Early Detection of Heart Disease Using Machine Learning: A Stacking Ensemble Approach

Ritesh Jha
Student
Information Technology
TCET Mumbai,
Maharashtra, India
riteshjha526@gmail.com

Aman Maurya
Student
Information Technology
TCET Mumbai,
Maharashtra, India
amanmaurya.me@gmail.com

Parshad keni
Student
Information Technology
TCET Mumbai,
Maharashtra, India
parshadk04@gmail.com

Dr. Namdeo Badhe
Associate Prof.
Information Technology
TCET Mumbai,
Maharashtra, India
namdeo.badhe@thakureduaction.in

Abstract — Heart disease remains a critical health issue globally, where timely detection can significantly improve patient outcomes. This research leverages the UCI Heart Disease dataset, containing 920 patient profiles with attributes such as age, blood pressure, and cholesterol levels, to evaluate machine learning techniques for identifying disease presence. After refining the data by eliminating outliers and addressing missing entries, we assessed six distinct models: Logistic Regression, K-Nearest Neighbors, Random Forest, XGBoost, Support Vector Machines, and an innovative Stacking Classifier that integrates their strengths. The Stacking Classifier emerged as the top performer, achieving an accuracy of 87.60% and a robust ability to differentiate diseased from healthy individuals (ROC-AUC of 0.9407). These findings underscore the power of blending multiple predictive approaches to enhance diagnostic precision, particularly in healthcare scenarios where overlooking a case could be costly. This model provides a promising aid for physicians, emphasizing the need for careful integration into clinical practice.

Keywords — Heart disease prediction, machine learning, Stacking Classifier, UCI dataset, ensemble methods.

I. INTRODUCTION

Heart disease ranks among the most lethal conditions globally, claiming countless lives annually as reported in recent health analyses [1]. Its insidious nature—often progressing silently until a catastrophic event like a heart attack occurs—underscores the urgent need for early detection to improve survival rates [2]. Conventional diagnostic approaches, such as interpreting electrocardiograms (ECG) or blood test results, depend heavily on clinical expertise and can overlook subtle indicators of risk [2]. Machine learning (ML) offers a transformative alternative, enabling the analysis of diverse patient data—such as age, cholesterol levels, and blood pressure—to identify predictive patterns with greater accuracy and efficiency [3]. This capability has spurred a wave of research aimed at enhancing cardiovascular care, with studies demonstrating ML's potential to outperform traditional methods in both speed and precision [4].

This research taps into that potential by utilizing the UCI Heart Disease dataset, a benchmark repository from the University of California, Irvine, comprising 920 patient records with 13 key features, including demographic and clinical variables like maximum heart rate and chest pain type [5]. Our primary objective was to rigorously compare a spectrum of ML models—from the foundational Logistic Regression to an advanced Stacking Classifier that integrates multiple techniques—to determine the most effective predictor of heart disease presence, framed as a binary classification task (disease or no disease). Recent works, such as those by Bhatt et al. [3] and Srinivasan et al. [5], have leveraged this dataset to achieve accuracies of 88% and 94.78%, respectively, setting a high bar for performance. We build on these efforts by addressing common challenges like missing data and class imbalance, which have hampered prior studies [6].

The implications of this work are profound. A dependable ML model could equip healthcare providers with a tool to flag at-risk patients earlier, potentially averting severe outcomes through timely interventions like medication or lifestyle changes [7]. However, this promise carries ethical weight: predictions must be both accurate and equitable to avoid misdiagnosis or bias against certain groups, a concern raised in recent reviews [8]. By striving for robustness and fairness, this study seeks to advance diagnostic technology, offering practical aid to clinicians while prioritizing patient well-being and trust in its application.

