JOURNAL OF BASIC SCIENCE AND ENGINEERING

Heart Disease Prediction Using Machine Learning: A Stacked Ensemble Approach with Imbalanced Data Handling

Aadarsh chaudhary[1], Aditi Sharma[2]

[1]Centre for Advanced Studies, Dr. A.P.J Abdul Kalam Technical University, Lucknow Uttar Pradesh 226031, India

[2] Department of Computer Science and Engineering, Institute of Engineering and Technology, Lucknow, Uttar Pradesh 226021, India

[2] Faculty of Engineering and Technology, Dr. A.P.J Abdul Kalam Technical University, Lucknow Uttar Pradesh 226031, India

DOI: https://doie.org/10.10399/JBSE.2025446080

ABSTRACT

Heart disease is a major cause of death globally, and hence there is a demand for precise and timely prediction models. In this research, a holistic machine learning model for heart disease classification is introduced based on a real dataset with 16 features derived from various clinical and physiological data,with total size of approximately 320,000 records . The dataset comprised demographic, behavioral, and clinical attributes and was characterized by substantial class imbalance. For this purpose, we have utilized a resampling pipeline that incorporates SMOTE, Borderline SMOTE, and Tomek Links. Various models were trained, such as Random Forest, XGBoost, Logistic Regression, and Deep Learning, all with correlation and mutual information-based feature selection. Optimal performance was realized by employing a stacking classifier with Random Forest and XGBoost as base learners and Logistic Regression as meta-learner, which resulted in 81% accuracy.The proposed ensemble demonstrated outstanding performance overall with accuracy being 80.97% and AUC being 0.8158. However, the heart disease instances' recall remained moderate at 54.68%, confirming the continuing challenge in detecting minority class instances. These findings highlight the strength of ensemble learning. Advanced resampling techniques can be used to enhance clinical risk prediction when there are imbalanced datasets.

Keywords—Heart Disease Prediction, Machine Learning, Ensemble Learning, Stacking Classifier, Class Imbalance, SMOTE, Feature Selection, Clinical Risk Prediction.

1. INTRODUCTION

Cardiovascular diseases (CVDs), especially heart disease, are the leading cause of death in the world, responsible for almost one-third of the total mortality every year as stated by the World Health Organization (WHO). Even with advances in medical science, prevention and early detection of heart disease are still biggest concerns, even in developing areas where healthcare is not readily accessible. The early detection of high-risk individuals would decrease the burden on the healthcare system and lead to better patient outcomes.

Over the past few years, machine learning (ML) has proven to be an effective tool for disease prediction as well as healthcare analytics. With the increasing repository of organized health data, ML models can recognize concealed patterns and couplings hidden among risk factors that may not be apparent with traditional statistical techniques. Real-world application of ML in medicine,

however, has been hindered by challenges such as quality of data, imbalance in classes, interpretability, and being well-validated.

The aim of this research work is to develop an effective heart disease prediction system based on using a variety of machine learning techniques. Used data contains over 320,000 patient records with demographic, behavioral, and clinical features. Categorical features were encoded, continuous attributes were scaled, and class imbalance was handled with a SMOTE - Borderline SMOTE - Tomek Links pipeline during preprocessing. We analyze and compare the performance of several classifiers, such as Random Forest, XGBoost, Logistic Regression, and a deep learning neural network. We also propose a stacking ensemble that synergizes the strengths of several models to achieve better generalization and predictive capability.

The fundamental aim of this research is to develop a strong, interpretable, and sensitive predictive model that can support clinicians in early detection of heart disease. In tackling issues such as data imbalance and decision threshold optimization, this work adds to the increasing domain of intelligent healthcare systems.

2. RELATED WORK

Machine learning algorithms have been used in several studies to predict heart disease using clinical and behavioral data. Basic models like Logistic Regression and Decision Trees have been widely used due to their interpretability and simplicity. Recent research includes ensemble models like Random Forest, Gradient Boosting, and XGBoost, which have the ability to model non-linear relationships and offer improved predictive performance.

