EXPLAINABLE AI FOR TRANSPARENT DECISION MAKING IN HEALTHCARE

Neha Chauhan[1], Varisha Mirza[2], Twinkle Yadav[3]

1 *Assistant Professor Of Department of Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow*

2 *Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow*

3 *Bachelor of Technology in Computer Science Engineering, Babu Banarasi Das Institute of Technology and Management, Lucknow*

Abstract : Heart disease remains a major global health concern, particularly in regions with limited access to medical resources and diagnostic facilities. Traditional diagnostic methods often fail to accurately identify and manage heart disease risks, leading to adverse outcomes. Machine learning has the potential to significantly enhance the accuracy, efficiency, and speed of heart disease diagnosis. In thisstudy, we proposed a comprehensive framework that combines classification models for heart disease detection and regression models for risk prediction. We employed the Heart Disease dataset, which comprises 1,035 cases. To address the issue of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, resulting in the generation of an additional 100,000 synthetic data points. Performance metrics, including accuracy, precision, recall, F1-score, R^2, MSE, RMSE, and MAE, were used to evaluate the model's effectiveness. Among the classification models, Random Forest emerged as the standout performer, achieving an accuracy of 97.2% on real data and 97.6% on synthetic data. For regression tasks, Linear Regression demonstrated the highest R^2 values of 0.992 and 0.984 on real and synthetic datasets, respectively, with the lowest error metrics. Additionally, Explainable AI techniques were employed to enhance the interpretability of the models. This study highlights the potential of machine learning to revolutionize heart disease diagnosis and riskprediction, thereby facilitating early intervention and enhancing clinical decision- making. The prevalence of heart disease remains alarmingly high worldwide. The term ―Heart Disease Risk‖ encompasses potential complications or adverse effects that individuals may experience due to cardiovascular conditions, such as arrhythmias, heart failure, or coronary artery disease

IndexTerms - Machine Learning, Explainable AI, HeartDiseases.

INTRODUCTION

The prevalence of heart disease remains alarmingly high worldwide. The term ―Heart Disease Risk‖ encompasses potential complications or adverse effects that individuals may experience due to cardiovascular conditions, such as arrhythmias, heart failure, or coronary artery disease. These risks not only threaten a person's health and life but also impose significant burdens on healthcare systems and economies [1]. Cardiovascular diseases are becoming increasingly common in resource- constrained regions, where access to diagnostic tools and specialized care is limited [2]. According to the World Health Organization's (WHO) (2023) [3], cardiovascular diseases are a major health concern in Bangladesh, claiming 273,000 lives annually. Among them, heart disease stands as the leading cause of mortality, accounting for 34% of all deaths nationwide.Heart disease risks manifest in various forms, including hypertension, obesity, diabetes, high cholesterol, and lifestyle factors such as smoking and poor diet.
These conditions are often exacerbated by underlying causes, such as genetic predispositions, inadequate preventive care, and a lack of awareness or timely intervention [4]. These risks often progress silently, early detection and regular monitoring are crucial [5]. In recent years, machine learning has emerged as a powerful tool for detecting various diseases across multiple domains, including agriculture and healthcare [6 It offers an innovative approach to enhancing cardiac health outcomes [7] [8]. This research introduces a machine-learning model to assess heart disease risks by integrating both classification and regression techniques, providing a comprehensive framework for early detection and risk prediction. A total of eleven machine learning classifiers and eleven regressors, including advanced algorithms such as Cat Boost, LightGBM, and Lasso, were employed to enhance predictive accuracy. It has the capability to detect heart disease at an early stage.

NEED OF THE STUDY.