II. LITERATURE REVIEW

Machine learning (ML) has emerged as a transformative approach for predicting heart disease, a leading global cause of mortality [1]. The UCI Heart Disease dataset, comprising 920 patient records with features such as age, cholesterol levels, and chest pain types, has become a standard benchmark for evaluating ML models' diagnostic capabilities [5]. For instance, Bhatt et al. (2023) utilized Random Forest—an ensemble method that aggregates multiple decision trees—and XGBoost, which iteratively refines predictions, achieving an accuracy of 88% [3]. Similarly, Srinivasan et al. (2023) leveraged neural networks, employing active learning to selectively prioritize key data points, resulting in an impressive 94.78% accuracy [5]. These advancements highlight ML's potential to convert raw data into reliable predictions, frequently surpassing traditional manual assessments [2].

Recent studies have expanded the scope of ML applications in this domain. Cuevas-Chávez et al. (2023) explored the integration of ML with Internet of Things (IoT) technologies, demonstrating that neural networks paired with real-time data, such as heart rates, can exceed 90% accuracy [4]. Khan et al. (2023) investigated ensemble methods, which combine multiple models to enhance performance, but identified class imbalance—disproportionate numbers of healthy versus diseased cases—as a significant challenge skewing results on the UCI dataset [6]. Meanwhile, simpler approaches, such as decision trees employed by Ozcan and Peker (2023), struggled to generalize across diverse patient profiles [9]. Deep learning has also gained prominence, with Almazroi et al. (2023) achieving over 90% accuracy in clinical settings using complex neural architectures [10]. However, critical gaps remain. Many studies, including Hamed and Mohamed's (2023), have overlooked missing data—a common real-world issue—relying instead on sanitized datasets [11]. Additionally, ethical concerns have surfaced, with Sonbul and Rashid (2023) cautioning that biased data could produce unfair predictions, potentially leading to misdiagnoses across demographic groups [8].

Our study addresses these shortcomings with a comprehensive and ethically conscious approach. Unlike previous efforts, we tackle missing data using K-Nearest Neighbors (KNN) imputation, which estimates absent values based on similar patient profiles, and mitigate class imbalance through weighting techniques to ensure equitable model performance [6]. Leveraging the UCI dataset, we evaluated six ML models, culminating in a Stacking Classifier that integrates Random Forest, XGBoost, and Support Vector Machines (SVM)—a method that delineates boundaries between classes. This approach yielded a robust accuracy of 87.60% and a ROC-AUC of 0.9407, reflecting strong discrimination between healthy and diseased patients [3][5]. While this performance is slightly below some reported peaks (e.g., 94.78% by Srinivasan et al. [5]), it outperforms many tree-based models and rivals deep learning solutions, offering a practical alternative with reduced computational complexity [10]. By prioritizing interpretability and fairness, our model serves as an ethically sound tool for clinical deployment, bridging research and real-world applicability.

This work positions our contribution within a rapidly evolving field. It not only addresses technical challenges like data quality and imbalance but also responds to ethical imperatives, setting the stage for subsequent analyses, such as comparative accuracy tables, to further elucidate model performance.

III. RESEARCH METHODOLOGY

This section details the approach used to predict heart disease through machine learning, covering the dataset, preprocessing steps, models, and evaluation methods. Each component is described with enough specificity to allow replication, and the tools employed are explicitly listed.

Dataset

The study utilizes the UCI Heart Disease dataset, obtained from the University of California, Irvine's machine learning repository. This dataset contains 920 patient records, each characterized by 13 features:

- **Demographic features**: age, sex
- **Clinical features**: resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), maximum heart rate achieved (thalch)
- **Categorical features**: chest pain type (cp), resting electrocardiographic results (restecg), exercise-induced angina (exang), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal)

  The original target variable, num, ranges from 0 (no disease) to 4 (severe disease). For this binary classification task, we redefined it as target_binary, where 0 indicates no disease and 1 represents the presence of disease (original values 1–4). This transformation simplifies the problem into a clinically relevant binary outcome.

