International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 01 | Jan - 2026 SJIF Rating: 8.586 ISSN: 2582-3930

Multimodal Heart Disease Prediction System Using Tabular and Image Data

Koustav Podder[1], Shubhradip Saha[2], Sudipta Kumar Dutta[3]

[1,2,3] *BP Poddar Institute of Management and Technology, Kolkata, India*

Abstract– Cardiovascular diseases remain one of the leading causes of mortality worldwide, making early and reliable risk prediction a critical requirement in modern healthcare systems. Traditional diagnostic approaches often rely on either clinical parameters or medical imaging in isolation, which may limit the predictiveness of the disease. This project aims to build a multimodal heart disease prediction framework that integrates clinical tabular data and echocardiography-based image information to estimate the probability of heart disease in a robust way. The tabular dataset includes demographics and clinical risk factors such as age, sex, chest pain type, blood pressure, cholesterol, electrocardiographic findings, and exercise-related parameters. Multiple classical machine learning models were trained and evaluated, including Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and XGBoost. Among all these, XGBoost achieved the best performance, with a test accuracy of 89.13% and an F1-score of 0.9029, demonstrating strong predictive capability. For the imaging modality, an echocardiography video-based model was trained using the EchoNet Dynamic dataset, where cardiac abnormality was determined based on an ejection fraction threshold of 50%. The image model achieved a ROC-AUC of 0.915, indicating excellent discriminative performance. Threshold optimization was performed to improve clinical sensitivity, increasing recall from 66.77% to 72.70% and reducing false negative cases. A late fusion approach using weighted averaging of predicted probabilities was adopted to combine both modalities, leveraging their complementary strengths. The system outputs a probabilistic risk score rather than a definitive diagnosis, making it suitable as a clinical decision support tool. Future enhancements include the integration of a RAG-based assistant for user guidance and the adoption of explainable AI techniques to improve transparency and interpretability.

Key Words: Heart Disease Prediction, Multimodal Learning, Echocardiography, Machine Learning, Deep Learning, Late Fusion, Clinical Decision Support, Prediction Threshold

1.INTRODUCTION

1.1 Overview

Cardiovascular diseases (CVDs) are among the leading causes of mortality worldwide, accounting for approximately 17.9 million deaths annually. Heart disease significantly contributes to long-term disability and premature death, making early risk detection essential for timely intervention and effective patient management. In clinical practice, diagnosis relies on a combination of patient history, laboratory investigations, and medical imaging.

Recent advances in machine learning have enabled large-scale analysis of clinical datasets to identify complex patterns associated with cardiovascular risk [4,5]. Simultaneously, medical imaging—particularly echocardiography—provides direct insight into cardiac structure and function and has demonstrated strong capability in assessing cardiac performance using AI-based approaches [2]. Despite this, most automated systems continue to rely on a single data modality, reducing predictive robustness. Integrating heterogeneous data sources offers a promising pathway to improve reliability and support more informed clinical decision-making [1,3].

1.2 Motivation

The motivation for this work stems from the need for a more reliable heart disease prediction framework. Clinical tabular data captures demographic and physiological risk factors, while echocardiography provides functional evidence such as ejection fraction, a key indicator of cardiac performance [2]. Combining these complementary sources can yield more informative and clinically meaningful predictions than either modality alone [1,3].

Machine learning techniques are well suited for multimodal integration due to their capacity to process heterogeneous data and produce probabilistic outputs, enabling joint reasoning across structured clinical variables and unstructured imaging information [1].

1.3 Problem Statement

This project aims to design a multimodal machine learning framework for estimating heart disease risk by integrating clinical tabular data with echocardiography-based imaging information. Independent predictive models are developed for each modality and their outputs are combined using a late fusion strategy based on weighted probability averaging. Such fusion-based frameworks are widely adopted in multimodal medical AI systems to improve robustness and generalization [1,3]. The system produces a probabilistic risk score intended to support clinical decision-making rather than deliver a definitive diagnosis.