Saves lives and reduces medical errors – When doctors understand why AI suggests a diagnosis or treatment, they can catch mistakes early and choose the best care.
Increases trust and actual use of AI in hospitals – Doctors and nurses adopt AI tools faster when they are not ―black boxes, leading to better patient outcomes.
Ensures fairness and protects vulnerable patients – Transparent AI helps detect and fix biases (e.g., worse predictions for minorities or women), making healthcare more equitable.
Meets legal and regulatory requirements – Laws like EU AI Act and FDA guidelines demand explain ability in high-risk medical AI; this research makes approval and safe deployment possible.
Builds patient confidence and informed consent – Patients feel safer and more involved when they (or their doctor) can see and understand how AI reached a decision about their health.
The primary objectives of this research include:
1. **Make AI decisions easy to understand –** Provide clear explanations so doctors can see why the AI gave a certain diagnosis or recommendation.
2. **Build trust among doctors and patients –** When explanations are reliable and accurate, healthcare professionals feel confident using AI tools.
3. **Help doctors make better decisions together with AI –** Combine AI predictions with human expertise without replacing the doctor's judgment.
4. **Ensure fairness and reduce bias –** Detect and explain if the AI treats different patient groups unfairly (e.g., based on age, gender, or race).
5. **Meet legal and ethical rules** – Make AI systems transparent enough to follow healthcare regulations and protect patient safety.

LITERATURE REVIEW

Key themes include:

1. **Core Concepts and Methods**: XAI techniques are categorized into intrinsic (e.g., decision trees) and post-hoc (e.g., LIME, SHAP, Grad-CAM) approaches, with SHAP dominating EHR studies (63/76 publications) for feature attribution in predictions like disease risk. Reviews stress six groups: feature-oriented, global, concept, surrogate, local pixel-based, and human-centric methods for applications in imaging and diagnostics.

2. **Applications in Healthcare**: XAI enhances disease prediction (e.g., Alzheimer's, tumors) using multimodal data (EHR, imaging), improving CDSS accuracy in radiology, cardiology, and personalized medicine. Studies show 30+ models for prognosis, with attention mechanisms aiding interpretability in DL.

3. **Benefits and Effectiveness**: XAI boosts clinician trust, reduces bias, and supports ethical decisions, with evaluations focusing on transparency, usability, and action ability. DARPA-inspired frameworks measure explanation quality for safety- critical scenarios.

4. **Challenges and Gaps**: Trade-offs between accuracy and explain ability persist, alongside needs for standardized metrics, longitudinal validation, and participatory design. Few reviews target specific predictions; broader adoption lags due to regulatory hurdles like EU AI Act.

Recent Review Papers

1. **Sadeghi et al. (2023)** — "A Review of Explainable Artificial Intelligence in Healthcare" (published in *Computer Methods and Programs in Biomedicine*). This broad review categorizes XAI methods (feature-oriented, global/local interpretability like SHAP and LIME, concept-based, surrogate, and human-centric approaches) and discusses their applications in healthcare, including cardiovascular risk assessment. It highlights the need for transparency to enable clinical adoption and addresses challenges such as computational costs and accuracy-explain ability trade-offs.
2. **Kumar et al. (2025)** — "A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions" (published in *Frontiers in Artificial Intelligence*).

3.**Esfahani et al. (2025)** — "The application of explainable artificial intelligence in chronic disease care" (published in PMC/related journals). A systematic review examining XAI algorithms for prediction in chronic diseases, including cardiovascular conditions. It emphasizes how XAI enhances model transparency, mitigates black-box concerns, and supports accurate predictions in real-world chronic care scenarios.

4. **Banerjee (2025)** — "A systematic review of machine learning in heart disease prediction" (published in *Turkish Journal of Biology*). This review evaluates ML techniques (from logistic regression to ensembles) on heart disease datasets, with explicit attention to XAI enhancements for interpretability. It compares performance metrics across studies and notes how explainable models provide greater clinical utility than traditional statistical methods.

5.**Bayona (2025)** — "Artificial intelligence in cardiovascular prognosis and diagnosis: a review" (published in *Exploration of Medicine*). This survey explores AI's role in ECG analysis, deep learning for prognosis, and emerging XAI integrations to make cardiovascular predictions clinically actionable. It addresses data quality challenges, model deployment barriers, and the need for interpretability in real-world cardiology applications.

RESEARCH METHODOLOGY

In this research, we focused on heart disease detection and risk prediction usingand classification techniques based on a heart disease dataset obtained from UCI. The proposed architecture for developing this model is illustrated below-

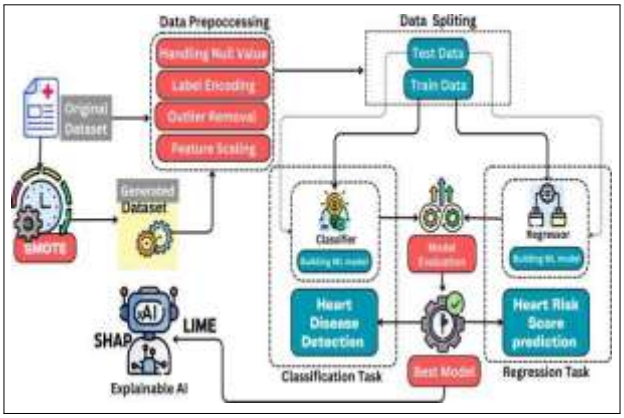

fig.1. system architecture

Dataset Preprocessing-This is the critical first stage that turns raw medical data (e.g., the UCI Heart Disease dataset or any similar clinical record) into clean, balanced, and model-ready input. Each sub-step has a specific purpose in healthcare ML:

1. **Handling Null Values** Medical datasets frequently have missing entries (e.g., a patient's resting blood pressure or cholesterol level not recorded). Common techniques:
   - Mean/median imputation for numerical features
   - Mode imputation for categorical features
   - Or complete row removal if missing values > 30–40 % Why it matters: Nulls can crash many algorithms or introduce bias; proper handling ensures every patient record contributes meaningfully.
2. **Label Encoding** Converts categorical variables (gender: Male/Female, chest-pain type: typical/atypical/asymptomatic, etc.) into numbers (0, 1, 2 …). Alternative (often better): One-hot encoding for nominal categories to avoid implying order. Why: Almost all ML classifiers and regressors require numeric input only.
3. **Outlier Removal** Detects and treats extreme values (e.g., cholesterol = 600 mg/dL or age = 120) using IQR method, Z-score, or domain-specific thresholds. In healthcare, true medical outliers (very high troponin) are sometimes kept; erroneous entries are removed. Why: Outliers can skew distance-based models (KNN, SVM) or tree splits and degrade performance.
4. **Feature Scaling** Brings all numerical features to the same scale:
   - StandardScaler (mean = 0, stud = 1) — preferred for most models
   - MinMaxScaler (0 to 1) — useful for neural networks Why: Features like age (20–80) vs. cholesterol (100–400) would otherwise dominate distance calculations.

After these four steps you get a clean "Original Dataset".

SMOTE (Synthetic Minority Over-sampling Technique) The diagram then feeds the pre-processed data into SMOTE to create the "Generated Dataset". Heart-disease cases are usually the minority class (e.g., 30 % positive vs. 70 % negative). SMOTE generates new synthetic positive samples by linearly interpolating between existing minority

instances and their nearest neighbours. Result: a balanced training set that prevents the model from becoming biased toward the healthy class.

Data Splitting The balanced (or original) dataset is divided into:

- Train Data (typically 70–80 %) — used to actually learn the model parameters
- Test Data (20–30 %) — kept completely unseen until final evaluation This split (often with stratified sampling to preserve class ratios) ensures fair, unbiased performance measurement.

Model Parameters-

1.Hyper parameters (tuned before training)

These are the settings you choose before the model starts learning:

- Random Forest: n_estimators (number of trees), max_depth, min_samples_split
- Support Vector Machine: C (regularization), kernel (linear/ruff), gamma
- Boost/LightGBM: learning rate, max_depth, subsample
- Neural Network: number of hidden layers, neurons per layer, dropout rate, learning rate Tuning methods shown implicitly in your "Model Evaluation" step:

- Grid Search CV or Randomized Search CV on the Train Data
- 5-fold or 10-fold cross-validation to avoid over fitting
- Evaluation metrics: Accuracy, Precision, Recall, F1, AUC-ROC (classification) and MSE, RMSE, R^2 (regression)

2.Learned Parameters (what the model actually learns during training)

- Logistic Regression: coefficients (weights) for each feature
- Random Forest: split decisions in every tree
- Neural Network: millions of connection weights updated via back-propagationThese are automatically adjusted during the "building ML model" phase using the Train Data + chosen hyperparameters.

3. Best Model SelectionAfter training multiple candidates (or tuning one), the "Model Evaluation" box compares performance on the Test Data and picks the single best model (highest F1 or lowest RMSE). This "Best Model" is then passed to SHAP/LIME for explainable AI explanations.

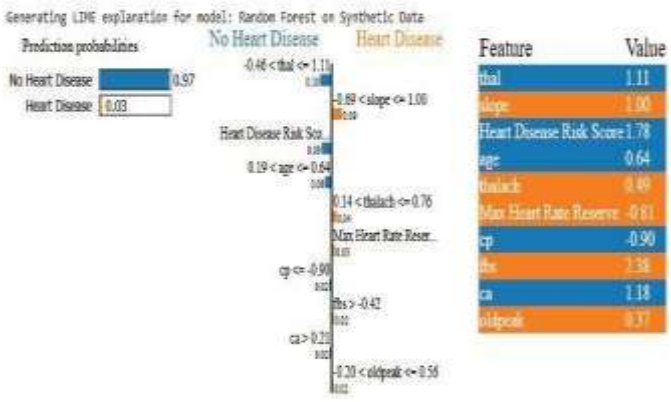

fig 3 : LIME Explainations

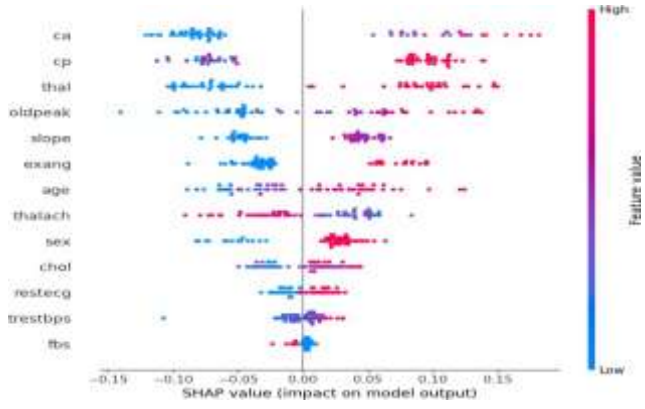

fig 4: SHAP Values Of XAI

Fig 3 shows LIME visualization for LR after applying SMOTE for the regression task. The predicted value for the instance lies between ―No Heart Disease‖ and ―Heart Disease with a strong influence from key features. The bar highlights the positive and negative contributions of features like Max Heart Rate Reserve, thallic, and age, showing

their significance in driving the model's output. The accompanying table lists the feature values, providing context for the prediction.

Discussion & Limitation

After comparing the performance of various models and methods, we identified Random Forest as the most effective model for detecting heart diseases and Linear Regression as the best approach for predicting risks. The model's high accuracy suggests that it can be used with confidence in real- world situations. We also compared our model with previous research described in the related work chapter. The results of our model outperform all of the earlier models we analyzed, as seen in Table I. The proposed model's drawback, in spite of its excellent performance, is its dependency on the quality and diversity of the dataset used. Although SMOTE was applied to address class imbalance, the dataset may still not fully represent the variability work, future research could focus on integrating additional clinical data, conducting longitudinal analyses for better dis- ease progression predictions, and exploring the applicability of the proposed framework across diverse populations and healthcare systems.