In order to overcome the problem of class imbalance—a prevalent issue in medical datasets—research has employed resampling methods such as SMOTE and Borderline SMOTE. Tomek Links have also been employed to remove borderline instances, improving the performance of classifiers. Deep learning algorithms, however robust, tend to be plagued by interpretability issues in clinical applications.

Stacking classifiers have proved to be promising through the combination of several algorithms for enhancing accuracy and generalization. Our research expands on these efforts through the integration of balanced resampling, feature selection, and ensemble modeling to attain greater sensitivity in the prediction of heart disease.

3. LITERATURE REVIEW

The application of machine learning for the prediction of heart disease has matured considerably, progressing from basic statistical modeling to higher-level ensemble and neural network-based techniques. Logistic regression, a conventional approach, has given initial understanding of how risk factors such as age, BMI, blood pressure, and outcomes relate to each other. They, however, are generally less effective in capturing non-linear and high-dimensional interactions, particularly in situations where datasets are imbalanced.

A number of research works have investigated decision tree and support vector machines (SVM) for the detection of heart disease. Such algorithms are capable of learning sophisticated patterns

but overfit, especially when learned with noisy or unbalanced data.Ensemble techniques like Random Forest, AdaBoost, and Gradient Boosting improved generalization by reducing bias and variance. Random Forest stood out in particular as particularly stable and tolerant of large feature sets, with boosting methods such as XGBoost and LightGBM competing strongly for the detection of non-linearity and optimization of the recall.

Class imbalance is still the most critical challenge in heart disease classification. The vast majority of real-world health data contains much more negative (non-disease) instances than positive ones, which makes the models biased toward the majority class. Methods like SMOTE (Synthetic Minority Oversampling Technique) and its extensions (e.g., Borderline SMOTE) have been researched heavily for synthesizing minority class instances. These methods, particularly when used with under-sampling methods like Tomek Links, assist in enhancing recall and F1-score by class distribution balancing.

Furthermore, there have been recent deep learning models that learn hierarchical features without the need for manual selection. Though formidable, these models are computationally expensive and demand large amounts of data, and their black-box nature renders them less interpretable in clinical contexts.

Current studies highlight ensemble and hybrid methods such as stacking to take advantage of the best model. This work takes such a direction further by combining resampling, ensemble learning, and threshold adjustment to enhance both accuracy and sensitivity in the prediction of heart disease.

4. DATA DESCRIPTION

The dataset used in this study comprises 16 features derived from various clinical and physiological data,with total size of approximately 320,000 records samples with a mix of continuous and categorical features. The target variable is binary (HeartDisease: 1 for presence, 0 for absence). The features include:

1. Demographic: AgeCategory, Sex, Race
2. Behavioral: Smoking, AlcoholDrinking, PhysicalActivity
3. Medical History: Stroke, Diabetic, Asthma, KidneyDisease, SkinCancer, DiffWalking
4. Health Metrics: BMI, MentalHealth, PhysicalHealth, SleepTime
5. Self-reported Status: GenHealth

Binary categorical features were label encoded, and multi-class categorical variables were one-hot encoded. Continuous features were standardized using StandardScaler. The dataset exhibited a class imbalance, with heart disease positive cases comprising ~8.5% of total observations. This imbalance was addressed using a resampling pipeline of SMOTE → Borderline SMOTE → Tomek Links applied only to the training set. The final dataset was split into 80% training and 20% testing, with proper stratification.

5. METHODOLOGY

This work suggests an end-to-end machine learning pipeline for predicting heart disease that includes data preprocessing, statistical feature selection, class balancing, model creation, ensemble

learning, and performance evaluation using cross-validation.

