22º Congresso Latino-americano de Software Livre e Tecnologias Abertas
LATINOWARE 2025
Latin.Science 2025

Optimizing Cardiovascular Disease Risk Prediction Using Ensemble Learning Techniques

Abner Gabriel Dias dos Santos
São Paulo State Univ. of Tech. (FATEC)
Ourinhos, Brazil
abner.santos@fatecourinhos.edu.br

Bruno Eduardo Marcolino
São Paulo State Univ. of Tech. (FATEC)
Ourinhos, Brazil
bruno.marcolino@fatecourinhos.edu.br

Thiago José Lucas
São Paulo State Univ. of Tech. (FATEC)
Ourinhos, Brazil
thiago@fatecourinhos.edu.br

Alessandra de Souza Lopes
São Paulo State University (UNESP)
Bauru, Brazil
alessandra.lopes@unesp.br

Carlos Alexandre Carvalho Tojeiro
São Paulo State University (UNESP)
Bauru, Brazil
carlos.tojeiro@unesp.br

Kelton Augusto Pontara da Costa
São Paulo State University (UNESP)
Bauru, Brazil
kelton.costa@unesp.br

***Abstract*—The heart is one of the most essential organs, responsible for ensuring blood circulation throughout the human body. A wide range of cardiovascular diseases (CVDs) poses severe risks to human health. According to data from the World Health Organization (WHO), CVDs account for approximately 17.9 million deaths annually, representing 31% of all global fatalities. Machine learning techniques have demonstrated significant potential in effectively predicting the risk of CVD occurrence. In this study, ensemble learning algorithms were employed to optimize predictive analysis for cardiovascular events. The dataset Indicators of Heart Disease (2022 UPDATE), provided by the Centers for Disease Control and Prevention (CDC), served as the basis for model training and evaluation. The best-performing models were Gradient Boosting and Logistic Regression, both achieving an accuracy and precision rate of 94.9%.**

***Index Terms*—Cardiovascular Diseases; Machine Learning; Predictive Analysis; Gradient Boosting.**

I. Introduction

The rising incidence of cardiovascular diseases has been a major concern among health researchers over the past decades. According to [1] all age groups, particularly young adults, are susceptible to such conditions, often triggered by atherosclerosis, which refers to the accumulation of fat on the inner walls of blood vessels. This condition primarily stems from the lack of physical activity and excessive consumption of fast food, foods rich in cholesterol, sodium, and other harmful substances. Among these, refined salt, processed meats, ultra-processed products, fried foods, sugars, and alcoholic beverages stand out, as they are widely associated with atherosclerosis in contemporary medicine. In this context, controlling and preventing these diseases has become a priority, given the high number of reported cases.

The healthcare field has increasingly adapted to technological advancements, as such resources enable more accurate research outcomes, resulting in significant discoveries that shape medical history and continue to emerge with the evolution of available techniques. This progress is essential for achieving higher precision in medical diagnoses. Various technologies have been employed to collect the largest possible volume of data in order to produce more reliable results. Among these technologies, Machine Learning (ML) stands out as a critical tool for predicting diseases and their diffusion patterns, as it is capable of analyzing real-time data and handling large datasets in a robust, organized, and precise manner.

The application of ML has become indispensable for managing the vast amount of health-related information in a structured way, providing an efficient alternative that outperforms human processing capacity. Its use enables improved monitoring of population health conditions and supports the prevention of potential disease outbreaks. According to Ibrahim and [2] numerous ML algorithms have been applied in the medical domain to enhance diagnostics and facilitate early predictions. The most commonly employed algorithms include Naïve Bayes, Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), K-Means Clustering, Decision Tree (DT), and Random Forest (RF).

Over time, researchers have identified a range of factors contributing to cardiovascular disease incidence, beyond inadequate diet, such as obesity, smoking, and physical inactivity. According to [3] , multiple ML techniques are currently applied to pattern analysis with the aim of making predictions and supporting decision-making. Prominent examples include Linear Regression, Decision Tree, and Random Forest, among others. This raises an important research question: how can the techniques that achieve the highest accuracy in meeting the study's objectives be identified? Furthermore, can the implementation of Ensemble Learning (EL) improve the precision and effectiveness of cardiovascular disease diagnostics?

The general objective of this study was to implement Ensemble Learning (EL) algorithms for the early diagnosis of cardiovascular diseases (CVDs). The specific objectives were defined as follows:

- Conduct a systematic literature review to identify related studies;
- Extract from the related studies the main variables of interest, as well as the algorithms and datasets most commonly used by other researchers;
- Perform data preprocessing to adapt the dataset to the classification algorithms;
- Evaluate the obtained results using metrics such as True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Accuracy, Precision, Recall, and Execution Time (s).

Regarding its structure, this paper is organized as follows: the first section presents the introduction, including the study context, the rationale for topic selection, the research problem, and both general and specific objectives. The second section covers the literature review, addressing theoretical concepts, ML techniques, and related studies. The third section describes the adopted methodology, including the algorithms and processed data. The fourth section discusses the results obtained from the applied methods. Finally, the fifth section provides the conclusions and final considerations.

II. Theoretical Background

According to [3], Ensemble Learning algorithms have been producing excellent results in solving problems across various domains, with widespread global adoption. The authors provide a brief discussion on the origins of Ensemble Learning, which began to gain prominence in the 1990s to reduce variance and improve accuracy in automated decision-making processes by combining multiple classifiers to minimize overall variance, thereby creating a more optimized single classifier. Their study emphasizes that Ensemble Learning plays a critical role in addressing several challenges associated with Machine Learning, such as estimation, confidence, error correction, incremental learning, and handling missing features, among others. The algorithms employed in their research included Naïve Bayes, Logistic Regression, Support Vector Machine, k-Nearest Neighbors, K-Means Clustering, Decision Tree, and Random Forest.

In [4] Machine Learning is highlight as a highly versatile technology for the development and evolution of disease prediction frameworks, with Ensemble Learning being one of its most effective techniques due to its ability to combine multiple classifiers, thereby enhancing performance and surpassing the capabilities of individual models. Given the variety of studies that utilized ensembles for disease prediction, their research aimed to provide a more comprehensive understanding of ensemble applications in commonly studied diseases. The study focused on conditions such as diabetes, skin diseases, kidney disease, liver disease, and heart disease, employing techniques such as Stacking, Bagging, Boosting, and Voting. The results indicated that although Bagging was the most frequently used technique among the analyzed studies, it exhibited the lowest performance. Conversely, Stacking produced the best results despite being less common, achieving a peak accuracy of 82.6%.

According to [5] the heart disease has been one of the leading causes of fatalities over the past decade, resulting in an overwhelming number of deaths. With the support of Machine Learning, predicting this disease has become significantly more efficient due to its ability to analyze indicators that may signal potential occurrences in advance, thus enabling more accurate and timely diagnoses. The study noted that ensemble models, a subset of Machine Learning, integrate multiple models to enhance prediction accuracy. Among the models evaluated, Support Vector Machine (SVM) and Decision Tree (DT) yielded the best individual performance scores, with SVM achieving an accuracy of 85.4% and DT reaching 80.1%. When combined using a Stacking approach, these models achieved a predictive accuracy of 98.12%, clearly demonstrating that ensemble models substantially outperform individual classifiers.

In [6] the authors emphasize that the early diagnosis of cardiovascular diseases remains a significant challenge in the medical field, one that must be addressed urgently given the alarming global rise in annual mortality rates. Considering the complexity of diagnosing cardiovascular conditions due to the numerous contributing factors, their research sought to enable early treatment through the application of artificial intelligence. The study employed an ensemble-based method within the Machine Learning domain, utilizing five categorization approaches with hyperparameter tuning for model optimization. Voting techniques were incorporated primarily to reduce false positive rates. Ultimately, the Random Forest (RF) model achieved the highest cross-validation score with 93.12%, while Logistic Regression (LR) demonstrated the best accuracy with 91.96%.

In a study conducted by [7], Coronary Heart Disease (CHD), one of the leading causes of mortality worldwide, was examined. The approach involved predictive methods such as K-Nearest Neighbors (KNN), Binary Logistic Classification, and Naïve Bayes, evaluated using accuracy curves, recall, and Receiver Operating Characteristic (ROC), along with ensemble modeling strategies like Bagging and Stacking. The study utilized a total of 80,000 patient records diagnosed with CHD to assess the performance of the modeling techniques. Analytical methods, including cross-validation and K-Folds, were also applied. The highest accuracy achieved in the study was 76.5%.

In work by [8] are conducted research introducing data mining techniques aimed at improving the predictive accuracy of cardiovascular diseases. The study focused on optimizing cardiovascular disease prediction through a stacking ensemble model based on a combination of heterogeneous classifiers. Initially, the research analyzed base classifiers and meta-classifiers, selecting the ensemble, followed by the implementation of heterogeneous classifiers, including Support Vector Machine, Naïve Bayes, K-Nearest Neighbor, XGBoost, and Random Forest, to optimize predictions. Ensemble performance was measured by accuracy, precision, recall, and F1-score, with the stacking ensemble achieving the highest accuracy at 90.16%.

In [9] are conducted a study on Stroke, a cardiovascular disease with increasing global prevalence. The research examined the implementation of machine learning algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest (RF) for stroke prediction. Data were collected from the Harvard Dataverse Repository, including clinical, physiological, behavioral, demographic, and historical information. Techniques addressing class imbalance, such as Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Random Oversampling (ROSE), were applied to improve the prediction of minority classes. The study introduced a combination of RF with ADASYN, forming the ADASYN RF model, which achieved an outstanding accuracy of 99% in stroke detection.

According to [10] highlighted machine learning as an essential tool in medicine due to its precision in disease prediction and diagnosis, emphasizing the use of ensemble learning to enhance predictive performance. The study focused on optimizing cardiovascular disease prediction through feature extraction methods, including Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), which select the most relevant data from the dataset. Subsequently, a comparative analysis of machine learning algorithms and ensemble methods was conducted. Evaluation metrics included accuracy, recall, precision, F-score, and ROC. The best-performing approach was Bagging combined with Decision Tree (DT), achieving 98.6% accuracy and demonstrating excellent predictive performance for cardiovascular diseases.

The research by [11] emphasizes that cardiovascular diseases have the highest mortality rates, with initially mild symptoms that can progress to conditions such as atrial arrhythmias, myocardial infarction, and cardiomyopathy. Predictive analysis using models such as Logistic Regression and neural networks achieved accuracies between 60% and 70%. Ensemble methods, however, reached an impressive 91% accuracy. The study utilized a massive dataset of 130,000 patients provided by MIT's GOSSIS community initiative, encompassing individuals from Argentina, Australia, the United States, and Brazil.

Diagnostic techniques for cardiomyopathy using multiple classifiers are presented by [12]. The study emphasized early and precise diagnosis of cardiac diseases through magnetic resonance imaging (MRI) analysis, leveraging multiple classifiers to improve predictive performance. The classifiers used included Logistic Regression, Naïve Bayes, Support Vector Machine, and neural networks, with Logistic Regression achieving the highest accuracy at 85.6%.

In [13] the authors addressed predictive systems using neural networks, such as RapidMiner, a tool for data-driven decision-making widely applied in predictive analysis. RapidMiner demonstrated superior results compared to similar tools, such as Weka and MATLAB, which are applied for related purposes but lack equivalent automation capabilities. The training phase yielded an accuracy of 87.3%.

A study by [14] discussed ensemble learning methods applied to large cardiovascular datasets containing variables such as weight, sex, age, and hypertension. Machine learning and neural network models were utilized to enable comprehensive comparisons. Electrocardiogram data were employed for detection, with the GA-ANN model (Genetic Algorithm-Artificial Neural Network) achieving 82.5% accuracy in distinguishing normal and anomalous cases. GA-ANN optimizes parameters to maximize predictive performance.

In [15] the authors are focused on frameworks, defined as collections of concepts and tools supporting software development, in this case applied to ensemble learning for cardiovascular disease prediction. Ensemble learning combines the analyses of multiple models to reduce variance and irrelevant information, enhancing accuracy and reliability. The study reported an accuracy of 93.01%, demonstrating strong performance compared to other research. Outlier identification, through statistical or expert methods, was emphasized to ensure robust results.

A study conducted by [16] proposes a cardiovascular disease detection method based on optimal feature selection combined with Ensemble Learning techniques. The study applies data preprocessing, including handling missing values, normalization, and outlier removal, followed by testing with individual classifiers (Random Forest, Gradient Boosting, XGBoost, SVM, Logistic Regression) and ensemble schemes such as Voting and Stacking. Results obtained from public CVD datasets indicate that ensemble models consistently outperform individual classifiers in metrics such as accuracy, F1-score, and AUC. The study emphasizes that integrating dimensionality reduction with ensemble methods provides greater robustness, computational efficiency, and predictive accuracy, demonstrating promise for clinical support applications.

A. Taxonomy of Related Works

The taxonomy compiled in Table 1 objectively presents the main variables extracted from the related works identified through the systematic literature review. Several methods for cardiovascular disease prediction were identified, utilizing diverse datasets.

The studies obtained from the systematic literature review demonstrate the effectiveness of Ensemble Learning techniques in predicting cardiovascular disease incidence. The results reported in these studies support the implementation of Machine Learning techniques due to their improvement in detection rates.

The results columns indicate Accuracy (ACC), Precision (PR), True Positives (TP), False Positives (FP), and Area under the Curve (AUC).

III. Methodology

This section highlights the methodological process adopted in conducting the experiments. This study was conducted entirely with free software: all experiments were run on a Debian 12 Linux server using the scikit-learn library in Python. The dataset was preprocessed to remove null or duplicate values and adjust discretization and normalization using the min/max technique. All classifiers were implemented using 10-fold cross-validation.

1) Dataset: The first step was to select an appropriate dataset for the study. The search was conducted on Kaggle, a platform and community focused on learning in the field of data science, where the dataset *Indicators of Heart Disease (2022 UPDATE)*, distributed by the CDC (Centers for Disease Control and Prevention), was located. This dataset contains all essential data to achieve the primary objective of the study.

The CDC is a disease control and prevention center located in Atlanta, Georgia. The dataset used was created based on telephone surveys conducted by the BRFSS, a risk factor surveillance system that collected data across all 50 U.S. states over 38 years, up to the 2022 update. The dataset includes demographic variables, habits such as tobacco and alcohol consumption, among others, and is also used in public health programs. Due to its reliability, it was selected for this study.

Some highlighted variables from the dataset include:

- **Chest Pain Type:** Indicates possible types of diseases, including cardiovascular disease (CVD).
- **Resting BP:** Blood pressure measured at rest, without the influence of exercise or stress.
- **Cholesterol:** Cholesterol levels, which when elevated, may be harmful to health.
- **Electrocardiogram (ECG):** Measures the electrical activity of the heart to identify cardiovascular issues.
- **Heart Rate (HR):** Can vary according to age, exercise routine, and presence of CVD.
- **Presence of Angina:** Indicates reduced blood flow in the coronary arteries, causing cardiac ischemia.
- **Heart Disease:** Target variable indicating presence (1) or absence (0) of heart disease.

2) Machine Learning Classifier Algorithms: Machine Learning classifiers analyze patterns to categorize data according to the task. The selected models include:

- **K-Nearest Neighbour (KNN):** A method that uses training data directly to generate predictions based on the proximity of examples.
- **Support Vector Machine (SVM):** Focused on binary problems, it creates a hyperplane that separates classes by maximizing the margin between neighboring vectors.
- **Logistic Regression (LR):** Uses statistical formulas to predict probabilities of dependent variables, efficient for large datasets.
- **Naive Bayes (NB):** Based on Bayes' theorem, it provides fast and accurate predictions when data are properly provided.
- **Neural Network (NN):** Neural networks process data in multiple stages, performing repeated predictive analyses to generate results.

3) Ensemble Learning: Ensemble Learning combines multiple Machine Learning models to increase system accuracy and robustness. The methods used include:

- **Random Forest (RF):** Generates multiple decision paths and combines the results for high accuracy.
- **Gradient Boosting:** Adjusts successive predictions based on the difference between actual and predicted values, which may lead to overfitting if not properly regularized.
- **AdaBoost:** Combines classifiers by weighting training examples to produce a more accurate final classifier.

4) Performance Evaluation: Performance evaluation allows the analysis of the effectiveness of the trained algorithms. The metrics used include:

- **TP (True Positive):** Correct prediction of a positive class.
- **FN (False Negative):** Incorrect prediction of a positive class.
- **FP (False Positive):** Incorrect prediction of a negative class.
- **TN (True Negative):** Correct prediction of a negative class.

Table I
RELATION OF TECHNIQUES, CLASSIFIERS, AND RESULTS OBTAINED BY THE AUTHORS FROM SECTION 2. SOURCE: PREPARED BY THE AUTHOR.

Article	Dataset	Classifier	Feature Selection	Ensemble Algorithm	ACC	PR	TP	FP	AUC
[3]	–	iForest, SCiForest	–	Bagging, Boosting, Stacking, Voting	–	–	–	–	–
[4]	UCI Cleveland Heart, UCI Chronic Kidney, UCI Dermatology, UCI Indian Liver, Pima Diabetes	RF, LR, NB, SMO	SMO	Bagging, Boosting (AdaBoost, XGBoost), Stacking, Voting	82,6% (Stacking)	99,67%	–	–	–
[5]	UCI	SVM, KNN, DT, NB, RF, AdaBoost	SMOTE, ADASYN	StackFed	98,12% (Stacking)	–	–	–	–
[6]	–	KNN, DT	–	RF, LR, SVM	92,93%	–	46,20%	9,78%	–
[7]	–	SVM, RF, AdaBoost, GradBoost, XGBoost	KNN	Bagging, Boosting, Stacking	–	76,5%	–	–	73,93%
[8]	UCI	SVM, NB, KNN, XGBoost, RF	–	Stacking	90,16% (Stacking)	90,16%	–	–	0,88
[9]	Harvard Dataverse	LR, RF, KNN	–	ADASYN RF	99%	99%	–	–	99%
[10]	–	KNN, SVM, NB, DT, RF	PCA	Bagging	98,7% (RF)	98,3%	98,7%	–	99,8%
[11]	Harvard Privacy Lab	LR, RF, NN	PCA	–	91%	91%	10254	947	–
[12]	ACDC	LR, GNB, SVM, MLP, CNN	–	–	97%	97%	–	–	–
[13]	Cleveland Clinic	NB, LR, KNN, RF, DT, SVM	–	Stacking	88,1%	–	–	–	–
[14]	UCI	NB, SVM, KNN, XGB	–	Voting	87%	–	–	–	–
[15]	–	LR, KNN, RF, XGB, MLP, AdaBoost, ExtraTree, Stacking, CatBoost	–	Stacking	93,1%	–	–	–	–
[16]	UCI	RF, AdaBoost, NB, KNN, NN, SVM	PSO	Voting, Stacking, Bagging, Boosting	88% (XGB)	88% (XGB)	–	–	–

- **Accuracy:** Proportion of correct predictions by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** Proportion of correct positive predictions among all positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** The model's ability to correctly identify positive occurrences.

$$Recall = \frac{TP}{TP + FN}$$

- **Time:** Time required to train the classifier, important for real-time applications.

IV. RESULTS

The analysis of the results obtained from the application of the classifiers to the CDC dataset shows expressive performance, with accuracy (ACC) rates above 87% for all tested models. Overall, the observed precision (PR) rates remained above 89%, demonstrating the effectiveness of machine learning techniques in predicting cardiovascular diseases.

Among the evaluated classifiers, Gradient Boosting and LR achieved 94.9% ACC and 93.7% PR, standing out in building robust predictive models. NB showed the lowest performance, with 87.7% ACC and 83.6% PR, due to the independence assumption between attributes, which is limited given the complexity of medical data.

KNN demonstrated consistent results with 94.4% ACC, 91.6% PR, and 94.4% REC, with low training time (35.475 s), making it advantageous for real-time clinical decision support systems.

NN and RF presented high metrics (NN: 94.8% ACC, 93.6% PR; RF: 94.7% ACC, 93.2% PR), although NN required higher computational cost (621.010 s). Gradient Boosting combined high ACC, PR, and REC despite higher training time (1115.082 s), while AdaBoost presented slightly lower metrics (91.8% ACC, 92.1% PR), confirming the effectiveness of ensemble approaches.

Table II
RESULTS OBTAINED FROM THE CDC DATASET. SOURCE: AUTHOR.

Classifier	TP	FP	FN	TN	ACC	PR	REC	F1-Score	Training (s)
KNN	231976	611	13051	384	94.4%	91.6%	94.4%	97.1%	35.475
SVM	224428	8159	12727	708	91.5%	89.9%	91.5%	95.5%	869.925
LR	229986	2601	10051	3384	94.9%	93.7%	94.9%	97.3%	692.764
NB	208062	24525	5735	7700	87.7%	83.6%	87.7%	93.2%	5.465
NN	229903	2684	10090	3345	94.8%	93.6%	94.8%	97.3%	621.010
RF	230143	2444	10675	2760	94.7%	93.2%	94.7%	97.2%	326.040
Gradient Boosting	230241	2346	10157	3278	94.9%	93.7%	94.9%	97.3%	1115.082
AdaBoost	221971	10616	9512	3923	91.8%	92.1%	91.8%	95.6%	151.082

V. CONCLUSION

The main objective of this study was to apply machine learning classifiers to a dataset aimed at predicting cardiovascular diseases, with a focus on comparative performance analysis between different techniques, particularly Ensemble Learning methods. The relevance of the topic is undeniable, considering that cardiovascular diseases represent one of the leading causes of morbidity and mortality worldwide, accounting for high hospitalization and death rates across various social and geographic contexts. In this scenario, the use of advanced computational techniques for early prediction and detection of these conditions emerges as a promising alternative to support more accurate diagnoses and clinical decision-making.

The proposed solution relied on a dataset containing relevant clinical attributes, enabling the construction of models capable of identifying subtle patterns and correlations among variables, thereby enhancing the clarity and accuracy in analyzing the incidence of cardiovascular diseases. The results showed that all tested classifiers achieved accuracy above 87%, demonstrating the effectiveness of machine learning techniques in the investigated context.

As expected, classifiers based on Ensemble Learning outperformed individual models. Gradient Boosting stood out, reaching 94.9% ACC and PR, followed by Logistic Regression with equivalent results. These findings are consistent with existing literature, highlighting Ensemble Learning as one of the most effective approaches in complex scenarios due to its ability to reduce variance and bias while increasing predictive robustness.

It is worth noting that the strong performance reinforces the potential of these techniques as clinical support tools, particularly in contexts where early diagnoses can significantly reduce the impact of cardiovascular diseases on public health. The capability of algorithms to identify patterns in large volumes of clinical data can contribute to treatment personalization and the development of more reliable and scalable predictive systems.

The limitations of this research are due to the lack of treatment of missing or unbalanced data, and the method did not analyze the risk of bias based on the demographic characteristics of the dataset (since it is centered in the USA). For future studies, it is recommended to use multiple datasets from different regions and populations to evaluate model generalization in heterogeneous scenarios. It is also necessary to apply XAI techniques to understand the results and support medical professionals, as well as the development of a clinical support tool (such as an application). Additionally, incorporating more advanced Ensemble algorithms, such as Stacking, recognized for its high performance in combining heterogeneous classifiers and efficiently handling large, high-dimensional datasets, is suggested. This methodological expansion could further enhance the robustness of results, consolidating machine learning as a strategic tool for preventing and mitigating the impact of cardiovascular diseases.

ACKNOWLEDGMENTS

The authors are grateful to State Center for Technological Education “Paula Souza” (CEETEPS) for research conditions at the Defensive Cybersecurity and Artificial Intelligence Laboratory (Detect.AI), and to Research Support Foundation of São Paulo State (FAPESP), Brazil grants #2023/12830-0.

REFERENCES

[1] L. K. Maceno and M. d. S. Garcia, “Fatores de risco para o desenvolvimento de doenças cardiovasculares em jovens adultos/risk factors for the development of cardiovascular diseases in young adults,” *Brazilian Journal of Health Review*, vol. 5, no. 1, pp. 2820–2842, 2022.

[2] I. Ibrahim and A. Abdulazeez, “The role of machine learning algorithms for diagnosing diseases,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 10–19, 2021.

[3] T. N. Rincy and R. Gupta, “Ensemble learning techniques and its efficiency in machine learning: A survey,” in *2nd International Conference on Data, Engineering and Applications (IDEA)*, 2020, pp. 1–6.

[4] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, “Ensemble learning for disease prediction: A review,” *Healthcare*, vol. 11, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2227-9032/11/12/1808

[5] A. Vinora, E. Lloyds, R. Nancy Deborah, M. Anandha Surya, V. Krithik Deivarajan, and M. MuthuVignesh, "Heart disease prediction using ensemble model," in *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, 2023, pp. 1–6.

[6] A. Ramachandran, S. K. Burra, and S. S. Verma, "Early cardiovascular disease detection using ensemble methods," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, vol. 6, 2023, pp. 2373–2379.

[7] K. P. Joshi, M. L. Prasad, R. Natchadalingam, P. C. S. Reddy, S. Mukherjee, and G. C. Babu, "An accurate prediction of coronary heart disease using ensemble algorithms," in *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, 2023, pp. 1–5.

[8] V. K. HM and D. Suresh, "Performance analysis of base and meta classifiers and the prediction of cardiovascular disease using ensemble stacking," in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 2023, pp. 584–589.

[9] R. Mia, S. Khanam, A. Mahjabeen, N. H. Ovy, D. Ghimire, M.-J. Park, M. I. A. Begum, and A. S. M. S. Hosen, "Exploring machine learning for predicting cerebral stroke: A study in discovery," *Electronics*, vol. 13, no. 4, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/4/686

[10] X.-Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, pp. 1–10, 2021.

[11] B. S. K. Jayasudha, P. N. Sudha, K. Keshav, and N. Ramesh, "Ensemble learning as a prerogative method of predicting mortality of patients with cardiovascular diseases," in *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2021, pp. 01–05.

[12] Q. Ye, L. Qiao, H. Chen, Q. Tao, and J. Xiao, "Automatic cardiomyopathy diagnosis with a cost-sensitive ensemble classifier," in *2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT)*, 2021, pp. 775–779.

[13] D. R. S. Victoria, K. Mayuri, N. Ramalakshmi, and M. G. Kumar, "An ensemble stacking methodology for identifying heart disease," in *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, 2023, pp. 1139–1143.

[14] T. Mahmud, A. Barua, M. Begum, E. Chakma, S. Das, and N. Sharmen, "An improved framework for reliable cardiovascular disease prediction using hybrid ensemble learning," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2023, pp. 1–6.

[15] S. Ambade and D. Chikmurge, "Enhancing cardiovascular disease prediction using ensemble learning," in *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE, 2023, pp. 1–6.

[16] A. Ashfaq, A. Imran, I. Ullah, A. Alzahrani, K. M. A. Alheeti, and A. Yasin, "Multi-model ensemble based approach for heart disease diagnosis," in *2022 International Conference on Recent Advances in Electrical Engineering Computer Sciences (RAEE CS)*, 2022, pp. 1–8.