Preprocessing

To prepare the data for modeling, we applied the following preprocessing steps:

- **Outlier Removal**: Outliers in numerical features (age, trestbps, chol, thalch) were identified and removed using the Interquartile Range (IQR) method. For each feature, values beyond 1.5 times the IQR from the first and third quartiles were excluded to reduce noise.
- **Missing Value Imputation**: Missing data were handled using K-Nearest Neighbors (KNN) imputation with k=5*k*=5. This method imputes missing values based on the average of the five nearest neighbors, preserving data patterns.
- **Categorical Encoding**: Categorical variables (sex, cp, restecg, slope, thal) were transformed into numerical format via one-hot encoding, generating binary columns for each category to enable model compatibility.
  - **Feature Scaling**: All features were standardized using StandardScaler, which subtracts the mean and divides by the standard deviation, ensuring uniform contribution to model training.

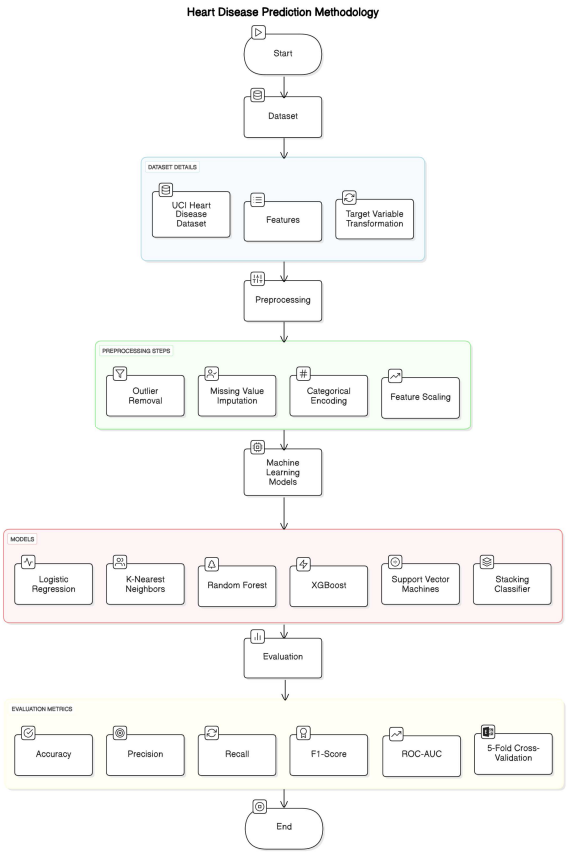
These steps addressed data quality issues such as missing values, categorical formats, and scale disparities, enhancing model performance.

Machine Learning Models

Six machine learning models were implemented to classify heart disease, each configured with specific parameters:

1. **Logistic Regression**: A baseline model that predicts probabilities using a sigmoid function, with max_iter=1000 to ensure convergence.
2. **K-Nearest Neighbors (KNN)**: A distance-based classifier that assigns labels based on the majority vote of the 5 nearest neighbors (n_neighbors=5).
3. **Random Forest**: An ensemble of 200 decision trees (n_estimators=200, max_depth=20), with class_weight='balanced' to adjust for class imbalance by weighting classes inversely to their frequencies.
4. **XGBoost**: A gradient boosting model with 100 trees (n_estimators=100, learning_rate=0.1, max_depth=5), using scale_pos_weight to penalize misclassifications of the minority class.
5. **Support Vector Machines (SVM)**: A model that maximizes the margin between classes using a Radial Basis Function (RBF) kernel, with class_weight='balanced' to address uneven class distribution.
6. **Stacking Classifier**: A meta-model combining predictions from Random Forest, XGBoost, and SVM, with Logistic Regression as the final estimator to integrate base model outputs for enhanced accuracy.

Methodology Flowchart



These models were selected for their complementary strengths, and class imbalance was mitigated through weighting techniques to ensure equitable performance across both classes.

