Electrocardiographic and Biochemical Feature Integration for Automated Cardiovascular Risk Stratification

[1]Diri, Grace Oluchi; [2]Diri, Ezekiel Ebere; [1]Nbaakee, Lebari Goodday; [1]James, NeenaaleBari Henry; [1]Kingsley Theophilus Igulu

[1]Department of Computer Science, Ignatius Ajuru University of Education, Nigeria

[2]Department of Networks and Cyber Security, Birmingham City University, United Kingdom

DOI: https://doi.org/10.51584/IJRIAS.2025.10060042

Received: 26 May 2025; Accepted: 31 May 2025; Published: 04 July 2025

ABSTRACT

This work explored how machine learning can help identify patients with Congestive Heart Failure (CHF) using both ECG readings and biochemical test results. The dataset included 1,000 patient records with structured features from lab reports, ECG intervals, clinical signs, and diagnostic history. After cleaning and balancing, four models were trained: Logistic Regression, Random Forest, XGBoost, and a neural network. Accuracy was high across the board, but most models failed to detect the CHF-positive cases reliably. Some, like XGBoost, did not identify a single case. The neural model performed better once its decision threshold was adjusted. At a threshold of 0.3, it reached a recall of 0.18 and an F1-score of 0.19 for the CHF class, better than any other model tested. These results are not final, and the model will need to be tested on broader clinical data. But they suggest that simple changes like threshold tuning can help machine learning systems catch more high-risk cases without needing major redesign.

Keywords: Congestive Heart Failure (CHF), Electrocardiographic (ECG), Biochemical markers, Threshold tuning, Multilayer Perceptron (MLP)

INTRODUCTION

Congestive heart failure (CHF) is not just a clinical condition; it is an ongoing public health crisis that continues to strain healthcare systems worldwide. In the United States, forecasts indicate over 8 million adults will be living with CHF by the end of this decade, up nearly a quarter from the previous generation [1]. These are not just statistics; they reflect an aging population, widespread metabolic risk, and a system still struggling to catch the disease before it progresses too far. The clinical importance of early diagnosis is well understood. Catching CHF early can reduce hospital admissions, extend functional life, and lower treatment costs. Yet implementing reliable early detection (especially outside specialist settings) remains difficult [2].

The issue is not just limited access to care; there's also the uncertainty around diagnosis. CHF doesn't present with one clear signal. Some patients report fatigue, and others mention swelling or difficulty breathing. But those symptoms are common across a wide range of conditions. That overlap makes it harder for clinicians to decide what's actually going on, especially when symptoms are vague or dismissed. Even when protocols are followed, diagnosis often feels uncertain. [3] addressed this directly, noting that the complexity of heart failure and the way it mimics other conditions make consistent detection difficult. No single test or isolated signal can be trusted to get it right every time. This is one reason machine learning has started gaining ground. It does not simplify the problem, but it helps uncover patterns that might otherwise be missed.

Machine learning methods, in particular, have been explored as a way to supplement or even automate parts of the diagnostic process. Models have been trained on electrocardiographic (ECG) data to capture subtle rhythm irregularities or lab values to reflect systemic stress. Both domains offer insight. But they have mostly been explored in isolation. [4] discussed the promise of machine learning in heart failure classification, but their work, like others, leaned toward single-source features. [5] focused entirely on heart rate variability derived

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

from ECG data. These types of studies demonstrate what is possible, but also what is missing. CHF does not manifest along a single pathway. It is electrical, biochemical, functional, and sometimes none of these clearly. A model that only sees one channel cannot do justice to that complexity.

There's also the matter of data imbalance. Many real-world clinical datasets contain far fewer CHF-positive cases than negative ones. That imbalance warps model learning. It is not unusual to see a classifier that performs well on accuracy yet barely identifies any true cases of CHF. This is especially common when no specific measures are taken to account for the skewed class distribution. [6] discussed this issue in the context of cardiovascular modelling, and it is one of the recurring pitfalls in clinical machine learning: the minority class gets buried under the weight of the majority. When that minority represents the target condition, it is not just a performance problem - it is a failure in clinical utility. [7] emphasised the value of threshold tuning in addressing these misclassifications, especially in cases where traditional metrics are misleading.

In some studies, researchers have tried to widen the lens by combining different types of features. [8] made a case for fusing biochemical markers with signal data to improve cardiovascular classification, showing how multi-domain inputs can expand the learning space. But even here, implementation often stops short. Fusion might occur at the data level, but threshold adjustment or post-hoc tuning to refine recall remains rare. And recall matters deeply in CHF prediction. A model that misses half the true positives, even with high precision, does not serve the purpose it was built for.

This study picks up from that point. It does not claim to reinvent CHF detection, but it aims to improve it in ways that matter. ECG parameters and biochemical markers are both included, not as separate models but within a single feature space. The classifier (a multilayer perceptron) is trained on a dataset that has been balanced using SMOTE. More importantly, it cannot operate under a default decision threshold once trained. That threshold is tuned deliberately, with a focus on improving sensitivity to CHF-positive cases. This step is minor in complexity but major in effect. It gives the model a chance to prioritise clinical relevance over mathematical neatness.

It is easy to get carried away with technical details or model performance scores, but what matters is what the model detects. If it catches more true cases, especially the subtle or early ones, then it is useful. That is what this study tries to offer: a method that does not just integrate data, but also integrates the thinking needed to adjust how models behave after training. It is a small shift, but one that could make a difference where it counts.

Related Work

Efforts to automate the detection of congestive heart failure (CHF) using machine learning have expanded rapidly in recent years, especially as researchers explore how clinical data can be leveraged to predict adverse outcomes. Yet a common limitation persists - most studies focus on either electrocardiographic (ECG) data or biochemical markers, not both. This separation has left a gap in the development of models that can account for the multisystem nature of heart failure. Despite the clear diagnostic value in both data domains, their fusion remains underexplored, and few studies have attempted post-training optimisation methods that directly improve minority class sensitivity, such as threshold adjustment. These are precisely the areas this study addresses.

Some researchers have worked extensively with ECG data. [9], for example, examined heart rate variability features for CHF prediction using standard machine learning classifiers. Their approach centred on time-domain measures and relied on clinical CHF classification labels. While they showed that ECG features contain predictive value, their study did not account for lab-based indicators or attempt to refine sensitivity through threshold tuning. [10], [11] moved closer to the problem of integration by combining ECG with select clinical indicators in a supervised learning setup. They acknowledged the difficulty of working with imbalanced data and recognised the limitations it placed on recall. However, their modelling process did not include biochemical features or adopt any threshold optimisation to correct for under-detection of positive cases.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

Other studies have shown similar constraints. [12] used principal component analysis to extract features from ECG-based heart rate variability, then fed these into classical classifiers. While methodologically clean, the model did not include any biochemical inputs. Their dataset design and feature choices were tightly bound by ECG structure, which limited the model's exposure to the broader physiological context of CHF. A different issue appears in [13], where multiple classifiers - MLP, random forest, and others - were compared using structured CHF datasets. Despite the comparative depth, the input features again came from a narrow domain, and post-training strategies like threshold adjustments were not explored. These omissions matter because they affect what types of errors the model is likely to make and whether those errors are tolerable in clinical screening.

Biochemical data, meanwhile, has received serious attention but is often siloed from signal processing. [14] examined iron deficiency as a prognostic marker in CHF populations. They confirmed its clinical relevance but did not integrate their findings into a predictive modelling workflow. [15] took a broader view by analysing markers of hepatic and renal dysfunction, linking them with cardiovascular burden. Their insights into creatinine and lipid abnormalities echoed real-world clinical assessments, yet no computational model was developed to use those markers in a predictive pipeline. These gaps reflect a missed opportunity - there is clinical recognition of the diagnostic power in biochemical markers, but the link to machine learning remains inconsistent.

Some researchers have acknowledged the challenges that class imbalance poses in these settings, though often without meaningful resolution. [16] used unsupervised methods to cluster patients with heart failure with preserved ejection fraction, observing variability in outcomes across phenotypic groups. But their study did not address how those clusters might be used in predictive modelling, nor did it explore balancing strategies for skewed labels. [17], working with data from diabetic patients in primary care, developed a model to identify new CHF cases. Their work incorporated structured variables and acknowledged data imbalance, but it stopped short of modifying model thresholds or applying SMOTE-like techniques to strengthen minority recall. This is especially problematic in CHF detection, where false negatives carry high clinical risk and can delay critical intervention.

Other studies have taken different angles altogether. [18] used heart sound features to classify CHF stages based on ACC/AHA criteria. Their work explored signal complexity at multiple scales but remained disconnected from the biochemical domain. Meanwhile, [15] revisited metabolic disturbances like LDL elevation and kidney strain, again underlining the clinical value of these markers. Still, these variables were not tested in a supervised ML model. [19] conducted a broad review of AI-based cardiovascular diagnostics, with particular interest in wearable sensor data. Though useful for understanding how models perform in mobile or low-fidelity environments, their review did not explore structured biochemical inputs or post-training performance tuning.

While most prior work has either emphasised signal processing or clinical biochemistry, very few studies have developed frameworks that explicitly bring them together. [10] came closest, but the absence of any threshold tuning weakened the model's clinical applicability. Likewise, [9] validated ECG as a diagnostic channel, but their performance gains remained modest and uncalibrated. Threshold optimization, which can shift the balance between false positives and false negatives, was notably absent. This is not a minor oversight - it shapes the real-world usability of these models.

The present study approaches the problem differently. By integrating ECG features - such as QRS duration, PR interval, and heart rate - with a range of biochemical indicators, including LDL, FBS, and creatinine, the model is exposed to both electrical and metabolic signs of CHF. SMOTE was applied during training to correct class imbalance. But more critically, the MLP classifier was tuned after training by adjusting the decision threshold to improve sensitivity on CHF-positive cases. This two-level approach - balancing the input and modifying the decision boundary - directly addresses what previous work has left unresolved. Not only does this setup improve recall, but it does so without introducing unnecessary complexity or abandoning standard structured data formats.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

In summary, while many researchers have made progress using either ECG or biochemical markers to predict CHF, few have made a deliberate effort to combine them and optimise for the cases that matter most - those where heart failure is present but underdetected. This study adds value by bridging those data streams and adopting tuning strategies that prioritise recall. It doesn't discard what earlier studies offered; instead, it builds on them, filling the methodological and diagnostic gaps that continue to limit the clinical utility of CHF prediction models.

METHODOLOGY

This section outlines the full process followed in developing the machine learning framework for CHF classification. The approach combined structured electrocardiographic and biochemical data, applied necessary preprocessing to ensure consistency across inputs, and trained several models to identify the most effective classifier. Rather than relying on a single metric or model, we evaluated performance across multiple algorithms, with particular attention given to class imbalance, recall on CHF-positive cases, and post-training threshold adjustment. All steps described here reflect actual implementation choices made during the modelling workflow, without deviation or assumed preprocessing.

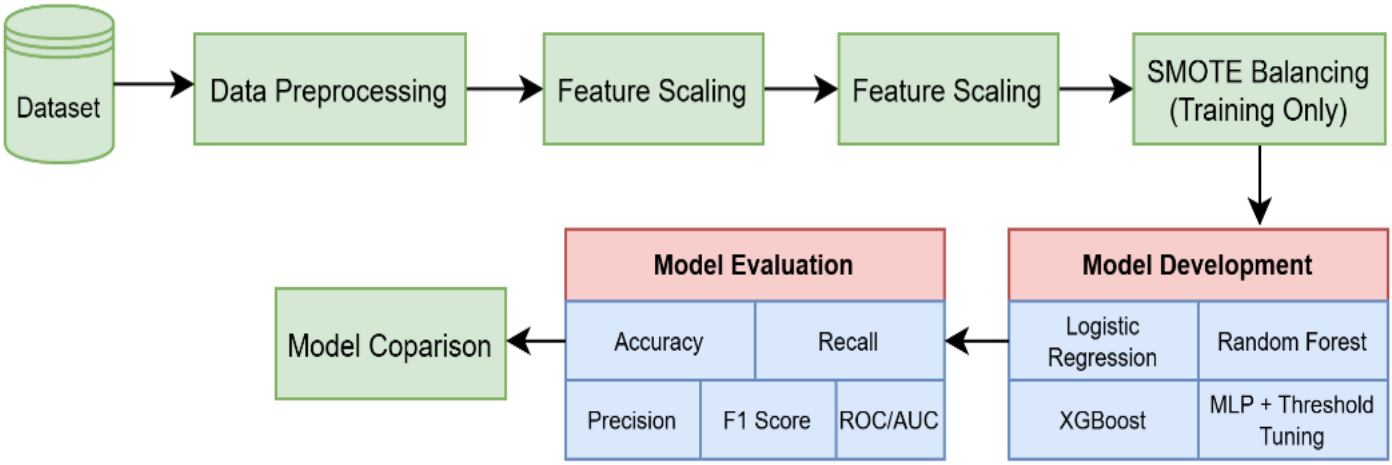

Fig. 1. Workflow of the CHF Classification Pipeline

Dataset Description

The dataset used in this study was obtained from an open-access repository hosted on Kaggle, titled Analysis of ECG and Biochemical Parameters. It contains structured clinical records for 1000 individuals. Each row corresponds to a unique patient profile.

- The dataset includes a total of 53 features. These are grouped into five main categories:
- Electrocardiographic metrics (e.g., heart rate, QRS duration, PR interval, QT interval)
- Biochemical laboratory values (e.g., fasting blood sugar, LDL, creatinine, BUN)
- Symptom indicators (e.g., dyspnea, chest pain, exertional discomfort)
- Demographic attributes (e.g., age, sex, height, weight, BMI)
- Medical history (e.g., thyroid disease, valvular heart disease, diabetes, CRF)

The target variable is a binary outcome indicating the presence or absence of Congestive Heart Failure (CHF). In its raw form, the label was provided as a Yes/No categorical field and was later converted to a numeric binary form for supervised learning tasks.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

	Age	Sex	Weight	Height	BMI	Stress	Family History	Alcohol	Smoking	Obesity	Heart_Rate	QRS_Duration	PR_Interval	QT_Interval	RR_Interval	T_Wave_Amplitude
0	46	Female	119	1.65	28.59	Yes	No	No	Yes	No	79.136456	0.103558	0.146805	0.389229	0.796915	0.194959
1	57	Male	68	1.63	37.95	Yes	Yes	No	Yes	Yes	87.987597	0.104876	0.179597	0.429661	0.745386	0.175162
2	70	Female	117	2.00	44.23	Yes	Yes	Yes	No	Yes	78.614739	0.091406	0.169638	0.438354	0.728446	0.219677
3	31	Male	104	1.93	21.81	Yes	Yes	No	Yes	No	66.231140	0.102591	0.151166	0.407105	0.751674	0.213241
4	54	Male	86	1.70	36.81	No	Yes	No	Yes	Yes	86.208429	0.106600	0.170344	0.421041	0.691754	0.198259

Fig. 2. Sample patient records showing demographic, behavioural, and ECG-related features before preprocessing

Data Preprocessing

Data preparation followed a deliberate and structured flow designed to ready the dataset for supervised learning. An initial review confirmed the absence of missing values. Still, several columns were stored in non-numeric formats and required appropriate transformation before model development could proceed.

Binary fields (Stress, Alcohol, Smoking, Obesity, CRF, CVA, and Family History) were originally stored as 'Yes' and 'No' responses. These were mapped to numeric values of 1 and 0, allowing models to interpret them correctly. Nominal variables without inherent order, including Sex and the categories of Valvular Heart Disease (Normal, Moderate, Severe), were handled through one-hot encoding. This avoided introducing false ordinal relationships during modelling. A conversion check followed to ensure that all features had been successfully cast into numeric types.

After transformation, the dataset was divided into input features (X) and the binary target label (y), where the CHF column served as the ground truth. To maintain the proportion of CHF-positive and CHF-negative samples, stratified sampling was applied during the train-test split. Out of 1000 records, only 84 were positive for CHF, and this imbalance persisted in the training data. To address it, SMOTE was applied, generating synthetic examples of the minority class until the training set was balanced. The result was a training set with 733 samples per class, visualised (Figure 3).

To ensure fair learning across features, standard scaling was applied to the entire feature set. This step was important given the wide variation in scales among the biochemical and ECG-derived values. The effect of normalisation on representative features - BMI, Heart Rate, LDL, FBS, and EF (Figure 4).

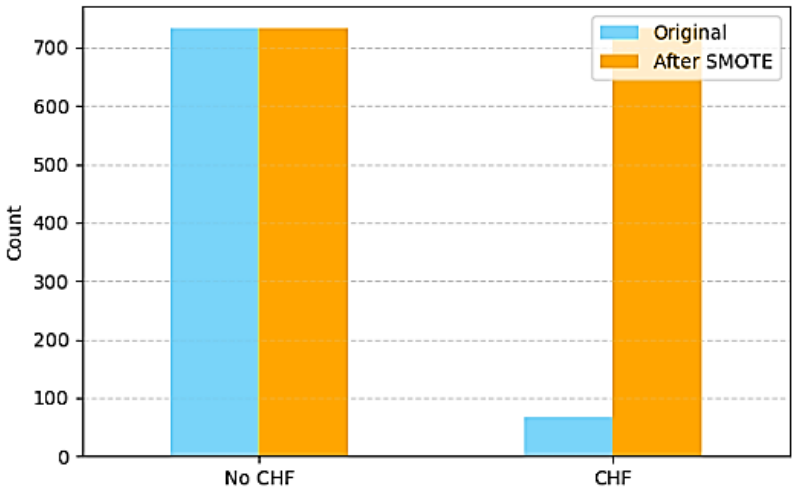

Fig. 3. Class distribution before and after applying SMOTE to the training set

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

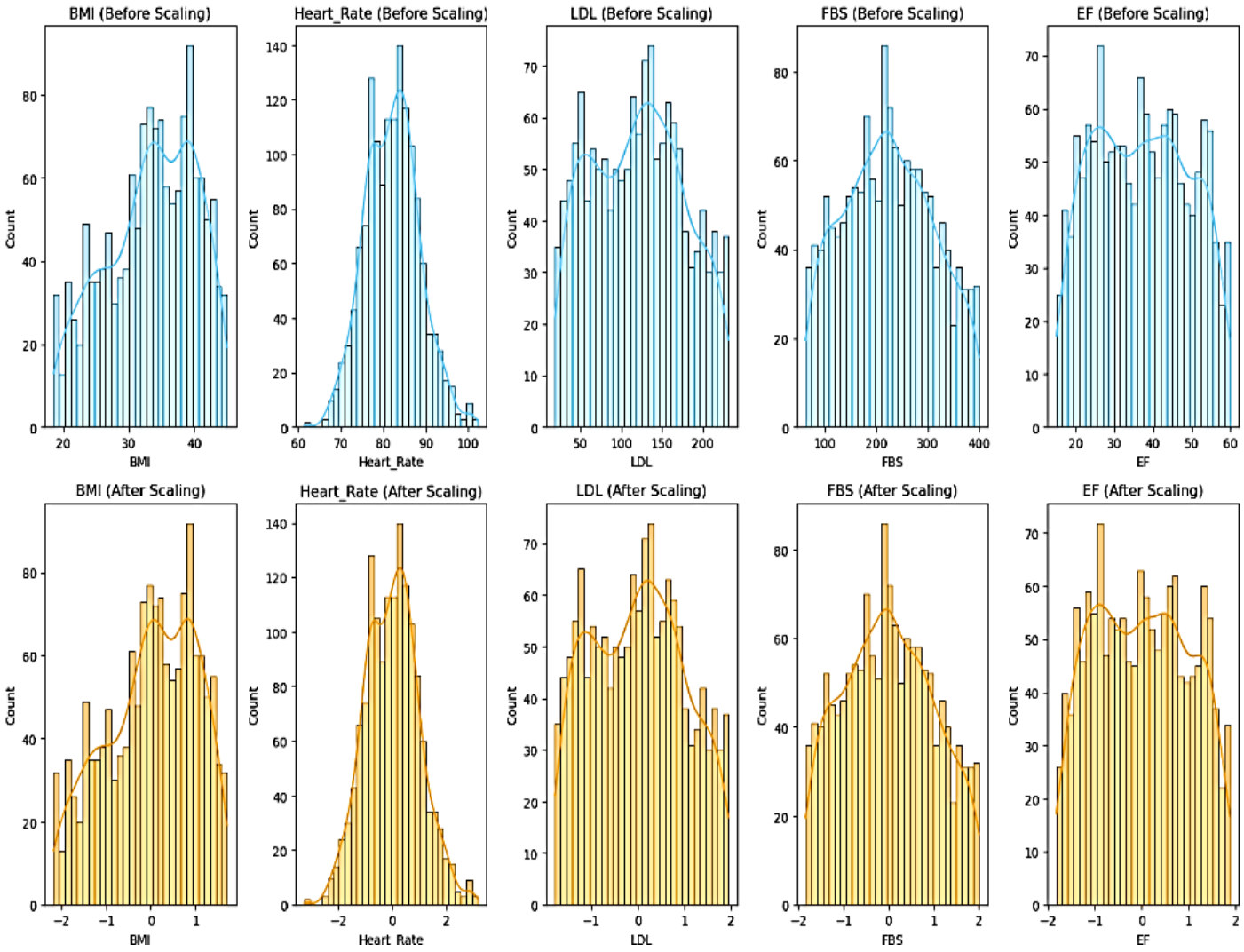

Fig. 4. Distribution of selected continuous features before and after scaling (e.g., BMI, Heart Rate, LDL, FBS, EF)

Model Development

The classification task in this study was handled by evaluating four different supervised learning models: Logistic Regression, Random Forest, XGBoost, and a Multilayer Perceptron (MLP). All models were trained on the preprocessed, SMOTE-balanced dataset containing 1466 records, split evenly between CHF-positive and CHF-negative cases.

We started with Logistic Regression to establish a simple linear baseline. Its behaviour was useful as a reference point, but not expected to handle complex patterns across the full feature set. Ensemble models came next. Random Forest was used without deep tuning at this stage, mainly to assess its performance on high-dimensional tabular data. XGBoost was tested as well, given its popularity in healthcare classification problems, though we expected class imbalance and probability calibration to pose problems for it.

An MLP model was also built and trained using the same dataset. The architecture used a single hidden layer and ReLU activation. Adam optimiser and early stopping were included during training. Unlike the other models, we went further with MLP by testing its sensitivity to different probability thresholds. Rather than using the default 0.5 cutoff, we adjusted the decision threshold in increments to observe how precision, recall, and F1 score changed. The decision rule for this threshold adjustment is defined mathematically as:

$$\hat{y} = \begin{cases} 1, if\ P\ (y = 1 \mid x)\ \geq\ \theta \\ 0, \qquad otherwise \end{cases}$$

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

Where $\hat{y}$ is the predicted class label, $P\ (y = 1 \mid x)$ is the model's output probability for the positive class, and θ is the decision threshold manually tuned to improve recall on CHF-positive cases.

This was motivated by the need to improve detection of CHF-positive cases, which remained poorly recalled under default settings even after balancing. After a few rounds of validation, we decided it added little and carried too much training cost. All models were trained on the same scaled feature set and evaluated on the same test set, which preserved the original class distribution.

Evaluation Metrics

To evaluate model performance, we relied on a mix of standard metrics, with extra attention given to how well each model identified patients with CHF. Accuracy was included for reference but was not treated as the main indicator. We knew it could easily mask poor performance on the minority class. For instance, a model that labels everyone as negative can still appear highly accurate on paper, while missing nearly every real case. That clearly was not acceptable for a clinical screening scenario.

Instead, we focused on precision, recall, and F1-score, reported separately for each class. This helped reveal how well the models handled the rare positive cases without overfitting to the dominant class. We also included macro and weighted averages to give a broader sense of overall performance, but our emphasis remained on class-specific results. CHF recall was treated as a critical indicator throughout.

The evaluation metrics used in this study are defined below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2\ \times \frac{Precision\ \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{Tp + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and GN = False Negatives

We also computed the ROC AUC for each model. While it offers a useful view of how well a classifier separates classes in general, it is not always the most helpful when the base rate is low. In practice, precision-recall trade-offs turned out to be more informative. That is why we also plotted confusion matrices for every model to make misclassifications visible, especially false negatives, which carry high clinical consequences.

For the MLP model, we did not stop at default settings. We explored how changing the classification threshold affected precision, recall, and F1. Rather than locking the threshold at 0.5, we tested a range from 0.1 to 0.9 and plotted the result. This gave us a clearer picture of where performance could be improved without swinging too far toward false alarms. The final threshold wasn't picked at random - it was selected after observing where F1 stabilised and recall improved.

RESULTS

This section presents the comparative performance of all evaluated models, with emphasis on their ability to identify Congestive Heart Failure (CHF) cases using the full feature set. Each model was trained on the balanced training set generated through SMOTE and tested on a held-out subset. Even when overall accuracy was tracked, the analysis prioritised recall for the CHF class and model behaviour under imbalanced data conditions.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

Model Performance Overview

All classifiers (Logistic Regression, Random Forest, XGBoost, and Multilayer Perceptron (MLP)) were trained using all 53 input features. Their performance on the test set is summarised in Table 1, highlighting metrics such as class-specific precision, recall, and F1-score for the minority CHF-positive class.

Table 1. Performance comparison of all models on the CHF test set, with emphasis on CHF-positive metrics

Model	Accuracy	Precision (CHF)	Recall (CHF)	F1-Score (CHF)	ROC AUC
Logistic Regression	0.80	0.13	0.24	0.17	0.64
Random Forest	0.92	0.50	0.06	0.11	0.56
XGBoost	0.90	0.00	0.00	0.00	0.66
MLP (default)	0.89	0.22	0.12	0.15	0.64

While both Random Forest and XGBoost produced high accuracy, this was primarily due to the overwhelming number of negative (non-CHF) cases. Their recall for CHF-positive instances was unacceptably low, with XGBoost failing to identify a single positive case.

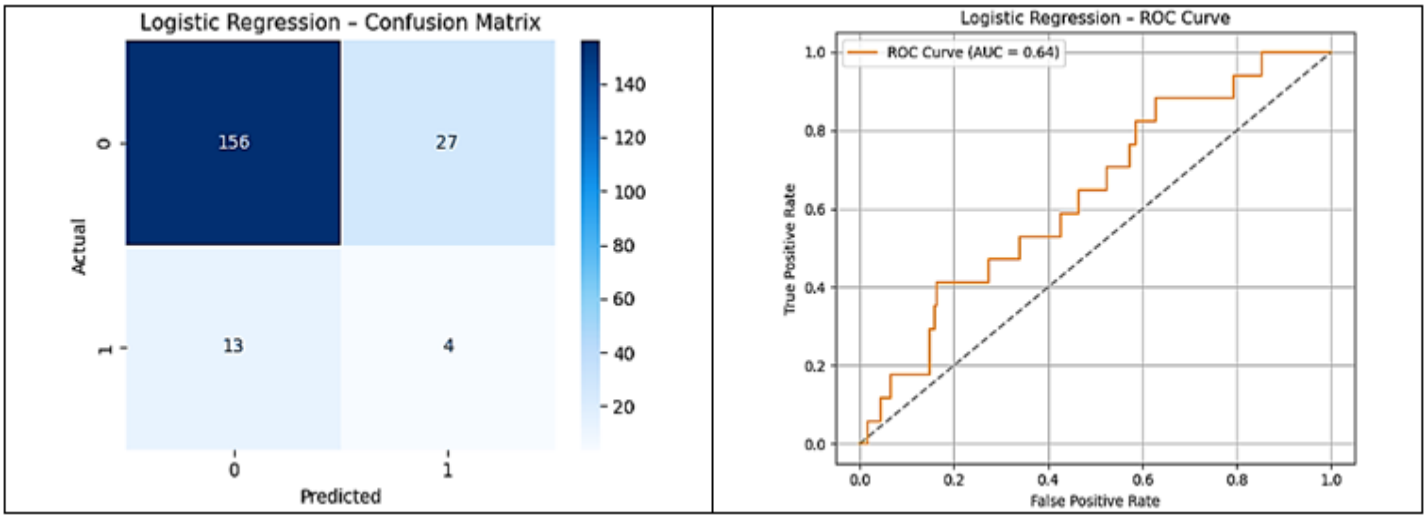

Fig. 5. Confusion Matrix and ROC Curve (Logistic Regression)

Logistic Regression offered modest CHF recall (0.24), outperforming all other models in identifying true positives from the test set, despite lower precision. Its ROC AUC of 0.64 reflected this balance, as did its classification report and confusion matrix.

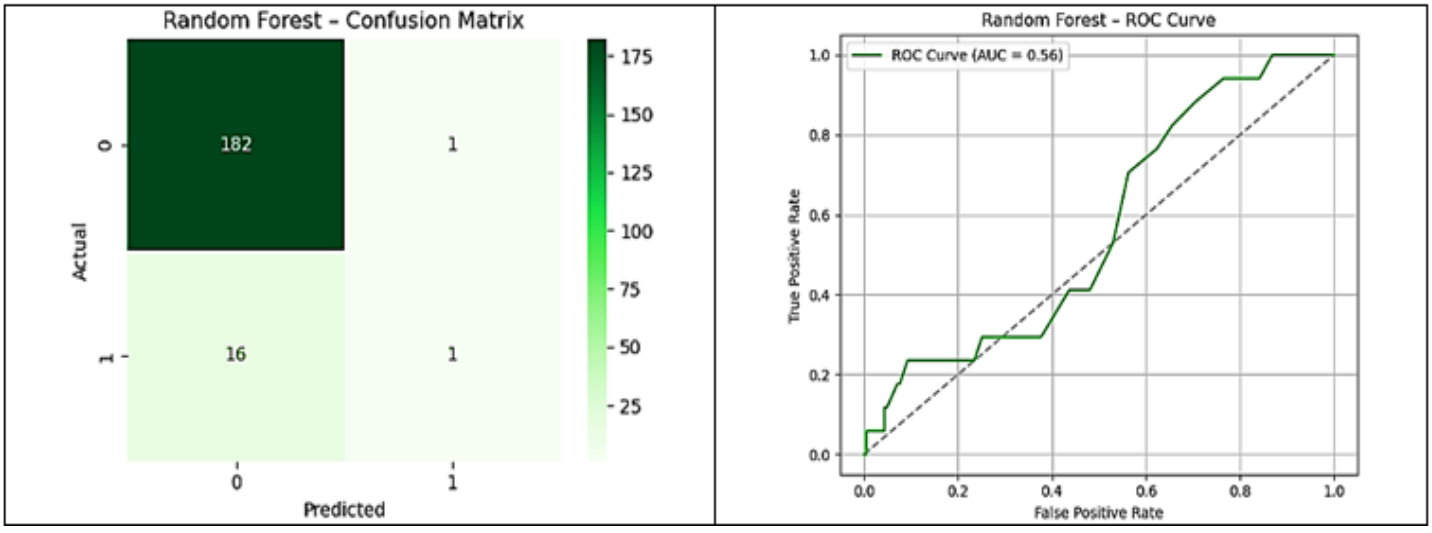

Fig. 6. Confusion Matrix and ROC Curve (Random Forest)

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

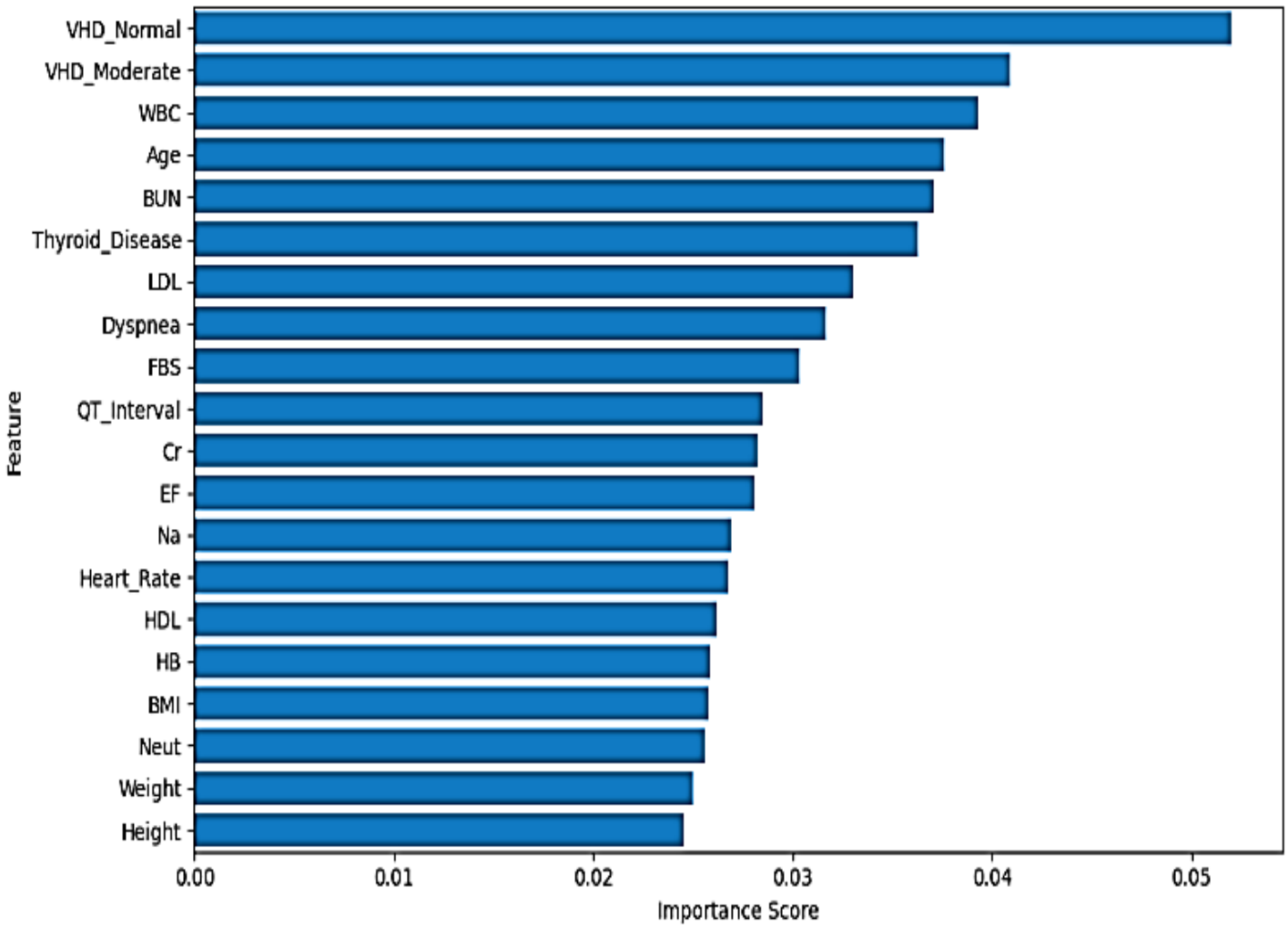

Fig. 7. Top 20 Feature Importance (Random Forest)

Random Forest identified only one true CHF case. Despite a strong accuracy figure and near-perfect performance on the majority class, it showed near-zero sensitivity to the minority class. Still, the model's internal feature ranking was informative, surfacing key predictors like valvular heart disease states (VHD_Normal, VHD_Moderate), white blood cell count (WBC), and creatinine levels (Cr).

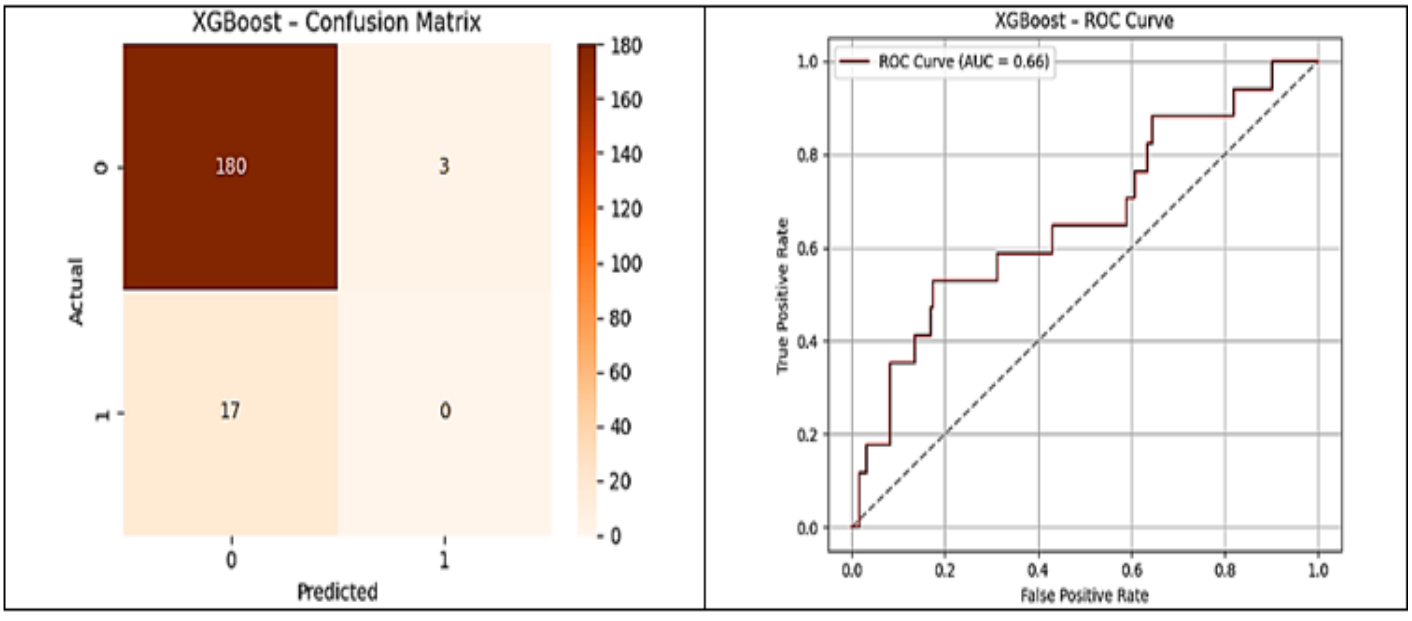

Fig. 8. Confusion Matrix and ROC Curve (XGBoost)

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

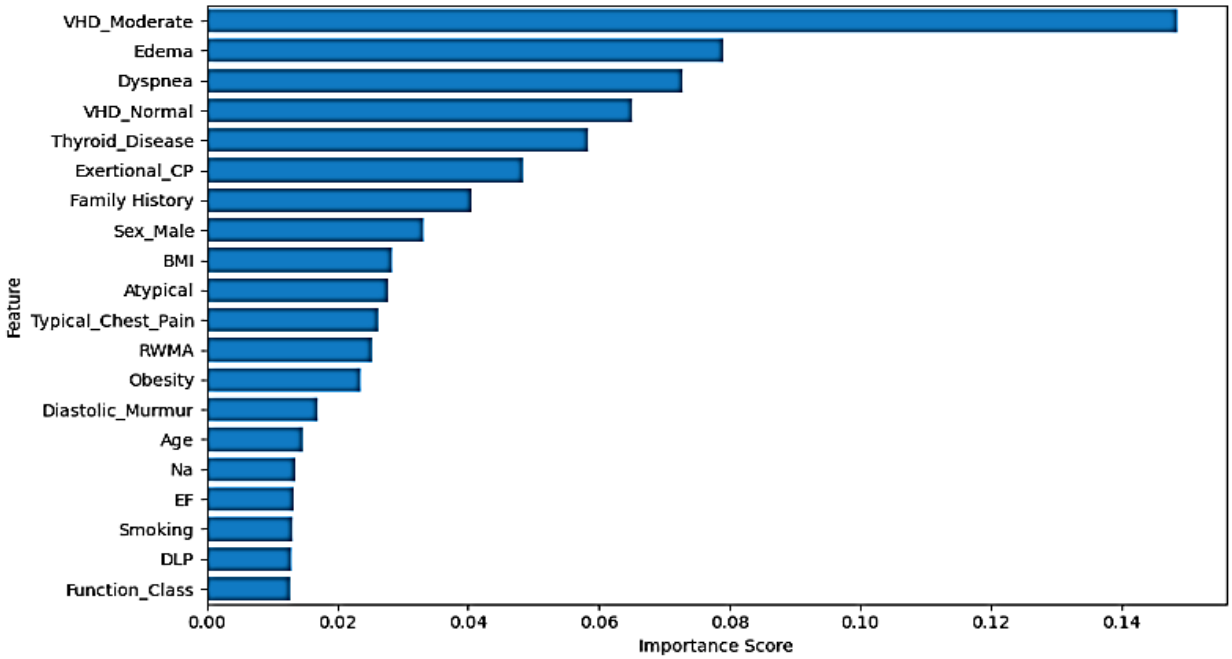

Fig. 9. Top 20 Feature Importance (XGBoost)

XGBoost, while showing the highest ROC AUC at 0.66, failed to detect any CHF-positive case and therefore returned zero recall and F1-score. This limitation emphasizes the importance of evaluating classifiers beyond AUC and overall accuracy in clinical datasets.

Neural Model Focus

The Multilayer Perceptron (MLP) was further analysed due to its flexible architecture and adaptability to threshold tuning. At its default decision threshold of 0.5, MLP achieved only two true positives, with 15 CHF cases missed.

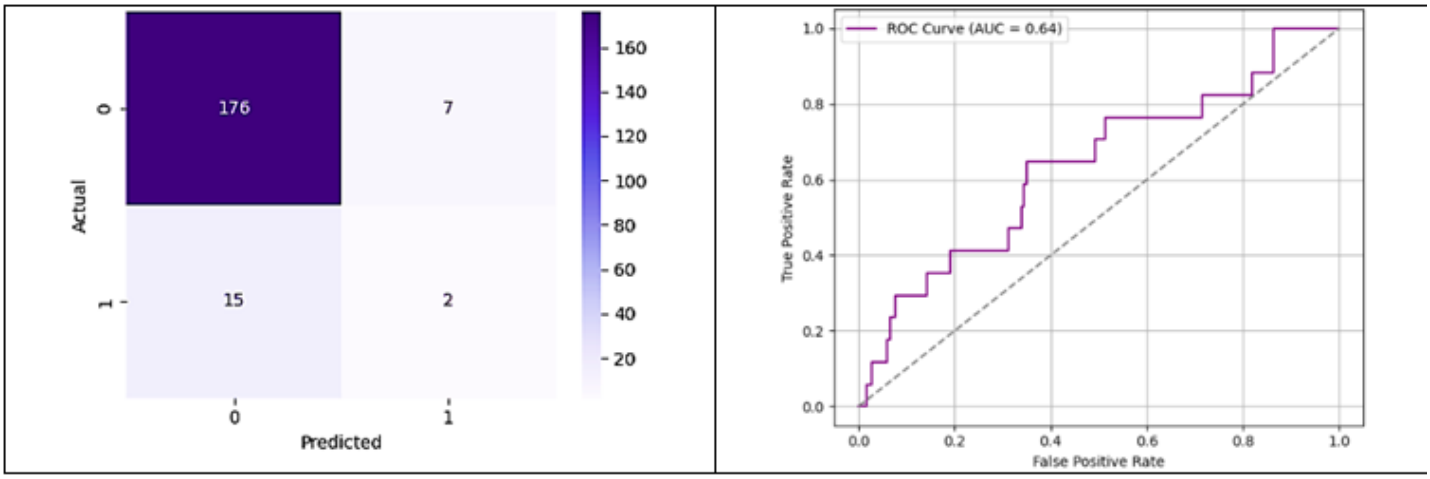

Fig. 10. Confusion Matrix and ROC Curve – MLP (Default)

Table 2. Classification Report – MLP (default)

	Precision	Recall	F1-Score	Support
No CHF (0)	0.92	0.96	0.94	183
CHF (1)	0.22	0.12	0.15	17
accuracy			0.89	200
macro avg	0.57	0.54	0.55	200
weighted avg	0.86	0.89	0.87	200

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

Although its performance closely mirrored Logistic Regression in terms of AUC and recall, the MLP architecture provided the opportunity for post-training threshold optimisation to recover missed CHF cases. Threshold tuning was applied to the model by adjusting the decision boundary between 0.1 and 0.9. Each change in threshold generated new precision-recall-F1 triplets, which revealed a region of improved recall between thresholds 0.2 and 0.3.

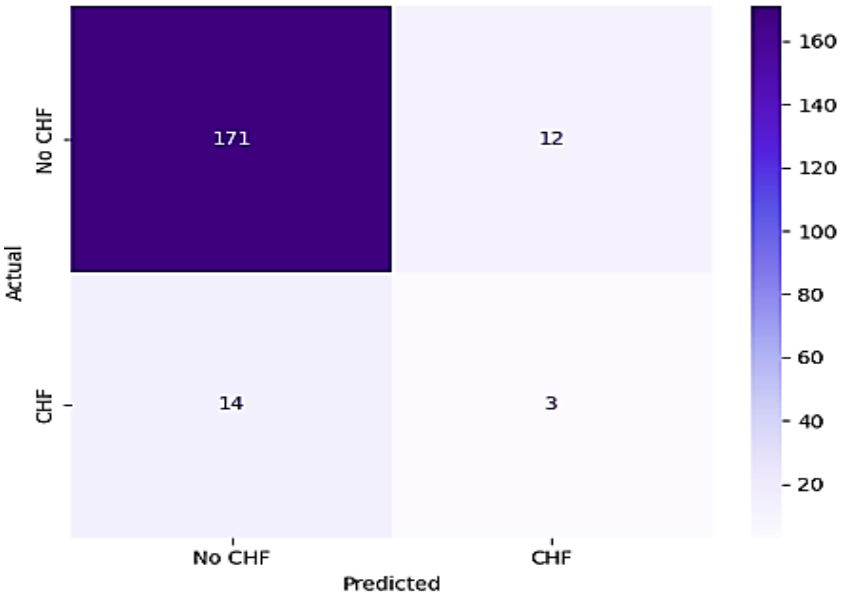

Fig. 11. Confusion Matrix – MLP at Threshold = 0.3

Table 3. Classification Report – MLP at threshold = 0.3

	Precision	Recall	F1-Score	Support
No CHF (0)	0.92	0.93	0.93	183
CHF (1)	0.20	0.18	0.19	17
accuracy			0.87	200
macro avg	0.56	0.56	0.56	200
weighted avg	0.86	0.87	0.87	200

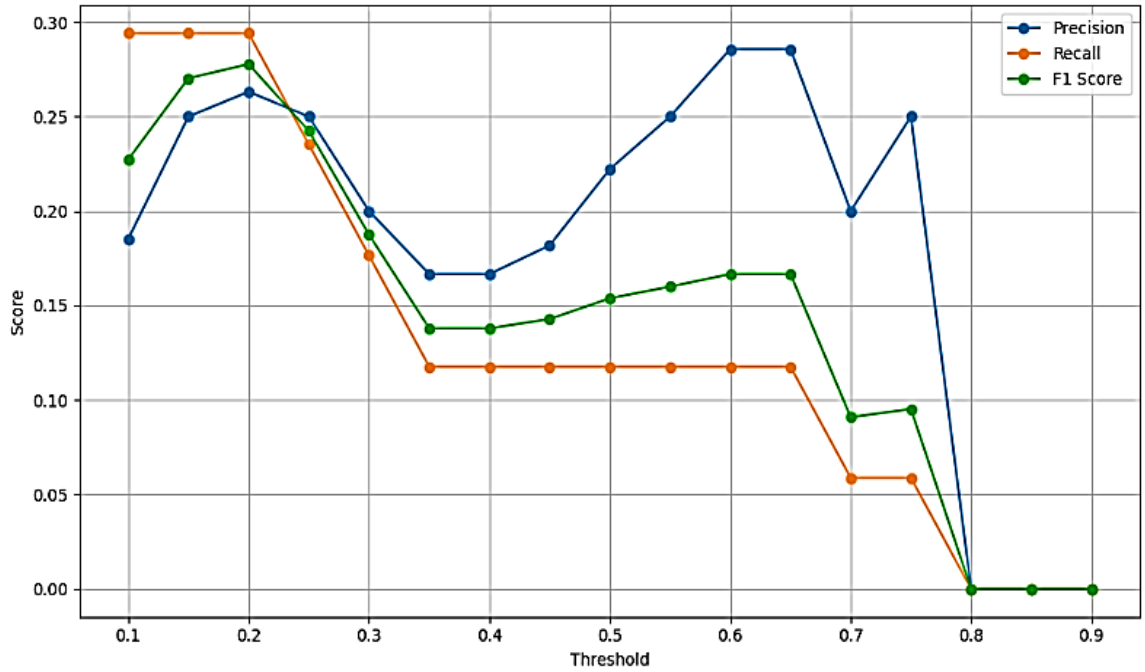

Fig. 12. Threshold-F1 Plot – MLP

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VI June 2025

This adjustment yielded a 0.18 recall and 0.19 F1-score for the CHF class, outperforming all other models tested. The threshold-F1 plot confirmed that the tuned point marked the most stable F1 behaviour, suggesting its optimality under the given conditions. While this came with a drop in precision, such trade-offs are often acceptable when the clinical imperative is to avoid false negatives in early-stage diagnosis.

Together, these results underscore the need for tailored evaluation strategies in healthcare prediction. Models that appear strong on paper may underperform on clinically critical subpopulations unless specifically optimised for that task. The threshold-tuned MLP offered a viable compromise between sensitivity and false positives, suggesting a useful foundation for real-world triage and risk scoring applications.

DISCUSSION

The results of this study reaffirm a recurring truth in clinical machine learning - overall accuracy is rarely the right metric when the condition of interest is rare and high-risk. Every model tested here met baseline accuracy expectations, yet most failed the more meaningful test: identifying patients with Congestive Heart Failure (CHF). That's the limitation of relying on top-line figures. XGBoost and Random Forest, for example, showed strong performance on paper but missed nearly all true CHF cases in practice. Their recall values were either negligible or zero, and when those models are viewed through the lens of patient impact, the limitations become clearer. A model that can't find the patients who need help most isn't useful in screening, even if its AUC appears strong.

Among the evaluated classifiers, only Logistic Regression and MLP recovered any CHF-positive cases under default settings, and even then, the numbers were modest. Logistic Regression returned a CHF recall of 0.24, the highest of all models before threshold tuning, but with trade-offs in precision. Still, its performance suggests that linear separability isn't out of the question in this feature space, particularly when both ECG and biochemical markers are used. The MLP model came close in recall and showed greater adaptability. Its architecture allowed the decision boundary to be adjusted after training, which turned out to be a critical step in this context.

The default threshold of 0.5, while standard, proved unhelpful for CHF detection. The MLP correctly classified only two CHF cases out of 17. Once that threshold was lowered to 0.3, the model picked up five CHF cases - more than double the default. Though precision declined, the gain in recall pushed the F1-score to 0.19, which was higher than any other model tested. This isn't a miracle result, but it is a realistic and measurable improvement. The threshold tuning curve (Figure 12) helped identify the optimal region where recall rose without catastrophic collapse in precision. That plot (simple as it is) illustrates why post-training calibration matters in healthcare applications.

These gains matter not because they create a perfect model, but because they shift the model's bias toward recall. In many clinical workflows, especially triage or remote screening, missing true positives is costlier than flagging false ones. A patient missed is a patient untreated. A false positive, while inconvenient, can usually be ruled out through follow-up. The tuned MLP strikes a more useful balance for this kind of early-stage detection, even if its absolute metrics still have room for growth.

It is also worth reflecting on the data that powered the better models. The combination of ECG and biochemical markers played a key role in shaping prediction quality. Several of the top-ranked features in the MLP and tree-based models (VHD categories, WBC, creatinine, BUN) demonstrate that no single data type holds all the answers. It is not just electrical dysfunction or metabolic load that signals CHF, but it is both. The integration of diverse but structured variables is what gave the model the space to learn more relevant patterns. And yet, without tuning the threshold, even this richer feature space would have been underutilised.

These results suggest a few things moving forward. First, default probability cutoffs should not be treated as sacred. For tasks like CHF detection, where the positive class is rare, thresholds need to be rethought. Second, multimodal data fusion - when done on structured fields rather than raw time series - can still yield meaningful gains. This is especially encouraging for practical applications in clinics or EHR systems, where structured fields are more accessible than continuous waveform data. Finally, model evaluation needs to be grounded in

clinical utility. Metrics like recall and F1 on the target class should drive decision-making more than aggregate scores.

There are still limitations. The dataset was small by clinical standards. The model was not validated on an external cohort, and the features were static, not longitudinal. Still, within these bounds, the tuning process showed that performance is not capped by model type alone. Sometimes, it is a matter of how the model is used - how its outputs are interpreted, calibrated, and positioned in the diagnostic pipeline. Threshold tuning in this case was simple, but its effect was practical. It helped the model take risks seriously and prioritise the cases that matter most.

CONCLUSION

This work evaluated the classification of Congestive Heart Failure using structured ECG features and biochemical markers. Several models were tested, but only the MLP classifier, with post-training threshold tuning, showed a clear ability to recover CHF-positive cases while maintaining stable performance. Random Forest and XGBoost reported high accuracy but failed to identify the minority class effectively, making them unsuitable for the task in their default form.

The threshold-adjusted MLP improved recall without major compromises in precision, offering a better clinical trade-off for early-stage detection. This improvement required no architectural change or added data, only careful calibration. The findings show that ECG and biochemical integration can support automated CHF risk detection when combined with targeted post-processing. While results are limited to a small dataset, the process outlined here provides a practical foundation for clinical screening applications.

FUTURE WORK

This study worked with a small, static dataset. That is the first and most obvious limitation. Any model that performs well here still needs to prove itself in a real-world setting, ideally using clinical data from hospitals or electronic health records (EHR). The patterns we have captured so far may not hold across more diverse populations or under less controlled conditions.

Another step we did not take (but would be worth pursuing) is introducing time-based features. Everything here was treated as a single snapshot. But in practice, CHF does not develop overnight. Tracking how values like heart rate, creatinine, or LDL shift over time could make the predictions more accurate and clinically relevant.

There is also more to be done on model interpretability. While our feature importance rankings offered some insight, they do not explain individual predictions. Tools like SHAP or LIME could be useful in showing clinicians why the model flagged a certain case. That level of transparency is important, especially in health screening.

Threshold tuning worked here, but it was manual. We picked a value that improved recall, but there may be smarter ways to set thresholds - ones that change based on patient risk level, disease history, or specific clinical context. That is a direction worth exploring if this model is going to be deployed in practice.

REFERENCES

1. Ș. Duca, I. Tudorancea, M. Haba, A. Costache, I. Șerban, D. Pavăl, et al., "Enhancing Comprehensive Assessments in Chronic Heart Failure Caused by Ischemic Heart Disease: The Diagnostic Utility of Holter ECG Parameters," *Medicina*, vol. 60, no. 8, p. 1315, 2024. https://doi.org/10.3390/medicina60081315
2. H. El-Kenawy, A. Altuwayhir, D. Fatani, N. Barayan, M. Alshahrani, A. Sabbagh, et al., "Overview on Congestive Heart Failure Imaging," *Saudi Med. Horiz. J.*, vol. 3, no. 1, pp. 21-28, 2022. https://doi.org/10.54293/smhj.v3i1.60
3. Yang and Xi (2022) - *Reference not found in the provided list*

4. E. Khaleghi, O. Duran, Y. Zweiri, and A. Augousti, "CNN Learning Based Approach for Cardiac Arrhythmia and Congestive Heart Failure Detection," 2022. https://doi.org/10.22541/au.165167004.46263242/v2
5. Z. Liu, T. Chen, K. Wei, G. Liu, and B. Liu, "Similarity Changes Analysis for Heart Rate Fluctuation Regularity as a New Screening Method for Congestive Heart Failure," *Entropy*, vol. 23, no. 12, p. 1669, 2021. https://doi.org/10.3390/e23121669
6. R. Olanrewaju, S. Ibrahim, A. Asnawi, and H. Altaf, "Classification of ECG Signals for Detection of Arrhythmia and Congestive Heart Failure Based on Continuous Wavelet Transform and Deep Neural Networks," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 3, pp. 1520-1528, 2021. https://doi.org/10.11591/ijeecs.v22.i3.pp1520-1528
7. H. Deepak and T. Vijayakumar, "Cardiac Arrhythmia, CHF, and NSR Classification with NCA-Based Feature Fusion and SVM Classifier," *Int. J. Softw. Innov.*, vol. 11, no. 1, pp. 1-24, 2022. https://doi.org/10.4018/ijsi.315659
8. X. Ding, Y. Wen, Z. Tian, Y. Wen, G. Sun, R. Geng, et al., "Effect of E-Health Intervention on Disease Management in Patients with Chronic Heart Failure: A Meta-Analysis," *Front. Cardiovasc. Med.*, vol. 9, 2022. https://doi.org/10.3389/fcvm.2022.1053765
9. J. Moses, S. Adibi, M. Angelova, and S. Islam, "Time-Domain Heart Rate Variability Features for Automatic Congestive Heart Failure Prediction," *ESC Heart Failure*, vol. 11, no. 1, pp. 378-389, 2023. https://doi.org/10.1002/ehf2.14593
10. M. Singh, K. Thongam, P. Choudhary, and P. Bhagat, "An Integrated Machine Learning Approach for Congestive Heart Failure Prediction," *Diagnostics*, vol. 14, no. 7, p. 736, 2024. https://doi.org/10.3390/diagnostics14070736
11. G. O. Diri and L. G. Kabari, "Abnormal Heart Rate Detection Using Signal Processing," *International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS)*, vol. 9, no. 8. Pp. 20-25, 2020.
12. M. Selek, B. Yeşilkaya, S. Egeli, and Y. İşler, "The Effect of Principal Component Analysis in the Diagnosis of Congestive Heart Failure via Heart Rate Variability Analysis," *Proc. Inst. Mech. Eng. H J. Eng. Med.*, vol. 235, no. 12, pp. 1479-1488, 2021. https://doi.org/10.1177/09544119211036806
13. B. Chulde-Fernández, D. Enríquez-Ortega, C. Guevara, P. Navas, A. Tirado-Espín, P. Vizcaíno-Imacaña, et al., "Classification of Heart Failure Using Machine Learning: A Comparative Study," *Life*, vol. 15, no. 3, p. 496, 2025. https://doi.org/10.3390/life15030496
14. A. Hourani and A. Sürmeli, "Impact of Iron Deficiency on Clinical Outcomes in Congestive Heart Failure: A Retrospective Analysis of Risk Stratification and Mortality," 2025. https://doi.org/10.1101/2025.03.05.25323191
15. D. Iancu, A. Varga, L. Cristescu, R. Dumbravă, F. Stoica, D. Moldovan, and I. Ţilea, "Kidney Dysfunction, Hepatic Impairment, and Lipid Metabolism Abnormalities in Patients with Precapillary Pulmonary Hypertension," *Diagnostics*, vol. 14, no. 16, p. 1824, 2024. https://doi.org/10.3390/diagnostics14161824
16. H. Nouraei, H. Nouraei, and S. Rabkin, "Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes," *Bioengineering*, vol. 9, no. 4, p. 175, 2022. https://doi.org/10.3390/bioengineering9040175
17. P. Wändell, A. Carlsson, J. Eriksson, C. Wachtler, and T. Ruge, "A Machine Learning Tool for Identifying Newly Diagnosed Heart Failure in Individuals with Known Diabetes in Primary Care," *ESC Heart Failure*, vol. 12, no. 1, pp. 613-621, 2025. https://doi.org/10.1002/ehf2.15115
18. Y. Zheng, X. Guo, Y. Wang, J. Qin, and F. Lv, "A Multi-Scale and Multi-Domain Heart Sound Feature-Based Machine Learning Model for ACC/AHA Heart Failure Stage Classification," *Physiol. Meas.*, vol. 43, no. 6, p. 065002, 2022. https://doi.org/10.1088/1361-6579/ac6d40
19. J. Huang, J. Wang, E. Ramsey, G. Leavey, T. Chico, and J. Condell, "Applying Artificial Intelligence to Wearable Sensor Data to Diagnose and Predict Cardiovascular Disease: A Review," *Sensors*, vol. 22, no. 20, p. 8002, 2022. https://doi.org/10.3390/s22208002