5.1 Data Preprocessing

The data set with 3,20,000 records was preprocessed and cleaned by label-encoding categorical variables. Binary categorical variables such as Smoking, AlcoholDrinking, Stroke, and Sex were labeled encoded while multi-class variables such as AgeCategory, GenHealth, and Race were one-hot encoded. Initial correlation analysis was performed in an attempt to identify and eliminate features that were not very relevant to the target variable (HeartDisease).

5.2 Feature Evaluation

A feature evaluation on multiple metrics was performed with Pearson correlation, Point-Biserial correlation, ANOVA F-value, and Mutual Information. These metrics provided the quantification of the interaction between each feature and target variable. Features with minimum correlations were eliminated to avoid noise and enhance model efficiency.

5.3 Feature Scaling

A few continuous features (SleepTime, PhysicalHealth, MentalHealth, and BMI) were standardized by the aid of StandardScaler for equal feature scaling and support of convergence to features' magnitudes by sensitive models.

5.4 Class Imbalance Treatment

To correct extreme class imbalance in the dataset, a resampling approach in three stages was adopted:

1. SMOTE (Synthetic Minority Over-sampling Technique) is employed to create synthetic samples for the minority class.
2. BorderlineSMOTE was utilized to create synthetic instances on the decision boundary to ensure improved class separation.
3. Tomek Links were employed for undersampling the majority class by eliminating borderline overlapping instances.

The sequential resampling gave a balanced training dataset that boosted minority class learning.

5.5 Model Building

Classification models were trained using the resampled dataset:

1. Traditional Machine Learning Models: Logistic Regression, Decision Tree, Random Forest, XGBoost, Gradient Boosting, AdaBoost, and LightGBM.
2. Deep Learning Model: Three hidden layers neural network, batch normalization, dropout regularization, and ReLU activation. Optimized using Adam optimizer and binary cross-entropy loss.

5.6 Cross-Validation

For guaranteeing robustness and combating overfitting, Stratified K-Fold Cross-Validation with 5 folds was utilized throughout training. This method maintained the class proportion in all the folds and allowed trusted approximation of model generalization. The mean and standard deviation of the F1-score over the folds were noted for purposes of comparison in performance.

5.7 Model Evaluation and Threshold Tuning

Models were tested on a hold-out test set with a fixed classification threshold of 0.6. Performance measures were Accuracy, Precision, Recall, F1 Score, ROC AUC, and confusion matrix. ROC and Precision-Recall curves were also drawn, and F1-score was taken as the main measure since the classes were imbalanced.

5.8 Ensemble Learning

A Stacking Classifier was implemented using:

1. Also known as base learners: Random Forest and XGBoost
2. Meta-Learner: Logistic Regression

6. IMPLEMENTATION DETAILS

The pipeline for machine learning was run in Python using libraries such as Pandas, NumPy, Scikit-learn, XGBoost, LightGBM, imbalanced-learn, and TensorFlow. The following steps list the running of the entire process:

6.1 Environment

1. Programming Language: Python 3.10+
2. Platform: Google Colab / Jupyter Notebook
3. Libraries used: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost, lightgbm, tensorflow, imblearn, scipy

6.2 Data Preprocessing

1. Categorical features such as Smoking, AlcoholDrinking, Stroke, and Sex were label-encoded.
2. Multi-class features such as AgeCategory, GenHealth, and Race were one-hot encoded via pd.get_dummies.
3. Continuous features (BMI, PhysicalHealth, MentalHealth, SleepTime) were scaled via StandardScaler.
4. Highly uncorrelated features with the target variable (HeartDisease) were eliminated through statistical analysis.

6.3 Feature Selection

Statistical methods were employed to measure feature importance:

1. Pearson Correlation
2. ANOVA F-value
3. Mutual Information

These values were averaged to rank features that make the prediction. The strength of feature correlation with the target variable is graphically summarized in **Figure 1**.