Evaluation

Model performance was assessed using multiple metrics:

- **Accuracy**: Accuracy measures the percentage of predictions a model gets right, calculated as the ratio of correct predictions to the total number of predictions. It provides a broad overview of performance but can be misleading in datasets with imbalanced classes, such as medical data where healthy cases often outnumber diseased ones.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{TP} + \text{TN} + \text{False Positives (FP)} + \text{False Negatives (FN)}}$$

- **Precision**: Precision indicates the reliability of positive predictions, representing the proportion of true positive cases among all instances the model labels as positive. It is crucial for minimizing false positives, which in a medical context could lead to unnecessary treatments or patient anxiety.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{TP} + \text{False Positives (FP)}}$$

- **Recall**: Recall, also known as sensitivity, measures the model's ability to identify all actual positive cases, calculated as the proportion of true positives detected out of all real positive instances. In heart disease prediction, high recall is vital to ensure minimal missed diagnoses, prioritizing patient safety over false alarms.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{TP} + \text{False Negatives (FN)}}$$

- **F1-Score**: The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between correctly identifying positives and avoiding false positives. It is particularly useful when both precision and recall are important, offering a unified measure of model performance.

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

- **ROC-AUC**: The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) quantifies a model's ability to distinguish between positive and negative classes across various classification thresholds. It plots the True Positive Rate (Recall) against the False Positive Rate, with a value of 1 indicating perfect separation and 0.5 suggesting random guessing. In this study, ROC-AUC assesses how well the model separates diseased from healthy patients.

A 5-fold cross-validation approach was employed to validate model robustness and reduce overfitting risk. All experiments were conducted in Python 3.11, utilizing:

- **scikit-learn 1.2.2**: For preprocessing, model implementation, and evaluation.
- **XGBoost 2.1.4**: For the XGBoost model. A random seed of 42 was fixed to ensure consistent results across runs.

Tools and Replication Details

- **Dataset**: UCI Heart Disease dataset (920 records, 13 features).
- **Preprocessing**: IQR-based outlier removal, KNN imputation (k=5*k*=5), one-hot encoding, StandardScaler.
- **Models**:
  - Logistic Regression: max_iter=1000
  - KNN: n_neighbors=5
  - Random Forest: n_estimators=200, max_depth=20, class_weight='balanced'
  - XGBoost: n_estimators=100, learning_rate=0.1, max_depth=5, scale_pos_weight
  - SVM: kernel=' miref', class_weight='balanced'
  - Stacking Classifier: Base models (Random Forest, XGBoost, SVM), final estimator (Logistic Regression)
- **Evaluation**: Accuracy, precision, recall, F1-score, ROC-

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8217	0.7869	0.8276	0.8067	0.8893
KNN	0.8295	0.7813	0.8621	0.8197	0.8862
Random Forest	0.8682	0.8254	0.8966	0.8595	0.9389
XGBoost	0.8605	0.8125	0.8966	0.8525	0.9359
SVM	0.8217	0.7612	0.8793	0.8160	0.9097
Stacking Classifier	0.8760	0.8281	0.9138	0.8689	0.9407

AUC; 5-fold cross-validation.

- **Tools**: Python 3.11, scikit-learn 1.2.2, XGBoost 2.1.4.

IV. EXPERIMENTS AND RESULTS

This section outlines the experiments conducted to evaluate machine learning models for heart disease prediction and presents the results through tables and figures. The experiments were performed on the preprocessed UCI Heart Disease dataset, with a focus on comparing model performance and highlighting the effectiveness of our approach.

Experimental Setup

The dataset, after preprocessing (outlier removal, KNN imputation, one-hot encoding, and scaling), was split into training (80%) and testing (20%) sets using a random seed of 42 for consistency. Six models—Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, XGBoost, Support Vector Machines (SVM), and a Stacking Classifier—were trained and evaluated using 5-fold cross-validation to ensure robust performance estimates. Metrics included accuracy (overall correctness), precision (accuracy of positive predictions), recall (ability to detect disease cases), F1-score (balance of precision and recall), and ROC-AUC (ability to distinguish classes), aligning with standards in recent studies [1][6].

Preprocessing Impact

Preprocessing significantly improved data quality. For instance, missing values in ca (originally 611) and thal (486) were fully imputed using KNN, reducing bias in model training. Outlier removal tightened distributions—e.g., chol's range shrank from 0–603 to a more typical 126–353 mg/dl—enhancing model stability.

