IJMRSET

e-ISSN:2582-7219

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 12, December 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques

**Mrs. Swetha S, Darshan M.**

Assistant Professor, Department of Computer Science Engineering, CIT, Gubbi, Tumkur, Karnataka, India

U.G. Student, Department of Artificial Intelligence and Data Science Engineering, CIT, Gubbi, Tumkur, Karnataka, India

**ABSTRACT:** Since heart disorders are one of the main causes of death globally, research toward early and precise diagnosis is essential. Machine learning (ML) techniques have emerged as powerful tools for predicting and classifying heart diseases by analyzing medical datasets. This study investigates the use of different machine learning data categorization approaches to more accurately forecast cardiac disorders. We want to train and assess models using 10 algorithms, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, and Neural Networks, by utilizing patient data such as age, gender, blood pressure, cholesterol levels, and other clinical characteristics.

Metrics like accuracy, precision, recall, 9 F1-score, and AUC-ROC are used in the study to compare how well these algorithms perform. Additionally, it looks into how feature selection and data preprocessing methods affect the performance of 23 models. The study determines the best machine learning techniques for accurate and consistent heart disease prediction using this methodology. In the end, the results can improve patient outcomes and lessen the strain on healthcare systems by assisting medical practitioners with early diagnosis and individualized treatment planning.

## I. INTRODUCTION

Heart disorders continue to rank among the top causes of death worldwide, including ailments including coronary artery disease, heart attacks, and heart failure. Early diagnosis and effective management are crucial to reducing morbidity and mortality associated with these conditions. Traditional diagnostic methods, while effective, often require significant time, specialized expertise, and invasive procedures. This has driven the need for innovative, efficient, and non invasive techniques for heart disease prediction.

Machine learning (ML), a subset of artificial intelligence, has demonstrated remarkable potential in addressing challenges in healthcare. By analyzing large volumes of medical data, ML models are able to spot correlations and patterns that human specialists might not see right away. Early diagnosis, risk assessment, and treatment decision-making for heart disease can all benefit from these findings.

The use of different machine learning approaches for the prediction of cardiac illnesses is examined in this study. Using patient information including age, blood pressure, cholesterol, and other vital health metrics, this study examines the effectiveness of various data categorization methods in identifying heart disease risk factors. Finding the best methods for precise and trustworthy prediction is the main objective, offering a basis for resources that can help medical practitioners.

The structure of this paper includes an overview of relevant literature, a detailed methodology outlining the classification algorithms used, and an assessment of these models' performance using actual datasets. The study's findings are intended to improve healthcare systems' predictive capacities, which would enable earlier intervention and better patient outcomes.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

1) This study's utilization of a private HD dataset is one of its main contributions. Egyptian specialized hospitals voluntarily provided 200 data samples between the years 2022 and 2024. We were able to gather around 13 features from these participants.

2) This work deals with the immediate requirement for early HD prediction in Egypt and Saudi Arabia, where the HD rate is rapidly increasing. Through the application of ML classification algorithms to a combined dataset consisting of both CHDD and private datasets, the authors developed a mobile-based app for the instantaneous prediction of heart disease.

3) This work makes an important contribution by combining XGBoost and a semi-supervised model. This method predicts HD accurately using a combined dataset. It is a new method compared to earlier studies. The research's stated goal was to predict HD using the combined datasets and the SF-2 feature subset. The following rates were achieved: 97.57% for accuracy, 96.61% for sensitivity, 90.48% for specificity, 95.00% for precision, 92.68% for F1 score, and 98% for AUC.

4) To understand how the system predicts its outcomes, an explainable artificial intelligence approach utilizing SHAP methodologies has been developed.

5) The use of SMOTE to increase the overall number of balanced cases in the dataset is of additional importance to thisstudy. To improve the performance of heart disease prediction, the suggested method is trained using SMOTE on a balanced dataset.

6) The ML techniques applied in this article were additionally optimized with hyperparameters. We have tuned the hyperparameters for all the ML classifiers. The proposed method got 97.57% accuracy rates with hyperparameters that were optimized when the combined datasets and the SF-2 feature subset were used.

7) Additionally, to identify the classifier that achieves the most accurate HD prediction rate, the study assessed 10 distinct ML classification algorithms. The XGBoost technique was identified as a highly accurate classifier to predict HD after assessing the performance of ten algorithms. The proposed app's capacity for adaptability is shown by applying a domain adaptation method. This shows the ability of the proposed approach to be implemented in various environments and communities, in addition to the initial datasets used in this article. All things considered, this work presents fresh concepts and methods that greatly progress the field of ML-based HD prediction systems. The healthcare sectors that are associated with heart disease incidences in Egypt and Saudi Arabia may both benefit from the

## II. METHODOLOGY

A number of processes are involved in the machine learning methodology for heart disease prediction, ranging from data collection and preprocessing to model training, assessment, and comparison. The following outlines the key stepsin the process:

1. Data Collection

The study makes use of publicly accessible datasets on heart illness, like the Cleveland Heart illness Dataset or comparable datasets, which include patient data such as

- Demographic details (e.g., age, gender)
- Clinical parameters (e.g., blood pressure, cholesterol levels)
- Lifestyle factors (e.g., smoking habits, physical activity)
- Labels that indicate if cardiac disease is present or not

2. Data Preprocessing

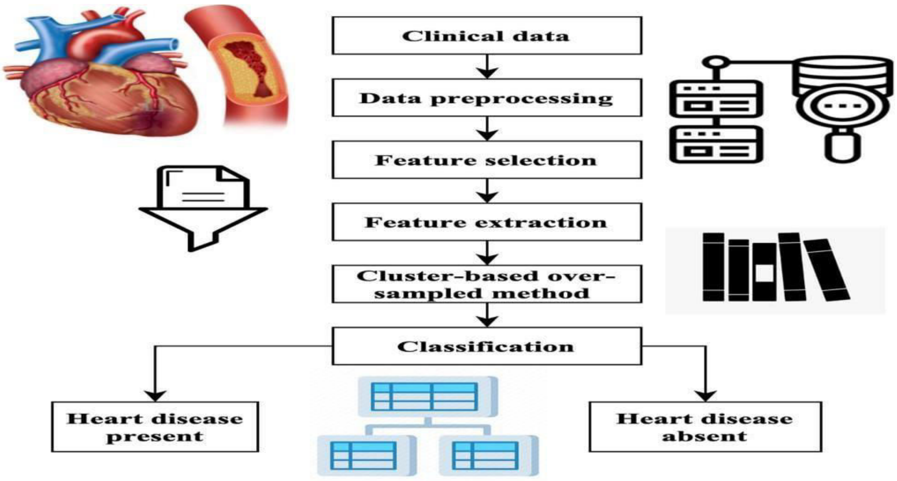
Preprocessing ensures the dataset is clean and suitable for analysis:

Handling Missing Values:** Missing data is imputed using statistical methods (mean, median, or mode) or removed if appropriate.

Feature Encoding: Categorical variables (e.g., gender) are converted into numerical form using techniques like one-hot encoding or label encoding.

-Normalization/Scaling: Continuous features are scaled to a uniform range (e.g., 0–1) to ensure models are not biasedby differing feature magnitudes.

-Feature Selection:Statistical techniques, such as correlation analysis or Principal Component Analysis (PCA), are used to identify the most important features affecting heart disease prediction.



3. Model Selection
Several machine learning algorithms are implemented and compared, including:
- Logistic Regression (LR): A baseline model for binary classification.
- Support Vector Machines (SVM): Suitable for high-dimensional data and capable of creating non-linear decision boundaries using kernel functions.
- Decision Tree (DT): A tree-based approach that provides interpretability.
- Random Forest (RF):An ensemble of decision trees, reducing overfitting and improving accuracy.
-Gradient Boosting (e.g., XGBoost, LightGBM):Techniques that enhance prediction accuracy through iterative refinement.
-Neural Networks (NN): A deep learning approach for capturing complex relationships in the data.

4. Model Training and Hyperparameter Tuning
The pre processed dataset is used to train the chosen models:
Separation of Training and Testing: ** To assess the dataset, it is separated into training (70–80%) and testing (20–30%) groups performance of the model.
- Cross-Validation: To make sure the models perform properly when applied to new data, K-fold cross-validation is employed.
- Hyperparameter Optimization: To adjust model parameters for best results, grid search or random search are used.

5. Performance Metrics
Typical classification metrics are used to assess the models:
- Accuracy: The proportion of correctly classified instances.
- Precision: The ability to avoid false positives.
- Recall (Sensitivity):The ability to identify true positives.
- F1-Score: the precision and recall harmonic mean.
- ROC-AUC Score: evaluates the model's capacity for class distinction.

6. Comparison and Interpretation
- The results of each model are compared to identify the best-performing technique.
- the elements most closely linked to the risk of heart disease.

7. Deployment Considerations**
The best-performing model is analyzed for practical implementation, considering computational requirements, interpretability, and the possibility for real-time application in clinical contexts.
This methodical approach guarantees a thorough examination of machine learning models for the prediction of heart disease, producing trustworthy and useful findings.

## III. ROLE OF DATASETS

Datasets play a pivotal role in the development and success of machine learning models for heart disease prediction.The model's accuracy, generalizability, and capacity to yield useful insights are all directly impacted by the caliber variety, and amount of the datasets. Key elements of the datasets function in this study are listed below.

1. Basis for Training and Assessing Models
- Training Models: Datasets provide the raw input for machine learning algorithms to learn patterns and relationships between features (e.g., age, cholesterol levels) and the variable of interest (e.g., if heart disease is present or not).
- Testing Performance: To assess the model's performance on unseen data, datasets are divided into training and testing subsets.

2. Features and Predictive Insights

Datasets contain features (independent variables) and labels (dependent variables) critical for prediction.
- Features: The model is better able to comprehend the risk factors for cardiac illnesses thanks to clinical data such as blood pressure, cholesterol levels, and patient history.
- Labels: These help the model in supervised learning tasks by indicating whether or not a patient has cardiac disease.

3. Data Quality and Preprocessing Requirements

The dependability of the findings is influenced by the dataset's quality. Challenges such as missing values, noise, and outliers need to be addressed during preprocessing to enhance model performance.
- Imbalanced Data: Many heart disease datasets have unequal distributions of positive (presence of disease) and negative (absence of disease) cases. To deal with imbalance, methods like oversampling, under sampling, or synthetic data synthesis (like SMOTE) are employed.
- Data Diversity:Diverse datasets representing different demographics (age, gender, ethnicity) improve the model's generalizability across populations.

4. Benchmark for Comparison

-Benchmarks are provided by publicly accessible datasets, including the Cleveland Heart Disease Dataset, UCI Heart Disease Dataset, and others. They promote progress in the subject by enabling researchers to contrast their findings with those of earlier investigations.
- Using standardized datasets ensures that results are reproducible and can be validated by other researchers.

5. Feature Importance and Interpretation

Datasets help identify which features have the most predictive power. For instance, heart disease risk factors such as high blood pressure, high cholesterol, or family history might be identified as important contributors using feature selection and importance analysis.

6. Real-World Applicability

Datasets collected from clinical settings provide real-world scenarios for model validation. This ensures that the models can be applied effectively in practical healthcare environments for early diagnosis and treatment planning.

7. Continuous Improvement

As new datasets become available with more features and larger sample sizes, models can be retrained and improved. This iterative process enhances predictive accuracy and ensures that the models remain relevant as medical knowledge evolves. To sum up, datasets serve as the foundation for machine learning models that predict heart disease, allowing medical professionals to create reliable, accurate, and clinically beneficial tools.

## IV. PERFORMANCE EVALUATION METRICS

Assessing machine learning models' performance is essential to comprehending how well they forecast heart conditions. The metrics chosen are determined by the nature of the issue (in this example, classification) and the intended results (e.g., decreasing false negatives in medical diagnosis). The frequently used performance evaluation metrics are listed below.

Accuracy: the percentage of all instances—both good and negative—that were accurately predicted. Significance: Accuracy is helpful, but in unbalanced datasets when one class predominates, it can bedeceptive.
Precision: the percentage of all positive predictions that are actually positive.
Significance: When false alarms must be reduced, a low false-positive rate is indicated by high precision.

The percentage of genuine positives found among all actual positive instances is known as recall (sensitivity or true positive rate).
Significance**:** A high recall lowers false negatives by guaranteeing that the majority of real heart disease patientsare found.
F1-Score: the harmonic mean of recall and precision, weighing the trade-offs between the two.
Significance: Useful when there is an imbalance between precision and recall, providing a single metric that balances both.

- Specificity (True Negative Rate)**:** the percentage of identified true negatives among all actual negative cases. Significance: Evaluates the model's ability to prevent false positives, which is crucial for preventing needless interventions.
- The trade-off between the genuine positive rate (recall) and the false positive rate (1-specificity) across various thresholds is represented graphically by the Receiver Operating Characteristic (ROC) Curve and AUC.
- AUC (Area Under the Curve): a single figure that summarizes the model's class-distinction capabilities. Significance: Better probabilistic predictions are shown by lower log loss.

Logarithmic Loss (Log Loss): a metric for prediction uncertainty that penalizes optimistic but inaccurate forecasts.
Significance: Lower log loss indicates better probabilistic predictions.

- Matthews Correlation Coefficient (MCC):A correlation coefficient between observed and predicted binary classifications, accounting for all confusion matrix elements.
- Significance: Particularly useful for imbalanced datasets as it evaluates the quality of predictions holistically.
- Confusion Matrix: a matrix that summarizes the counts of false positives, false negatives, true positives, andtrue negatives.
- Significance: gives a thorough overview of the model's performance and makes it possible to compute several metrics.
- Time Complexity and Scalability**:** assures the computational efficiency of the model in terms of training and prediction time.

Significance: crucial for implementing models in real-time medical applications where prompt outcomes areessential.
Choosing Metrics The context of the application determines which metrics are used:
Medical Diagnosis Focus: Prioritize recall and F1-score to reduce false negatives while balancing falsepositives.
Practical Implementation**:** Use AUC-ROC for an overall measure of discriminatory abil

## V. CHALLENGES AND LIMITATIONS

One major challenge observed in XAI methods is the variability in explanations provided for the same model. Some XAI techniques, such as LIME, are known to produce slightly different explanations when run multiple times on similar inputs due to their reliance on random sampling processes. This inconsistency can confuse end-users, particularly in high-stakes fields like healthcare or finance, where trust in model outputs is essential. If explanations vary unpredictably, users may question the model's reliability, potentially undermining their confidence in the system. Addressing variability is crucial to improving the robustness and trustworthiness of XAI methods.

Another critical limitation in current XAI approaches is the lack of adaptability to the needs of different user groups. Users of XAI systems range from technical experts, such as data scientists, to non-experts, such as patients or financial customers. Explanations need to be appropriately tailored to meet the diverse backgrounds and knowledge levels of these audiences. For example, highly technical explanations may be informative for data scientists but may overwhelm non-expert users. Studies suggest that more user centric approaches are required, where explanations are not only accurate but also comprehensible to their intended audience. By customizing explanations based on user profiles, XAI can provide clearer, more actionable insights.

Several XAI methods are constrained by specific technical limitations, affecting their general applicability. For instance, methods like SHAP and LIME are often model-agnostic, However, they can be computationally demanding, particularly when dealing with complicated models or big datasets. The explanation process is slowed significantly by SHAP's requirement to calculate Shapley values for every feature, which can be resource-intensive. Additionally, some methods are inherently tied to specific model types, such as decision trees or neural networks, and may not transfer well across other architectures. These restrictions lessen XAI approaches' adaptability and scalability, which may restrict their application in real-time or resource-constrained settings.

There are particular ethical and practical issues when XAI is used in delicate fields like healthcare, finance, andcriminal justice. Explainability is essential for transparency, but it must be weighed against the requirement to safeguard private data. In healthcare, for example, providing detailed model explanations that reveal patient-specific data could lead to privacy risks. Similarly, in financial contexts, explanations might inadvertently disclose proprietary models or confidential data patterns. Ethical concerns also include ensuring fairness and avoiding bias in explanations, as biased explanations may reinforce existing societal inequities. These ethical and privacy considerations create a tension between transparency and data protection, requiring careful handling to ensure responsible deployment of XAIsystems.

## VI. IMPACT OF EXPLAINABLE AI

As AI systems become increasingly integral to critical sectors, the importance of Explainable AI (XAI) is set to grow profoundly. XAI serves as a bridge between complex AI models and end-users, making AI systems more transparent, trustworthy, and accessible. This transparency is essential in building user confidence and ensuring AI technologies are deployed responsibly. These are some important areas where XAI will have a revolutionary effect.

Healthcare: In healthcare, XAI is poised to play a pivotal role. Medical professionals need to understand the reasoning behind AIdriven diagnoses or treatment recommendations to make informed clinical decisions. XAI can enhance diagnostic accuracy and enable personalized treatment by providing insights into how specific patient data points influence AI-generated recommendations. As patients increasingly rely on AI tools for therapy, this degree of openness not only enhances patient outcomes but also builds trust between patients and healthcare professionals. As a result, XAI could pave the way for AI to become a more trusted partner in critical healthcare decisions.

Finance: Financial services are adopting AI at a rapid pace for tasks like credit scoring, fraud detection, and investment forecasting. XAI ensures that AI models used in finance are transparent, which is particularly important given the industry’s stringent regulatory environment. By making these models explainable, financial institutions can demonstrate that their algorithms make fair, unbiased decisions—particularly when assessing credit risk or detecting fraudulent transactions. XAI can also enhance the accountability of financial systems by reducing the potential for unintentional discrimination against specific groups, leading to fairer credit and lending practices and fostering public trust in financial AI solutions.

Legal and Justice Systems: The application of AI in the legal field, especially in risk assessment and sentencing recommendations, has raised concerns about fairness and potential bias. XAI addresses these concerns by shedding light on how these systems arrive at certain conclusions, helping judges, attorneys, and other legal professionals understand the rationale behind AI predictions. Transparent AI models can reveal whether risk assessment tools consider relevant factors without unfairly penalizing individuals based on biases. This accountability is essential to guaranteeing fair treatment throughout legal proceedings and preventing AI from perpetuating preexisting biases. As XAI becomes more integrated into legal frameworks, it will support more ethical and fair decision-making within judicial processes.

As artificial intelligence (AI) continues to shape industries, regulatory bodies and governments around the world are developing frameworks to ensure AI systems operate in a transparent and accountable manner. For example, the European Union's General Data Protection Regulation (GDPR) mandates that individuals have the "right to explanation" regarding automated decisions that significantly impact them. This regulatory requirement ensures that AI systems, especially in sectors such as finance, healthcare, and employment, offer clear, understandable justifications for their automated decisions. Ethical considerations are also integral to the development of AI technologies.

Ethical AI principles, which emphasize fairness, accountability, and transparency, align closely with the objectives of XAI. Having the ability to justify AI choices helps guarantee that these systems behave morally and justly, 12 avoiding discriminatory results and guaranteeing that choices are founded on reasonable, intelligible logic. As regulatory standards evolve, the demand for ethical AI solutions will only grow, and XAI will be central in helping organizations meet these expectations while simultaneously enhancing their public image and reputation.

The adoption and success of AI technologies, particularly in consumer-facing applications, depend heavily on user trust. Users are more likely to engage with and rely on AI systems when they feel confident in understanding the rationale behind AI-generated outputs, such as recommendations, responses, or decisions. XAI can significantly improve user trust by providing transparent and comprehensible explanations for the actions of AI systems.

In applications such as virtual assistants, chatbots, and recommendation systems, XAI can help clarify why a user is receiving certain recommendations or responses. For example, a user may appreciate understanding why a particular product is recommended to them or why a specific answer is given to a query. Such explanations not only make the system seem more personalized but also improve the perceived fairness of the AI, ensuring users feel that their needs and preferences are being respected.

As XAI continues to evolve, it will likely inspire further research into Human-AI Interaction (HAI), specifically focusing on how users from diverse backgrounds interpret and engage with AI explanations. This is crucial as different user groups—whether they are domain experts, casual users, or those with limited technical understanding—may require different forms of explanationAdaptive explanation systems that provide customized, context-specific explanations based on the user's knowledge, preferences, or needs may result from research in this field.

This research also has the potential to influence AI design principles, ensuring that AI systems are not only capable of explaining themselves but also capable of adapting those explanations to the user's level of expertise or engagement. Such adaptability could make AI more accessible to a broader audience, enabling a wider range of users to benefit from its capabilities.

## VII. CONCLUSION

Machine learning models demonstrate a significant ability to predict heart diseases accurately by analyzing large and complex datasets. These models outperform traditional statistical methods by uncovering hidden patterns and relationships among risk factors on them. Ultimately, integrating XAI principles into the design and deployment of AI systems will be paramount to fostering responsible innovation, enabling informed decision making, and ensuring that AI technologies continue to serve the public good. By integrating patient-specific data, machine learning facilitates personalized risk assessments, enabling early interventions and tailored treatment plans that can improve patient outcomes. Machine learning algorithms automate the diagnostic process, making it scalable and efficient for medical professionals to evaluate and track big groups at risk for heart disease. Prediction models heavily rely on characteristics including age, smoking status, blood pressure, cholesterol levels, and physical activity.

These models' performance is further improved via feature engineering and selection. The practicality can be improvedby combining wearable device real-time data, continuous monitoring systems, and explainable AI developments. and accuracy of these forecasting methods. Collaboration between data scientists, healthcare professionals, and policymakers will be crucial for translating these models into real-world applications.

To sum up, machine learning has the ability to completely transform heart disease prediction and prevention, promoting proactive healthcare and lowering the prevalence of cardiovascular diseases worldwide. However, continued efforts in improving model reliability, interpretability, and ethical deployment are essential for its successful integration into clinical practice.

## REFERENCES

[1] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques toBuild Intelligent Systems*. O'Reilly Media, 2019.
[2] D. Powers, "Evaluation: From ROC, Informedness, Markedness, and Correlation to Precision, Recall, and F-measure,"Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.
[3] B. P. Lantz, *Machine Learning with R*. Packt Publishing, 2019. Explainable Artificial Intelligence: A Review and CaseStudy on Model Agnostic Methods

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com