Journal of Pioneering Medical Sciences
Received: February 28, 2025 | Accepted: April 5, 2025 | Published: May 5, 2025
Volume 14, Issue 04, Pages 60-66
DOI https://doi.org/10.47310/jpms2025140408

Comparing some Machine Learning Models for Cardiovascular Disease

Sara Noori Mohammad Ali[1*] and Nawzad Muhammed Ahmed[2]
[1,2]Department of Statistics and Informatics, College of Administration and Economics, University of Sulaimani- Sulaymaniyah, Iraq

Author Designation: [1]Assistant Lecturer, [2]Professor

*Corresponding author: Sara Noori Mohammad Ali (e-mail: sara.mohammad@univsul.edu.iq).

Abstract **Background and Aim:** Cardiovascular disease remains a leading cause of morbidity and mortality worldwide, necessitating the development of accurate predictive models for early diagnosis. Therefore, this study aimed to evaluate and compare the performance of three machine learning models-Random Forest, Decision Tree, and K-Nearest Neighbors-in predicting cardiovascular disease based on key risk factors. **Method:** This retrospective study utilized patient data from Shar Hospital in Sulaimaniyah City. The dataset included demographic and clinical risk factors such as age, smoking status, diabetes, hypertension, and family history of cardiovascular disease. The three machine learning models were trained and tested using various data-splitting ratios, and their performance was assessed using accuracy, F1-score, recall, precision, and specificity. Statistical analysis and model validation were conducted using Python in Jupyter Notebook. **Results:** A total of 300 patient records were included in the study. The Random Forest model demonstrated the highest predictive accuracy compared to Decision Tree and K-Nearest Neighbors, consistently outperforming the other models across different training-testing configurations. Feature importance analysis revealed that age and family history were the most influential predictors of cardiovascular disease, whereas gender and marital status had minimal impact. The confusion matrix further confirmed the reliability of the Random Forest model, showing a high number of correctly classified cases with minimal false positives and false negatives. **Conclusions:** The findings indicate that Random Forest is the most effective model for cardiovascular disease prediction, with strong classification performance and high accuracy. The study also highlights the importance of age and family history as dominant risk factors. These results support the application of machine learning in clinical settings for early detection and risk assessment, enabling better-informed medical interventions.

Key Words Cardiovascular Disease Prediction, Machine Learning Models, Risk Factors Analysis, Medical Data Classification

INTRODUCTION

Heart disease remains the leading cause of mortality worldwide, accounting for 31% of all deaths globally (World Health Organization, 2021). Despite advancements in medical science, the burden of cardiovascular diseases continues to rise, with an estimated 23.6 million people projected to die from CVDs by 2030 [1]. Early detection and accurate prediction of heart disease risk factors are crucial for timely intervention and improved patient outcomes. In recent years, Machine Learning (ML) techniques have emerged as powerful tools for predictive modeling in healthcare, particularly in the realm of cardiovascular disease diagnosis [2,3]. ML algorithms can effectively analyze complex, high-dimensional medical data and uncover hidden patterns that may elude traditional statistical methods [4]. By leveraging the predictive capabilities of ML, healthcare professionals can make more informed decisions, stratify risk, and personalize treatment plans for patients at risk of heart disease.

Among the various ML algorithms, Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN) have gained significant attention for their potential in predicting heart disease. RF is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting [5]. DT algorithms, such as C4.5 and CART, create a flowchart-like tree structure to model decisions and their possible consequences [6]. KNN is a non-parametric algorithm that classifies new instances based on the majority class of their k-nearest neighbors in the feature space [7]. While these algorithms have shown promising results in heart disease prediction, their comparative performance and optimal implementation remain areas of active research.

The selection of appropriate ML algorithms and their optimal configuration are critical factors in developing accurate and reliable heart disease prediction models. Comparative studies have been conducted to evaluate the performance of RF, DT, and KNN in this context. For instance, Amin *et al.* [8] compared seven ML algorithms, including RF, DT, and KNN, for heart disease prediction using the Cleveland Heart Disease dataset. Their results showed that RF achieved the highest accuracy of 89.7%, followed by DT (86.9%) and KNN (84.8%). Similarly, Mohan *et al.* [9] evaluated the performance of RF, DT, and KNN on the Statlog Heart Disease dataset, with RF demonstrating the best accuracy of 88.7%. These studies highlight the potential of RF as a robust algorithm for heart disease prediction. However, the performance of ML algorithms can vary depending on the dataset, feature selection techniques, and hyperparameter tuning. Feature selection is a crucial preprocessing step that identifies the most informative attributes for heart disease prediction, reducing dimensionality and improving model efficiency [10].

Wrapper, filter, and embedded methods have been employed for feature selection in heart disease datasets [11]. Hyperparameter tuning involves optimizing the algorithm's parameters to enhance its performance on a specific dataset [12]. Techniques such as grid search, random search, and Bayesian optimization have been utilized to find the optimal hyperparameters for RF, DT, and KNN in heart disease prediction [13,14]. The integration of feature selection and hyperparameter tuning with ML algorithms has shown improved predictive performance in various studies. Yet, there is a need for further research to investigate the optimal combination of these techniques for heart disease prediction using RF, DT, and KNN. A comprehensive comparative analysis will provide valuable insights for practitioners and researchers looking to develop efficient and accurate heart disease prediction models.

Moreover, the interpretability and explainability of ML models are crucial considerations in healthcare applications. While RF, DT, and KNN have shown promising results in heart disease prediction, their interpretability varies. DT algorithms produce easy-to-understand decision rules that can be visualized as a tree structure, making them more interpretable compared to other ML algorithms [15]. RF, being an ensemble of decision trees, provides a measure of feature importance, indicating the relative contribution of each attribute to the prediction [15]. However, the complex structure of RF models can make them less interpretable than individual decision trees. KNN, on the other hand, is a simple and intuitive algorithm that makes predictions based on the similarity of instances, but its interpretability may be limited in high-dimensional feature spaces [16]. Developing interpretable and explainable heart disease prediction models is essential for building trust among healthcare professionals and facilitating clinical decision-making.

Notwithstanding the expanding corpus of research on machine learning-based cardiac disease prediction, many gaps in the literature need to be addressed. Although several research have evaluated the performance of RF, DT, and KNN algorithms, there is a deficiency of thorough assessments that examine the influence of feature selection methods and hyperparameter optimization on their prediction accuracy. This study seeks to evaluate the efficacy of three machine learning models-Random Forest, Decision Tree, and K-Nearest Neighbors-in forecasting cardiovascular illnesses.

Research Question

Which machine learning model-Random Forest, Decision Tree, or K-Nearest Neighbors-achieves the highest predictive accuracy?

METHODS

Study Design, Setting, and Data Source

This study is a retrospective analysis using patient data from Shar Hospital in Sulaimaniyah City, Iraq. The dataset includes records of patients diagnosed with cardiovascular disease, incorporating 10 common risk factors (age, gender, marital status, smoking, diabetes, hypertension, chronic cardiovascular disease, dyslipidemia, hypothyroidism, and family history) and one target variable (cardiovascular disease status).

Study Sample and Data Preprocessing

The dataset included 300 patient records, selected based on completeness and relevance to cardiovascular disease. Before analysis, the data underwent preprocessing to ensure its suitability for machine learning applications. The dataset was examined for missing values, and no missing data were found. Additionally, outlier detection methods were applied to identify extreme distributions, but no significant outliers were observed. To assess the significance of each predictor variable, the chi-square test was conducted, ensuring that the selected features contributed meaningfully to cardiovascular disease prediction.

Sample Size

The study utilized a dataset consisting of patient records, which were divided into training and testing sets to evaluate model performance. Various sample sizes, including 100, 200, 300, 400, and 500 records, were tested to determine the most appropriate size for optimal prediction accuracy. After analyzing the performance across different sample sizes, it was observed that using 300 records provided the most balanced and reliable results.

Machine Learning Models

This study employed three supervised machine learning models for predictive analysis: Random Forest, Decision Tree, and K-Nearest Neighbors. The Random Forest model, an ensemble learning method, was utilized for its ability to combine multiple decision trees, thereby enhancing predictive accuracy and minimizing overfitting. The Decision Tree algorithm was applied due to its interpretable structure, which allows data to be split into branches based

on decision rules derived from feature values. The K-Nearest Neighbors algorithm, a non-parametric method, classified new data points by evaluating their proximity to existing data points in the feature space, relying on similarity measures to determine predictions.

Model Training and Evaluation

The dataset was partitioned into training and testing sets using multiple configurations, including training-to-testing ratios of 95-5, 90-10, down to 10-90 and 5-95, to determine the optimal model performance. The models were evaluated using key classification metrics, including accuracy, F1-score, recall, precision, and specificity. A confusion matrix was used to assess the classification performance, allowing for the identification of true positives, true negatives, false positives, and false negatives. To further interpret model behavior, feature importance analysis was conducted for the Random Forest model, identifying the most influential predictors in cardiovascular disease classification.