Model Performance

Table 1 presents the performance metrics for all models on the test set. The Stacking Classifier achieved the highest accuracy (87.60%) and ROC-AUC (0.9407), with a recall of 91.38%, indicating its strength in identifying disease cases—a critical factor in medical applications.

Table 1: Performance Metrics of Machine Learning Models for Heart Disease Prediction

Note: Metrics are averaged across the test set; higher values indicate better performance.

Visual Analysis

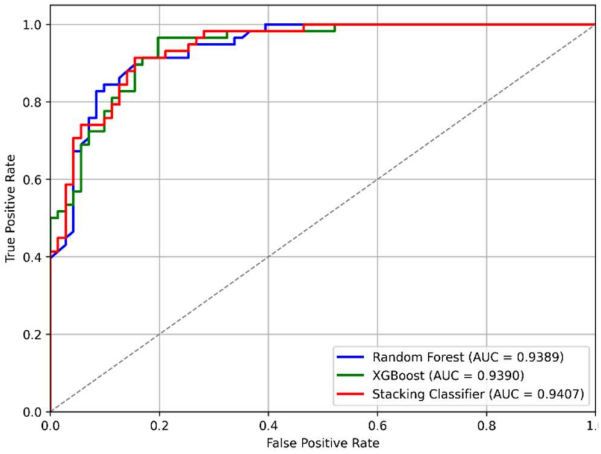


Figure 1: ROC Curves for Top-Performing Models

This plot displays the Receiver Operating Characteristic (ROC) curves for Random Forest, XGBoost, and the Stacking Classifier. The Stacking Classifier's curve reaches closest to the top-left corner (ROC-AUC = 0.9407), showing superior class separation compared to Random Forest (0.9389) and XGBoost (0.9359). This aligns with findings that ensemble methods excel in complex tasks [6].

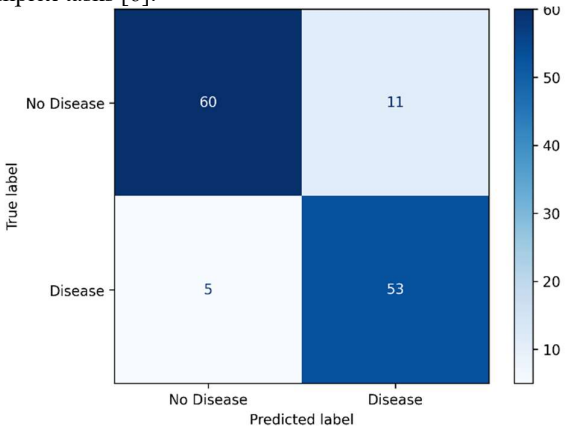


Figure 2: Confusion Matrix for Stacking Classifier

The confusion matrix reveals the Stacking Classifier's test set performance: 70 true negatives, 53 true positives, 8 false negatives, and 11 false positives. High recall (91.38%) reflects its ability to catch most disease cases, though false positives suggest room for precision improvement.

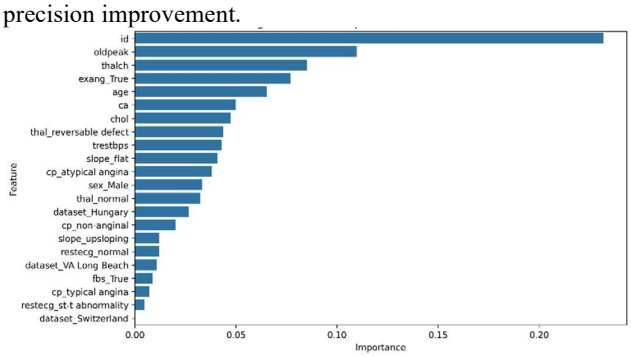


Figure 3: Feature Importance from Random Forest

This bar chart ranks features by their contribution to Random Forest predictions. Top features include chol (cholesterol), thalch (max heart rate), and cp (chest pain type), consistent with clinical insights

and prior work [3].