Vol. 22, No. 1, (2025)
ISSN: 1005-0930

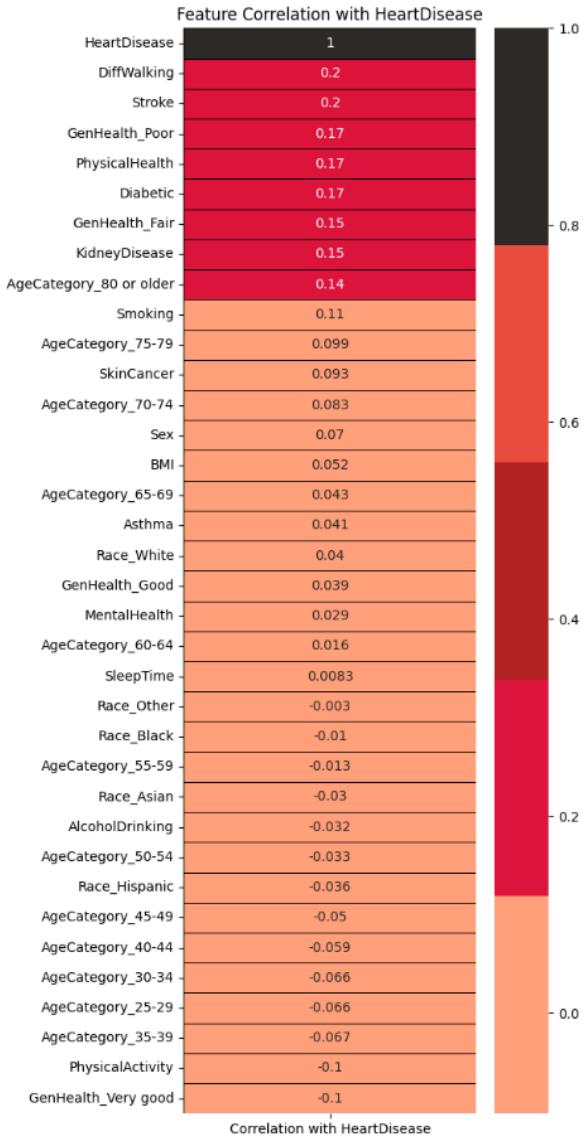

Figure 1. Correlation of features with the presence of heart disease.

6.4 Class Balancing

Because of severe class imbalance (heart disease is the minority class), the following hybrid resampling approach was taken:

1. SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic samples of minority class.
2. Borderline SMOTE to create samples along the decision boundary.
3. Tomek Links to delete borderline majority class examples.
4. This gave a balanced dataset for model training.

6.5 Training the Model

Various machine learning models were trained on the resampled data:

1. Traditional Models: Random Forest, Decision Tree, Logistic Regression, XGBoost, Gradient Boosting, AdaBoost, LightGBM.

2. Deep Learning Model: Three hidden layers feedforward neural network with ReLU activation, Batch Normalization, Dropout, and sigmoid output layer. Trained with Adam optimizer and binary cross-entropy loss.

6.6 Cross-Validation

For robustness ensuring, all the models were tested on Stratified 5-Fold Cross-Validation. Mean and standard deviation of F1 scores were used to measure stability over various data splits.

6.7 Threshold Tuning

A fixed threshold of 0.6 was used to trade-off between precision and recall for the minority class. The threshold was chosen by plotting F1-score vs. threshold using precision_recall_curve. As evident from **Figure 2**, precision-recall curve over thresholds facilitated the visualization of the trade-off, whereas **Figure 3** shows the peak F1-score at the best threshold of 0.57, informing the selection of a threshold.

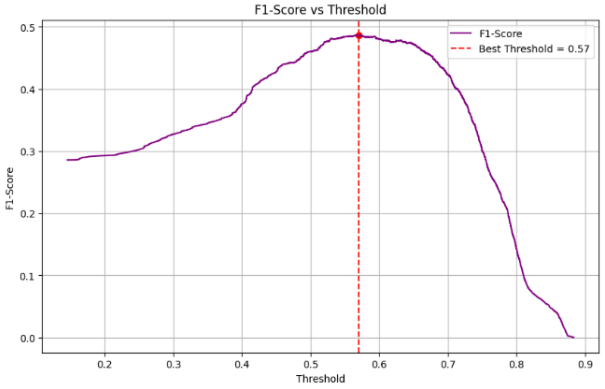

Figure 2. Precision and recall vs. threshold.