Statistical Analysis

All analyses were performed using Python (Jupyter Notebook 6.5.4), with Anaconda serving as the primary data science platform. Cross-validation techniques were implemented to validate the model performance, ensuring the reliability of the results. Additionally, statistical significance testing was conducted to confirm the robustness of the predictive models and to determine the impact of different feature variables on classification outcomes.

Ethical Considerations

This study adhered to established ethical research standards, ensuring the confidentiality and anonymity of patient data throughout the analysis. All personally identifiable information was removed before data processing to maintain privacy and comply with ethical guidelines. Since the dataset was obtained directly from the hospital and did not involve direct interaction with patients, obtaining an ethical approval code was not required. The research was conducted in accordance with institutional and regulatory standards for handling sensitive medical data, ensuring that no ethical breaches occurred during the study.

RESULTS

The results of this study provide a comparative evaluation of three machine learning models-RF, DT, and KNN-in predicting cardiovascular disease. The findings are presented in three sections: model performance metrics, feature importance analysis, and demographic distribution of key risk factors.

Model Performance Metrics

The accuracy of each model was assessed using different training-testing splits. As shown in Table 1, the RF model consistently outperformed the other two models across all training-testing ratios, achieving the highest average accuracy of 90.78% when sum accuracy was divided by 10.

Similarly, as presented in Table 2, when accuracy was calculated over 20 trials, RF remained the best-performing model with an average accuracy of 88.94%. The DT model showed moderate accuracy, whereas KNN had the lowest predictive performance across all trials. The results confirm that RF is the most effective machine learning algorithm for cardiovascular disease prediction within this dataset.

Feature Importance Analysis

To understand the impact of individual risk factors on cardiovascular disease prediction, a feature importance analysis was conducted using the RF model. The results, visualized in Figure 1, highlight that age and family history were the two most influential predictors, followed by

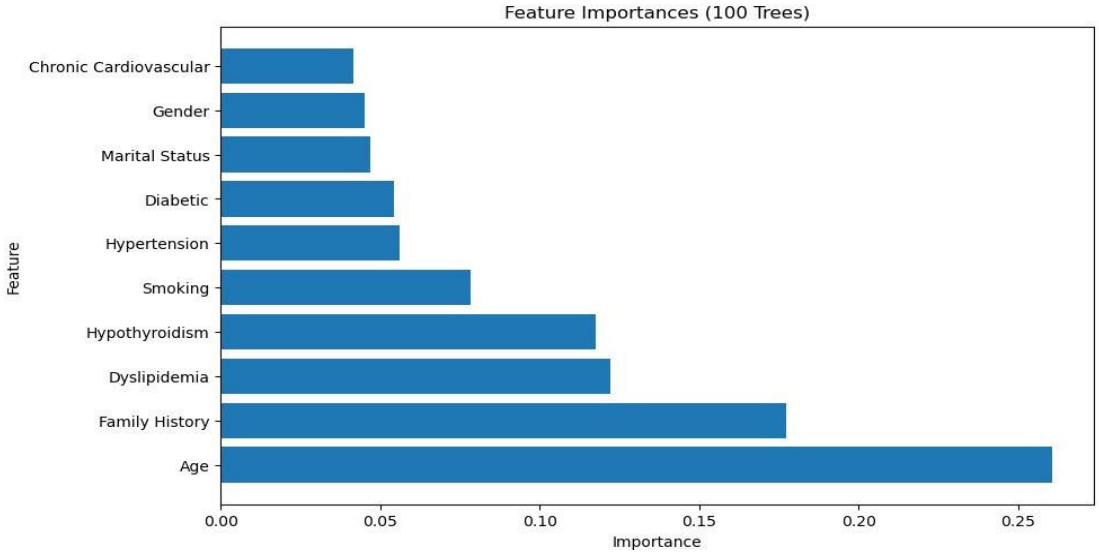

Figure 1: Feature importance analysis of risk factors in cardiovascular disease prediction

Table 1: Accuracy comparison of machine learning models across 10 training-testing splits

N		95-5	90-10	85-15	80-20	75-25	70-30	65-35	60-40	55-45	50-50	average	%
100	DT	1	0.8	0.8666	0.85	0.88	0.9	0.9428	0.875	0.9111	0.84	0.88655	88.655
	KNN	0.8	0.9	0.8667	0.9	0.88	0.9	0.8857	0.9	0.8889	0.88	0.88013	88.013
200	DT	0.9	0.9	0.9333	0.95	0.92	0.8833	0.9	0.9	0.9	0.89	0.90766	90.766
	KNN	0.9	0.85	0.8667	0.85	0.86	0.85	0.8143	0.8125	0.8222	0.82	0.84457	84.457
300	DT	0.8666	0.9	0.9333	0.9333	0.8933	0.8888	0.7905	0.9083	0.8148	0.82	0.87489	87.489
	KNN	0.8	0.8	0.8222	0.8333	0.8133	0.8	0.7619	0.7583	0.7852	0.78	0.79542	79.542
400	DT	0.85	0.8	0.8333	0.8625	0.85	0.8583	0.85	0.8563	0.8667	0.875	0.85021	85.021
	KNN	0.85	0.8	0.8	0.8125	0.83	0.8	0.8286	0.825	0.8167	0.815	0.81778	81.778
500	DT	0.88	0.84	0.8667	0.86	0.856	0.86	0.8514	0.865	0.8177	0.808	0.85048	85.048
	KNN	0.8	0.8	0.8	0.79	0.8	0.7933	0.8057	0.805	0.7689	0.792	0.79549	79.549
		RF											
N		95-5	90-10	85-15	80-20	75-25	70-30	65-35	60-40	55-45	50-50	average	%
100	10 DT	0.8	0.7	0.8667	0.85	0.88	0.9	0.9429	0.9	0.8667	0.88	0.85863	85.863
	100 DT	0.8	0.8	0.8667	0.9	0.92	0.9333	0.9429	0.925	0.8889	0.9	0.88768	88.768
200	10 DT	1	0.85	0.8667	0.875	0.9	0.8833	0.8429	0.8875	0.9	0.88	0.88854	88.854
	100 DT	1	0.85	0.8667	0.9	0.88	0.8667	0.8429	0.8875	0.8778	0.86	0.88316	88.316
300	10 DT	0.7333	0.8667	0.8889	0.9167	0.8533	0.8556	0.819	0.8667	0.837	0.8933	0.85305	85.305
	100 DT	0.9333	0.9	0.9556	0.95	0.8933	0.8889	0.8762	0.9	0.8741	0.9067	0.90781	90.781
400	10 DT	0.8	0.825	0.8667	0.825	0.82	0.85	0.85	0.8438	0.8444	0.88	0.84049	84.049
	100 DT	0.75	0.8	0.8333	0.8625	0.87	0.8833	0.8714	0.875	0.8944	0.885	0.85249	85.249
500	10 DT	0.8	0.8	0.88	0.87	0.864	0.8867	0.8514	0.84	0.8533	0.864	0.85094	85.094
	100 DT	0.84	0.84	0.8667	0.88	0.88	0.8733	0.8629	0.86	0.8578	0.856	0.86167	86.167

Table 2: Accuracy comparison of machine learning models across 20 training-testing splits

N		95-5	90-10	85-15	80-20	75-25	70-30	65-35	60-40	55-45	50-50	45-55	40-60	35-65	30-70	25-75	20-80	15-85	10-90	5-95	average	%
100	DT	1	0.8	0.8666	0.85	0.88	0.9	0.9428	0.875	0.9111	0.84	0.8393	0.85	0.8615	0.8286	0.8267	0.8375	0.7177	0.8778	0.8842	0.86257	86.2568
	KNN	0.8	0.9	0.8667	0.9	0.88	0.9	0.8857	0.9	0.8889	0.88	0.8929	0.8333	0.8308	0.8286	0.7733	0.8875	0.8942	0.8864	0	0.82254	82.2542
200	DT	0.9	0.9	0.9333	0.95	0.92	0.8833	0.9	0.9	0.9	0.89	0.9009	0.9083	0.9077	0.90714	0.9067	0.9125	0.9177	0.5944	0.7211	0.88174	88.1739
	KNN	0.9	0.85	0.8667	0.85	0.86	0.85	0.8143	0.8125	0.8222	0.82	0.8288	0.8333	0.8462	0.85	0.8533	0.8563	0.8529	0.8556	0.8564	0.84624	84.6237
	DT	0.8666	0.9	0.9333	0.9333	0.8933	0.8888	0.7905	0.9083	0.8148	0.82	0.8242	0.8944	0.8974	0.78095	0.7956	0.8917	0.8745	0.637	0.814	0.85046	85.0455
300	KNN	0.8	0.8	0.8222	0.8333	0.8133	0.8	0.7619	0.7583	0.7852	0.78	0.806	0.8278	0.83077	0.82857	0.8311	0.8292	0.8314	0.8296	0.8316	0.81054	81.0539
400	DT	0.85	0.8	0.8333	0.8625	0.85	0.8583	0.85	0.8563	0.8667	0.875	0.8688	0.875	0.8308	0.8357	0.84	0.8406	0.84706	0.7583	0.8263	0.8434	84.3403
	KNN	0.85	0.8	0.8	0.8125	0.83	0.8	0.8286	0.825	0.8167	0.815	0.8145	0.8083	0.8	0.8036	0.8267	0.7906	0.8	0.8056	0.8342	0.81375	81.3753
500	DT	0.88	0.84	0.8667	0.86	0.856	0.86	0.8514	0.865	0.8177	0.808	0.8109	0.8167	0.8185	0.8229	0.824	0.8225	0.84	0.82	0.6968	0.83037	83.0374
	KNN	0.8	0.8	0.8	0.79	0.8	0.7933	0.8057	0.805	0.7689	0.792	0.8073	0.8067	0.8185	0.8143	0.8213	0.8275	0.7977	0.8044	0.8253	0.8041	80.41
		RF																				
N		95-5	90-10	85-15	80-20	75-25	70-30	65-35	60-40	55-45	50-50	45-55	40-60	35-65	30-70	25-75	20-80	15-85	10-90	5-95	average	%
100	10 DT	0.8	0.7	0.8667	0.85	0.88	0.9	0.9429	0.9	0.8667	0.88	0.875	0.8667	0.7846	0.8857	0.8133	0.85	0.7412	0.8778	0.8842	0.85078	85.0779
	100 DT	0.8	0.8	0.8667	0.9	0.92	0.9333	0.9429	0.925	0.8889	0.9	0.8929	0.8833	0.8769	0.8571	0.8667	0.8625	0.7882	0.8778	0.8842	0.87718	87.7179
200	10 DT	1	0.85	0.8667	0.875	0.9	0.8833	0.8429	0.8875	0.9	0.88	0.8649	0.8417	0.8692	0.9	0.8533	0.8375	0.8235	0.8778	0.8579	0.87427	87.4274
	100 DT	1	0.85	0.8667	0.9	0.88	0.8667	0.8429	0.8875	0.8778	0.86	0.8739	0.875	0.8692	0.9	0.8933	0.8938	0.8941	0.8722	0.8526	0.88188	88.1879
	10 DT	0.7333	0.8667	0.8889	0.9167	0.8533	0.8556	0.819	0.8667	0.837	0.8933	0.897	0.9111	0.8718	0.8857	0.88	0.8667	0.8392	0.8333	0.8526	0.86147	86.1468
300	100 DT	0.9333	0.9	0.9556	0.95	0.8933	0.8889	0.8762	0.9	0.8741	0.9067	0.903	0.8889	0.8923	0.881	0.8622	0.875	0.8392	0.8407	0.8386	0.88942	88.9421
400	10 DT	0.8	0.825	0.8667	0.825	0.82	0.85	0.85	0.8438	0.8444	0.88	0.8281	0.825	0.8654	0.8464	0.8567	0.8656	0.8618	0.8306	0.8316	0.84295	84.2953
	100 DT	0.75	0.8	0.8333	0.8625	0.87	0.8833	0.8714	0.875	0.8944	0.885	0.8507	0.8583	0.8615	0.875	0.8633	0.8688	0.8676	0.8444	0.8447	0.85575	85.5747
500	10 DT	0.8	0.8	0.88	0.87	0.864	0.8867	0.8514	0.84	0.8533	0.864	0.8509	0.8567	0.8369	0.84	0.8267	0.8475	0.8259	0.8444	0.8253	0.84546	84.5458
	100 DT	0.84	0.84	0.8667	0.88	0.88	0.8733	0.8629	0.86	0.8578	0.856	0.8727	0.85	0.8462	0.8457	0.8427	0.855	0.8471	0.86	0.8337	0.85631	85.6305

dyslipidemia, hypothyroidism, and smoking. In contrast, gender and marital status contributed minimally to the model's predictions. These findings indicate that older individuals and those with a family history of cardiovascular disease are at a significantly higher risk, reinforcing existing clinical knowledge on heart disease risk factors.

Classification Performance of the Best Model

The RF model, which demonstrated the highest accuracy, was further evaluated using a confusion matrix, as shown in Figure 2. The matrix reveals that RF had a strong classification performance, with high true positive and true negative rates. Specifically, out of 60 test cases, 50 were correctly identified as having cardiovascular disease (true positives), and 7 were correctly classified as not having the disease (true negatives). The low number of false positives (2) and false negatives (1) indicates the model's robustness in distinguishing between affected and non-affected individuals. The precision, recall, and F1-score values further confirm that RF is the most reliable model for cardiovascular disease prediction in this study.