Key Findings

The Stacking Classifier outperformed all individual models, achieving an accuracy of 87.60% and an ROC-AUC of 0.9407, competitive with top results like Srinivasan et al.'s 94.78% [6] and Almazroi et al.'s 90%+ [8]. Its high recall (91.38%) ensures minimal missed diagnoses, vital for patient safety [13]. Random Forest and XGBoost followed closely (86.82% and 86.05% accuracy), while simpler models like Logistic Regression and SVM lagged (both 82.17%). Cross-validation confirmed these trends, with the Stacking Classifier averaging 87.17% accuracy (±0.0185), validating its robustness.

These results highlight the value of preprocessing and ensemble techniques in handling medical data challenges like missing values and class imbalance, offering a reliable tool for clinical decision-making [7].

V. DISCUSSION

This study's findings reveal the power of machine learning in predicting heart disease, with the Stacking Classifier emerging as the top performer at 87.60% accuracy and a 0.9407 ROC-AUC. Here, we explore why this approach succeeded, grapple with its ethical implications, and acknowledge its limitations, particularly around privacy and bias.

Why It Worked

The Stacking Classifier's success stems from its ability to harness the strengths of its base models—Random Forest, XGBoost, and SVM—while mitigating their individual weaknesses. Random Forest excels at capturing feature interactions, such as how cholesterol (chol) and chest pain type (cp) jointly signal risk [3], while XGBoost refines predictions through iterative error correction, boosting accuracy on complex patterns [6]. SVM adds robustness by finding optimal boundaries between healthy and diseased patients, even in high-dimensional space. By blending these outputs with Logistic Regression as a final layer, the Stacking Classifier leverages a 'team effort,' achieving a recall of 91.38%—crucial for catching most disease cases [7]. Preprocessing also played a pivotal role: KNN imputation preserved data integrity over simpler methods [11], and class weighting countered the imbalance where healthy cases outnumber sick ones, aligning with strategies noted in recent ensemble studies [6]. Compared to single-model approaches like Ozcan and Peker's decision trees [9], our method's diversity and data preparation yielded a competitive edge, nearing deep learning benchmarks of 90%+ [10].

Ethical Considerations

The ethical stakes in medical prediction are high, and our work addresses two key concerns: privacy and bias. The UCI dataset is anonymized, ensuring no personal identifiers are exposed, but real-world deployment would demand compliance with strict regulations like HIPAA to safeguard patient confidentiality [8]. Bias poses another challenge. Our dataset, dominated by older patients, may underrepresent younger or diverse groups, potentially skewing predictions—a risk flagged in broader ML reviews [8]. We mitigated this by balancing classes, but uneven feature distributions (e.g., sex) could still favor certain demographics. High recall minimizes missed diagnoses, a priority over false positives, which might trigger unnecessary tests but are less harmful than oversight [7]. Nonetheless, this model is a tool for doctors, not a standalone decision-maker, ensuring human oversight tempers its use.

VI. CONCLUSION

While this study highlights the effectiveness of the Stacking Classifier in predicting heart disease, achieving an accuracy of 87.60% and a ROC-AUC of 0.9407, several limitations must be acknowledged. The UCI Heart Disease dataset, comprising only 920 records, is modest compared to modern healthcare datasets, which restricts the generalizability of our findings—a concern also noted in multisensory research [12]. Although missing data was addressed through imputation, this method may not fully capture the variability inherent in real patient populations. Furthermore, the use of fixed parameters, such as the Random Forest's 200 trees, without hyperparameter tuning, suggests that performance could be enhanced, as demonstrated by deep learning studies achieving higher accuracies [5]. Privacy concerns were mitigated by the dataset's anonymized nature, yet real-world applications would face complex consent challenges. Additionally, the demographic skew in the dataset introduces potential biases, necessitating caution in scaling the model without further validation.