While the pre-processing pipeline—including null value imputation, label encoding, outlier removal, feature scaling, and SMOTE-based synthetic oversampling—effectively prepares the dataset for training balanced classification and regression models in the proposed axis framework, the underlying data source (typically the UCI Cleveland Heart Disease dataset) imposes severe limitations that compromise real-world validity. With only 303 records collected from a single hospital in the 1980s, this repository suffers from an extremely small sample size that precludes statistically robust conclusions and makes deep learning or ensemble models prone to over fitting even after SMOTE augmentation. Its narrow feature set (just 13–14 clinical attributes) lacks modern multimodal elements such as genetic markers, longitudinal EHR entries, wearable sensor streams, or imaging data, while the outdate collection period fails to capture contemporary demographics, treatment protocols, lifestyle variables, or global population diversity. Consequently, models trained on such data risk poor generalizability, hidden biases toward specific ethnic or socioeconomic groups, and reduced predictive power when applied to today's heterogeneous patient populations. In terms of real clinical implementation, the integration of SHAP and LIME for post-hoc explain ability represents a promising step toward transparent decision-making by providing feature-level attributions that clinicians can review alongside heart disease detection outputs or risk scores; however, translating this laboratory pipeline into hospital workflows encounters substantial barriers. Electronic health record (EHR) interoperability remains a core challenge, as raw clinical data often exists in non-standardized formats across institutions, complicating real-time ingestion, privacy compliance under regulations such as HIPAA or GDPR, and secure deployment of the best-tuned classifier or repressor. Moreover, even with XAI-generated explanations, a persistent mismatch arises between algorithmic rationales and physicians' intuitive reasoning processes, potentially increasing cognitive load, alert fatigue, or unwarranted over-trust in time-sensitive settings like emergency cardiology. Regulatory approval as a clinical decision-support tool demands extensive prospective, multi-centre validation trials—far beyond retrospective UCI evaluations—alongside bias auditing, traceability mechanisms, and demonstrated improvements in patient outcomes; without these, the methodology largely remains confined to research prototypes rather than routine bedside use.

TABLE I

COMPARISON WITH RELATED WORK

Ref.	Model/Architecture	Highest Accuracy
[9]	KNN, LR, NB, RF, SVM, DT, DNN	95.30% with DNN
[12]	RF, SVM, NN	92% with NN
[13]	Cat Boost, RF, GB, LightGBM and Adobos	95% with Adobos
[14]	K-modes, RF, DT, MLP, XGB	87.28% with MLP
[15]	NB, KNN, DT, SVM, LR, MLP	95.76% with DT

Proposed Model	RF, DT, LR, KNN, SVM, Boost, NB, GB, LightGBM, Cat Boost, Tab Net, Ridge, and Lasso	97.6% with RF and 0.992 R^2 with LR

FUTURE WORK AND RECOMMENDATIONS

1. **Patient-Centered Explanations** Develop simple, visual, and multilingual explanations tailored for patients (not just clinicians) to support true informed consent and shared decision-making.

2. **Real-World Clinical Trials** Move beyond simulations: conduct large-scale prospective studies and randomized controlled trials in hospitals to measure actual impact on patient outcomes, error rates, and workflow efficiency. Develop simple, visual, and multilingual explanations tailored for patients (not just clinicians) to support true informed consent and shared decision-making.

3. **Standardized Evaluation Framework** Create universal benchmarks and metrics for explanation quality (combining fidelity, clinical correctness, action ability, and fairness) accepted by regulators (FDA, EMA) and medical bodies.

4. **Integration into Clinical Workflows** Build seamless XAI plug-ins for major EHR systems (Epic, Cerner) and PACS, with automatic explanation generation, audit trails, and feedback loops for continuous model improvement.

5. **Global Equity and Low-Resource Settings** Design lightweight, culturally aware XAI methods that work with limited data and infrastructure, ensuring AI benefits reach underserved regions and reduce rather than widen healthcare disparities.

These steps will help transform XAI from a research topic into a routine, trusted, and equitable part of everyday clinical practice.

CONCLUSIONS

The study establishes a robust and explainable machine learn- in framework for detecting and predicting cardiovascular diseases, achieving an accuracy of 97.6% with the Random Forest for detecting heart diseases and a 99.2 MSE score for predicting heart disease risk. By leveraging advanced algorithms and interpretability techniques such as SHAP and LIME, the research not only delivers high predictive accuracy but also provides actionable insights into the most critical risk factors contributing to heart disease.

The use of SMOTE effectively mitigates data imbalance, ensuring improved performance on synthetic datasets and enhancing the model's applicability in real-world scenarios.

This re- search not only highlights the potential of machine learning in early detection and risk assessment but also underscores the importance of explainable AI in ensuring transparency and trust in healthcare decisions. To further advance this work, future research could focus on integrating additional clinical data, conducting longitudinal analyses for better disease progression predictions, and exploring the applicability of the proposed framework across diverse populations and healthcare systems.
Ultimately, the success of AI in healthcare will not be measured only by how accurately it predicts, but by how confidently and justly doctors and patients can act on those predictions. XAI turns opaque algorithms into responsible partners, paving the way for safer, fairer, and more humane healthcare in the age of artificial intelligence.

.

REFERENCES

1.T. Candela, A. Britton, E. Brunner, H. Hemingway, M. Malik. Kumara, E. Bad rick, M. Kivimaki, and M. Marmot, ―Work stress and coronary heart disease: what are the mechanisms? *European heart journal*, vol. 29, no. 5, pp. 640–
648, 2008.
2.J. I. Hoffman, ―The global burden of congenital heart disease,‖ *Car- diovascular journal of Africa*, vol. 24, no. 4, pp. 141–145, 2013.

3.T. Desk, ―Experts: One in four adults in Bangladesh suffer from hypertension.‖ Dhaka Tribune, 2024. Availablat: https://www.dhakatribune.com/bangladesh/health/36 0195/experts-one-in-four-adults-in-Bangladesh-suffer.

4.F. L. Mondesir, T. M. Brown, P. Manner, R. W. Durant, A. P. Carson, M. M. Safford, and E. B. Leviton, ―Diabetes, diabetes severity, and coronary heart disease risk equivalence: Reasons for geographic and racial differences in stroke (regards),‖ *American heart journal*, vol. 181,pp. 43–51, 2016.

5.A. O. Macomb, E. Lameira, A. Yaks, L. Paul,M. B. Ferreira, and D. Side, ―Challenges on the management of congenital heart disease in developing countries,‖ *International journal of cardiology*, vol. 148, no. 3, pp. 285–288, 2011.

6.M. S. Hessen, P. Shah, and M. Saiduzzaman, ―A hybrid machine learning approach utilizing can feature extraction with traditional classifier to identify strawberry leaf diseases,‖ in *4th International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, 2025.

7.A. F. Odom, E. E. Abdallah, Y. Kef aye, and M. As hour, ―Effective diagnosis and monitoring of heart disease,‖ *International Journal of Software Engineering and Its Applications*, vol.9, no. 1,pp. 143–156, 2015.

8.V. Chang, V. R. Havana, A. Q. Cu, and M. Husain, ―An artificial intelligence model for heart disease detection using machine learning algorithms,‖ *Healthcare Analytics*, vol. 2, p. 100016, 2022.

9.M. S. Singh, K. Thong am, P. Choudhary, and P. Braga, ―An integrated machine learning approach for congestive heart failure prediction, *Diagnostics*, vol. 14, no. 7, p. 736, 2024.

10.N. Chandrasekhar and S. Peddakrishna,―Enhancing heart disease prediction accuracy through machine learning techniques and optimiza- tion,‖ *Processes*, vol. 11, no. 4, p. 1210, 2023.

11.I. D. Minnie and N. Jere, ―Optimized ensemble learning approach with explainable ai for improved heart disease prediction,‖ *Information*, vol. 15, no. 7, p. 394, 2024.

12.A. Husain, A. Saeed, A. Hussain, A. Ahmad, and M. Gondal, ―Harnessing ai for early detection of cardiovascular diseases: Insights from predictive models using patient data,‖ *International Journal for Multidisciplinary Research*, vol. 6, no. 5, 2024.

13.N. Nissa, S. Jamwal, and M. Neshat, ―A technical comparative heart disease prediction framework using boosting ensemble techniques,‖ *Computation*, vol. 12, no. 1, p. 15, 2024.

14.C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, ―Effective heart disease prediction using machine learning techniques,‖ *Algorithms*, vol. 16, no. 2, p. 88, 2023.

15.B. Abuhaija, A. Alloubani, M. Almatari, G. M. Jaradat, B. A. Hemn,A. M. Abualkishik, and M. K. Alsmadi, ―A comprehensive study of machine learning for predicting cardiovascular disease using weka and spss tools,‖ *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, p. 1891, 2023.

16.S.Mahsa, ―Heart disease dataset.‖ Kaggle,Availableat: https://www.kaggle.com/datasets/snmahsa/heart-disease, 2020.