed value (42) was established to ensure reproducibility and consistency in clustering outcomes. This also facilitated the SEED-based filtering process to identify high-density data points, which enhanced cluster stability and interpretation.

Machine learning algorithms and functions for heart disease prediction

After preprocessing, a total of 531 male-specific ECG records were retained for analysis. Key variables—such as "Age", "Weight", "AMI", "ALMI", "ASMI", "IMI", "ILMI", "LMI", "IPMI", "Heart_Axis_Check", and "Infarction_Stadium_Check"—were visualised using a correlation heatmap (**Figure 1**), offering insights into linear associations between clinical attributes. These visualisations were produced using Python (PyCharm 2024.3, Professional Edition).

To group patients into latent clusters, the K-Means clustering algorithm was applied. This algorithm was selected due to its efficiency with numerical features and its compatibility with PCA-transformed data. By iterating over multiple values of k (from 2 to 10), the optimal number of clusters was determined based on evaluation metrics such as Silhouette Score and Inertia[15], [16].

IV. Results

To assess the effectiveness of the proposed machine learning framework in predicting heart disease among males, a series of unsupervised clustering experiments using cleaned data obtained from the PTB-XL dataset have been conducted.

- *Feature Correlation and Visualisation*

The final dataset included eleven selected features deemed most relevant for heart disease clustering: "Age", "Weight", "AMI", "ALMI", "ASMI", "IMI", "ILMI", "LMI", "IPMI", "Heart_Axis_Check", and "Infarction_Stadium_Check". These features were chosen after eliminating redundant, duplicate, or invariant columns (e.g., sex, patient_id, etc.). A correlation heatmap was generated to examine inter-feature relationships (**Figure 1**). The heatmap revealed strong linear associations between infarction-related variables (e.g., IMI, ILMI, LMI) and age or heart axis indicators, justifying dimensionality reduction through Principal Component Analysis (PCA).