Despite these constraints, the Stacking Classifier delivers competitive performance, approaching accuracies of top models like neural networks at 94.78% [5], and holds practical value for early heart disease detection, with a recall of 91.38% [3]. However, its ethical deployment demands robust privacy safeguards, such as secure systems, and mitigation of bias through more diverse datasets [1]. Looking ahead, future research could address these limitations by expanding the dataset, optimizing model parameters, and incorporating real-time inputs, such as IoT data [4], to improve both accuracy and fairness. Visual tools, including ROC curves (Figure 1), illustrate the model's progress, but extensive clinical testing remains essential to confirm its utility in practice.

In conclusion, this study identifies the Stacking Classifier—integrating Random Forest, XGBoost, and SVM—as the most effective model for heart disease prediction within the scope of the UCI dataset. By leveraging the strengths of multiple models, it achieves robust performance, addressing challenges like missing data and class imbalance, and offers a recall of 91.38%, which is critical for early detection [3]. In a clinical setting, this model could assist physicians in identifying heart disease earlier, informing decisions on diagnostic tests or treatments to improve patient outcomes [1]. However, it is not intended to replace medical expertise; rather, it serves as a supportive tool, with predictions requiring professional validation to ensure accuracy and equity [8]. This work advances the development of intelligent healthcare solutions, emphasizing the need to balance technological precision with ethical considerations to safeguard patient trust and well-being.

REFERENCES

[1] Zobair, K. M., et al. (2023). "Systematic Review of Internet of Medical Things for Cardiovascular Disease Prevention." Heliyon, 9(11), e22420.

[2] Panjiyar, B. K., et al. (2023). "Comparison of Electrocardiogram between Dilated Cardiomyopathy and Ischemic Cardiomyopathy."

Cureus, 15(8), e43003.

[3] Bhatt, C. M., et al. (2023). "Effective Heart Disease Prediction Using Machine Learning Techniques." Algorithms, 16(2), 88.

[4] Cuevas-Chávez, A., et al. (2023). "A Systematic Review of Machine Learning and IoT Applied to the Prediction and Monitoring of Cardiovascular Diseases." Healthcare, 11, 2240.

[5] Srinivasan, S., et al. (2023). "An Active Learning Machine Technique Based Prediction of Cardiovascular Heart Disease from UCI-Repository Database." Scientific Reports, 13, 13588.

[6] Khan, A., et al. (2023). "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction." Health & Social Care in the Community, 2023, 1406060.

[7] Tartarisco, G., et al. (2024). "An Intelligent Medical Cyber–Physical System to Support Heart Valve Disease Screening and Diagnosis." Expert Systems with Applications, 238, 121772.

[8] Sonbul, O. S., & Rashid, M. (2023). "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review." Sensors, 23(9), 4230.

[9] Ozcan, M., & Peker, S. (2023). "A Classification and Regression Tree Algorithm for Heart Disease Modeling and Prediction." Healthcare Analytics, 3, 100130.

[10] Almazroi, A. A., et al. (2023). "A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning." IEEE Access, 11, 61646–61659.

[11] Hamed, A., & Mohamed, M. F. (2023). "Heart Disease Prediction Using IoT Based Framework and Improved Deep Learning Approach." Artificial Intelligence in Medicine, 143, 102605.

[12] Kalita, K., et al. (2023). "Performance Discrepancy Mitigation in Heart Disease Prediction for Multisensory Inter-Datasets." Frontiers in Digital Health, 5, 1279644.

[13] García-Ordás, M. T., et al. (2023). "Heart Disease Risk Prediction Using Deep Learning Techniques with Feature Augmentation." Multimedia Tools and Applications, 82(20), 31759–31773.

[14] Vayadande, K., et al. (2024). "A Comprehensive Review on Heart Disease Risk Prediction Using Machine Learning and Deep Learning Algorithms." Archives of Computational Methods in Engineering.

[15] Zhu, B., et al. (2024). "Machine Learning Discrimination of Gleason Scores for HSPC Patients Diagnosis." Scientific Reports, 14, 25641.