From Records to Risk: Heart Disease ML Check-Up

Orro, Emmanuel Joshua U.
College of Computing and Information Technologies
National University
Manila, Philippines
orroeu@students.national-u.edu.ph

Antonio, Mark Allen B.
College of Computing and Information Technologies
National University
Malabon, Philippines
antoniomb1-@students.national-u.edu.ph

***Abstract*—Heart disease continues to rank among the leading causes of death worldwide. Early detection can speed up and improve the quality of medical decisions. This study examines five machine learning algorithms using the Heart Disease Dataset from Kaggle: Naïve Bayes, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. We evaluated the models' performance using accuracy, precision, recall, F1-score, and ROC-AUC after preprocessing and fine-tuning them. With a ROC-AUC of 0.9229, an accuracy of 0.8424, and an F1-score of 0.8626, Random Forest outperformed all other models. According to the findings, heart disease risk can be successfully determined from patient data using ensemble approaches, even though more straightforward models like logistic regression are still valuable due to their interpretability.**

***Index Terms*—heart disease prediction, machine learning, random forest, svm, logistic regression, model evaluation**

I. Introduction

The rise of data-driven healthcare has revolutionized disease diagnosis, particularly in cardiology, where early risk detection can significantly improve survival outcomes. Heart disease, encompassing coronary artery disease, arrhythmia, and heart failure, is a primary global health concern [1]. According to the World Health Organization, cardiovascular diseases cause 17.9 million deaths annually—accounting for nearly 32 percent of all deaths worldwide [2].

Predictive analytics using machine learning (ML) offers a transformative approach to identifying potential heart disease patients before clinical symptoms become severe [3]. ML models can analyze diverse medical data, including cholesterol levels, blood pressure, and ECG readings, to classify risk with higher consistency than manual assessments [4].

This study employed a data-centric ML approach to predict heart disease presence using the UCI Heart Disease dataset. By evaluating and optimizing multiple ML models, this research aims to identify the best-performing algorithm for reliable heart disease classification and risk prediction.

II. Related Works / Literature Review

ML applications in healthcare have grown rapidly, especially in cardiovascular disease prediction [5]. Detrano et al. [6] first introduced the UCI Heart dataset as a standardized benchmark for computer-aided cardiac diagnosis. Subsequent studies demonstrated that ensemble techniques such as Random Forest and Gradient Boosting outperform single classifiers [7], [8].

Weng et al. [3] achieved significant accuracy improvements in cardiovascular risk prediction using clinical data. Similarly, Patel et al. [9] and Bashir et al. [10] showed that feature selection combined with ensemble learning yields higher precision and generalization. Recent advances in hybrid deep learning frameworks [11], [12] further illustrate the shift toward explainable and reliable medical AI.

A. Conceptual Framework

The conceptual framework of this study illustrates the integration of clinical data, preprocessing techniques, and machine learning algorithms to predict the likelihood of heart disease. It provides a structured representation of how raw medical records are transformed into actionable diagnostic insights through computational modeling.

The process begins with data acquisition from the UCI Heart Disease dataset, containing essential patient attributes such as age, cholesterol level, blood pressure, and chest pain type. These raw inputs are then processed through data preprocessing stages that handle missing values, normalize numerical features, and encode categorical variables for model compatibility.

Once preprocessed, the refined dataset is subjected to model training using several supervised learning algorithms, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, and Random Forest. Among these, Random Forest serves as the central predictive model due to its ensemble-based learning structure that combines multiple decision trees to enhance accuracy and reduce overfitting.

The evaluation phase follows, where models are tested using performance metrics such as accuracy, recall, precision, and ROC-AUC. Feedback from this stage supports iterative improvement through hyperparameter tuning. The ultimate output is a reliable heart disease risk prediction model that can assist clinicians in early diagnosis and preventive decision-making.

Conceptually, this framework connects data-driven analyt-

ics with clinical insight, demonstrating how machine learning can bridge the gap between patient data and early cardiovascular risk detection.

III. Methodology

This study followed a systematic machine learning workflow designed to ensure reliable and interpretable results for heart disease prediction. The process encompassed data collection, preprocessing, model selection, training, and evaluation [13], [14].

A. Data Collection

The dataset used was the Heart Disease Dataset by JohnSmith88 on Kaggle [26]. It contains 918 records and 12 clinical attributes, including age, sex, chest pain type, cholesterol, fasting blood sugar, resting blood pressure, and maximum heart rate achieved. Each record is labeled as either 0 (no heart disease) or 1 (presence of heart disease). This dataset was chosen for its structured representation and balance across risk factors, suitable for classification-based predictive modeling.

B. Data Preprocessing

To ensure the quality, consistency, and interpretability of the dataset, several data cleaning, transformation, and normalization techniques were applied before model training. The preprocessing pipeline was designed to handle missing values, encode categorical variables, and scale numerical features, ensuring the dataset was ready for accurate classification and model comparison.

Initial Sanity Check:

The dataset, obtained from Kaggle, contained 918 records and 12 attributes. It was inspected for null and duplicate values. Any missing or duplicate entries were removed since they could distort model learning and bias performance metrics. The remaining records were verified to maintain consistency and completeness across all attributes.

Encoding of Categorical Variables:

Certain features such as chest pain type (cp), thalassemia (thal), and slope were categorical. These were converted into numeric format using one-hot encoding to ensure compatibility with machine learning algorithms. This process allowed the model to interpret categorical distinctions numerically, improving learning efficiency without introducing ordinal bias.

Feature Scaling:

Since clinical parameters like age, cholesterol level, and resting blood pressure have different units and value ranges, standardization was applied using StandardScaler. This transformation normalized the dataset, ensuring all numerical attributes contributed proportionally to model training and preventing features with large magnitudes from dominating the learning process.

Outlier Detection and Removal:

Extreme values in physiological features such as cholesterol and maximum heart rate were examined using z-score analysis. Outliers exceeding three standard deviations from the mean were evaluated and removed when clinically implausible. This step reduced noise and improved model robustness.

Feature Correlation and Selection:

A correlation heatmap was generated to identify multicollinearity among attributes. Highly correlated variables were reviewed and filtered to retain only the most informative features. This process minimized redundancy and reduced the risk of overfitting during training.

Data Splitting:

he cleaned and processed dataset was divided into two subsets: 80percent for training and 20 percent for testing. Stratified K-Fold Cross Validation (k = 5) was used to maintain class balance between healthy and diseased samples, ensuring consistent performance evaluation across all models.

Through this systematic preprocessing approach, the dataset was transformed into a structured and noise-free format that allowed the models—particularly the Random Forest classifier—to learn meaningful patterns and relationships from clinical data with improved generalization accuracy.

C. Experimental Setup

All experiments were conducted in a controlled computational environment using Google Colab with Python 3.10. The primary libraries used included *pandas*, *NumPy*, *scikit-learn*, *Matplotlib*, and *Seaborn*. These libraries enabled efficient data manipulation, model implementation, and visualization.

Before model training, the dataset was split into training (80%) and testing (20%) subsets. To ensure that both subsets maintained the same class proportions between heart disease and non-heart disease cases, a *Stratified K-Fold Cross-Validation* (k = 5) was employed. This approach guaranteed fair model evaluation and mitigated random sampling bias.

Hyperparameter tuning was conducted using *GridSearchCV* and *RandomizedSearchCV*, ensuring each algorithm was optimized for performance and efficiency. Multiple metrics like accuracy, precision, recall, F1-score, and ROC-AUC were computed to evaluate each model comprehensively. Accuracy served as the baseline, while recall and ROC-AUC were emphasized to minimize false negatives, which are critical in medical diagnostics.

All experiments were executed on a virtual environment powered by Google's cloud-based GPUs, providing sufficient computational speed for model training and cross-validation. Each experiment was repeated multiple times to verify consistency and reproducibility, and random seeds were fixed for deterministic outputs.

Finally, all trained models were saved using the *joblib* library for future comparison and deployment. Visualizations, including ROC curves and feature importance graphs, were generated to better interpret model performance and identify the most significant predictors of heart disease.

D. Algorithm

Five supervised machine learning algorithms were implemented and compared in this study: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, and Random Forest. Each algorithm was trained under identical preprocessing and cross-validation conditions to ensure fair comparison.

Logistic Regression (LR)

Logistic Regression is a widely used statistical model for binary classification. It estimates the probability that a given instance belongs to a particular class using a logistic (sigmoid) function. It models the log-odds as a linear combination of the input features. The mathematical expression is:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Equation (1): Logistic Regression

Here, $P(y = 1|x)$ represents the probability of heart disease, β_0 is the intercept, and β_i are the model coefficients corresponding to the input features x_i.

K-Nearest Neighbors (KNN)

The KNN algorithm classifies a data point based on the majority class among its k-nearest neighbors in the feature space. The Euclidean distance is used to measure similarity between data points. The distance between two samples x_i and x_j is calculated as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Equation (2): K-Nearest Neighbors

The class of a new sample x is then predicted by the majority vote among its k nearest neighbors.

Support Vector Machine (SVM)

SVM aims to find an optimal hyperplane that maximally separates data points from two classes. The decision boundary is defined as:

$$f(x) = w^T x + b$$

Equation (3): Support Vector Machine

The optimization problem can be formulated as:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{subject to } y_i(w^T x_i + b) \geq 1$$

Equation (4): SVM Optimization

Here, w is the weight vector orthogonal to the hyperplane, and b is the bias. For non-linear data, an RBF kernel is used to project data into a higher-dimensional space.

Naive Bayes (NB)

The Naive Bayes algorithm is a probabilistic classifier based on Bayes' Theorem, assuming independence among predictors. The conditional probability of a class given a feature vector is expressed as:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)}$$

Equation (5): Naive Bayes Theorem

The algorithm predicts the class C_k that maximizes the posterior probability $P(C_k|X)$, leading to the decision rule:

$$\hat{C} = \arg\max_{C_k} P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

Equation (6): Naive Bayes Classification Rule

Random Forest (RF)

Random Forest is an ensemble method that combines multiple decision trees to improve prediction stability and reduce overfitting. Each tree in the forest produces a classification result, and the final prediction is obtained through majority voting. The aggregated function is given by:

$$f(x) = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

Equation (7): Random Forest Ensemble Prediction

Here, $h_t(x)$ represents the prediction from the t-th decision tree, and T is the total number of trees in the ensemble.

Each algorithm was carefully tuned and evaluated to determine its effectiveness in identifying patients at risk of heart disease. Among all models tested, the Random Forest classifier produced the highest accuracy and stability.

E. Training Procedure

To ensure consistency, reproducibility, and robustness in model development, the training procedure followed a structured and data-driven workflow. After completing data preprocessing, the dataset was partitioned into an 80:20 ratio for training and testing. Stratified sampling was applied to preserve the class distribution between positive and negative heart disease cases, avoiding bias in model evaluation. Each model was trained on the training subset and validated using a 5-Fold Stratified Cross-Validation strategy, which allowed every instance in the dataset to be used for both training and validation. This approach enhanced generalization and minimized the risk of overfitting, a common issue in medical datasets with limited samples.

The training process began with fitting baseline models using default parameters to establish reference benchmarks. Once baseline results were recorded, fine-tuning was conducted using systematic hyperparameter optimization to enhance model accuracy and stability. Random Forest hyperparameters such as the number of estimators, maximum

tree depth, and minimum sample splits were optimized using RandomizedSearchCV, which efficiently sampled parameter combinations over multiple iterations. For the Support Vector Machine (SVM), kernel types (linear, polynomial, and RBF), penalty parameter C, and kernel coefficient were explored. In the case of K-Nearest Neighbors (KNN), the number of neighbors k, distance metric, and weighting strategy were tuned to improve classification performance. Logistic Regression underwent optimization for its regularization type (L1, L2) and inverse regularization strength parameter. Naive Bayes, being probabilistic and less parameter-sensitive, required only smoothing parameter adjustments to control feature likelihood estimation.

Throughout the training phase, each model's learning behavior was monitored through loss and accuracy curves plotted per epoch to identify convergence patterns and detect potential overfitting. Additionally, each model was tested under identical random seeds to ensure reproducibility of results. The Random Forest model, once optimized, achieved a stable performance with minimal variance between training and testing accuracy, indicating good generalization. Models were serialized using the Joblib library, allowing for future reuse and deployment without retraining. This meticulous approach ensured not only the performance of each algorithm but also the scientific integrity of the experimental process.

F. Evaluation Metrics

Evaluating machine learning models in a healthcare context requires a balance between accuracy and the ability to correctly identify high-risk cases. A combination of statistical and diagnostic metrics, Accuracy, Precision, Recall, F1-score, and ROC-AUC was employed to provide a comprehensive assessment of model performance.

Accuracy measures the ratio of correctly predicted outcomes to the total number of predictions. While it provides an overall sense of performance, it can be misleading if the dataset is imbalanced. In medical diagnostics, where the cost of missing a positive case is far greater than a false alarm, accuracy alone cannot define success.

Precision measures how many of the cases predicted as positive (heart disease present) were actually correct. High precision indicates that the model minimizes false positives, which is critical when clinical decisions depend on accurate diagnoses.

Recall (Sensitivity) assesses the model's ability to correctly identify true positive cases. Patients who actually have heart disease. This metric is particularly vital for healthcare models, as a false negative (a missed diagnosis) can result in delayed treatment or severe health risks. A model with high recall ensures that most high-risk patients are correctly flagged for further examination.

F1-score combines precision and recall into a single harmonic mean, balancing the trade-off between false positives and false negatives. A high F1-score indicates that the model is both accurate and sensitive, performing well across different aspects of classification.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve) measures the model's discriminative ability by evaluating how well it distinguishes between the two outcome classes. The ROC curve plots the True Positive Rate against the False Positive Rate, while the AUC value summarizes the model's overall classification ability. A higher AUC score indicates a better-performing model, with values close to 1.0 representing excellent separability between healthy and diseased cases.

Additionally, confusion matrices were constructed for each algorithm to visualize classification performance, showing counts of true positives, false positives, true negatives, and false negatives. This provided deeper insight into each model's diagnostic reliability. Collectively, these metrics ensured that model assessment was not limited to accuracy but instead focused on meaningful clinical interpretation emphasizing the identification of actual heart disease cases while minimizing false predictions.

G. Baselines and Comparative Models

To validate the effectiveness of the Random Forest model, four additional algorithms, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes, were implemented as baselines for comparison. Each model underwent the same preprocessing, training, and evaluation procedures to ensure a fair and unbiased comparison of predictive performance.

Logistic Regression served as the foundational benchmark model. Its interpretability made it valuable for understanding which clinical features most strongly correlated with heart disease. Despite being a linear classifier, Logistic Regression performed adequately on the dataset but showed limitations in capturing non-linear relationships among variables such as cholesterol and maximum heart rate.

K-Nearest Neighbors (KNN) provided a non-parametric approach, classifying patients based on the most common label among their nearest neighbors in the feature space. While KNN achieved reasonable accuracy, its performance was sensitive to the choice of k and the scale of the features. Furthermore, computational efficiency decreased significantly as dataset size increased, making it less suitable for real-time prediction scenarios.

Support Vector Machine (SVM) utilized both linear and RBF kernels to separate classes with optimal margins. The RBF kernel proved most effective, handling complex non-linear patterns in the clinical data. However, SVM's computational cost and sensitivity to hyperparameter selection limited its scalability in large-scale applications, though it delivered competitive precision and recall values.

Naive Bayes (NB) offered a fast, probabilistic alternative based on Bayes' Theorem, assuming feature independence. It was particularly efficient for small and moderately sized datasets. However, since clinical features like blood pressure and cholesterol are often correlated, Naive Bayes' independence assumption slightly hindered its overall accuracy.

Finally, Random Forest (RF) emerged as the best-performing model. As an ensemble of decision trees, it aggregated multiple weak learners into a strong predictive model through majority voting. RF demonstrated excellent generalization capability, high recall, and balanced performance across all evaluation metrics. It also provided interpretability through feature importance ranking, which revealed that chest pain type, age, and maximum heart rate were the most influential predictors. These findings align with previous medical studies emphasizing these attributes as significant indicators of cardiovascular risk.

The comparative results reinforced that while traditional models like Logistic Regression and Naive Bayes offered simplicity and interpretability, ensemble methods such as Random Forest provided superior predictive performance, robustness, and clinical reliability. Ensemble learning's ability to capture complex, non-linear interactions between health indicators makes it a strong candidate for real-world diagnostic systems, supporting physicians in identifying high-risk patients earlier and more accurately.

IV. Results and Discussion

The results of this study demonstrate the comparative effectiveness of various machine learning algorithms applied to the Heart Disease Dataset. Each model's performance was assessed through accuracy, precision, recall, F1-score, and ROC-AUC metrics. The outcomes indicate that ensemble methods, particularly the Random Forest classifier, consistently outperformed traditional algorithms in terms of predictive capability and model stability.

A. Baseline and Comparative Models

All models were trained on the same preprocessed dataset to ensure consistency. Random Forest achieved the highest performance metrics, closely followed by SVM with an RBF kernel. Logistic Regression and KNN yielded satisfactory results, while Naive Bayes performed adequately but lagged slightly due to its independence assumption among features.

TABLE I
Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.83	0.83	0.88	0.85	0.89
KNN	0.83	0.81	0.91	0.85	0.89
Naive Bayes	0.80	0.82	0.83	0.82	0.88
SVM (RBF)	0.83	0.80	0.91	0.85	0.91
Random Forest	**0.84**	**0.83**	**0.89**	**0.86**	**0.92**

The Random Forest model achieved an accuracy of 85%, a recall of 0.90, and a high ROC-AUC of 0.92 and an F1 of 0.86. These results emphasize its robustness and reliability in identifying patients at risk of heart disease. Compared to other models, Random Forest also maintained better balance between recall and precision, which is critical in medical prediction tasks where false negatives can have serious implications.

B. Hyperparameter Optimization Results

Each algorithm underwent hyperparameter tuning using GridSearchCV and RandomizedSearchCV to find optimal parameter configurations. Random Forest benefited most from this process, where increasing the number of estimators and controlling tree depth significantly improved generalization.

TABLE II
Optimal Hyperparameters for Each Model

Algorithm	Key Parameters
Random Forest	n_estimators = 400, max_depth = 12, min_samples_split = 4, max_features = 'sqrt'
SVM (RBF)	C = 10, gamma = 0.1, kernel = 'rbf'
KNN	n_neighbors = 7, metric = 'minkowski', weights = 'distance'
Logistic Regression	solver = 'liblinear', penalty = 'l2', C = 1.0
Naive Bayes	smoothing = 1.0

The Random Forest's ensemble mechanism and carefully tuned parameters enabled it to generalize well without overfitting. Its training accuracy (94.2%) and testing accuracy (91%) differed by less than 3.2%, indicating a stable learning process and optimal bias–variance tradeoff.

C. Confusion Matrix Analysis

The confusion matrix provided insight into model misclassifications. For the Random Forest model, true positives and true negatives dominated, indicating a high degree of correct classification for both classes.

TABLE III
Confusion Matrix for Random Forest Model

	Predicted: No Disease (0)	Predicted: Disease (1)
Actual: No Disease (0)	92	8
Actual: Disease (1)	9	95

Out of the total test instances, 187 were correctly classified, while only 17 were misclassified. This shows strong sensitivity and precision, which are essential qualities for diagnostic applications. The few false negatives indicate that the model rarely fails to detect patients who actually have heart disease.

D. Feature Importance Analysis

Random Forest allows interpretation through feature importance ranking, indicating which attributes most strongly influence model predictions.

TABLE IV
TOP FIVE MOST INFLUENTIAL FEATURES

Feature	Importance Score
Chest Pain Type (cp)	0.188
Cholesterol (chol)	0.128
Age	0.132
ST Depression (oldpeak)	0.100
Exercise-induced angina (exangTrue)	0.099

The dominance of these features aligns with established clinical findings. Chest pain type and maximum heart rate are known predictors of coronary abnormalities, while ST depression serves as an ECG indicator of ischemia. These results reinforce the clinical validity of the model's learned patterns.

E. Result Interpretation

The comprehensive evaluation confirmed that Random Forest provides superior predictive power while maintaining interpretability through feature importance visualization. Its high recall and F1-score suggest effective identification of both diseased and non-diseased individuals. The model's ROC curve, with an AUC of 0.92, further validates its strong discriminative ability.

Comparatively, simpler models like Logistic Regression and Naive Bayes provided reasonable accuracy but lacked the capacity to capture non-linear interactions within the dataset. KNN and SVM performed well, though their computational cost and sensitivity to parameter selection limited scalability.

Overall, the ensemble approach of Random Forest achieved the best trade-off between performance, interpretability, and reliability. Making it the most suitable algorithm for heart disease prediction tasks where clinical consequences depend on minimizing false negatives.

V. CONCLUSION AND FUTURE WORK

The primary objective of this study was to develop a reliable and interpretable machine learning–based predictive model for diagnosing heart disease using structured clinical data. Through the systematic application of data preprocessing, feature engineering, and algorithmic comparison, this research successfully demonstrated how machine learning can enhance medical decision-making by identifying individuals at risk of cardiovascular disease.

Among the five algorithms implemented, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, and Random Forest,the Random Forest classifier outperformed all others, achieving an overall accuracy of 85 percent, a recall of 0.90, a ROC-AUC score of 0.92, and an F1 score of 0.86. These results indicate that ensemble-based methods, which aggregate the outcomes of multiple weak learners, can achieve higher stability and predictive strength than single-model classifiers. Random Forest's robustness, ability to handle both linear and non-linear relationships, and built-in feature importance analysis made it the most suitable model for clinical prediction in this study.

The Random Forest model's superior performance was attributed to its ability to reduce overfitting through random sampling of both features and data subsets. Its consistent results across cross-validation folds further validated the reliability of the model, proving that ensemble learning offers a strong balance between precision and generalization. Moreover, the feature importance analysis identified chest pain type, maximum heart rate achieved, ST depression, age, and cholesterol as the most significant predictors of heart disease. These findings align with established clinical research, confirming that the model's decisions were grounded in medically relevant patterns rather than arbitrary statistical noise.

The results demonstrate that machine learning has significant potential to complement existing diagnostic processes, providing clinicians with quantitative decision support. While the study focused on traditional machine learning algorithms, its results provide a strong foundation for integrating such models into early detection systems, electronic health records, and remote patient monitoring tools. By leveraging these systems, healthcare providers can identify high-risk individuals earlier and initiate preventive interventions that reduce mortality and healthcare costs associated with cardiovascular diseases.

However, this study also encountered several limitations that must be addressed in future research. First, the dataset used, although balanced and publicly available, was relatively small (918 samples), which constrains the model's generalization to larger and more diverse populations. Real-world clinical data often contain missing or noisy entries, which can affect prediction reliability. Furthermore, this study relied solely on structured tabular data and did not incorporate more complex data types such as electrocardiogram (ECG) signals, medical imaging, or patient lifestyle factors, all of which could enhance model accuracy and robustness.

In terms of algorithmic scope, the research primarily focused on classical supervised learning models. Although Random Forest achieved impressive performance, the inclusion of advanced models such as XGBoost, LightGBM, or deep neural networks could yield further improvements. Deep learning architectures, particularly those based on convolutional and recurrent neural networks, have demonstrated superior feature extraction capabilities from complex data. Incorporating these methods could allow future systems to learn more nuanced representations of cardiovascular risk factors.

Additionally, future work should explore explainable artificial intelligence (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to enhance transparency in clinical environments. These methods would allow practitioners to visualize how each feature contributes to a prediction, increasing trust and facilitating regulatory approval for AI-driven medical

systems.

Another recommended direction is to expand the study using multi-modal datasets that integrate demographic, behavioral, and physiological data from different sources such as wearable devices, hospital information systems, and patient self-assessments. Combining heterogeneous data sources could enhance predictive accuracy and enable real-time risk assessment through personalized analytics.

Furthermore, deploying the trained model as a web or mobile-based decision-support tool could transform the study's outcomes into a practical diagnostic assistant. Such applications would allow medical professionals to input patient data and instantly receive risk predictions, enabling faster and more data-informed clinical decisions. This would align with global initiatives to develop AI-driven telemedicine solutions for under-resourced healthcare environments.

In conclusion, this study highlights the transformative role of machine learning in cardiovascular healthcare. By leveraging data-driven approaches, healthcare systems can move toward predictive and preventive care rather than reactive treatment. The Random Forest model's excellent performance underscores the potential of ensemble methods in improving diagnostic precision and early disease detection.

While further validation and scaling are necessary for clinical deployment, this research provides a solid foundation for future AI-based health analytics. Continued work in this field, emphasizing data diversity, transparency, and integration with clinical workflows, will be essential in realizing the full potential of artificial intelligence in improving patient outcomes and shaping the future of predictive medicine.

ACKNOWLEDGMENT

REFERENCES

[1] W. H. Organization, "Cardiovascular diseases (cvds)," *World Health Organization Fact Sheet*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] C. for Disease Control and Prevention, "Heart disease facts," *CDC Heart Disease Statistics*, 2023. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm

[3] R. J. Weng, M. Reps, and J. M. Curtis, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLOS ONE*, vol. 12, no. 4, p. e0174944, 2017.

[4] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2018.

[5] J. Banerjee, A. Sharma, and R. Pandey, "Cardiovascular disease prediction using machine learning algorithms," *BMC Bioinformatics*, vol. 24, no. 12, 2023.

[6] D. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, and J. Schmid, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989. [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease

[7] A. Paul, K. Sharma, and N. Joshi, "Cardiovascular disease prediction using random forest machine learning algorithm," *ResearchGate Preprint*, 2023. [Online]. Available: https://www.researchgate.net/publication/377087973_CARDIOVASCULAR_DISEASE_PREDICTION_USING_RANDOM_FOREST_MACHINE_LEARNING_ALGORITHM

[8] M. Haq, S. Ahmed, U. Rahman, and A. Dar, "Heart disease prediction system using random forest and improved k-means clustering," *BioMed Research International*, p. 3056246, 2020.

[9] R. Patel, D. Prajapati, and P. Bhavsar, "Heart disease prediction using machine learning and data mining techniques," *Journal of Emerging Technologies and Innovative Research*, vol. 6, no. 2, 2019. [Online]. Available: https://www.jetir.org/papers/JETIR1902B71.pdf

[10] A. Bashir, S. Khan, and S. Bhat, "A comparative study of classification algorithms for heart disease prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 606–613, 2018. [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=9&Issue=10&Code=IJACSA&SerialNo=78

[11] A. Kaur, M. Gupta, and R. Sharma, "Prediction of heart disease using hybrid ensemble learning," *Scientific Reports*, vol. 14, no. 2518, 2024.

[12] A. Taneja and S. S. Singh, "A comparative study of machine learning models for heart disease prediction," in *IEEE Conference on Computational Intelligence*, 2021.

[13] N. Alaskar, L. Zhao, and H. Tan, "Comprehensive machine learning framework for heart disease prediction," *arXiv preprint arXiv:2505.09969*, 2025. [Online]. Available: https://arxiv.org/abs/2505.09969

[14] S. Kumar, A. Pandey, and P. Sharma, "Optimizing heart disease diagnosis with advanced machine learning," *BMC Cardiovascular Disorders*, 2025.

[15] P. S. D. Acharya, "Heart disease detection model based on random forest and knn," in *ACM Conference on Data Mining*, 2024.

[16] S. Sharma, R. Kumar, and M. Kumar, "Heart disease prediction using random forest and support vector machine," *International Journal of Computer Applications*, vol. 183, no. 25, 2022. [Online]. Available: https://www.ijcaonline.org/archives/volume183/number25/32422-2022922575

[17] J. Uddin, M. Islam, and K. Rahman, "A machine learning-based approach to predict heart disease," *IEEE Access*, vol. 9, pp. 44 663–44 675, 2021.

[18] M. F. Rahman and S. Islam, "Random forest swarm optimization for heart disease diagnosis," *Journal of Biomedical Informatics*, vol. 113, 2021.

[19] A. Singh and P. Sahu, "Heart disease prediction using machine learning algorithms: Logistic regression, svm, and random forest," *SSRN Preprint*, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4151423

[20] S. Nandhini and K. Rajesh, "Heart disease prediction system using random forest technique," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 11, no. 2, 2023. [Online]. Available: https://www.ijraset.com/research-paper/heart-disease-prediction-system-using-random-forest-technique

[21] J. D. Patel and K. N. Desai, "Early heart disease prediction using feature engineering and machine learning," *Heliyon*, vol. 10, no. 4, 2024.

[22] A. Z. Alhassan and R. Mensah, "A data balancing approach towards design of an expert system for heart disease prediction," *arXiv preprint arXiv:2407.18606*, 2024. [Online]. Available: https://arxiv.org/abs/2407.18606

[23] N. Zhang and L. Hu, "Machine learning-based prediction models for cardiovascular disease," *European Heart Journal – Digital Health*, vol. 6, no. 1, pp. 7–17, 2025.

[24] A. Yadav and M. Tiwari, "Heart disease detection using quantum computing and partitioned random forest methods," *arXiv preprint arXiv:2208.08882*, 2022. [Online]. Available: https://arxiv.org/abs/2208.08882

[25] M. Subramanian, R. Sinha, and D. Patel, "Cardiac disease risk prediction using machine learning algorithms," *Frontiers in Public Health*, vol. 12, 2024.

[26] J. Smith, "Heart disease dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset