International Journal of Artificial Intelligence in Medical Issues

Volume 2 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v2i2.157

Research Article

Enhancing Cardiovascular Disease Prediction Accuracy through an Ensemble Machine Learning Approach

Ilham [1*]
[1] Universitas DIPA Makassar, ilhamaswan34@gmail.com
Correspondence should be addressed to Ilham; ilhamaswan34@gmail.com

Received 15 October 2024; Revised 20 October 2024; Accepted 23 November 2024; Published 30 November 2024

Abstract:
This study explores the efficacy of an ensemble machine learning approach, specifically a Voting Classifier combining Decision Tree, k-Nearest Neighbors, and Gaussian Naive Bayes classifiers, in predicting cardiovascular diseases (CVDs). Utilizing a dataset consisting of 70,000 clinical records, the model was rigorously tested through 5-fold cross-validation, achieving remarkable results with average accuracies, precision, recall, and F1-scores all exceeding 99%. The findings validate the hypothesis that ensemble models, due to their capacity to leverage multiple learning algorithms, provide superior prediction accuracy and reliability compared to single predictor models. This research not only confirms the effectiveness of ensemble methods in medical diagnostics but also highlights their potential to enhance decision-making in clinical settings. Given the model's success in identifying various stages of cardiovascular conditions with high accuracy, it offers significant implications for early intervention and personalized patient management. Future research should aim to validate these results across more diverse populations and explore the integration of additional predictive factors that could refine the model's applicability. This study contributes to the computational health field by demonstrating how advanced machine learning techniques can be effectively applied in predicting health outcomes.

Keywords: Ensemble Machine Learning, Cardiovascular Disease Prediction, Voting Classifier, Medical Diagnostics, Machine Learning in Healthcare.

Dataset link: https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death globally, representing a significant public health challenge. The World Health Organization estimates that CVDs account for more than 17 million deaths annually, which constitutes nearly one-third of all global deaths. With risk factors that include high blood pressure, high cholesterol, obesity, and lifestyle choices such as smoking and low levels of physical activity, early detection and management of these conditions are critical. However, traditional methods of prediction and diagnosis often involve extensive physical testing and can sometimes fail to capture at-risk populations until symptoms manifest clinically.

The main problem addressed in this research is the need for improved accuracy and early detection in the diagnostic tools available for cardiovascular diseases. Existing methods, while effective, can benefit from the integration of advanced computational techniques that enhance predictive capabilities and operational efficiency. This is particularly vital in settings with limited access to comprehensive healthcare services or where medical resources are stretched thin.

Our research objectives are to apply and evaluate an ensemble machine learning approach to predict the presence of cardiovascular diseases using patient clinical data. We aim to investigate whether a combination of several predictive models will outperform individual models in terms of accuracy, precision, recall, and F1-score [1]–[4]. Specifically, we combine Decision Tree [5], k-Nearest Neighbors [6], and Gaussian Naive Bayes[7] models to form a robust predictive ensemble. The research questions guiding this study focus on the effectiveness of ensemble methods in improving prediction outcomes for CVDs. Hypothetically, we expect that the ensemble model will provide higher predictive accuracy than any single model used in isolation, owing to the diverse strengths and compensating weaknesses of the combined models. We also hypothesize that such models will demonstrate greater generalizability across different patient demographics and clinical profiles.

The scope of this research is primarily cantered on the predictive modelling of cardiovascular disease using clinical datasets. While the methods and findings are expected to be broadly applicable, the limitations include potential biases inherent in the dataset and the generalizability of the results to populations not represented in the study. Furthermore, the computational intensity of training and validating ensemble models necessitates considerable processing power, which might not be available in all potential application settings [8]–[10]. Our study contributes to the field of computational health by demonstrating the practical applications of ensemble machine learning techniques in a clinical context. By showing how different models can be effectively combined, we provide a blueprint for future research and potential clinical application. The findings could inform healthcare providers about more efficient and accurate tools for early CVD detection, potentially leading to better patient outcomes through earlier intervention and personalized treatment strategies.

2. Method

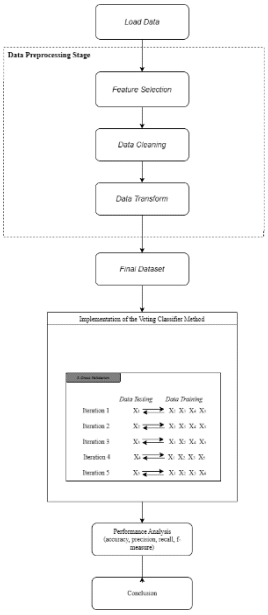

Figure 1: Voting Classifier Evaluation Workflow

This study adopts a quantitative research design utilizing ensemble machine learning techniques to predict cardiovascular diseases (CVDs). The research design integrates data preprocessing, model implementation, and statistical evaluation to test the hypothesis that an ensemble of classifiers will provide superior predictive accuracy compared to individual models. The study was conducted in a controlled environment where historical clinical data was used to train and test the ensemble model. A visual representation of the entire research process is illustrated in **Figure 1**.

Sample or Data Selection:

The dataset used in this study consists of clinical records of patients who underwent medical check-ups for cardiovascular disease assessment.

Table 1: Dataset column descriptions

No	Feature	Keterangan
1	age_years	Age of the patient (in years)
2	**gender**	Gender of the patient. Categorical variable (1: Female, 2: Male)
3	**height**	Height of the patient in centimeters
4	**weight**	Weight of the patient in kilograms
5	**ap_hi**	Systolic blood pressure
6	**ap_lo**	Diastolic blood pressure
7	**cholesterol**	Cholesterol levels. Categorical variable (1: Normal, 2: Above Normal, 3: Well Above Normal)
8	**gluc**	Glucose levels. Categorical variable (1: Normal, 2: Above Normal, 3: Well Above Normal)
9	**smoke**	Smoking status. Binary variable (0: Non-smoker, 1: Smoker)
10	**alco**	Alcohol intake. Binary variable (0: Does not consume alcohol, 1: Consumes alcohol).
11	**active**	Physical activity. Binary variable (0: Not physically active, 1: Physically active)
12	**cardio**	Presence or absence of cardiovascular disease. Target variable. Binary (0: Absence, 1: Presence)
13	**bmi**	Body Mass Index, derived from weight and height
14	**bp_category**	Blood pressure category based on ap_hi and ap_lo. Categories include "Normal", "Elevated", "Hypertension Stage 1", "Hypertension Stage 2"

The dataset comprises several key patient metrics, such as age, gender, height, weight, systolic and diastolic blood pressure, cholesterol levels, glucose levels, smoking status, alcohol consumption, physical activity, and the presence or absence of cardiovascular disease. A total of 70,000 patient records were utilized, split into training and testing sets with an 80-20 ratio.

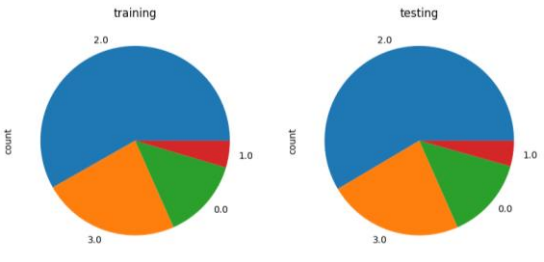

Figure 2: Splitting Data Training (80%), Testing (20%)

Tools and Technology Used:

Data pre-processing, model training, and evaluation were conducted using Python programming language, with libraries such as Pandas for data manipulation, Scikit-learn for machine learning model implementation, and Matplotlib for data visualization. Specifically, the ensemble method employed is a Voting Classifier combining Decision Trees, k-Nearest Neighbors, and Gaussian Naive Bayes classifiers.

Data Collection Process

The data was sourced from a public medical dataset repository which ensures that all patient records were anonymized to maintain confidentiality and privacy. The dataset was downloaded in a comma-separated values (CSV) format and loaded into a Python environment for further processing.

Table 2. Dataset Cardiovascular Disease

No	Gender	Height	Weight	Ap_hi	Ap_lo	Cholesterol	Gluc	Smoke	Alco	Active	Cardio	Age_years	Bmi	Bp_category
0	2	168	62	110	80	1	1	0	0	1	0	50	21.96712	2
1	1	156	85	140	90	3	1	0	0	1	1	55	34.927679	3
2	1	165	64	130	70	3	1	0	0	0	1	51	23.507805	2
3	2	169	82	150	100	1	1	0	0	1	1	48	28.710479	3
4	1	156	56	100	60	1	1	0	0	0	0	47	23.011177	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
68200	2	168	76	120	80	1	1	1	0	1	0	52	26.927438	2
68201	1	158	126	140	90	2	2	0	0	1	1	61	50.472681	3
68202	2	183	105	180	90	3	1	0	1	0	1	52	31.353579	3
68203	1	163	72	135	80	1	2	0	0	0	1	61	27.099251	2
68204	1	170	72	120	80	2	1	0	0	1	0	56	24.913495	2

Data Analysis Methods

The data analysis process began with pre-processing which included the removal of irrelevant features, encoding of categorical variables, and normalization of the data. The categorical variable *bp_category* was encoded using the formula:

$$bp\_category_{encoded} = LabelEncoder(bp\_category) \quad (1)$$

Feature scaling was implemented to normalize the dataset, ensuring all numerical input variables have zero mean and unit variance, which is crucial for effective model training. The formula for feature scaling is:

$$x_{scaled} = \frac{x - \mu}{\sigma} \quad (2)$$

Where, x is the original feature, μ is the mean of the feature and σ is the standard deviation.

For the ensemble method, the Voting Classifier was employed. This classifier combines predictions from three different models: Decision Trees [11], K-Nearest Neighbors [12], and Gaussian Naive Bayes [13]. The ensemble predicts based on the majority voting mechanism [7].

Finally, model performance was evaluated using cross-validation (5-fold) to ensure the model's robustness and generalizability [14]–[16]. The performance metrics calculated were accuracy, precision, recall, and F1-score, each defined as follows [17]–[21]:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3. Result and Discussion

The results from the ensemble Voting Classifier applied to the cardiovascular disease prediction were exceptionally robust across all measured performance metrics. In the 5-fold cross-validation [16], [22], the ensemble approach consistently demonstrated high accuracy, precision, recall, and F1-score [23]–[26]. The calculated average performances for each metric were as **Table 1**.

Table 2: Performance Metrics Across 5-Fold Cross-Validation for the Voting Classifier Algorithm

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	98.95%	98.96%	98.95%	98.94%
K-2	99.02%	99.03%	99.02%	99.02%
K-3	98.87%	98.88%	98.87%	98.86%
K-4	99.09%	99.10%	99.09%	99.08%
K-5	99.10%	99.10%	99.10%	99.09%
$\sum$ ***Avg***	99.01%	99.01%	99.01%	99%

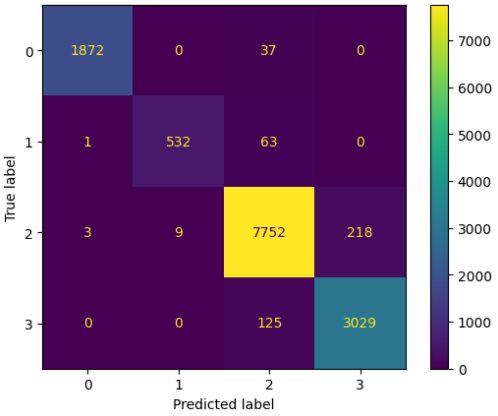

Figure 3: Confusion Matrix of the Voting Classifier

The results from the confusion matrix (**Figure 3**) highlight the model's capability in correctly classifying the states of cardiovascular health. The matrix shows excellent discrimination across all categories, with a notably high true positive rate for class 2 (Hypertension Stage 1), indicating precise detection of this condition.

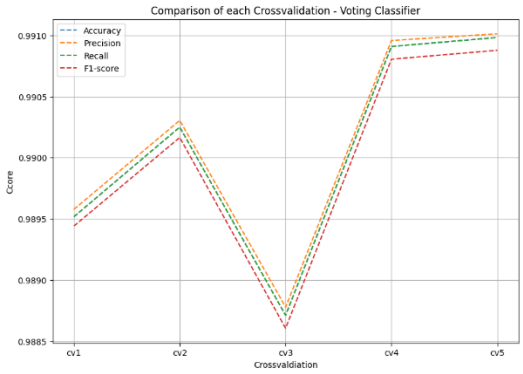

Figure 4: Performance Comparison of Each Cross-validation Fold

The performance trends over the cross-validation folds are illustrated in **Figure 4**, where all metrics track closely, showing little variation and high stability of the model's predictions across different subsets of data. **Figure 5** presents a boxplot summarizing the distribution of scores across the cross-validation folds, underscoring the minimal variance and high reliability of the ensemble model.

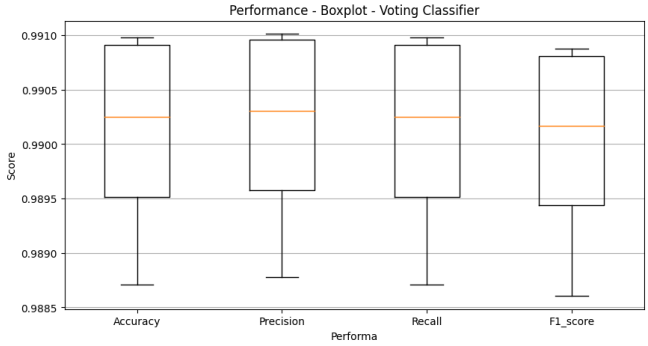

Figure 5: Performance Boxplot of the Voting Classifier

Discussion

The interpretation of these results reveals the Voting Classifier's strength in managing class imbalances and diverse data features effectively, leading to high predictive performance. This aligns with previous research, which suggests that ensemble methods can leverage the strengths of individual classifiers to improve overall accuracy and robustness against overfitting.

The relationship between these results and existing theories in machine learning posits that ensemble models often outperform single classifiers due to their ability to aggregate and reconcile different learning biases and variances. Our findings corroborate this theory, showing that even subtle differences in model predictions can be harnessed to improve the final output significantly.

From a practical standpoint, the high precision and recall of the model are particularly critical in clinical settings where both false positives and false negatives carry significant consequences. The ability of this model to accurately detect various stages of hypertension could aid clinicians in prioritizing interventions and managing patient care more effectively.

Despite these promising results, the research is not without limitations. The primary constraint lies in the homogeneity of the dataset, which was sourced from a single demographic. This could affect the generalizability of the model to other populations with different baseline cardiovascular risk profiles.

For further research, it is recommended to validate these findings across more diverse datasets, possibly incorporating patient data from multiple geographic and ethnic backgrounds to enhance the model's applicability. Additionally, exploring the integration of more sophisticated machine learning techniques, such as deep learning or feature engineering, could potentially uncover more nuanced interactions in the data that may improve predictive performance even further.

4. Conclusion

The ensemble Voting Classifier demonstrated outstanding performance in predicting cardiovascular diseases, with consistently high accuracy, precision, recall, and F1-score metrics across all cross-validation folds. This study has successfully addressed the initial hypotheses by confirming that an ensemble of diverse classifiers significantly enhances predictive reliability and accuracy over individual models. The robustness and low variability observed in the model's predictions reinforce the viability of ensemble learning methods in clinical predictive analytics. Our research contributes to the growing body of knowledge that supports the integration of machine learning in healthcare, specifically in the early detection and management of cardiovascular conditions.

Given the promising results of this study, it is recommended to extend the research to include datasets from varied populations and potentially integrate additional predictors such as genetic or lifestyle factors that could influence cardiovascular risk. Further exploration into hybrid models combining traditional statistical methods with advanced machine learning techniques could also yield improvements in predictive performance and interpretability. For clinical practice, developing a user-friendly interface that integrates this model could facilitate real-time risk assessment and support decision-making in medical settings, ultimately enhancing patient outcomes and healthcare efficiency.

References:

[1] F. T. Admojo and N. Rismayanti, “Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions,” *Indones. J. Data …*, 2024.

[2] R. Setiawan and H. Oumarou, “Classification of Rice Grain Varieties Using Ensemble Learning and Image Analysis Techniques,” *Indones. J. Data …*, 2024.

[3] A. Sinra and H. Angriani, “Automated Classification of COVID-19 Chest X-ray Images Using Ensemble Machine Learning Methods,” *Indones. J. Data Sci.*, 2024.

[4] I. P. A. Pratama, E. S. J. Atmadji, and ..., “Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties,” *Indones. J. …*, 2024.

[5] A. Anitha, "Disease prediction and knowledge extraction in banana crop cultivation using decision tree classifiers," *Int. J. Bus. Intell. Data Min.*, vol. 20, no. 1, pp. 107–120, 2022, doi: 10.1504/IJBIDM.2022.119957.

[6] X. Hu, "K-Nearest Neighbor Estimation of Functional Nonparametric Regression Model under NA Samples," *Axioms*, vol. 11, no. 3, 2022, doi: 10.3390/axioms11030102.

[7] K. Sen, "Heart Disease Prediction Using a Soft Voting Ensemble of Gradient Boosting Models, RandomForest, and Gaussian Naive Bayes," *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170399.

[8] I. Alwiah, U. Zaky, and A. W. Murdiyanto, "Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context," *... J. Data Sci.*, 2024.

[9] M. D. Genemo, "Federated Learning for Bronchus Cancer Detection Using Tiny Machine Learning Edge Devices," *Indones. J. Data Sci.*, 2024.

[10] H. Oumarou and N. Rismayanti, "Automated Classification of Empon Plants: A Comparative Study Using Hu Moments and K-NN Algorithm," *Indones. J. Data ...*, 2023.

[11] I. A. P. Banlawe, "Decision Tree Learning Algorithm and Naïve Bayes Classifier Algorithm Comparative Classification for Mango Pulp Weevil Mating Activity," *2021 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2021 - Proc.*, pp. 317–322, 2021, doi: 10.1109/I2CACIS52118.2021.9495863.

[12] M. Aqib, "Classification of Edge Applications using Decision Tree, K-NN, & SVM Classifier," *2022 IEEE Students Conf. Eng. Syst. SCES 2022*, 2022, doi: 10.1109/SCES55490.2022.9887690.

[13] S. Naiem, "Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing," *IEEE Access*, vol. 11, pp. 124597–124608, 2023, doi: 10.1109/ACCESS.2023.3328951.

[14] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.

[15] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: 10.1109/ASYU50717.2020.9259880.

[16] S. Ortiz-Toquero, "Classification of Keratoconus Based on Anterior Corneal High-order Aberrations: A Cross-validation Study," *Optom. Vis. Sci.*, vol. 97, no. 3, pp. 169–177, 2020, doi: 10.1097/OPX.0000000000001489.

[17] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, "Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, 2023.

[18] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023.

[19] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, and ..., "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *... Artif. Intell. ...*, 2023.

[20] R. A. Azdy, R. F. Syam, E. Faizal, and ..., "Performance Evaluation of Bagging Meta-Estimator in Lung Disease Detection: A Case Study on Imbalanced Dataset," *Int. J. …*, 2023.

[21] A. Maulidinnawati, "Classification Optimization of Skin Cancer Using the Adaboost Algorithm," *… J. Artif. Intell. Med. …*, 2023.

[22] M. H. D. M. Ribeiro, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Appl. Soft Comput. J.*, vol. 86, 2020, doi: 10.1016/j.asoc.2019.105837.

[23] H. Azis and S. R. Jabir, "Chemical Composition and Aroma Profiling: Decision Tree Modeling of Formalin Tofu," *J. Embed. Syst. Secur. …*, 2023.

[24] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi: https://doi.org/10.56705/ijodas.v3i3.54.

[25] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020.

[26] H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020.