1.4 Objectives

The primary objective is to develop an integrated multimodal framework for heart disease risk prediction. This involves preprocessing and modeling clinical features using multiple classical machine learning algorithms, as demonstrated in recent heart disease prediction studies [4,5], and designing an independent image-based model that assesses cardiac abnormality using ejection fraction–guided labeling and medically relevant evaluation metrics derived from echocardiography-based AI research [2].

The final objective is to combine both modalities through weighted late fusion to obtain a robust probabilistic risk estimate, following established multimodal fusion practices [1,3], while ensuring extensibility for future enhancements such as explainable AI integration and intelligent user assistance.

2. PROPOSED SYSTEM

2.1 System Overview

The proposed system is a multimodal heart disease prediction framework designed to assist in clinical decision support by integrating structured clinical risk factors with echocardiography video analysis. The system follows an end-to-end pipeline that processes heterogeneous inputs independently and combines their predictive outputs to generate a final, clinically interpretable risk score.

2.1.1 System Inputs

The system accepts two types of inputs:
• Tabular Clinical Data This includes patient-specific risk factors such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression (oldpeak), and ST slope. These parameters represent commonly used clinical indicators for cardiovascular risk assessment.
• Echocardiography Video Data The system accepts apical echocardiography videos (e.g., apical four-chamber view) that capture cardiac motion and structural dynamics. These videos provide direct visual information about heart function, particularly ventricular contraction patterns.

2.1.2 Independent Modality Processing

Each input modality is processed through a dedicated pipeline:
• Tabular Processing Pipeline: The clinical data is first preprocessed through categorical encoding and feature alignment based on the training schema. Multiple machine learning models trained on structured risk-factor data independently generate probability estimates for the presence of heart disease. These probabilities are then aggregated using an ensemble strategy to produce a single tabular-based risk probability.
• Echocardiography Video Processing Pipeline: The uploaded echocardiography video is preprocessed by extracting frames, resizing, and normalization. The processed video clip is passed through a deep learning model trained to analyze spatiotemporal cardiac motion patterns. The model outputs a probability score indicating the likelihood of abnormal cardiac function. Each modality operates independently, ensuring that the system remains robust even if one input source is noisy or partially informative.

Probability Generation Both pipelines produce probabilistic outputs rather than binary decisions.
• The tabular pipeline outputs a probability representing risk inferred from clinical factors.
• The echocardiography pipeline outputs a probability representing risk inferred from cardiac imaging. Using probabilities allows consistent comparison and combination across modalities.

2.1.3 Multimodal Fusion Strategy

The system employs a late fusion strategy at the decision level. The probabilities generated by the tabular ensemble and the echocardiography model are combined using a weighted averaging approach. Greater weight is assigned to the echocardiography-based probability due to its direct measurement of cardiac function, while the tabular probability contributes complementary population-level risk information. This fusion mechanism ensures that structural and functional evidence from imaging is prioritized, while still incorporating valuable clinical context.

2.1.4 System Output

The final output of the system is a single multimodal risk probability representing the overall likelihood of heart disease. This result is presented to the user along with:
• Individual modality predictions (clinical and echocardiography),
• The fused multimodal prediction,
• A qualitative risk interpretation (high or low risk).
The system is designed to support clinicians by providing transparent, interpretable predictions, and it does not replace professional medical judgment.

