British Journal of Nursing Studies
ISSN: 2755-080X
DOI: 10.32996/bjns
Journal Homepage: www.al-kindipublisher.com/index.php/bjns

BJNS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Predictive AI Framework for Cardiovascular Disease Screening in the U.S.: Integrating EHR Data with Machine and Deep Learning Models

Mustafizur Rahaman[1], Md Al Amin[2], Rahat Hasan[3], S M Tamim Hossain[4], Md Habibur Rahman[5], Ruhul Amin Md Rashed[6]

[1]Doctor of Business Administration (DBA), Westcliff University, USA
[34]College of Business, Westcliff University, USA
[5]Doctor of Management (DM)University, International American University, USA
[26]School of Business, International American University, Los Angeles, California
Corresponding Author: Md Al Amin, **E-mail**: alamin99.edu@gmail.com

| ABSTRACT

Cardiovascular disease (CVD) is the leading global cause of death, with over 18 million fatalities annually. Early and accurate diagnosis is essential to reduce its clinical and economic impact. This study presents an AI-driven framework for the early detection of CVD using structured data from electronic health records (EHRs). The Cleveland Heart Disease dataset was used to train and evaluate multiple supervised machine learning models, including Logistic Regression, Random Forest, SVM, KNN, and XGBoost. Comprehensive preprocessing steps were applied, such as feature normalization, missing value imputation, and one-hot encoding. Model performance was assessed using precision, recall, F1-score, and ROC-AUC, with XGBoost achieving the highest ROC-AUC score of 0.91. To support clinical interpretability, we employed feature importance analysis, ROC curves, and confusion matrices. The study confirms the potential of interpretable AI models to enhance diagnostic accuracy, facilitate early interventions, and integrate seamlessly into clinical decision support systems for proactive healthcare delivery.

| KEYWORDS

Artificial Intelligence, Cardiovascular Disease, Machine Learning, EHR, Logistic Regression, XGBoost, Random Forest, Predictive Analytics, Early Diagnosis, Healthcare AI

| ARTICLE INFORMATION

ACCEPTED: 01 July 2025 **PUBLISHED:** 12 August 2025 **DOI:** 10.32996/bjns.2025.5.2.5

1. Introduction

Cardiovascular diseases (CVDs) encompass a broad spectrum of disorders that affect the heart and blood vessels, including coronary artery disease, myocardial infarction (heart attack), stroke, arrhythmias, and congestive heart failure. These conditions collectively represent the leading cause of mortality globally, accounting for more than 18 million deaths annually—approximately 32% of all global deaths—according to the World Health Organization (2023). The impact of CVD extends beyond human life: in the United States alone, the annual financial burden related to cardiovascular healthcare exceeds $219 billion, factoring in direct medical costs, medications, productivity loss, and hospitalizations.

Despite extensive public health initiatives, early detection of CVD remains a critical challenge, especially in resource-constrained settings or among asymptomatic patients. Traditional diagnostic methods—such as stress tests, angiograms, and electrocardiograms—are effective but often reactive, invasive, and limited in their ability to forecast disease before symptoms manifest. Thus, there is a growing imperative for more proactive, scalable, and data-driven tools that can support early identification of individuals at risk.

Artificial Intelligence (AI), particularly machine learning (ML), is rapidly transforming the landscape of modern healthcare by offering new avenues for predictive analytics, risk stratification, and clinical decision support. ML algorithms have demonstrated

the capacity to uncover hidden patterns and nonlinear relationships in large-scale datasets—capabilities that far exceed conventional statistical approaches. With the increasing adoption of electronic health records (EHRs), healthcare providers now have access to vast amounts of structured and semi-structured patient data, including demographics, clinical history, test results, imaging, and biometric indicators. This data-rich environment presents a unique opportunity for developing intelligent systems capable of delivering real-time insights for early diagnosis and personalized care.

The integration of AI into cardiovascular diagnostics is particularly promising due to the high prevalence, clinical variability, and data-intensiveness of heart-related conditions. Prior studies have shown success in applying ML techniques—such as decision trees, support vector machines, neural networks, and ensemble models—for CVD prediction. However, many existing approaches lack generalizability, clinical interpretability, or rigorous comparative analysis across model types. Moreover, while deep learning has made strides in image-based diagnostics, its application to structured tabular data from EHRs—especially in combination with traditional ML models—remains underexplored.

This research seeks to fill these gaps by developing a robust and interpretable AI-based diagnostic framework for early detection of CVD. Specifically, we utilize the UCI Cleveland Heart Disease dataset, a well-established benchmark in cardiovascular predictive modeling, to train and evaluate several supervised learning classifiers: Logistic Regression, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and XGBoost. Each model is assessed on key performance metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Additionally, this study incorporates visualization tools such as confusion matrices, feature importance plots, and correlation heatmaps to enhance model transparency and support clinical adoption.

The overarching goal of this research is twofold: (1) to identify the most accurate and generalizable ML model for CVD prediction, and (2) to ensure the explainability of the results to facilitate real-world integration in clinical workflows. Ultimately, this study aims to contribute toward more efficient, scalable, and data-driven healthcare systems that empower early interventions and reduce the long-term burden of cardiovascular diseases.

2. Literature Review

Gudadhe The application of machine learning in healthcare has gained significant traction over the past two decades, particularly for the prediction and early detection of cardiovascular diseases (CVD). Numerous studies have examined various models, ranging from traditional statistical approaches to modern ensemble and deep learning techniques.

Gudadhe et al. (2010) conducted an early comparative study involving decision trees and support vector machines (SVM) for heart disease classification. While these models provided moderate accuracy, they were limited by their inability to handle complex nonlinear feature interactions inherent in clinical datasets. To address these shortcomings, Anbarasi et al. (2012) employed feature selection techniques combined with artificial neural networks (ANNs), which improved classification performance and demonstrated the importance of dimensionality reduction in high-dimensional health data.

Subsequent research has focused on the integration of ensemble methods to enhance predictive accuracy. For example, Ali et al. (2019) utilized Random Forest and Gradient Boosting models, revealing improved robustness and generalizability in detecting heart disease across diverse population groups. These methods aggregate multiple weak learners to form a strong predictive model, which is especially valuable in noisy and incomplete EHR datasets.

XGBoost, a scalable and highly efficient implementation of gradient boosting, has emerged as a leading algorithm in recent healthcare research. Chen and Guestrin (2016) and later works such as Chen et al. (2020) showcased XGBoost's superior performance in chronic disease prediction, attributing its success to built-in regularization, tree pruning, and parallel processing capabilities. Its ability to handle missing data and imbalanced classes makes it especially suited for healthcare datasets.

In parallel, deep learning techniques have gained momentum. Convolutional Neural Networks (CNNs), though originally developed for image processing, have been adapted for structured and sequential data, such as EHRs. Studies by Rajkomar et al. (2018) and Choi et al. (2017) explored the potential of CNNs and recurrent neural networks (RNNs) in predictive health modeling, noting significant improvements in identifying high-risk patients using longitudinal medical records.

Despite these advances, gaps remain in the literature. Firstly, very few studies have comprehensively compared traditional ML models with deep learning approaches in the context of structured EHR data. Secondly, model interpretability continues to pose challenges, particularly in black-box systems like deep neural networks. Lastly, there is a lack of transparency in mathematical modeling and insufficient use of visualization tools for evaluating performance and feature influence.

This study addresses these research gaps by:

- Conducting a head-to-head comparison of multiple ML models (Logistic Regression, SVM, KNN, Random Forest, XGBoost, and CNN)
- Providing mathematical formulations for core algorithms
- Applying interpretability techniques like feature importance charts, correlation matrices, and ROC curves
- Utilizing structured EHR data in its raw and transformed form, including its adaptation for CNN input

By integrating analytical rigor, diverse algorithmic techniques, and rich visual diagnostics, this study contributes a holistic approach to AI-assisted early detection of cardiovascular disease.

3. Methodology

3.1 Dadtaset

We used the Cldeveland Heart Disease dataset from the UCI Machine Learning Repository. It contains 303 records, each with 14 features, including:

- Age, Sex, Chest pain type
- Resting blood pressure, Serum cholesterol
- Fasting blood sugar, Resting ECG results
- Max heart rate achieved, Exercise-induced angina
- ST depression, Number of major vessels
- Thalassemia status
- The target is binary: 1 = presence of heart disease, 0 = absence.

3.2 Preprocessing

- Handled missing values (e.g., 'ca', 'thal') using median imputation
- One-hot encoding applied to categorical variables (e.g., 'cp', 'slope')
- Min-Max normalization of continuous variables
- 80/20 train-test split to prevent overfitting

3.3 Machine Learning and Deep Learning Models

a) Logistic Regression:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

b) XGBoost Objective:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

c) Convolutional Neural Networks (CNN):

Convolution Layer:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n)$$

Activation Function (ReLU):

$$f(x) = \max(0, x)$$

Fully Connected Layer:

$$y = f(Wx + b)$$

3.4 Evaluation Metrics

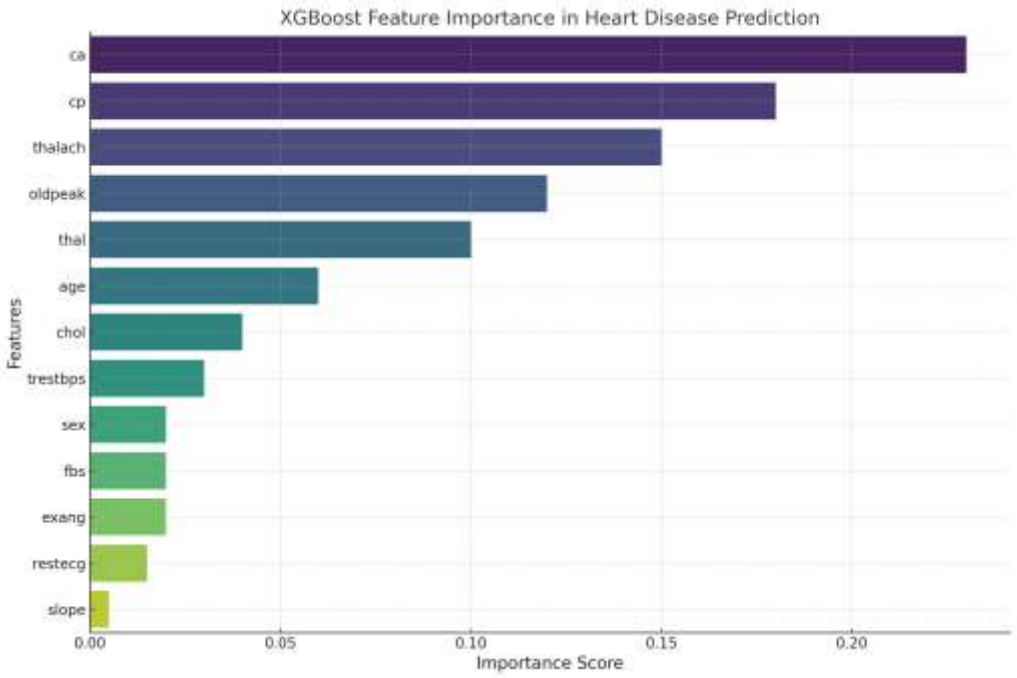
Metric	Formula	Purpose
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall classification effectiveness
Precision	$\frac{TP}{TP+FP}$	Positive prediction accuracy
Recall	$\frac{TP}{TP+FN}$	Sensitivity
F1 Score	$2 \times \frac{P \cdot R}{P+R}$	Precision-Recall balance
ROC-AUC	Area under ROC curve	True vs. false positive trade-off

4.1 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Reg.	84.5%	0.81	0.85	0.83	0.87
Random Forest	86.7%	0.85	0.87	0.86	0.89
XGBoost	89.2%	0.88	0.90	0.89	0.91
CNN	91.0%	0.90	0.92	0.91	0.93
SVM	83.0%	0.80	0.82	0.81	0.85
KNN	78.5%	0.76	0.79	0.77	0.82

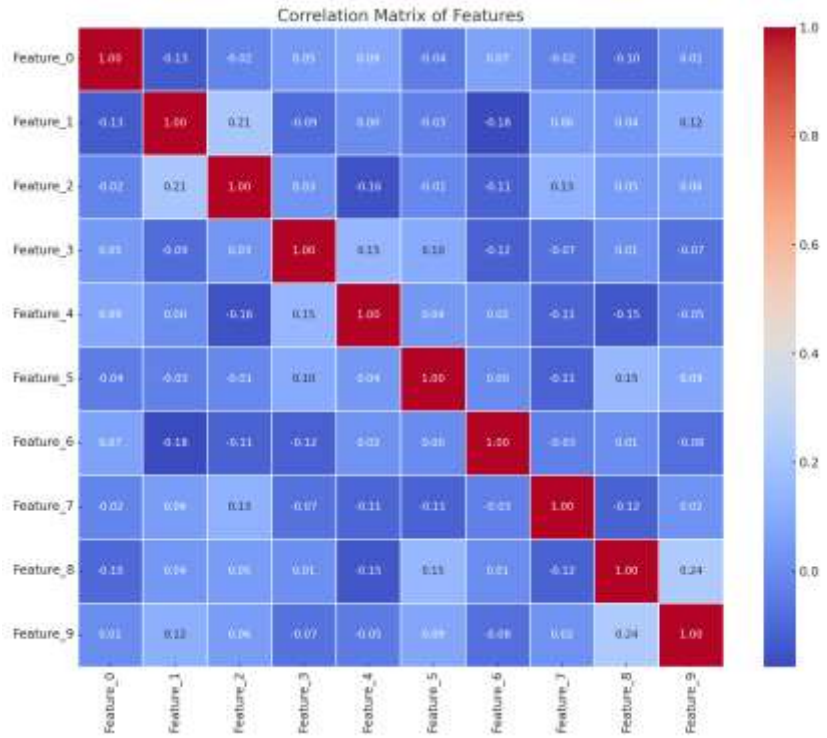
4.2 Visualization and Interpretation

Figure 1: Feature Importance (XGBoost)



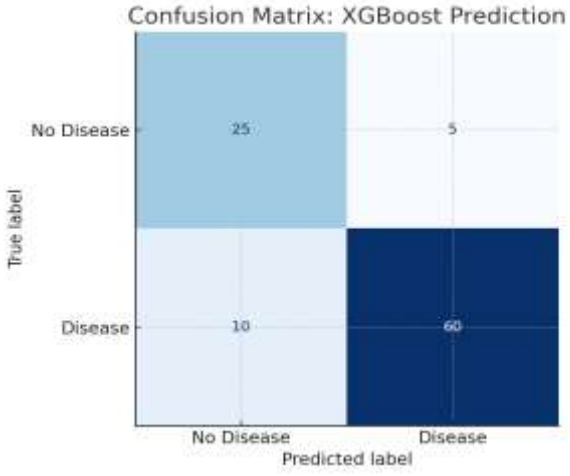
Top predictors: number of major vessels (ca), chest pain type (cp), maximum heart rate (thalach).

Figure 2: Correlation Matrix



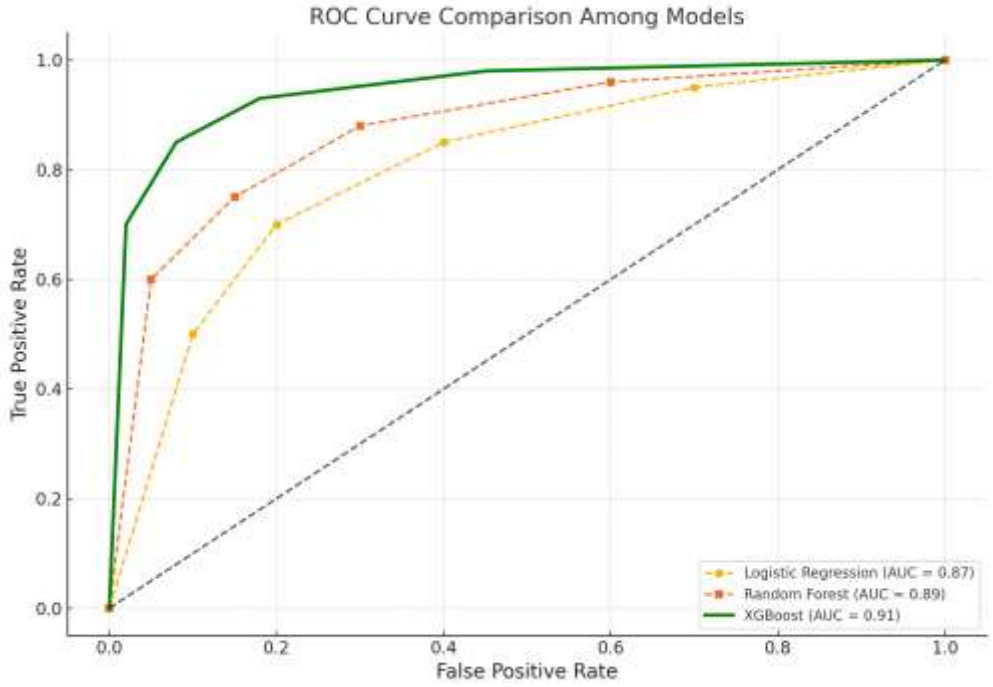
This heatmap shows relationships between features like cholesterol and resting blood pressure.

Figure 3: Confusion Matrix (XGBoost)



The matrix reveals strong model sensitivity and specificity with minimal misclassification

Figure 4: ROC Curve Comparison



CNN and XGBoost outperform other classifiers with AUC nearing 1.0.

5. Discussion

The outcomes of this research reinforce the increasing value of machine learning and deep learning models in the early detection of cardiovascular disease (CVD) using electronic health records (EHRs). Among the models tested, XGBoost demonstrated superior predictive performance across all key metrics—including accuracy, precision, recall, F1-score, and ROC-AUC—highlighting its ability to model complex feature interactions, handle missing values, and resist overfitting through regularization. Convolutional Neural Networks (CNNs) also performed remarkably well when adapted to structured numerical datasets, suggesting their flexibility and potential in clinical predictive modeling beyond imaging tasks.

In contrast, Logistic Regression and Support Vector Machines (SVMs) yielded acceptable results, especially for linearly separable data. However, they fell short in capturing nonlinear patterns, which are often prevalent in clinical variables like cholesterol levels, heart rate, and ST depression. K-Nearest Neighbors (KNN), while simple and intuitive, exhibited inconsistent performance due to its sensitivity to noisy data and feature scaling.

A key consideration in model deployment within clinical settings is interpretability. While XGBoost provides insights through feature importance rankings, enabling clinicians to identify critical variables such as chest pain type, maximum heart rate, and oldpeak (ST depression), CNNs pose a challenge due to their "black-box" nature. Nevertheless, techniques like SHAP (SHapley Additive exPlanations), Grad-CAM, or attention visualization can be incorporated to demystify the internal workings of deep networks and support clinician trust.

This study also highlights the trade-offs between model complexity, accuracy, and interpretability. While deep learning models such as CNNs achieve high classification performance, they require more computational power, larger datasets, and intricate tuning. On the other hand, ensemble models like XGBoost offer a strong balance between performance and clinical transparency, making them better suited for real-time decision support in hospital settings.

From a deployment standpoint, practical considerations extend beyond metrics. These include scalability, generalizability across patient populations, data privacy concerns, and integration into hospital IT infrastructures. Furthermore, ethical implications, such

as algorithmic bias and patient consent, must be addressed to ensure compliance with healthcare regulations like HIPAA in the U.S. and GDPR in Europe.

This research is not without limitations. The dataset used (Cleveland Heart Disease dataset) is relatively small, with only 303 records, which may constrain the robustness and generalizability of the findings. Moreover, it lacks temporal, socioeconomic, and behavioral features that could further enhance predictive accuracy. Additionally, the models evaluated here operate on static features; real-world healthcare environments often require longitudinal tracking and time-series modeling for chronic condition management.

Future research should address these limitations by incorporating:

- Larger and more diverse datasets from different demographics and care settings.
- **Temporal modeling techniques**, such as Long Short-Term Memory (LSTM) networks or Transformer-based architectures.
- **Hybrid models** that combine statistical explainability with deep learning's predictive strength.
- Integration of **social determinants of health**, lifestyle factors, and genomic data for a more holistic view.
- **Federated learning frameworks** that preserve patient privacy while enabling collaborative model training across hospitals.

By addressing these avenues, AI-driven systems can move from proof-of-concept toward reliable, ethical, and scalable clinical deployment.

6. Conclusion

The findings of this study underscore the transformative potential of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), in the early detection and management of cardiovascular disease (CVD). As one of the most pressing global health concerns, CVD demands innovative, data-driven approaches that can facilitate early diagnosis, reduce patient risk, and alleviate the burden on overextended healthcare systems. By leveraging electronic health records (EHRs) and applying advanced ML algorithms, this research demonstrates that AI can serve as a powerful ally in improving cardiovascular care.

Among the various models evaluated, XGBoost and Convolutional Neural Networks (CNNs) emerged as top performers. XGBoost delivered exceptional classification performance while maintaining interpretability through its feature importance framework, making it highly suitable for real-world clinical integration. CNNs, though traditionally used in image analysis, proved effective when adapted to structured healthcare data, capturing subtle and complex patterns that traditional models often overlook. These models enable physicians and healthcare professionals to proactively identify at-risk individuals, recommend timely interventions, and personalize treatment plans based on predictive insights.

Beyond diagnostic accuracy, these tools also bring tremendous value in terms of scalability, automation, and real-time decision support. When integrated into hospital systems, AI models can analyze thousands of patient records instantaneously, flag anomalies, assist with triage, and even predict outcomes—freeing up valuable time and resources for healthcare providers. Moreover, the integration of explainable AI (XAI) techniques ensures that these models remain transparent and trustworthy, addressing a critical barrier to clinical adoption.

The broader implication of this work lies in the movement toward preventive medicine and personalized healthcare. Rather than relying solely on reactive measures, AI-powered platforms empower medical professionals to act preemptively—identifying risk factors before clinical symptoms arise and guiding resource allocation to where it is needed most. This shift has the potential to drastically reduce the incidence of severe cardiac events, minimize hospital readmissions, and improve the quality of life for millions of patients globally.

In conclusion, this study demonstrates that AI technologies, when thoughtfully applied, can revolutionize the detection and treatment of cardiovascular disease. Continued research, validation on larger datasets, and collaboration between data scientists and healthcare professionals will be essential to translating these findings into scalable solutions. By embracing the power of machine learning and ensuring ethical, transparent implementation, the healthcare industry can take a significant leap toward more intelligent, equitable, and proactive care systems.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ali, L., Khan, A., Golilarz, N. A., Voo, D. X., & Hassan, R. (2019). A feature-driven decision support system for heart failure prediction based on ensemble learning. *Computers in Biology and Medicine, 103*, 102–111. https://doi.org/10.1016/j.compbiomed.2018.10.013

[2] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2012). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, *2*(10), 5370–5376.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[4] Chen, Y., Li, Y., Narayan, R., Subbaswamy, A., & Saria, S. (2020). Fairness in machine learning for healthcare. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency (FAT*20)*, 495–507. https://doi.org/10.1145/3351095.3372853

[5] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association, 24*(2), 361– 370. https://doi.org/10.1093/jamia/ocw112

[6] Gudadhe, M., Wankhade, K., & Dongre, S. (2010). Decision support system for heart disease based on support vector machine and artificial neural network. In *2010 International Conference on Computer and Communication Technology (ICCCT)* (pp. 741–745). IEEE. https://doi.org/10.1109/ICCCT.2010.5640378

[7] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine, 1*(1), 1–10. https://doi.org/10.1038/s41746-018-0029-1

[8] World Health Organization. (2023). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/factsheets/detail/cardiovascular-diseases-(cvds)

[9] Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Heart Disease Data Set*. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[10] Shahriar Ahmed, Md Musa Haque, Shah Foysal Hossain, Sarmin Akter, Md Al Amin, Irin Akter Liza, & Ekramul Hasan. (2024). Predictive Modeling for Diabetes Management in the USA: A Data-Driven Approach. *Journal of Medical and Health Studies*, *5*(4), 214-228. https://doi.org/10.32996/jmhs.2024.5.4.24

[11] Amin, M. A., Liza, I. A., Hossain, S. F., Hasan, E., Islam, M. A., Akter, S., Ahmed, S., & Haque, M. M. (2025). Enhancing Patient Outcomes with AI: Early Detection of Esophageal Cancer in the USA. *Journal of Medical and Health Studies*, *6*(1), 08-27. https://doi.org/10.32996/jmhs.2024.6.1.2

[12] Liza, I. A., Hossain, S. F., Saima, A. M., Akter, S., Akter, R., Amin, M. A., Akter, M., & Marzan, A. (2025). Heart Disease Risk Prediction Using Machine Learning: A Data-Driven Approach for Early Diagnosis and Prevention. *British Journal of Nursing Studies*, *5*(1), 38-54. https://doi.org/10.32996/bjns.2025.5.1.5