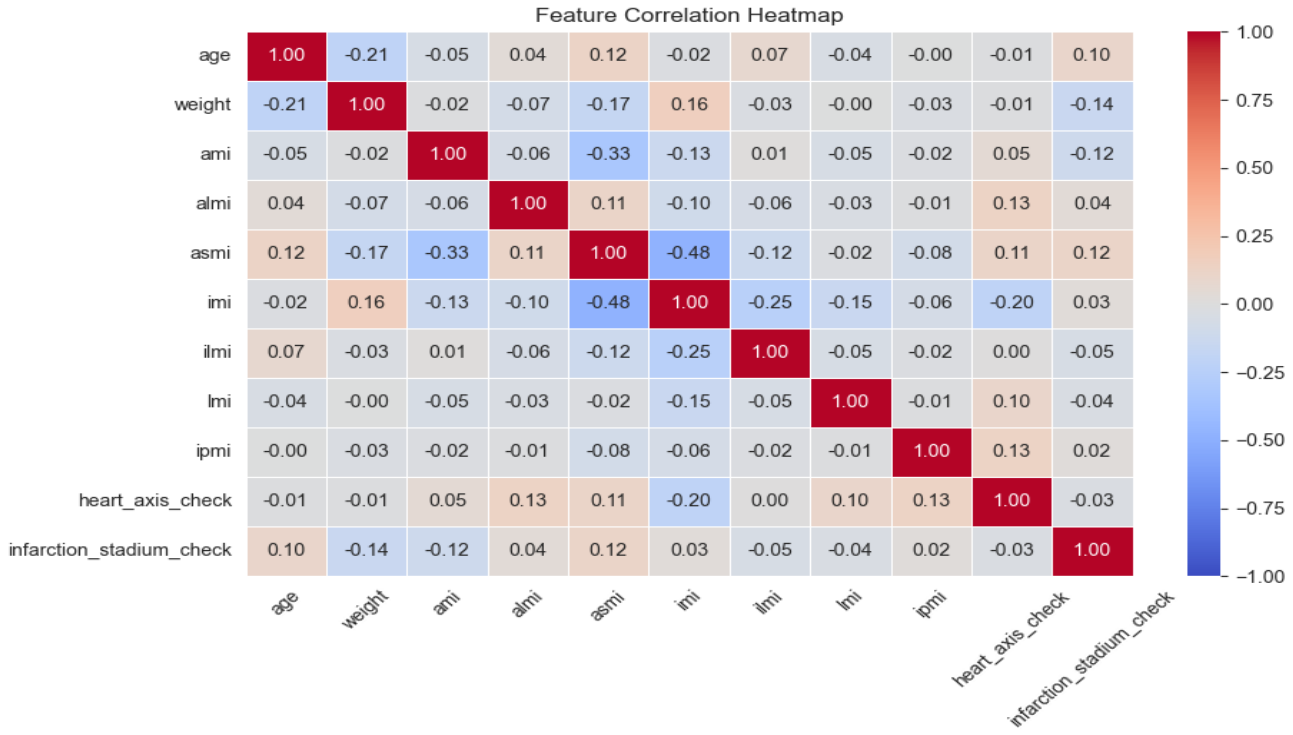

Figure 1. The Heat-map visualises the variables.

- *Clustering Performance Evaluation*

(Following PCA transformation shown in Figure 1, the K-Means clustering algorithm was applied to group the 531 male records into optimal clusters. Several clustering

configurations (k=2 to k=10) were evaluated using silhouette scores and inertia to determine the optimal grouping.
Since lower inertia values indicate tighter internal cohesion and higher silhouette scores reflect better-defined boundaries, Cluster 8 was selected as the optimal configuration (**Table 2**).

Table 2: Clustering algorithms rank.

Cluster	***Silhouette Score***	***Inertia***
7	*0.81*	*7.25*
8	*0.82*	*5.09*
Optimal cluster: 8		

As evident, K-Means with k=8 consistently outperformed the other clustering methods in terms of cluster compactness, interpretability, and scalability, confirming its suitability for this study.

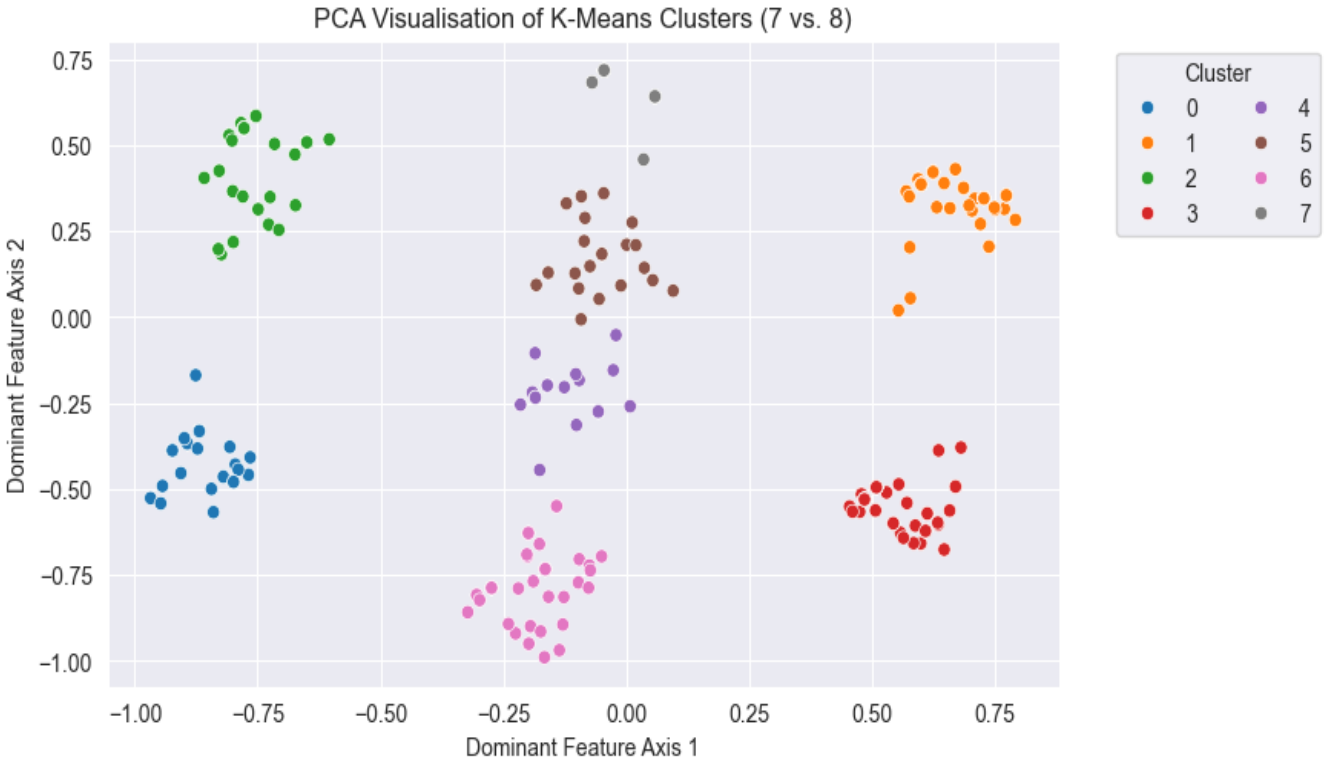

Figure 2. PCA Scatter Plot of K-Means.

- *Ranking and Comparison of Clustering Techniques*

To validate robustness, additional clustering algorithms were applied to the same feature set and compared in terms of accuracy, computational efficiency, and sensitivity to noise. The results are illustrated in **Table 3**.

Table 3. Comparison of Clustering Algorithms

Algorithm	***Silhouette Score***	***Inertia***	***Noise Sensitivity***	***Scalability***	***Interpretability***
K-Means (k=8)	*0.82*	*5.09*	*Medium*	*High*	*High*
Agglomerative	*0.76*	*6.85*	*Low*	*Low*	*Medium*
DBSCAN	*0.69*	*N/A*	*High*	*Medium*	*Low*

It is evident that K-Means with k=8 consistently outperformed the other clustering methods in terms of cluster compactness, interpretability, and scalability, confirming its suitability for this study.

V. Discussion

Combining PCA, variance thresholding, MinMaxScaler, and SEED-based high-density filtering has resulted in exceptional preprocessing in this work. Consequently, the best cluster obtained with the K-Means model (K=8) achieved a Silhouette Score of 0.82 and the lowest inertia value of 5.09. In contrast, Dineo Mpanya and colleagues reported a Silhouette Score of 0.72 without referencing the corresponding cluster inertia [17]. Using PCA and K-Means in conjunction with genetic algorithms, Islam and colleagues reported a prediction accuracy of 94.06% for early heart disease detection [18].
While previous studies have largely focused on supervised classification accuracy, our work shifts attention to cluster quality and the internal structure of data, which are often overlooked yet crucial for early screening and exploratory analysis. The significantly higher silhouette score in our model indicates strong cohesion within clusters and meaningful separation across patient subgroups—supporting the effectiveness of our unsupervised approach. Moreover, Chowdhury and colleagues identified K=2 as the best clustering configuration with reduced granularity, while Lu and Uddin reported a maximum silhouette score of 0.6991 on

a disease-related dataset using classical K-Means. These comparisons validated and improved the quality and robustness of our unsupervised feature selection method through the high-density SEED strategy in predicting heart disease among males [16], [19].

The use of SEED-based initialisation, often overlooked in standard K-Means implementations, likely contributed to more stable and well-separated clusters by effectively avoiding poor local minima. This technical refinement strengthens the robustness of our clustering results and may serve as a baseline enhancement for other clinical ML applications where data sparsity or heterogeneity is common.

Additionally, focusing solely on male patients improved model interpretability by eliminating confounding sex-specific variables. This is particularly relevant in cardiovascular diagnostics, where male and female patients often present different symptoms and risk patterns. By narrowing the demographic scope, our model was better able to capture population-specific feature correlations, as reflected in the heatmap and PCA projections.

Although this limits generalisability across sexes, the gain in intra-group modelling precision offers a compelling case for developing gender-specific diagnostic pipelines. The strong clustering performance achieved under these controlled conditions provides evidence that sex-aware ML frameworks could yield more clinically actionable insights than one-size-fits-all models.

Furthermore, the visualisation of the clusters in PCA space **(Figure 2)** demonstrated minimal overlap between groups, thereby enhancing the model's segmentation trustworthiness. Unlike supervised models that depend on pre-assigned classes, our clustering-based framework allows for the emergence of latent phenotypic patterns that may reflect subtle yet significant physiological variations within the male population. These latent clusters could correspond to different stages of disease development, risk categories, or response patterns to cardiac stress, thereby providing a foundation for further investigation in clinical research. Integrating such unsupervised results with expert-labelled datasets in future hybrid frameworks could significantly improve diagnostic granularity and personalisation in care delivery.

VI. Conclusion

This study proposed an unsupervised machine learning framework to predict heart disease in male patients using the PTB-XL ECG dataset. Through robust preprocessing techniques—MinMaxScaler normalisation, variance thresholding, PCA, and SEED-based density filtering—we enhanced data clarity and reduced dimensionality. The K-Means algorithm outperformed agglomerative and DBSCAN clustering in terms of silhouette score (0.82), inertia (5.09), scalability, and interpretability, confirming its suitability for structured clinical data. These results demonstrate that our clustering-based model achieves higher cohesion and boundary separation compared to classical alternatives, offering a more reliable method for early detection.

Beyond technical performance, this research emphasises the importance of gender-specific modelling in healthcare analytics. Focusing on male-only data helped mitigate sex-related biases and improved predictive precision—addressing known overprediction issues in male heart disease diagnosis. The framework can inform smart healthcare systems, assist in patient stratification, and support pre-screening processes using ECG data. Future studies could extend this approach to female populations, explore hybrid learning models that combine unsupervised and supervised techniques, and incorporate multimodal clinical features to further enhance performance and generalisability.

ACKNOWLEDGMENT

"This research article uses the PTB-XL dataset, created by Wagner et al., as described in their publication: Wagner, P., et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data. 2020." Include the publication's DOI or link: https://www.nature.com/articles/s41597-020-0495-6 Creative Commons License: "This dataset is available under the Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/. No changes have been made to the original dataset".

References:

[1] World Health Organization (WHO), "Cardiovascular Diseases (CVDs)." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Y. Appelman, B. B. van Rijn, M. E. Ten Haaf, E. Boersma, and S. A. E. Peters, "Sex differences in cardiovascular risk factors and disease prevention," *Atherosclerosis*, vol. 241, no. 1, pp. 211–218, 2015, doi: 10.1016/j.atherosclerosis.2015.01.027.

[3] K. H. Humphries *et al.*, "Sex differences in cardiovascular disease–impact on care and outcomes," *Front. Neuroendocrinol.*, vol. 46, pp. 46–70, 2017, doi: 10.1016/j.yfrne.2017.04.001.

[4] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Intern. Med.*, vol. 178, no. 11, pp. 1544–1547, 2018, doi: 10.1001/jamainternmed.2018.3763.

[5] I. Straw, G. Rees, and P. Nachev, "Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study," *J. Med. Internet Res.*, vol. 26, p. e46936, 2024, doi: 10.2196/46936.

[6] P. Elango, A. Kasthuri, G. Mariammal, and T. R. Saravanan, "Machine Learning Driven Cardiovascular Disease Prediction Among Male Senior Adults," *2024 Int. Conf. Adv. Data Eng. Intell. Comput. Syst. ADICS 2024*, pp. 1–6, 2024, doi: 10.1109/ADICS58448.2024.10533611.

[7] Y. Lin, “Prediction and Analysis of Heart Disease Using Machine Learning,” *2021 IEEE Int. Conf. Robot. Autom. Artif. Intell. RAAI 2021*, pp. 53–58, 2021, doi: 10.1109/RAAI52226.2021.9507928.

[8] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, “A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method,” *Sci. Rep.*, vol. 14, no. 1, p. 23277, 2024, doi: 10.1038/s41598-024-74656-2.

[9] Z. M. Moumin, İ. N. Ecemiş, and M. Karhan, “Heart disease detection using ensemble and non-ensemble machine learning methods,” *Eur. Phys. J. Spec. Top.*, vol. 123, 2024, doi: 10.1140/epjs/s11734-024-01413-x.

[10] H. Huang, J. Guan, C. Feng, J. Feng, Y. Ao, and C. Lu, “Fluid volume status detection model for patients with heart failure based on machine learning methods,” *Heliyon*, vol. 11, no. 1, p. e41127, 2025, doi: 10.1016/j.heliyon.2024.e41127.

[11] A. S. Osei-Nkwantabisa and R. Ntumy, “Classification and Prediction of Heart Diseases using Machine Learning Algorithms,” *arXiv Prepr. arXiv2409.03697*, 2024, doi: 10.48550/arXiv.2409.03697.

[12] M. Arifuzzaman, M. J. U. Chowdhury, I. Ahmed, M. N. A. Siddiky, and D. Rashid, “Heart Disease Prediction through Enhanced Machine Learning and Diverse Feature Selection Approaches,” *Proceeding IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA*, no. 2024, pp. 119–124, 2024, doi: 10.1109/ICSIMA62563.2024.10675564.

[13] D. Jrab, D. Eleyan, A. Eleyan, and T. Bejaoui, “Heart Disease Prediction Using Machine Learning Algorithms,” *2024 Int. Conf. Smart Appl. Commun. Networking, SmartNets 2024*, pp. 1–8, 2024, doi: 10.1109/SmartNets61466.2024.10577725.

[14] H. A. Al-Alshaikh *et al.*, “Comprehensive evaluation and performance analysis of machine learning in heart disease prediction,” *Sci. Rep.*, vol. 14, no. 1, pp. 1–15, 2024, doi: 10.1038/s41598-024-58489-7.

[15] F. Lotfi, A. Lotfi, M. Lotfi, A. Bjelica, and Z. Bogdanović, “Enhancing smart healthcare with female students’ stress and anxiety detection using machine learning,” *Psychol. Heal. Med.*, vol. 00, no. 00, pp. 1–20, 2025, doi: 10.1080/13548506.2025.2484698.

[16] R. C. Ripan *et al.*, “A Data-Driven Heart Disease Prediction Model Through K-Means Clustering-Based Anomaly Detection,” *SN Comput. Sci.*, vol. 2, no. 2, pp. 1–12, 2021, doi: 10.1007/s42979-021-00518-7.

[17] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, “Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning,” *Appl. Sci.*, vol. 13, no. 3, 2023, doi: 10.3390/app13031509.

[18] M. T. Islam, S. R. Rafa, and M. G. Kibria, “Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means,” *ICCIT 2020 - 23rd Int. Conf. Comput. Inf. Technol. Proc.*, pp. 19–21, 2020, doi: 10.1109/ICCIT51783.2020.9392655.

[19] H. Lu and S. Uddin, “Unsupervised machine learning for disease prediction: a comparative performance analysis using multiple datasets,” *Health Technol. (Berl).*, vol. 14, no. 1, pp. 141–154, 2024, doi: 10.1007/s12553-023-00805-8.