Demographic Distribution of Key Risk Factors

The demographic distribution of the most important risk factors-age and family history-is presented in Table 3. The cross-tabulation results indicate that cardiovascular disease prevalence increases with age, with the highest number of cases observed in the 70-80 age group, followed by the 50-60 age group. Additionally, individuals with a family history of heart disease were more likely to be diagnosed with cardiovascular disease across all age groups, reinforcing the genetic predisposition to heart conditions. The results emphasize the necessity for targeted screening and early intervention strategies for high-risk populations.

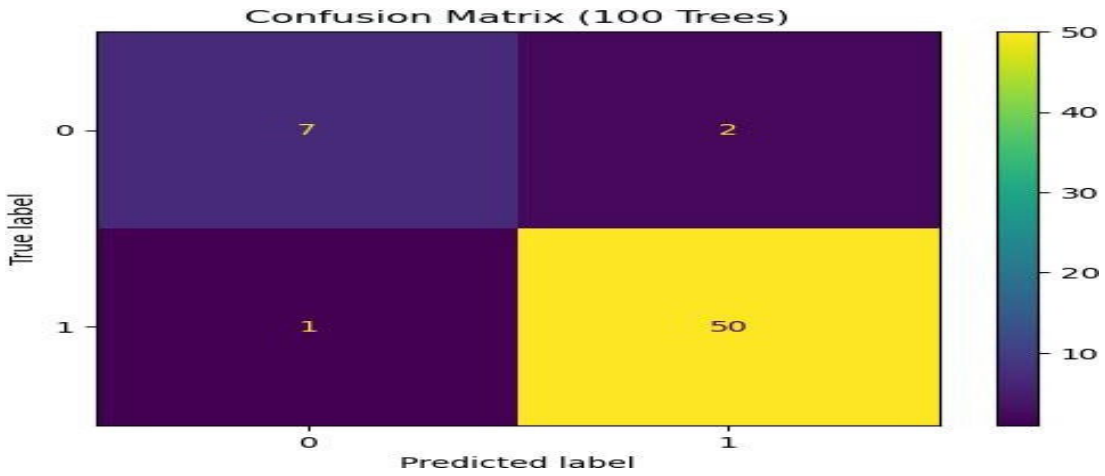

Figure 2: Confusion matrix of the best performing model (Random forest with 100 Decision Trees)

Table 3: Cross-tabulation of age and family history in cardiovascular disease patients

	Family history		
Age	NO	Yes	Total
20-30	2	3	5
30-40	5	7	12
40-50	9	15	24
50-60	30	35	65
60-70	30	25	55
70-80	29	48	77
80-90	16	29	45
90-100	7	10	17
Total	128	172	300

DISCUSSION

The present study aimed to compare the performance of three machine learning models-Random Forest, Decision Tree, and K-Nearest Neighbors-in predicting cardiovascular diseases. Overall, the Random Forest model consistently demonstrated superior accuracy and robustness compared to the other models, making it the most effective for predicting cardiovascular disease in this study population**.**

Cardiovascular diseases are a leading cause of mortality worldwide, and early prediction is crucial for timely intervention and improved patient outcomes [17]. However, the complex interplay of risk factors and the need for personalized risk assessment pose challenges in accurately identifying individuals at high risk. Given these challenges, the development of robust predictive models becomes essential in enhancing cardiovascular disease management. Therefore, we conducted this study to evaluate the performance of different machine learning algorithms in the context of our specific population.

The demographic characteristics of our study participants, with a higher prevalence of cardiovascular disease among older individuals and those with a family history of the condition, align with established risk factors reported in the literature [18]. The age distribution and family history prevalence in our sample are consistent with global epidemiological data, highlighting the representativeness of our study population [19]. Such representativeness enhances the applicability of our findings beyond our study cohort, reinforcing their clinical significance. Among the machine learning models evaluated, the Random Forest model consistently outperformed the Decision Tree and K-Nearest Neighbors models in terms of accuracy across all training-testing splits. This finding is in line with previous studies that have demonstrated the superiority of Random Forest in various healthcare prediction tasks [20,21]. The robustness and ability of Random Forest to handle complex interactions between variables likely contribute to its strong performance in cardiovascular disease prediction. Moreover, its reliability across different evaluation approaches underscores its potential for integration into clinical decision-making frameworks.

The strong classification performance of the Random Forest model, with a high number of correctly identified cases and minimal false positives and negatives, highlights its balanced precision and recall. This finding is particularly important in the context of cardiovascular disease prediction, where both sensitivity and specificity are crucial for effective screening and targeted interventions [22]. By achieving a balance between these metrics, the model minimizes misclassification risks, thereby improving patient outcomes. The ability of the Random Forest model to accurately distinguish between patients with and without the disease suggests its potential utility in risk stratification and personalized management strategies. Our analysis of risk

factor importance revealed that age and family history had the highest impact on cardiovascular disease prediction, while factors such as gender and marital status contributed minimally. This finding aligns with the well-established role of aging and genetic predisposition in the development of cardiovascular diseases [22,23]. This underscores the necessity of prioritizing high-risk individuals based on these key determinants, allowing for early interventions that could mitigate disease progression. The dominance of biological and hereditary factors over sociodemographic variables emphasizes the need for targeted screening and prevention efforts in high-risk populations based on age and family history. This insight can guide the development of risk assessment tools and interventions that prioritize these key risk factors**.**

While our study provides valuable insights into the performance of machine learning models for cardiovascular disease prediction, some limitations should be acknowledged. The retrospective nature of the data and the reliance on electronic health records may introduce potential biases and data quality issues. Additionally, the study was conducted in a single center, which may limit the generalizability of the findings to other populations with different demographic and clinical characteristics. Moreover, we have also discussed concerns regarding data imbalance, potential overfitting in the Random Forest model due to the limited dataset size, and the lack of external validation as key limitations of the study. These additions have been highlighted in yellow in the revised manuscript. To address these limitations, future studies should expand to diverse settings and incorporate real-time data collection methodologies. Future research should aim to validate these results in larger, multi-center cohorts and explore the integration of additional risk factors and biomarkers to further improve the predictive performance of the models.

CONCLUSIONS

Our study demonstrates the superior performance of the Random Forest model in predicting cardiovascular diseases compared to Decision Tree and K-Nearest Neighbors. The high accuracy, robustness, and balanced classification performance of Random Forest highlight its potential as a valuable tool for risk stratification and personalized management in cardiovascular disease. The dominant role of age and family history in predicting cardiovascular disease emphasizes the importance of targeted screening and prevention strategies for high-risk individuals. Further research is needed to validate these findings in diverse populations and explore the integration of machine learning models into clinical decision support systems to improve cardiovascular disease prevention and management.

Disclosures

This study forms part of the PhD thesis conducted within the Statistics and Informatics Department at the College of Administration and Economics, University of Sulaimani, Sulaimani, Iraq.

Acknowledgement

Thanks to all the peer reviewers and editors for their opinions and suggestions and for their support of this research.

Ethical Statement

Ethical approval was not required for this study as the data were obtained directly from Shar Hospital and did not involve direct interaction with patients. All data were anonymized to ensure confidentiality and compliance with ethical research standards.

REFERENCES

[1] Nowbar, Alexandra N., *et al.* "Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors From NCD Risk Factor Collaboration." *Circulation: Cardiovascular Quality and Outcomes,* vol. 12, no. 6, June 2019. https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.118.005375.

[2] Srinivasan, Satish Mahadevan, and Vinod Sharma. "Applications of AI in cardiovascular disease detection-A review of the specific ways in which AI is being used to detect and diagnose cardiovascular diseases." AI in Disease Detection: Advancements and Applications, edited by Singh, Rajesh, *et al.*, United states, IEEE Press, 2025, pp. 123-146. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781394278695.ch6.

[3] Ogunpola, Adedayo, *et al.* "Machine learning-based predictive models for detection of cardiovascular diseases." *Diagnostics,* vol. 14, no. 2, January 2024. https://www.mdpi.com/2075-4418/14/2/144.

[4] Mirza, Bilal, *et al.* "Machine learning and integrative analysis of biomedical big data." *Genes,* vol. 10, no. 2, January 2019. https://www.mdpi.com/2073-4425/10/2/ 87.

[5] Halabaku, Erblin, and Eliot Bytyçi. "Overfitting in machine learning: A comparative analysis of decision trees and random forests." *Intelligent Automation & Soft Computing,* vol. 39, no. 6, January 2019. https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=10798587&AN=182059750&h=VYgprPLJJREWnmXtDAknI55aYvFvoaRXsipPD04oL0gRF53FJ73F506TCevv0ngOPDjBD%2FW4TBg1Og10KvHwzQ%3D%3D&crl=c.

[6] Priyanka, and Dharmender Kumar. "Decision tree classifier: A detailed survey." *International Journal of Information and Decision Sciences,* vol. 12, no. 3, April 2020, pp. 246-269. https://www.inderscienceonline.com/doi/abs/10.1504/IJIDS.2020.108141.

[7] Zhang, Shichao. "Challenges in KNN classification." *IEEE Transactions on Knowledge and Data Engineering,* vol. 34, no. 10, January 2021, pp. 4663-4675. https://ieeexplore.ieee.org/abstract/document/9314060/.

[8] Amin, Mohammad Shafenoor, *et al.* "Identification of significant features and data mining techniques in predicting heart disease." *Telematics and Informatics,* vol. 36, March 2019, pp. 82-93. https://www.sciencedirect.com/science/article/abs/pii/S0736585318308876.

[9] Mohan, Senthilkumar, *et al.* "Effective heart disease prediction using hybrid machine learning techniques." *IEEE access,* vol. 7, June 2019, pp. 81542-81554. https://ieeexplore.ieee.org/abstract/document/8740989/.

[10] Pathan, Muhammad Salman, *et al.* "Analyzing the impact of feature selection on the accuracy of heart disease prediction." *Healthcare Analytics,* vol. 2, November 2022. https://www.sciencedirect.com/science/article/pii/S2772442522000235.

[11] Bashir, Saba, *et al.* "A novel feature selection method for classification of medical data using filters, wrappers, and embedded approaches." *Complexity,* vol. 1, August 2022. https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/8190814.

[12] Diaz, Gonzalo I., *et al.* "An effective algorithm for hyperparameter optimization of neural networks." *IBM Journal of Research and Development,* vol. 61, no. 5, September 2017. https://ieeexplore.ieee.org/abstract/document/8030298.

[13] Valarmathi, R., and T. Sheela. "Heart disease prediction using hyper parameter optimization (HPO) tuning." *Biomedical Signal Processing and Control,* vol. 70, September 2021. https://www.sciencedirect.com/science/article/pii/S1746809421006303.

[14] Rimal, Yagyanath, and Navneet Sharma. "Hyperparameter optimization: a comparative machine learning model analysis for enhanced heart disease prediction accuracy." *Multimedia Tools and Applications,* vol. 83, no. 18, November 2023, pp. 55091-55107. https://link.springer.com/article/10.1007/s11042-023-17273-x.

[15] Sarailidis, Georgios, *et al.* "Integrating scientific knowledge into machine learning using interactive decision trees." *Computers & Geosciences,* vol. 170, January 2023. https://www.sciencedirect.com/science/article/pii/S0098300422001972.

[16] Halder, Rajib Kumar, *et al.* "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications." *Journal of Big Data,* vol. 11, no. 1, August 2024. https://link.springer.com/article/10.1186/s40537-024-00973-y.

[17] Naser, Marwah Abdulrazzaq, *et al.* "A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges." *Algorithms,* vol. 17, no. 2, February 2024. https://www.mdpi.com/1999-4893/17/2/78.

[18] Ellis, Katrina R., *et al.* "Perceptions of rural African American adults about the role of family in understanding and addressing risk factors for cardiovascular disease." *American Journal of Health Promotion,* vol. 33, no. 5, September 2018, pp. 708-717. https://journals.sagepub.com/doi/abs/10.1177/0890117118799574.

[19] Parisi, Rosa, *et al.* "National, regional, and worldwide epidemiology of psoriasis: systematic analysis and modelling study." *BMJ,* vol. 369, May 2020. https://www.bmj.com/content/369/bmj.m1590.abstract.

[20] Speiser, Jaime Lynn, *et al.* "A comparison of random forest variable selection methods for classification prediction modeling." *Expert systems with applications,* vol. 134, November 2019, pp. 93-101. https://www.sciencedirect.com/science/article/pii/S0957417419303574.

[21] Fawagreh, Khaled, and Mohamed Medhat Gaber. "Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach." *Computing,* vol. 102, no. 5, January 2020, pp. 1187-1198. https://link.springer.com/article/10.1007/s00607-019-00785-6.

[22] Dodge, Kenneth A. "Annual Research Review: Universal and targeted strategies for assigning interventions to achieve population impact." *Journal of Child Psychology and Psychiatry,* vol. 61, no. 3, October 2019, pp. 255-267. https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/jcpp.13141.

[23] Dhiman, Paula, *et al.* "Availability and quality of coronary heart disease family history in primary care medical records: implications for cardiovascular risk assessment." *PLoS One,* vol. 9, no. 1, January 2014. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081998.