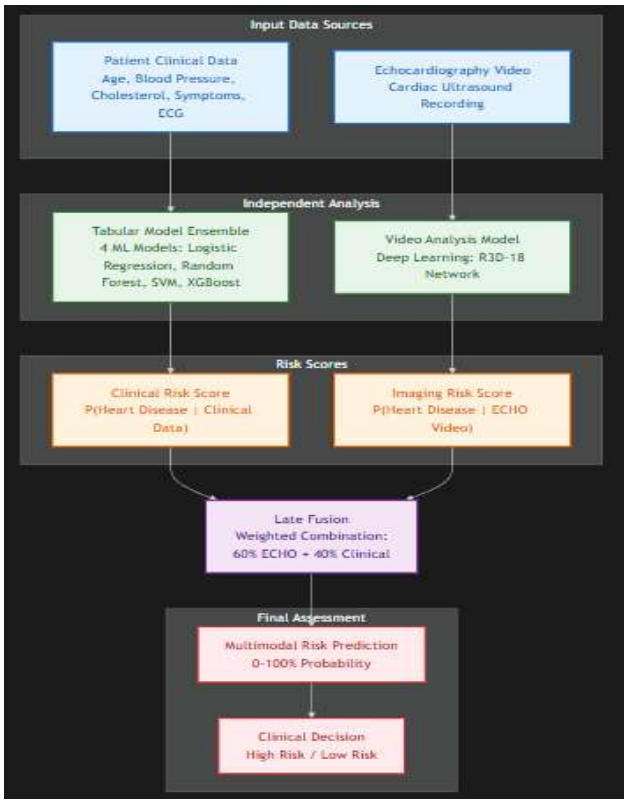

Fig -1: System Architecture

2.2 Dataset Description

This section describes the datasets used in the proposed multimodal heart disease prediction system. Two complementary data sources are employed: a clinical tabular

dataset and an echocardiography-based image dataset. Relevant exploratory analysis and visualizations are presented within each subsection to support understanding of the data characteristics.

2.2.1 Clinical Tabular Dataset

The clinical tabular dataset used in this study is sourced from Kaggle, a widely used platform for publicly available datasets in the healthcare and machine learning domain. The dataset consists of structured patient-level clinical records commonly used for heart disease prediction tasks. The data set name is fedesoriano. (September 2021). Each record represents an individual patient and includes demographic information, physiological measurements, and diagnostic test results. The target variable, HeartDisease, is a binary label indicating the presence or absence of heart disease. The dataset contains clinically relevant features such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, and ST segment slope. These features are widely used in cardiovascular risk assessment and are suitable for classical machine learning-based classification.
Attribute Information 1. Age: age of the patient [years] 2. Sex: sex of the patient [M: Male, F: Female] 3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic] 4. RestingBP: resting blood pressure [mm Hg] 5. Cholesterol: serum cholesterol [mm/dl] 6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise] 7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria] 8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202] 9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No] 10. Oldpeak: oldpeak = ST [Numeric value measured in depression] 11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping] 12. HeartDisease: output class [1: heart disease, 0: Normal] The dataset includes both numerical and categorical variables, requiring preprocessing steps such as encoding and normalization prior to model training. Fig -2 shows the HeartDisease class distribution.

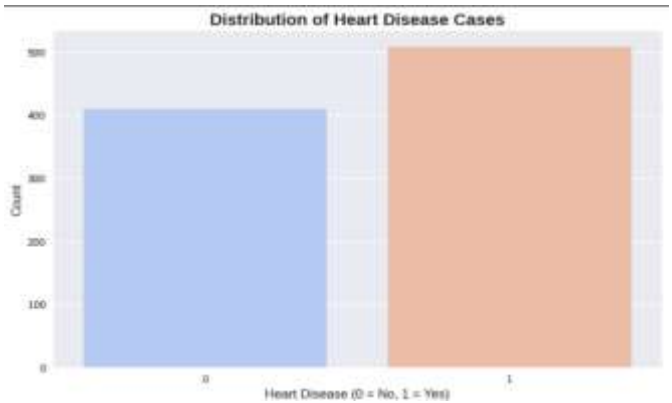

Fig – 2: Target class distribution

2.2.2 Echocardiography Image Dataset

The echocardiography dataset used in this project is the EchoNet-Dynamic dataset, a publicly available medical imaging dataset designed for the analysis of cardiac function using echocardiographic video data.
The dataset consists of video sequences capturing cardiac motion over complete cardiac cycles. Each video is represented as a sequence of frames that reflect the dynamic behavior of the heart, enabling assessment of ventricular function.
The dataset provides clinically measured ejection fraction (EF) values for each sample. In this project, EF is used as a functional indicator to categorize samples into normal and abnormal cardiac function based on a predefined threshold. This labeling strategy enables binary classification while preserving clinical relevance. The distribution of ejection fraction values across the dataset is analyzed to understand the spread of cardiac function. Fig -3 depicts the distribution of EF values, highlighting the separation between normal and abnormal cases.

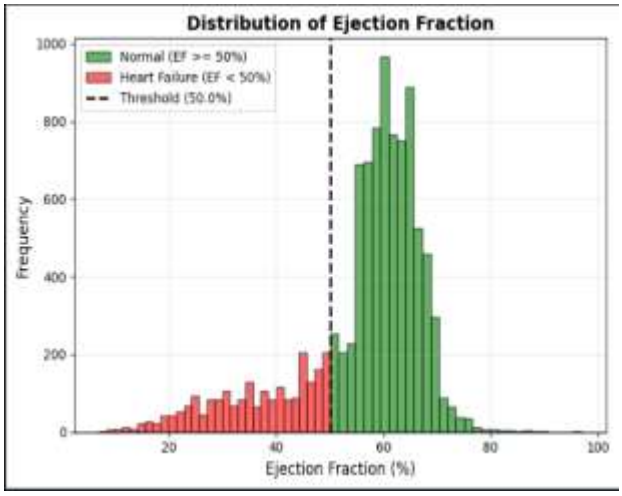

Fig – 3: EF distribution

2.3 Training Independent Modalities

The proposed system follows a modular design in which each data modality is developed, trained, and evaluated independently before multimodal fusion. This section describes the construction of the clinical tabular model, including data preprocessing, exploratory analysis, model training, and probability-based prediction. Developing independent modalities ensures interpretability, flexibility, and reliable probability estimation prior to fusion.

2.3.1 Clinical Tabular Model

The clinical dataset was first examined for duplicates, missing data, structural consistency and invalid data. The zero values of cholesterol and resting blood pressure were replaced by a K-Nearest Neighbors (KNN) imputation strategy. This approach estimates missing values based on the similarity between samples and helps preserve the underlying data distribution. Categorical features were identified and converted into numerical representations using one hot encoding, enabling their use in classical machine learning algorithms. All boolean and floating-point fields were converted to integer format to maintain consistency across the dataset.

Then correlation analysis was conducted to identify attributes with stronger associations to heart disease. Features such as age, maximum heart rate, ST depression, and fasting blood sugar showed notable correlations with the target label.

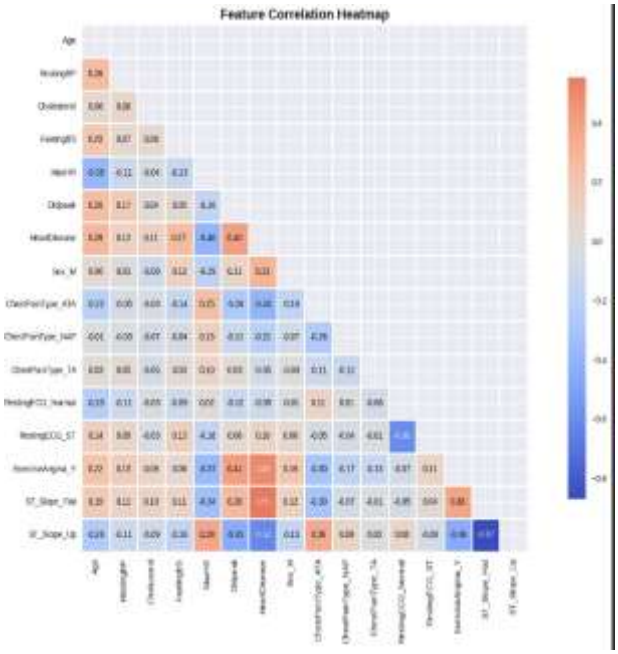

Fig – 4: Correlation heatmap

To ensure consistent preprocessing across different machine learning models, a unified preprocessing pipeline was constructed. Numerical features such as age, resting blood pressure, cholesterol, maximum heart rate, and ST depression were standardized using z score normalization. Remaining features were passed through without scaling.

Although the dataset was moderately balanced, class weights were computed to reduce bias during training. Balanced class weights were calculated using the training labels and supplied to applicable classifiers. This strategy helps penalize misclassification of minority class samples more strongly and improves generalization.

Multiple classical machine learning models were evaluated to determine the most suitable approach for clinical tabular data. The models considered include Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and XGBoost. Each model was trained using the same preprocessing pipeline to ensure fair comparison. Hyperparameter tuning was performed using systematic parameter exploration and cross validation. Model performance was evaluated using accuracy and cross-validation stability, and the best-performing configurations were selected for final evaluation.

2.3.2 Echocardiography Model

The echocardiography model uses Ejection Fraction (EF) as the supervisory signal for classification. EF is a clinically accepted quantitative measure of left ventricular systolic function and is widely used to assess heart failure severity.

To formulate the problem as a binary classification task, EF values provided in the EchoNet Dynamic dataset are converted into class labels using a fixed threshold. Samples with EF values below the threshold are labeled as heart failure (abnormal), while samples with EF values equal to or above the threshold are labeled as normal. This approach enables the model to learn visual patterns associated with impaired cardiac function while maintaining clinical interpretability. The EF distribution is visualized using histogram and bar plots to illustrate the separation induced by the threshold and to highlight class imbalance.

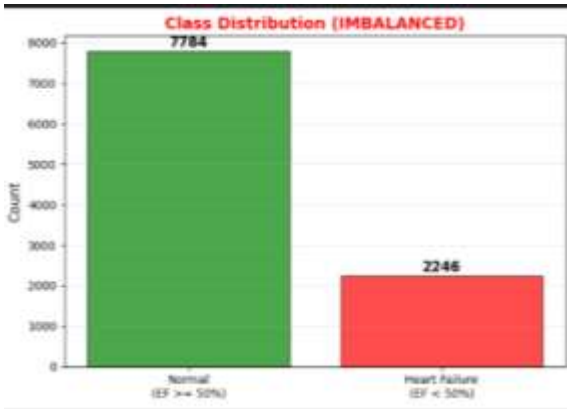

Fig – 5: Class distribution

Analysis of the labeled dataset reveals a significant class imbalance, with normal cases outnumbering heart failure cases. This imbalance poses a risk of biased learning, where the model may favor the majority class. To address this issue we used focal loss strategy to focus learning on hard-to-classify samples. The loss dynamically down-weights easy examples while emphasizing minority and misclassified samples, improving robustness under class imbalance.

Training is further optimized using:

• AdamW optimizer for stable convergence
• Learning rate scheduling based on validation performance
• Mixed precision training (AMP) to accelerate computation
• Gradient accumulation to simulate larger batch sizes within GPU memory limits
These strategies collectively improve training stability, convergence speed, and generalization.
To further mitigate class imbalance and improve generalization, data augmentation is applied to the training set, particularly benefiting minority class samples. Augmentation techniques include:
• Random horizontal flipping
• Brightness adjustment
• Contrast adjustment
These transformations introduce variability while preserving clinical structure, improving model robustness without altering diagnostic features.

Model Architecture is built using a 3D convolutional neural network (R3D-18) pretrained on large-scale video datasets. This architecture is well-suited for echocardiography data as it captures both spatial and temporal patterns across cardiac cycles. Key architectural design choices include:
• Use of a pretrained backbone for faster convergence
• Initial freezing of backbone layers to stabilize early training
• A custom classification head with dropout and fully connected layers 37
• Gradual unfreezing of the backbone for fine-tuning after initial epochs
This design balances representational power with computational efficiency and reduces overfitting on limited medical data

2.4 Multimodal Fusion Module

To integrate predictions from clinical and imaging data, a multimodal fusion framework is adopted. The system combines outputs from independently trained tabular and echocardiography models to produce a unified heart disease risk estimate.

2.4.1 Late Fusion Strategy

A late fusion approach is used, where each modality generates its own probability prediction before fusion. The clinical tabular model predicts heart disease probability based on structured patient features, while the echocardiography model predicts abnormal cardiac function using image-derived information supervised by EF-based labels. Late fusion is chosen because it preserves the independence of each modality, improves interpretability, and allows each model to be trained and optimized separately. This design also provides modularity, enabling future improvements or replacement of individual models without affecting the overall system.

2.4.2 Weighted Averaging

The final multimodal risk score is computed using weighted averaging of modality outputs. Let P_{tab} be the probability from the tabular model and P_{img} be the probability from the image model. The final prediction is given by:
$P_{final} = W_{tab}P_{tab} + W_{img}P_{img}$. $W_{tab} + W_{img} = 1$
Here the weights were taken 0.60 for image model prediction and 0.40 for tabular model prediction. Weights are assigned empirically based on validation performance and clinical relevance. The resulting probability represents the overall heart disease risk and can be used directly or thresholded for binary classification. This fusion method provides a simple and reliable baseline for multimodal integration, with scope for future enhancement using adaptive fusion techniques.

3. RESULTS

3.1 Tabular Model Results

A comparative analysis of all trained tabular models is presented using accuracy, F1-score, and recall on the test dataset. These metrics provide a balanced view of overall correctness, class-wise performance, and sensitivity toward heart disease cases.

Table -1: Tabular model results

No.	Model	Test Accuracy	CV Mean Accuracy	CV Std Dev
1	XGBoost	0.8913	0.8256	0.0418
2	Random Forest	0.8859	0.8376	0.0558
3	Support Vector Machine (SVM)	0.8859	0.8354	0.0441
4	Logistic Regression	0.8804	0.8409	0.0481
5	Decision Tree	0.8152	0.7885	0.0639

Best Performing Model

- **Model:** XGBoost
- **Test Accuracy:** 89.13%
- **Cross-Validation Mean Accuracy:** 82.56%
- **Cross-Validation Std Dev:** 0.0418

From the results, XGBoost achieved the highest test accuracy and F1-score among all models, indicating strong overall performance and balanced precision–recall behavior. Random Forest and SVM also demonstrated competitive performance, while Logistic Regression provided a stable baseline. The Decision Tree model showed comparatively lower

performance, likely due to overfitting and limited generalization.

The confusion Matrix and probability distribution is shown below for XGBoost. This analysis is particularly important in medical applications, where false negatives (missed disease cases) can have serious consequences.

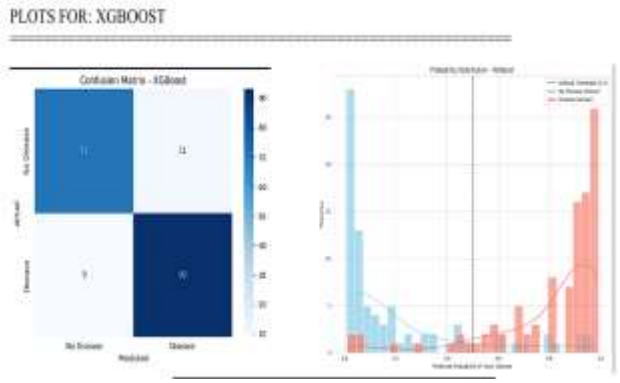

Fig – 6: xgboost plots

3.2 Image Model Results

The image model demonstrates strong discriminative capability in identifying abnormal cardiac function. The key performance metrics obtained on the test dataset are summarized below: Accuracy: 87.64% ROC–AUC: 0.9151

The high ROC–AUC value indicates that the model effectively separates normal and abnormal cases across a wide range of decision thresholds. This confirms that the learned representations capture clinically meaningful patterns from echocardiography data. While accuracy provides an overall correctness measure, ROC–AUC is emphasized as it is less sensitive to class imbalance and threshold choice.

FINAL TEST METRICS
Accuracy: 87.64%
Precision: 0.7525
Recall: 0.6677
F1-score: 0.7075
ROC-AUC: 0.9151

ROC curves are used to illustrate the trade-off between the true positive rate and the false positive rate across different decision thresholds. The consistently high area under the ROC curve (AUC) demonstrates the strong discriminative capability of the image model, indicating its ability to distinguish between normal and abnormal cardiac function independent of a fixed classification threshold.

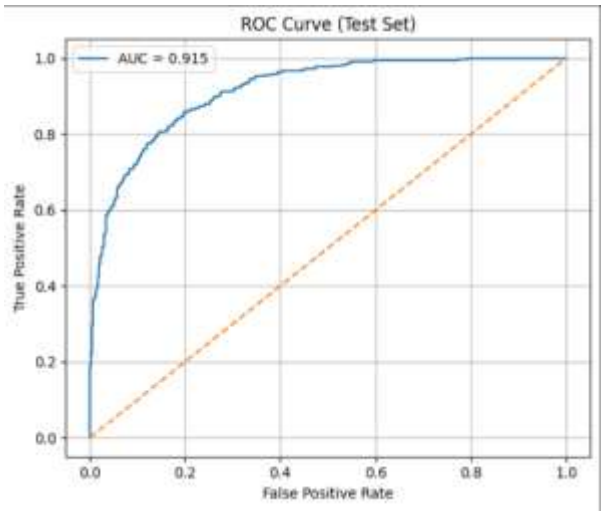

Fig -7: ROC Curve for image model

To analyze the effect of decision threshold selection, the model is evaluated using two thresholds: the default threshold of 0.5 and an optimized threshold of 0.45. At the default threshold (0.5), the model achieves balanced performance with higher specificity and precision. When the threshold is lowered to 0.45, recall (sensitivity) improves, resulting in fewer false negatives, while precision and specificity decrease slightly. This trade-off is particularly relevant in clinical screening scenarios, where minimizing missed disease cases is often prioritized over reducing false positives.

Confusion Matrix Comparison:
Default (0.5): TN=1094, FP=74, FN=112, TP=225
Optimal (0.45): TN=1052, FP=116, FN=92, TP=245

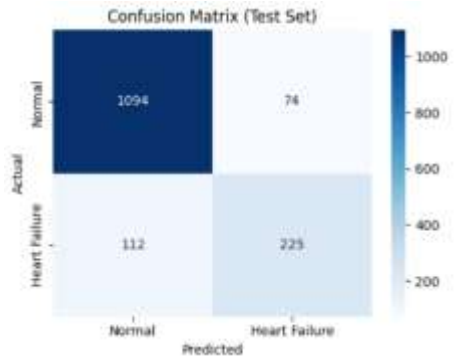

Fig -8: Confusion matrix with threshold 0.5

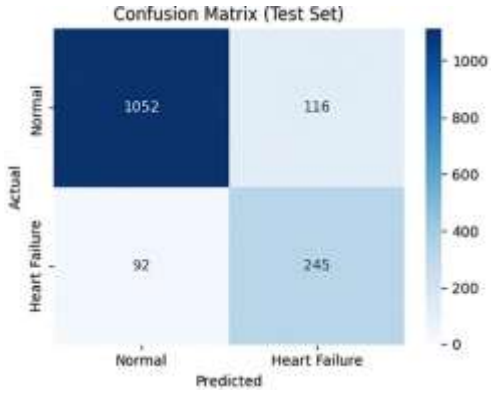

Fig -9: Confusion matrix with threshold 0.45

4. CONCLUSIONS

The proposed system demonstrates an effective modular approach to heart disease prediction by independently modeling clinical tabular data and echocardiography videos. The tabular models achieved strong predictive performance, with XGBoost providing the best balance of accuracy and F1-score, confirming the effectiveness of classical machine learning methods for structured clinical data. The echocardiography model, based on EF-derived labels, showed good discriminative ability, with ROC-AUC indicating reliable separation between normal and abnormal cardiac function. Threshold optimization further highlighted the trade-off between sensitivity and specificity, which is important for clinical screening applications.

The multimodal framework is implemented using a late fusion design, where probability outputs from individual modalities are jointly presented. This ensures modularity and allows independent evaluation of each model while enabling future integration of more advanced fusion strategies.

However, the study has certain limitations. The absence of paired multimodal data prevents direct evaluation of a unified multimodal model. Additionally, EF-based thresholding simplifies a continuous clinical measure and may lead to information loss in borderline cases.

5. Future Scope

The current "Multimodal Heart Disease Prediction System" establishes a strong foundation for integrating clinical and imaging data. However, several opportunities exist to enhance its accuracy, transparency, and real-world utility. The following areas define the roadmap for future development:

1. Advanced Multimodal Fusion: Currently, the system uses a Late Fusion approach (weighted averaging). Future work will explore Intermediate Fusion using Transformer-based architectures. By employing Cross Modal Attention mechanisms, the model could learn deeper connections between specific clinical risk factors (like high blood pressure) and subtle visual patterns in echocardiogram videos, potentially improving accuracy.
2. Explainable AI (XAI) for Clinical Trust To move from a "black box" prototype to a clinical tool, transparency is essential. We plan to integrate:

• SHAP/LIME: To quantify exactly how much each clinical feature (e.g., Age vs. Cholesterol) contributed to the final risk score.
• Grad-CAM Heatmaps: To visually highlight the specific regions of the heart muscle in the video that triggered an "Abnormal" prediction, allowing doctors to verify the AI's logic.

REFERENCES

1. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med. 2020 Oct 16;3:136. doi: 10.1038/s41746-020-00341-z. PMID: 33083571; PMCID: PMC7567861.
2. Ouyang, D., He, B., Ghorbani, A. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020)
3. Multimodal Fusion of Echocardiography and Electronic Health Records for the Detection of Cardiac Amyloidosis Zishun Feng, Joseph A. Sivak, Ashok K. Krishnamurthy, arXiv, 2024
4. Miah, M. A., Rahman, M. M., Uddin, M. M., Rahman, M. A., & Ahmed, M. R. (2023). Heart disease prediction using machine learning algorithms and feature importance ranking. Informatics in Medicine Unlocked, 42, 101251
5. "Machine-Learning Insights from the Framingham Heart Study," Kahouadji N, arXiv/DOAJ, 2024.

BIOGRAPHIES

Koustav Podder currently a final year B.Tech Computer Science and Engineering student at B.P. Poddar Institute of Management and Technology with an interest in Artificial Intelligence, Machine learning algorithms and generative AI applications. Motivated to leverage technical knowledge and problem-solving skills in innovative software development projects

Shubhradip Saha is a final-year Computer Science and Engineering student at the B.P. Poddar Institute of Management and Technology. With an interest in Full Stack Development and Generative AI. he has a proven track record in software engineering, having engineered scalable RESTful APIs and implemented MVC architectures during his tenure at Ardent Computech. He is an Oracle Cloud Infrastructure Certified Generative AI Professional.

Mr. Sudipta Kumar Dutta
Currently working as an Assistant Professor in the department of CSE, at B.P. Poddar Institute of Management and Technology. He has M.Tech degree in computer science from JIS college of Engineering and B.Tech degree in computer science from JIS college of Engineering. His research domain is Artificial Intelligence, Machine learning and Data Mining.