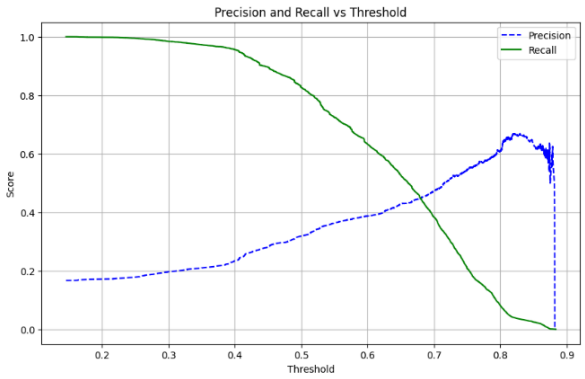

Figure 3. Precision and recall vs. threshold

6.8 Ensemble Model

A Stacking Classifier was built with:

1. Base Models: Random Forest and XGBoost
2. Meta-Learner: Logistic Regression

The stacking model was trained on the last resampled dataset and validated using the same threshold and metrics.

6.9 Evaluation Metrics

All the models were validated on the following metrics on a held-out test set:

1. Accuracy
2. Precision
3. Recall
4. F1 Score
5. ROC AUC Score
6. Confusion Matrix
7. ROC Curve

7. RESULT

Vol. 22, No. 1, (2025)
ISSN: 1005-0930

This work compared several machine learning and deep learning models for predicting heart disease based on a sample dataset of 320,000. These models were trained on a balanced dataset using SMOTE, Borderline SMOTE, and Tomek Links, and tested on a distinct test set. A threshold value of 0.6 was employed to maximize the F1-score for the minority class (heart disease patients).

7.1 Individual Model Performance

The performance of all models was assessed using accuracy, precision, recall, F1-score, and ROC AUC. The deep learning model and ensemble models showed competitive performance. A detailed comparison of evaluation metrics across all models is presented in **Table 1**, which highlights that Random Forest achieved the highest overall accuracy (81.20%), while Logistic Regression had the highest recall (57.51%), indicating its effectiveness in identifying the positive class. The comparative F1-scores of all models are presented in **Figure 4**, highlighting Logistic Regression as the top performer. The corresponding ROC curves with AUC values for each model are shown in **Figure 5**, providing insight into classification trade-offs. Additionally, **Figure 6** displays the confusion matrix for the Logistic Regression model, detailing true and false classifications. Together, these figures offer a comprehensive evaluation of model performance across multiple metrics.

To further evaluate the quality of predicted probability distributions, Binary Cross-Entropy (BCE) loss was computed for all models (**Figure 7**). Lower BCE corresponds to well-calibrated probability estimates. LightGBM had the lowest BCE (0.4066), implying high probabilistic confidence, followed by Gradient Boosting and XGBoost.

Table 1. Performance metrics of various classification models

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest	0.8120	0.4553	0.5496	0.4984	0.8215
XGBoost	0.8043	0.4359	0.5482	0.4866	0.8193
Logistic Regression	0.7687	0.3798	0.5751	0.4577	0.8036
Gradient Boosting	0.8001	0.4305	0.5370	0.4781	0.8146
AdaBoost	0.7922	0.4208	0.5364	0.4714	0.8134
LightGBM	0.8072	0.4390	0.5475	0.4874	0.8154
Decision Tree	0.7916	0.4110	0.5257	0.4617	0.7976
Deep Learning	0.7994	0.4261	0.5429	0.4779	0.8123

Vol. 22, No. 1, (2025)
ISSN: 1005-0930

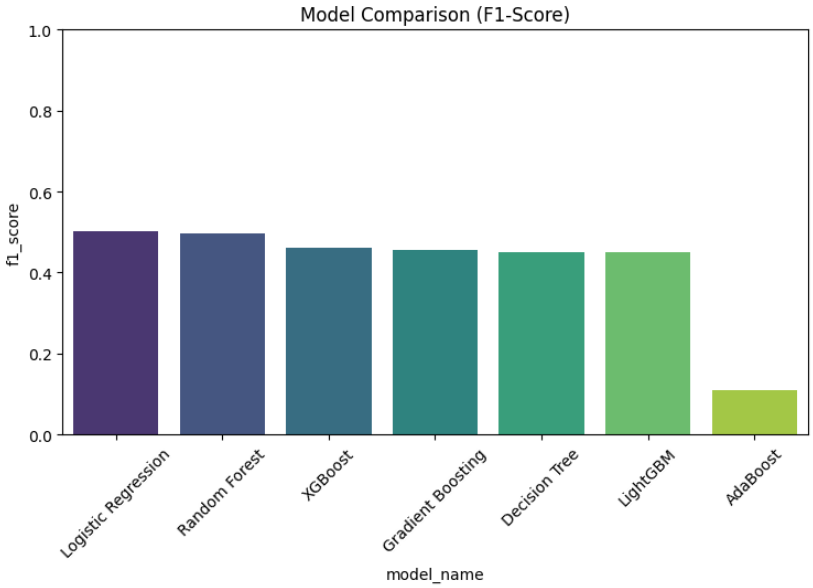

Figure 4. F1-Score Comparison of Classifiers

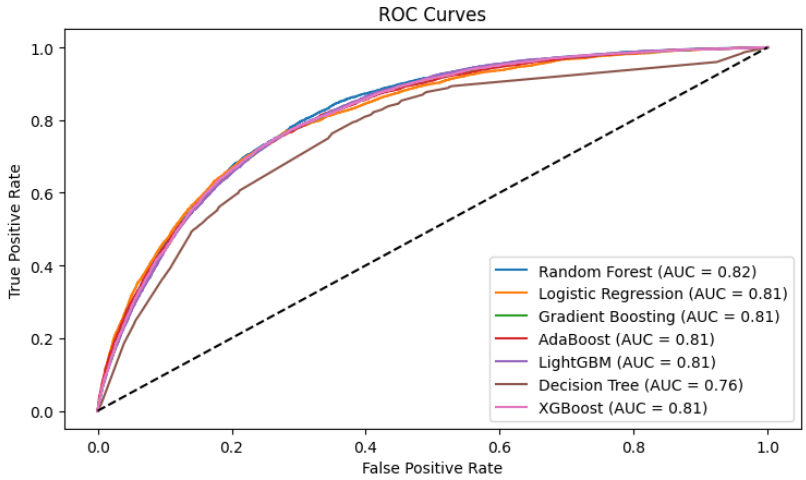

Figure 5. ROC Curves of Classifiers

Vol. 22, No. 1, (2025)
ISSN: 1005-0930

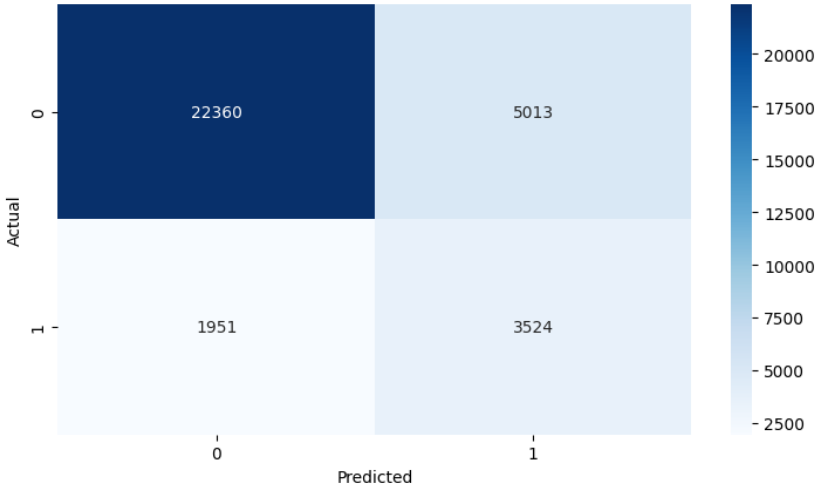

Figure 6. Confusion Metrics - Logistic Regression

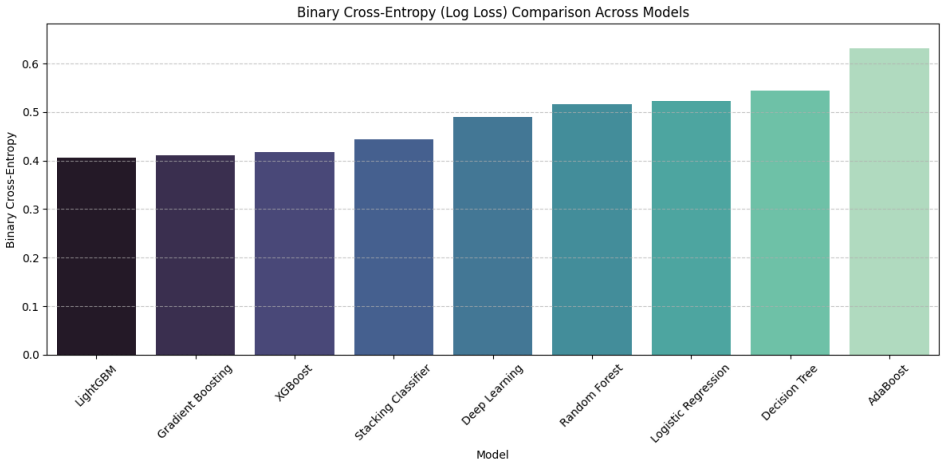

Figure 7. Binary cross-entropy (log loss) comparison across models

7.2 Stacking Classifier Performance

Best performance was achieved by the Stacking Classifier, with Random Forest and XGBoost as base learners and Logistic Regression as the meta-learner. It produced:

1. accuracy: 80.97%
2. precision: 44.27%
3. recall: 54.68%
4. f1-score: 48.93%
5. roc_auc: 0.8158

This model's classification report is presented in **Figure 8** displaying precision, recall, and f1-score for both classes and macro and weighted averages.

Vol. 22, No. 1, (2025)
ISSN: 1005-0930

	precision	recall	f1-score	support
0	0.90	0.86	0.88	27373
1	0.44	0.55	0.49	5475
accuracy			0.81	32848
macro avg	0.67	0.70	0.69	32848
weighted avg	0.83	0.81	0.82	32848

Figure 8. Classification report of the stacking model with 81% accuracy

7.3 Insights

1. Stacking model outperformed all the individual models, particularly on recall and AUC, which are crucial in heart disease detection.
2. While minority class recall (heart disease) improved significantly, it remains difficult due to class imbalance and overlapping features.
3. Ensemble learning and advanced resampling were both found useful for the clinical classification task.

8. MATHEMATICAL FORMULAS

Evaluation Metrics

These metrics are derived from the confusion matrix:

1. TP (True Positive): Correctly predicted heart disease instances
2. TN (True Negative): Correctly predicted non-heart disease instances
3. FP (False Positive): Non-heart disease instances incorrectly predicted as heart disease
4. FN (False Negative): Heart disease instances incorrectly predicted as non-heart disease

8.1 Accuracy

Accuracy measures the overall proportion of correctly classified instances among all predictions. It is defined as:

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

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

8.2 Precision

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It is calculated as:

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

Vol. 22, No. 1, (2025)
ISSN: 1005-0930

where TP is true positives and FP is false positives. A high precision indicates a low false positive rate.

