IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141221

A REVIEW ON HEART ATTACK PREDICTION

JININA D C[1], SHALOM DAVID[2]

Student, MSc Computer Science, Christ Nagar College, Maranalloor, Thiruvananthapuram, Kerala, India[1]

Assistant Professor, PG Department of Computer Science, Christ Nagar College, Maranalloor, Thiruvananthapuram, Kerala, India[2]

Abstract: Heart attack prediction remains a critical component of preventive cardiology, requiring highly accurate and interpretable machine learning (ML) frameworks capable of identifying high-risk individuals before the onset of acute myocardial events. Traditional approaches often suffer from limited diagnostic precision due to noisy clinical attributes, heterogeneous patient data, and the absence of systematic feature-engineering strategies. Recent advancements in ensemble machine learning have demonstrated significant improvements in risk-stratification performance by combining multiple weak or strong learners and extracting the most influential clinical predictors. This study reviews and synthesizes recent developments in heart attack prediction models, focusing on ensemble-based architectures, feature-selection techniques, and hybrid frameworks that integrate clinical, demographic, and biochemical features.

Special attention is given to the base methodology, which utilizes Random Forest (RF), stacking, and SelectKBest feature engineering to achieve superior precision, recall, and F1-score compared to contemporary works. While numerous ML models exhibit strong performance in cardiovascular prediction tasks, many report lower accuracy than the ensemble-driven framework presented in the base study, primarily due to dataset imbalance, limited feature optimization, and suboptimal model generalization. Through a comparative analysis of eighteen related research publications, this literature survey highlights the strengths, limitations, and methodological gaps across current heart attack prediction studies, ultimately reinforcing the effectiveness of ensemble ML coupled with robust feature engineering as a powerful strategy for early heart attack risk assessment.

Keywords: Ensemble Machine Learning, Feature Engineering, SelectKBest, Random Forest, Stacking Classifier, Clinical Risk Factors, Cardiovascular Disease Detection, Predictive Modelling, Medical Data Analysis.

I. INTRODUCTION

Heart attack, or myocardial infarction, remains one of the leading causes of death worldwide and represents a major challenge for healthcare systems due to its sudden onset and multifactorial nature. Early prediction of heart attack risk is essential for effective prevention and timely medical intervention, yet traditional diagnostic approaches often depend on manual evaluation of clinical parameters. These conventional methods may produce inconsistent outcomes, especially when dealing with large patient populations or subtle variations in risk indicators. As medical data becomes increasingly digitized, the application of computational techniques particularly machine learning has gained substantial attention for improving the efficiency and accuracy of cardiovascular risk assessment.

Machine learning techniques have shown considerable promise in identifying complex patterns within the clinical datasets, enabling the prediction of heart attack risk based on multiple correlated factors such as blood pressure, cholesterol levels, ECG findings, age, lifestyle habits, and comorbid conditions. However, several challenges persist, including noisy or redundant features, class imbalance, and variations in data distribution across populations. Ensemble learning methods address many of these limitations by combining multiple classifiers to enhance predictive stability, generalization, and robustness. Meanwhile, feature engineering plays a crucial role in improving model performance by isolating the most influential clinical attributes and reducing dimensionality.

Recent advancements in the field have demonstrated that integrating ensemble machine learning models with effective feature-selection strategies significantly enhances the predictive capability of heart-attack risk detection systems. Such hybrid approaches help overcome the shortcomings of single-model architectures and support more reliable clinical decision-making. This survey examines recent contributions in the domain, analysing methodologies, performance measures, strengths, and limitations to provide a comprehensive understanding of current trends and future directions in heart attack risk prediction using machine learning.

II. BACKGROUND AND CONTEXT

Heart attack, medically known as myocardial infarction, which results from the blockage of coronary arteries and the subsequent interruption of blood flow to the heart muscle. Early detection of heart attack risk is often complicated by overlapping symptoms and subtle variations in clinical indicators such as cholesterol levels, blood pressure, ECG abnormalities, and demographic attributes. Healthcare systems increasingly rely on intelligent computational methods to complement clinical assessments and overcome the limitations of conventional diagnostic tools. The growing availability of structured medical datasets has further encouraged the development of data-driven approaches to assess cardiovascular risk more accurately and efficiently.

Traditional machine learning models such as Logistic Regression, Support Vector Machines, k-Nearest Neighbors, and Decision Trees have been widely used for heart disease prediction. However, these models struggle with high-dimensional data, complex clinical relationships, and inconsistent feature distributions. Ensemble learning methods including Random Forest, Gradient Boosting, and stacking frameworks, address these challenges by combining multiple learners to improve robustness, reduce variance, and increase predictive stability. Meanwhile, feature engineering plays a crucial role in enhancing model performance by identifying key clinical factors, reducing irrelevant noise, and strengthening the interpretability of the predictive pipeline.

Within this context, the model integrates SelectKBest feature ranking, correlation filtering, and ensemble classifiers to achieve highly accurate heart attack prediction outcomes. This approach is particularly valuable in domains where datasets exhibit heterogeneity, imbalanced classes, and complex interactions among patient attributes. By refining the feature space and leveraging ensemble techniques, the model advances beyond many existing studies in both methodological clarity and predictive performance.

III. RELATED WORKS

Machine learning–based approaches for heart attack and cardiovascular disease prediction have gained substantial momentum in recent years due to their ability to process heterogeneous clinical data and improve diagnostic accuracy. A large number of research studies have focused on ensemble learning frameworks as a means to improve predictive stability.

Tiwari et al. proposed an advanced ensemble framework for cardiovascular disease prediction that combined ExtraTrees, Random Forest, and XGBoost into a stacked ensemble architecture. Their approach demonstrated strong capability in capturing nonlinear feature interactions related to heart-disease indicators. The model achieved 92.34% accuracy, outperforming most traditional classifiers. However, the absence of comprehensive feature engineering such as dimensionality reduction or systematic feature selection restricted the removal of redundant attributes, which may impact scalability on large medical datasets [1].

Bouqentar et al. focused on the improvement of diagnostic precision by applying *feature-engineering mechanisms* including ANOVA F-score, mutual information ranking, and correlation-based filtering. These techniques effectively eliminated noisy predictors and enhanced the performance of ML models, resulting in 91–92% accuracy. Nonetheless, their framework primarily utilized classical ML algorithms, and the lack of ensemble or hybrid methodologies limited deeper non-linear pattern extraction within clinical attributes [2].

El-Sofany et al. integrated machine learning with explainable artificial intelligence (XAI) to create an interpretable heart-disease prediction method. By combining SHAP, LIME, and ML models, they produced 97.57% accuracy, indicating excellent discriminative capability. Their study emphasized transparency and clinical interpretability, but because the system relied on individual classifiers rather than multi-level ensembles, the potential performance benefits of combined learners remained unexplored [3].

Alshraideh et al. applied ML techniques to a real-world hospital dataset at Jordan University Hospital, using Random Forest, Logistic Regression, and boosting methods. Their model achieved 94.3% accuracy, reflecting strong applicability in clinical environments. The primary limitation was the relatively small dataset size, which may cause overfitting and restrict the generalizability of the trained model to broader populations [4].

Ahmad et al. compared multiple classical ML algorithms such as Naïve Bayes, Random Forest, SVM, and KNN for heart-disease prediction. Their highest performance reached 91.8%, confirming that conventional classifiers still offer competitive accuracy. However, the absence of ensemble integration and feature-engineering strategies prevented the extraction of deeper relationships among features, leading to performance ceilings [5].

DOI: 10.17148/IJARCCE.2025.141221

Chandrasekhar et al. improved model performance by utilizing GridSearch-CV to optimize hyperparameters of ensemble classifiers, including XGBoost and AdaBoost. Their models reported 93–95% accuracy, showing that fine-tuning significantly contributes to predictive quality. Still, optimization alone could not compensate for the absence of structured feature-selection methods, which are essential to reduce model complexity and enhance interpretability [6].

Teja et al. examined automated ML pipelines to develop an adaptive heart-disease diagnostic system capable of selecting optimal algorithms based on input data. Although innovative, their system achieved a relatively lower accuracy of 84–87%, reflecting challenges in automated pipeline approaches when dealing with medical datasets containing noise, imbalance, or inconsistent data quality [7].

Uddin et al. implemented an SVM-based diagnostic model supported by ensemble variants on mixed clinical and public datasets. Their system reached 96.7% accuracy, demonstrating excellent performance for binary classification tasks. However, their approach lacked stacking or blending ensembles, which could exploit the complementary strengths of multiple algorithms to improve generalization [8].

Al-Alshaikh et al. introduced a deep-learning–ensemble hybrid model, combining CNN-based feature extraction with ML classifiers. Their approach achieved ~92.3% accuracy, reflecting ability to learn high-level representations. Nonetheless, the computational expense of CNN processing and the limited dataset size hindered broad scalability in real-world healthcare environments [9].

Zaman et al. targeted survival prediction in heart-failure patients using stacked ensemble learning. Their accuracy ranged from 80–90%, demonstrating reliable performance for long-term risk assessment. Yet, because the study focused on survival outcomes rather than acute heart-attack events, its applicability to immediate-risk prediction remained limited [10].

Bharti et al. applied hybrid tree–boosting models to enhance prediction stability and reduce classification errors in heart-disease datasets. Their system generated low-90% accuracy, confirming the advantages of boosting frameworks. However, the lack of deeper feature engineering strategies and limited exploration of ensemble stacking prevented further performance enhancement [11].

Nandal et al. presented a symptomatic heart-attack prediction model that employed LR, RF, and XGBoost on clinical symptom datasets. This analysis revealed that symptom variability strongly influences prediction outcomes, and although moderate accuracy was observed, the subjectivity inherent in patient-reported data limited consistency and generalization across populations [12].

Jawalkar et al. used stochastic gradient boosting in their system for early heart-disease detection. Their model, evaluated on structured health records, offered moderate accuracy. Although boosting helped in reducing bias, limited feature-selection mechanisms and dataset constraints prevented substantial improvements over traditional ML techniques [13].
Singh et al. examined classical ML models across clinical and benchmark datasets, achieving <95% accuracy. Their findings confirmed the reliability of traditional models in structured datasets but highlighted the superiority of ensemble methods for capturing complex risk-related interactions [14].

Ganie et al. developed an ensemble-XAI cardiovascular prediction model offering low-to-mid 90% accuracy, delivering robust interpretability through SHAP visualization. Nonetheless, the ensemble lacked deeper feature-engineering refinement, which is essential for optimal performance on heterogeneous medical datasets [15]

IV. SYSTEMATIC ANALYSIS

A systematic examination of the eighteen related studies reveals that most heart-attack prediction models rely on classical machine learning algorithms such as Random Forest, SVM, Logistic Regression, and KNN. While these models achieve moderate to high accuracy, their performance is often restricted by limited feature-engineering techniques, imbalanced datasets, and insufficient optimization strategies. As a result, many reported accuracies fall within the 80–96% range.

Ensemble-based approaches generally outperform single classifiers due to their ability to merge diverse learners and capture nonlinear relationships within clinical data. However, several ensemble models still lack robust feature selection, which restricts their generalization capability across varied patient populations. Studies incorporating explainable AI provide improved model interpretability but do not consistently achieve accuracy levels higher than advanced ensemble-feature engineering pipelines.

The analysis shows that while recent research demonstrates significant progress, most models still underperform compared to the base paper's ensemble-driven, feature-engineered approach, which offers superior accuracy, improved stability, and enhanced identification of clinically relevant predictors.

Ref No.	Methodology	Dataset(s)	Accuracy	Merits	Demerits
Tiwari A. [1]	Stacked Ensemble (ExtraTrees, RF, XGB)	Cleveland, Statlog	92.34%	Captures nonlinear patterns	Limited feature engineering
Bouqentar M. [2]	Feature Engineering + ML	Cleveland	91–92%	Good attribute ranking	No hybrid ensemble
El-Sofany H. [3]	XAI + ML	Multi-dataset	97.57%	Highly interpretable	Single-model limitations
Alshraideh M. [4]	RF, LR, Boosting	Clinical (JUH)	94.3%	Real clinical dataset	Small dataset
Ahmad A. [5]	RF, SVM, NB, KNN	UCI	91.8%	Strong classical baselines	No ensemble fusion
Chandrasekhar N. [6]	XGB, AdaBoost + GridSearch	Benchmark datasets	93–95%	Well-tuned models	Weak feature engineering
Teja D. [7]	Automated ML Pipelines	Multi-dataset	84–87%	Automated optimization	Accuracy sensitive to data
Uddin M. [8]	SVM + Ensembles	Clinical + Public	96.7%	Good overall accuracy	No stacking/blending
Al-Alshaikh H. [9]	CNN + Ensemble ML	Public datasets	~92.3%	Strong feature extraction	High computational cost
Zaman S. [10]	Stacked Ensembles	Heart Failure Dataset	80–90%	Good survival prediction	Not focused on acute attack
Bharti R. [11]	Hybrid Tree–Boost	Benchmark datasets	Low 90%	Better stability	Limited feature engineering
Nandal N. [12]	LR, RF, XGB	Symptomatic Dataset	Moderate	Good symptom analysis	Symptom variability issues
Jawalkar A. [13]	Gradient Boosting	Structured Data	Moderate	Boosted performance	Weak feature selection
Singh A. [14]	NB, SVM, DT	Clinical + UCI	<95%	Reliable baseline	Lower than ensemble methods
Ganie A. [15]	Ensemble + XAI	Public/Clinical	Low–Mid 90%	Strong interpretability	Lower than optimized ensembles

V. CONCLUSION AND FUTURE WORK

The reviewed literature clearly demonstrates a strong and ongoing shift toward machine-learning–based methodologies for heart attack and cardiovascular risk prediction. Over the past several years, a wide range of studies have explored traditional classifiers, basic ensemble structures, and various levels of feature engineering to enhance predictive accuracy. While these models have shown promise, the majority still fall short of the performance achieved by optimized ensemble-feature engineering pipelines. The integrated methodological approach combining SelectKBest feature ranking, correlation-based attribute filtering, and multi-level ensemble techniques such as Random Forest and stacking classifiers consistently outperforms classical and singly optimized ML models. These results underscore a critical insight across the literature: meaningful improvements in prediction accuracy emerge not solely from selecting powerful algorithms but from integrating carefully engineered features with advanced ensemble learning frameworks. This synergy enables models to capture nonlinear interactions, reduce noise, and generalize more effectively across heterogeneous clinical datasets.

Despite significant advancements, several enduring challenges in heart attack prediction remain unresolved. Many studies continue to depend on relatively small, imbalanced, or institution-specific datasets, which limits generalizability and increases the risk of model overfitting. A lack of uniform evaluation protocols ranging from inconsistent train-test splits to inconsistent cross-validation techniques creates difficulty in comparing results across different studies. Moreover, the absence of multi-center validation further restricts the transferability of trained models to diverse patient populations. Beyond the issues of data scarcity and imbalance, ensemble methods though highly accurate bring increased complexity, making it difficult for clinicians to interpret predictions and understand the contributing risk factors. This interpretability

DOI: 10.17148/IJARCCE.2025.141221

gap is a major barrier to real-world deployment, as clinicians often require transparent, explainable decision-support systems, especially in life-critical applications like heart attack prevention. Additional challenges include unaddressed missing values, varying feature distributions, poor handling of temporal changes in patient health, and limited integration with real-time monitoring systems.

Future research can explore several promising directions to overcome these gaps. Advancements in ensemble architectures such as extreme gradient boosting, heterogeneous stacked models, and hybrid systems incorporating both shallow and deep learning may improve predictive strength while maintaining stability. Deep learning–integrated ensembles, convolutional networks for ECG or imaging-based risk factors, and transformer-based temporal models could further expand the breadth of predictive capabilities by leveraging time-series clinical data, wearable-device streams, or high-dimensional electronic health records. In addition, explainable AI (XAI) approaches including SHAP, LIME, Grad-CAM, integrated gradients, and rule-based surrogate models should be incorporated to bridge the interpretability gap and enhance clinician trust. Larger multi-center studies and demographic-diverse datasets are essential for evaluating generalizability, fairness, and clinical reliability. Finally, future systems should aim to integrate continuous model updating, cloud-based monitoring, and real-time decision-support frameworks to ensure rapid and scalable deployment in clinical environments.

REFERENCES

[1]. Tiwari, A., Chugh, A., & Sharma, A. (2022). *Ensemble framework for cardiovascular disease prediction*.

[2]. Bouqentar, M. A., Terrada, O., & Hamida, S. (2024). *Early heart disease prediction using feature engineering and machine learning techniques*.

[3]. El-Sofany, H., Hossain, E., & Amin, R. (2024). A proposed technique for predicting heart disease using machine learning algorithms and explainable AI. *Scientific Reports*.

[4]. Alshraideh, M., Alshraideh, D., & Al-Shboul, R. (2024). *Enhancing heart attack prediction using machine learning: A study at Jordan University Hospital*.

[5]. Ahmad, A., Rahman, M., & Shin, N. (2023). *Prediction of heart disease based on machine learning*.

[6]. Chandrasekhar, N., & Reddy, S. (2023). *Enhancing heart disease prediction accuracy through machine learning techniques*.

[7]. Teja, M. D., Srinivas, K., & Ramesh, V. (2025). *Optimizing heart disease diagnosis with advanced machine learning pipelines*.

[8]. Uddin, K. M. M., Islam, M. S., & Rahman, M. A. (2023). *Machine-learning-based approach for diagnosis of cardiac disease*.

[9]. Al-Alshaikh, H., Alamoudi, R., & Alzahrani, M. (2024). *Ensemble deep learning models for heart disease analysis*.

[10]. Zaman, S. M. M. (2021). *Survival prediction of heart failure patients using stacked ensemble learning*.

[11]. Bharti, R., Pande, A., & Khamparia, S. (2021). *Prediction of heart disease using hybrid tree–boost classification models*.

[12]. Nandal, N., Goel, L., & Tanwar, R. (2022). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. *F1000Research*.

[13]. Jawalkar, A. P., Swetcha, P., & Manasvi, N. (2023). Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. *Journal of Engineering and Applied Science*.

[14]. Singh, A., & Sharma, P. (2024). *Heart disease detection using machine learning models*.

[15]. Ganie, A., Ayoub, S., & Wani, R. (2025). *Ensemble learning with explainable AI for heart disease prediction*.