8.3 Recall (Sensitivity / True Positive Rate)

Recall, also referred to as sensitivity or true positive rate, quantifies the proportion of actual positive cases correctly identified by the model. It is expressed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP is true positives and FN is false negatives. High recall is crucial when missing positive cases has serious consequences.

8.4 F1-Score

The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure that accounts for both false positives and false negatives, particularly useful in imbalanced datasets. It is calculated as:

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

8.5 Binary Cross-Entropy (Log Loss)

Binary Cross-Entropy (BCE), also known as Log Loss, measures the dissimilarity between predicted probabilities and the actual binary class labels. It evaluates how well a model's predicted probability distribution matches the true labels. The formula is:

$$\text{Binary Cross} - \text{Entropy} = -N1i = 1\sum N[yi \cdot log(y\text{^}i) + (1 - yi) \cdot log(1 - y\text{^}i)]$$

where:
N is the total number of samples
yi is the actual label (0 or 1) for sample i
$y\text{^}i$ is the predicted probability for class 1 for sample i

9. DISCUSSION

This work illustrates the efficiency of ensemble and traditional machine learning methods for prediction of heart disease on an imbalanced data. The optimal performance was obtained by a stacking classifier with Random Forest and XGBoost as the base learners and Logistic Regression as the meta-learner and it had an accuracy of 80.97%, an AUC of 0.8158, and recall of 54.68% for the positive class. Despite application of advanced resampling methods (SMOTE, BorderlineSMOTE, Tomek Links) and threshold tuning, recall performance was not improved, an indication of the lingering issue of minority class instance discovery.

Conventional models like Logistic Regression and Random Forest provided interpretable and relatively balanced accuracy. Deep learning offered slight recall improvement but at the expense of interpretability, which is important in clinical use. The future study must explore the integration

of domain knowledge, cost-sensitive learning, and explainability techniques such as SHAP values to enhance the transparency and reliability of the models towards better clinical applicability of machine learning-based predictive systems.

10. CONCLUSION

This research delivers an end-to-end machine learning pipeline for heart disease prediction using a combination of classical models, deep learning, and ensemble-based methods. The pipeline entailed rigorous data preprocessing, including label encoding, one-hot encoding, feature correlation analysis, and class imbalance remedy by SMOTE, Borderline-SMOTE, and Tomek Links. These assisted in offering balanced information and facilitating learning in the model.

Out of all the models compared, stacking classifier with Random Forest and XGBoost as base learners and Logistic Regression as the meta-learner performed the best with an accuracy of 80.97% and AUC of 0.8158. While that reflects high generalization, positive instances of heart disease were still 54.68% recalled, reflecting the consistent issue with identifying minority class instances important to clinical diagnosis.

Earlier techniques such as Logistic Regression and Random Forest were interpretable and stable, whereas deep learning resulted in small gains in recall and lost explainability.Resampling and threshold adjustment improved sensitivity but the trade-off in specificity suggests that more optimization is necessary.

Future research will involve combining domain knowledge, cost-aware learning, and explainability AI methods like SHAP values. These methodologies can greatly enhance the real-world applicability and reliability of AI systems in actual healthcare settings.

REFERENCES

[1] Hong, S., Lee, S., Lee, J., Cha, W. C., & Kim, K. (2020). "Prediction of cardiac arrest in emergency department based on machine learning and sequential characteristics: Model development and retrospective clinical validation"
[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
[4] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS) (pp. 4765–4774).
[5] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).
[7] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
[8] Lee, H., Yang, H. L., Ryu, H. G., et al. (2023). "Real-time machine learning model to predict in-hospital cardiac arrest using heart rate variability in ICU". npj Digital Medicine, 6, 215.
[9] Data source]: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease