Grenze International Journal of Engineering and Technology, June Issue

Deep Meta- Ensemble Learning with SMOTE for Heart Disease Prediction: A Hybrid Approach

Chaithra C S[1] and Siddesha S[2]
[1,2]Department of Computer Applications, JSS Science and Technology University, Mysuru, India
Email: chaithracs@jssstuniv.in

***Abstract*—heart disease is still a global challenge in health sector as a major reason for death. American Heart Association (AHA) reported statistics in the year 2024 about 19.91 million global deaths is due to cardiovascular disease. Hence early prediction helps in diagnosis which required reliable predictive models helps in assisting early diagnosis and preventive care helps in reducing the risk of death. This study investigates Machine Learning (ML) models and gradient boosting models performance in predicting the disease accurately. The study includes the standard dataset, which is available in data repositories like UCI and Mendeley. The study evaluates base algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector classification, Naive Bayes and gradient boosting models like XG Boost, LightGBM, CatBoost with SMOTE and without SMOTE for prediction of heart disease which show importance of data balancing in machine learning. This study gives comparison analysis of traditional ML algorithms and gradient boosting algorithms and hybrid ensembled meta learning neural network framework integrating SMOTE for handling imbalance class and meta learning with Neural Networks for efficient performance for heart disease prediction. The hybrid ensembled meta-learning framework outperformed of all other traditional methods by achieving accuracy of 94.49% and AUC-ROC of 95.44%. This study highlights that ensembled classification model when integrated with deep learning model improves their performance on balanced datasets.**

***Keywords*— cardiovascular disease, ensemble learning, SMOTE, prediction, neural network, meta learning, hybrid model, meta learning, voting classifier.**

I. INTRODUCTION

Heart disease is one of the major causes of increase in death rate globally. According to the global statistics released by American Heart Association (AHA) the death rate is increased in the year 2022 [1]. Maintaining the patient's data and monitoring the patients' health is one of the major concerns for health care sectors. To help the healthcare practitioner in this regard there is a necessitate of machine intelligence which plays a major role in all the sectors mainly in healthcare. The intervalence of ML algorithms in accurate prediction of patient's heart condition will reduce the risk of severance of mortality which helps healthcare practitioner to identify for further enhanced treatment.

The major issue with respect to dataset of cardiac disease is the data availability is minimal. Machine Learning algorithms helped in various aspects of clinical decision making, furthermore ML algorithms reduce the need of expensive clinical laboratories for investigations and reduces the financial burden of individuals. Many research study has been proposed on available standard datasets.

Grenze ID: 01.GIJET.11.2.713

This study focuses on comparative analysis of traditional ML algorithms, boosting algorithms and hybrid meta learning approach with Neural Network model in heart disease prediction. The study also emphasis on comparison of ML and gradient boosting models with SMOTE and without SMOTE to show case the importance of data balancing. We fused two standard datasets and evaluate how accurately the Machine Learning, boosting methods and Neural Network algorithms performs to predict the disease. Three standard datasets available on UCI (1191 records) and Mendeley dataset (1000 records) have been used for study [3,20]. We employed ML algorithms, boosting algorithms, ensembled based hybrid meta learning Neural Network model for evaluating the model's results and given a comparison study for each algorithm.
The study gives detailed description of dataset in section 2, elaborates the related work in section 3, section 4 emphasizes about data preprocessing techniques applied in this study, section 5 gives insights of ML models and Proposed model results and finally conclusion is given in the last section.

II. DATASET

The major risk factors for heart disease are of two types one is controllable, and the other is non-controllable factors. Controllable factors like obesity, stress, eating habits, sedentary lifestyle and smoking, non-controllable factors like age, gender, genetics and family history [3]. The dataset contains clinical features, patient history and lifestyle features which majorly play important roles in predicting the heart disease [19]. Our study combined two different standard datasets available on UCI repository [2] which contains 1191 records, and Mendeley dataset contains 1000 records of each about 12 features. Combined dataset totally contains 2191 records about 12 features including target variable. After deleting duplicate values, the dataset contains 1918 records. The table 2.1 represents clinical data of a person which is considered as attributes for heart disease prediction.

TABLE I: DATASET DESCRIPTION

Sl.No	Attributes	Term Given	Description
1	Age	Age	In years
2	Sex	Sex	1-Male, 0-Female
3	Chest Pain Type	1, 2,3,4	1-typical angina, 2-atypical angina, 3- non-anginal pain, 4- asymptomatic
4	Resting Blood Pressure	Resting_bp	In mm Hg
5	Serum cholesterol	Cholesterol	In mg/dl
6	Fasting Blood Sugar	Fast_Blood_Sugar	Fas_blood_sugar>120 mg/dl 1=true, 0=false
7	Resting Echocardiogram	Resting_ECG	0=normal, 1=having ST wave abnormality, 2= showing probable or definite left ventricular hypertrophy
8	Maximum Heart Rate	Max_heart_rate	71-202
9	Exercise Induced Angina	Exercise_angina	0= no,1=yes
10	Oldpeak=ST	Oldpeak	Numeric
11	The slope of the peak exercise ST segment	ST slope	1=upsloping, 2= flat, 3=downsloping
12	Class	Target	1=heart disease, 0=normal

III. RELATED WORK

Machine Learning (ML) models' usage has played significant role in healthcare specially in cardiovascular disease prediction [18]. Different approaches of ML is used to predict the cardiovascular disease. Korial et al. ensemble based system for cardiovascular disease detection has given 92.11% of accuracy in which chi-square feature selection method is used and voting mechanism to predict the heart disease [4]. A hybrid feature selection and ensemble model achieved 98.91% of accuracy by stacking the ML models [3]. Alshaikh et al. performed performance analysis on heart disease prediction model with sampling method using SMOTE achieved 95.5% of accuracy [5].
A study was performed a study to enhance the heat attack prediction using Jordan university hospital private dataset, their model performed well of accuracy 94.3% which is enhanced by particle swarm optimization technique and data preprocessing [6]. A method was devised to evaluated ML algorithms like Navie bayes (NB), decision tree (DT), random forest (RF) and K-nearest neighbour (KNN) with standard datasets out of which KNN outperformed with accuracy of 90.78% [7]. A method was proposed in which uses Grid Search CV for

hyper parameter tuning on 70k dataset to optimise the classifier models such as RF, DT, multilayer perceptron (MP), XG-Boost (XGB) in which XGB, MP, RF achieved 0.95 of accuracy [8].

A combination of stacked genetic algorithm and ensemble learning algorithm with feature selection was reported which achieved 97.57% of accuracy, the model uses filter algorithm to reduce the features with low value and uses genetic algorithm to select features with high values [9]. A different approach done using deep learning model and feature augmentation for prediction of heart disease, here the proposed model is trained with augmented data with sparse autoencoder and neural network and classifier network which takes the maximized input for the classification which achieved about 90.09% of accuracy [10]. ML algorithms were examined on two different datasets namely Cleveland and IEEE Dataport where Grid Search CV with five-fold cross validation for optimization is applied to determine the best parameters and a soft voting ensemble classifier resulted in accuracy of 93.4% for Cleveland and 95% for IEEE dataset [11]. The performance of ML models analysed varying in training and testing ratio comparing the results of their novel ensemble based multilayer dynamic system on 70k dataset which achieved about 92.7% of accuracy for the ratio 80:20 [12].

An ensembled stacking deep model was proposed to enhance the heart disease prediction accuracy by integrating two hybrid deep learning model to ensemble model in which the model outperformed on Cleveland dataset about 98.41% of accuracy [13]. A model was developed incorporating convolutional neural network (CNN) for heart disease failure early prediction and model accuracy was about 0.86 [14].

A work reported using different feature selection strategies like chi square, ANOVA to predict heart disease in early stages, three feature subsets were employed with different cross validation for prediction with ten ML algorithms out of all XG-Boost algorithm outperformed about 97.57% of accuracy [15]. Self-attention-based transfer model was proposed to predict cardiovascular disease risk with Cleveland and UCI datasets and achieved about 96.51% accuracy [16]. Many research study employed different ML techniques to produce efficient model for earliest heart disease prediction. Reinforcement and transfer learning were also employed for accurate disease predictions. Even though of several research studies have been done, there is some shortfall which can be implemented in future research studies. Talking about dataset availability of heart disease prediction in most of the research studies the dataset is limited to western or developed nations also socio economic and environmental factors can be taken into consideration. Available dataset is static which is used to train the model, temporal data patterns can be considered for more accurate results. There is an opportunity for multimodal data integration which is crucial to validate the result.

IV. Proposed Method

A. Data Preprocessing

Data preprocessing is done to clean the data so that the model can perform better. The data preprocessing includes removal of duplicates, checking for null values, checking the corelation of the features to identify features which is redundant. We check for null values and corelation and removing the duplicates from the dataset as a step of preprocessing. When combined there is 2191 records was available after deleting duplicates and missing values the dataset reduced to 1918 rows with 12 columns including target variable. Now the dataset is cleaned with no duplicates or missing values which is suitable for evaluation. The dataset has 56.7% positive class and 43.3% of negative class which is slightly imbalance.

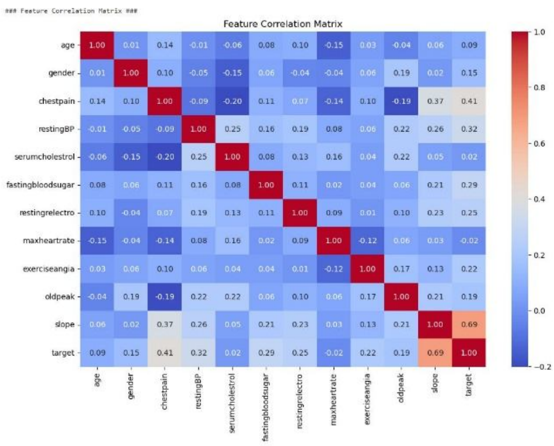


Figure 1: Corelation Matrix

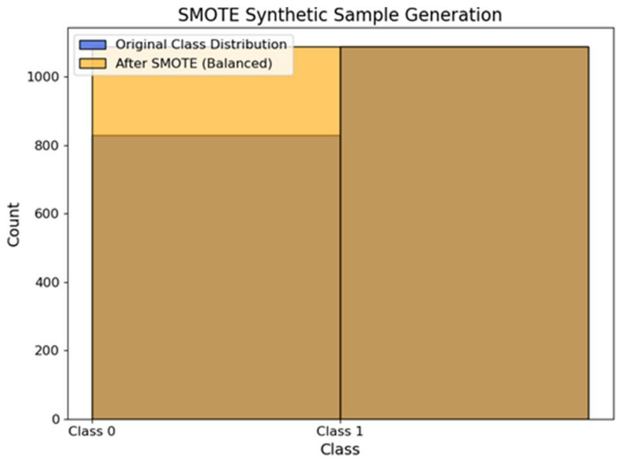
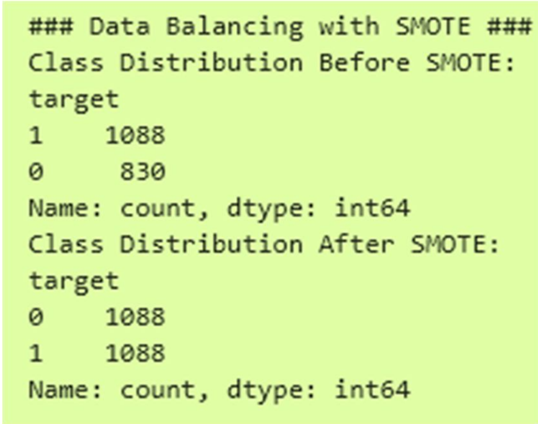


Figure 2a: Data Balance with SMOTE



```
### Data Balancing with SMOTE ###
Class Distribution Before SMOTE:
target
1    1088
0     830
Name: count, dtype: int64
Class Distribution After SMOTE:
target
0    1088
1    1088
Name: count, dtype: int64
```

Figure 2b: Data Balance with SMOTE

The corelation matrix Figure 1 Corelation matrix shows that slope and chestpain is strongly related means they are strong predictors. restingBp and fastingbloodsugar shows moderate positive corelation. The features like chestpain (0.37) and slope shows potential multicollinearity. The maxheartrate feature shows negative corelation. The dataset has been split into training data of 60% and testing data of 40% respectively. SMOTE is applied to balance the training data to ensure both the classes has equal distribution in Figure 2a and Figure 2b. Data normalization is taken care by scaling the features which improves the model performance [17].

SMOTE

Synthetic Minority Oversampling Technique is a method applied for maintain the balance in the dataset. It generates the synthetic samples for the minority class and ensures both classes have equal distribution. It creates the samples using existing class and nearest neighbours; hence imbalance dataset creates a bias and leads to poor performance of the models. Applying SMOTE improves the model generalization and improves model's performance with minority classes [17].

Feature selection

SelectKBest feature selection technique is applied to retain the 15 most significant features. SelectKBest with ANOVA F-statistic is used to measure the linear dependency between the features and target variable which reduces the irrelevant and redundant features which may increase noise. This method enhances the model efficiency by selecting most important features for accurate predictions.

Feature Standardization

Features are standardized applying StandardScaler. The model, which is sensitive to the scale of input data, this Scalar Standardization is applied so that all features equally contribute to increase the model's performance, and all the features are on the similar scale which is crucial for a few algorithms.

V. Experimentation and Discussion

A. Machine Larning

Traditional Machine Learning algorithms has been trained to evaluate the model performance like Logistic Regression, Random Forest, Naive Base Classifier, Decision Tree, Support vector classifier and LightGBM, XGBoost and CatBoost with 20%, 30% and 40% test split ratio to verify the best model. SVM and CatBoost outperformed with 91% accuracy for 20% test split, RF and CatBoost outperformed with 91% accuracy, for 40% test split SVM and CatBoost outperformed with 92% accuracy. The base ML models and Gradient Boosting models are trained for 30% testing ratio with SMOTE and without SMOTE to evaluate the model performance when applied SMOTE.

XGBoost, LightGBM, RF outperformed out of all the base models and boosting models about the accuracy of 93% and AUC-ROC of about 97%. SVM also outperformed but less when compared to other which performed well. The observation with SMOTE and without SMOTE is SMOTE improved Recall, AUC-ROC and Sensitivity which is good for better prediction of positive cases, without SMOTE model favoured majority classes which leaded to high specificity but low sensitivity as seen in Table 1. the impact of SMOTE on model's accuracy is shown in figure 3.

TABLE II: TRADITIONAL ML MODEL COMPARISON WITH AND WITHOUT SMOTE

Dataset	Model	Accuracy	AUC-ROC	Precision	F1-score	Sensitivity	Specificity
With SMOTE	Random Forest	0.923	0.973	0.925	0.925	0.925	0.921
	Decision Tree	0.862	0.862	0.881	0.862	0.844	0.880
	SVM	0.934	0.968	0.934	0.935	0.937	0.930
	Naive Bayes	0.903	0.965	0.9	0.906	0.913	0.893
	XGBoost	0.931	0.971	0.923	0.933	0.943	0.918
	LightGBM	0.923	0.974	0.927	0.925	0.922	0.924
	CatBoost	0.912	0.970	0.926	0.913	0.901	0.924
Without SMOTE	Random Forest	0.916	0.969	0.932	0.930	0.927	0.900
	Decision Tree	0.855	0.853	0.889	0.877	0.866	0.840
	SVM	0.913	0.950	0.934	0.926	0.918	0.905
	Naive Bayes	0.901	0.959	0.920	0.916	0.912	0.883
	XGBoost	0.907	0.952	0.916	0.923	0.930	0.875
	LightGBM	0.911	0.955	0.922	0.926	0.930	0.883
	CatBoost	0.897	0.963	0.922	0.913	0.904	0.887

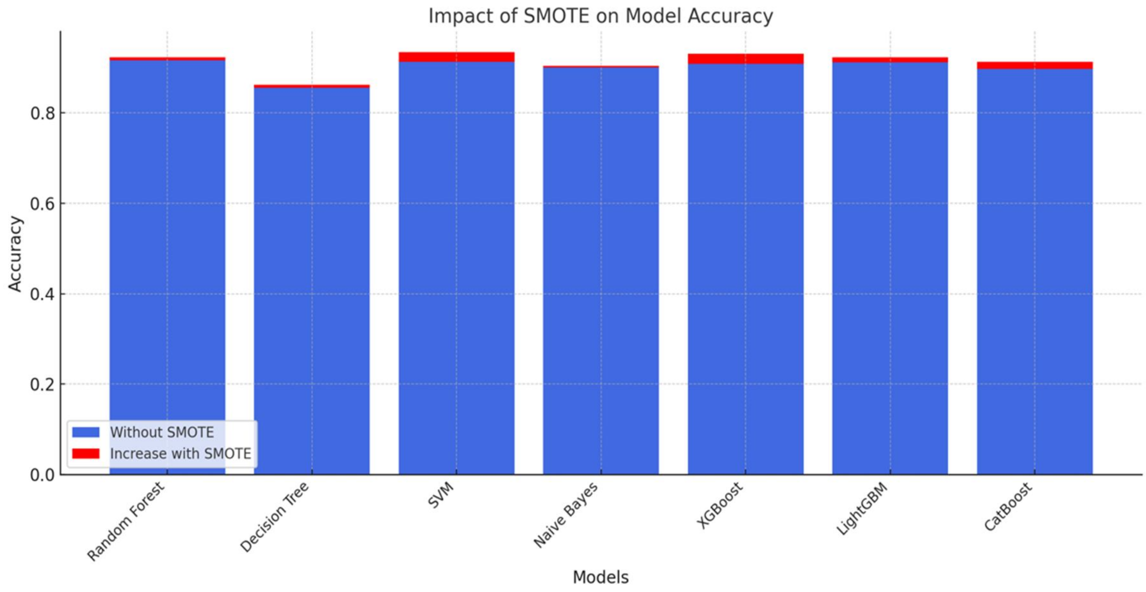


Figure 3: Impact of SMOTE on Model Accuracy

The ensembled model by combining five algorithms RF, SVM, XGBoost, LightGBM and CatBoost which performed well in previous evaluation with SMOTE was trained where is takes outputs for base models to make accurate predictions with hard majority voting in which majority of classifiers is selected for final predictions. All the ensembled model (Table 2) improved their performance when SMOTE is applied.

TABLE III: RESULTS OF ENSEMBLED MODEL WITH AND WITHOUT SMOTE

Dataset	Model	Accuracy	AUC-ROC	Precision	F1-score	Sensitivity	Specificity
With SMOTE	Stacking Ensemble	0.928	0.968	0.933	0.929	0.925	0.930
	Bagging Ensemble	0.912	0.974	0.921	0.914	0.907	0.918
	AdaBoost Classifier	0.863	0.864	0.884	0.864	0.844	0.883
	Gradient Boosting	0.932	0.976	0.936	0.934	0.931	0.933
	Voting Ensemble	0.926	0.974	0.928	0.928	0.928	0.924
Without SMOTE	Stacking Ensemble	0.916	0.958	0.927	0.930	0.933	0.892
	Bagging Ensemble	0.909	0.968	0.929	0.923	0.918	0.896
	AdaBoost Classifier	0.852	0.851	0.891	0.874	0.857	0.844
	Gradient Boosting	0.918	0.964	0.932	0.931	0.930	0.900
	Voting Ensemble	0.914	0.964	0.927	0.928	0.930	0.892

B. Proposed Model

For better study in heart disease prediction, we employed a hybrid and ensembled meta learning approach (Figure 4) which uses both ML and Deep Learning meta learning approach. The dataset is processed in the beginning and SMOTE is applied for class balance distribution, SlectKBest ANOVA F-test is used to select the most relevant features which improved model efficiency. The dataset is splitted into 60% training and 40% testing ratios to evaluate the model. We used four high performing ML amodels RF, XGBoost, LightGBM and CatBoost which is trained on dataset and each model predictions are collected as meta-features. These meta-features aggregated using soft voting classifier. The predictions of voting classifier are used as additional features in meta learning phase.

A neural network meta model is trained with meta features which is the aggregated output of voting ensemble of five base models. Deep learning model has multiple dense layers with batch normalization, dropout regularization and L2 weight constrains to prevent the model from overfitting and improves generalization. The model uses Adam optimizer with low learning rate of 0.00005 for 50 epochs with 20% validation split to monitor the performance. 0.5 threshold is applied for meta model predictions for classification. The proposed model is evaluated by accuracy, sensitivity, specificity, AUC-ROC, Recall, Precision metrics. The proposed model outperformed well with hight accuracy of 94.49% and AUC-ROC of 95.44% (Table 3) which has high predictive strength. The proposed model outperforms all the traditional model which makes highly suitable for heart disease predictions and medical applications.

To ensure the stability of model's performance stability test is performed using 5-fold cross validation. Low deviation in F1-score of 1.07% and accuracy of 1.53% confirms that the model is relatively stable and performed well across different training subsets. Overall, the model is stable and consistent.

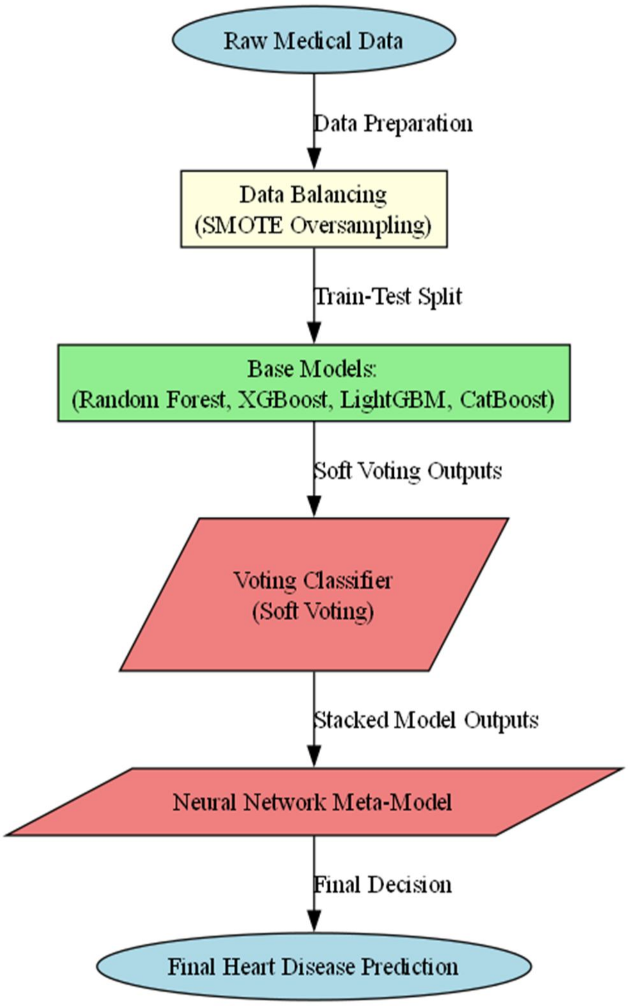


Figure 4: Neural Network Meta-Model Pipeline

TABLE IV: META-MODEL RESULTS WITH AND WITHOUT SMOTE

Meta-Model	Accuracy	AUC-ROC	Precision	Recall	F1-Score	Specificity
With SMOTE	0.9449	0.9544	0.9333	0.9612	0.9471	0.9277
Without SMOTE	0.9253	0.9413	0.9251	0.9450	0.9349	0.8996

C Performance metrices

The following metrics were considered to determine the model's efficiency in prediction.

Accuracy

Total number of correct predictions made by the model both true positive and true negatives.
Accuracy =TP+TN/TP+TN+FP+FN ---------------------------------1

Recall/Sensitivity

The actual positive cases which is predicted by model also called as true positive rate.
Recall = TP/TP+FN--2

Specificity

The actual negative instances which is correctly predicted by model also called as True Negative Rate.
Specificity = TN/TN+FP---3

Precision

The number of predicted instances which are actually correct.
Precision=TP/TP+FP--4

F1 Score

It is the mean of precision and recall.
F1 Score=2*[Precision*Recall]/ [Precision +Recall] ------5

AUC-ROC

it is used to well distinguish between positive class and negative class.

TABLE V: RESULTS OF ML MODELS AND NEURAL NETWORK MODEL.

Model	Accuracy	AUC-ROC	Precision	F1-Score	Sensitivity	Specificity
Optimized Meta-Model	**0.944**	**0.954**	**0.933**	**0.947**	**0.961**	**0.927**
Random Forest	0.923	0.973	0.925	0.925	0.925	0.921
Decision Tree	0.862	0.862	0.881	0.862	0.844	0.880
SVM	0.934	0.968	0.934	0.935	0.937	0.930
Naive Bayes	0.903	0.965	0.90	0.906	0.913	0.893
XGBoost	0.931	0.971	0.923	0.933	0.943	0.918
LightGBM	0.923	0.974	0.927	0.925	0.922	0.924
CatBoost	0.912	0.970	0.926	0.913	0.901	0.924

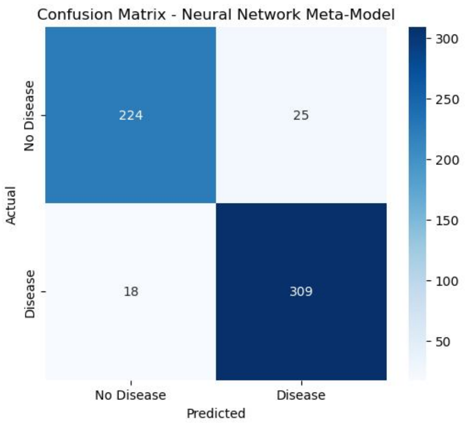


Figure 5.a: Neural Network Model without SMOTE.

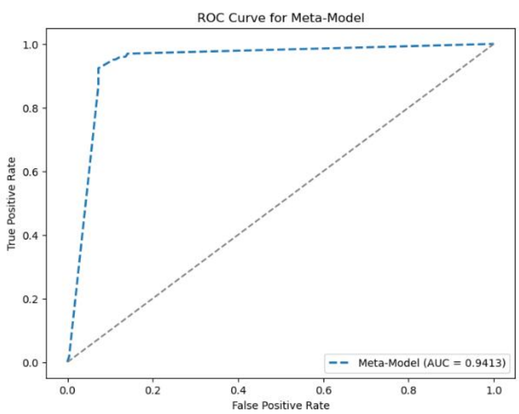


Figure 5.b: Training and Validation Accuracy Plot

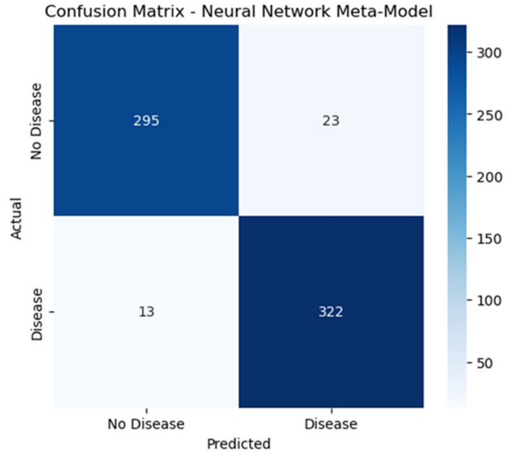


Figure 6.a: Neural Network Model with SMOTE.

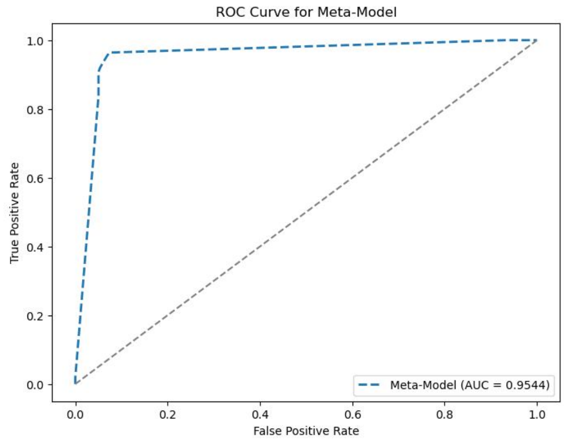


Figure 6.b: Training and Validation Accuracy Plot

VI. CONCLUSION

The proposed model reported good performance by successfully integrating ML models with Neural Network meta classifier for better classification and generalization. The model which implied feature section method using SelectKBest and SMOTE for data balancing and meta learning using deep neural network which resulted in high accuracy of 94.48% and AUC-OC of 95.44% which shows stronger predicting capabilities. Precision of 93.33% and Recall of 96.12% shows reliability of model minimizing false positive cases. Sensitivity of 96.12% and specificity of 92.77% which showcase balance in classification. The meta model uses voting classifier aggregate outputs to optimize decision making. The study has the feature scope of where hyper tuning the parameters to improve stability, threshold optimization and feature engineering can be employed in future studies.

REFERENCES

[1] Heart and stroke statistics. www.heart.org. (n.d.). https://www.heart.org/en/about-us/heart-and-stroke-association-statistics
[2] Doppala, B. P., & Bhattacharyya, D. (2021). Cardiovascular_Disease_Dataset. Mendeley Data. https://doi.org/10.17632/dzz48mvjht.1
[3] Mahajan, Abhigya & Kaushik, Baij & Rahmani, Mohammad Khalid Imam & Banga, Abdulbasid. (2024). A Hybrid Feature Selection and Ensemble Stacked Learning Models on Multi-Variant CVD Datasets for Effective Classification. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3412077.
[4] Korial, A. E., Gorial, I. I., & Humaidi, A. J. (2024). An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection. Computers, 13(6), 126. https://doi.org/10.3390/computers13060126
[5] Al-Alshaikh, Halah & P, Prabu & Poonia, Ramesh & Saudagar, Abdul & Yadav, Manoj & Alsagri, Hatoon & AlSanad, Abeer. (2024). Comprehensive evaluation and performance analysis of Machine Learning in heart disease prediction. Scientific Reports. 14. 10.1038/s41598-024-58489-7.
[6] Mohammad Alshraideh, Najwan Alshraideh, Abedalrahman Alshraideh, Yara Alkayed, Yasmin Al Trabsheh, Bahaaldeen Alshraideh, and Honggang Chen. 2024. Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital. Appl. Comp. Intell. Soft Comput. 2024 (2024). https://doi.org/10.1155/2024/5080332
[7] Devansh Shah, Samir Patel, and Santosh Kumar Bharti. 2020. Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci. 1, 6 (Nov 2020). https://doi.org/10.1007/s42979-020-00365-y
[8] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088
[9] Abdollahi, Jafar & Nouri-Moghaddam, Babak. (2022). A hybrid method for heart disease diagnosis utilizing feature selection-based ensemble classifier model generation. Iran Journal of Computer Science. 5. 10.1007/s42044-022-00104-x.
[10] García-Ordás, M.T., Bayón-Gutiérrez, M., Benavides, C. et al. Heart disease risk prediction using deep learning techniques with feature augmentation. Multimed Tools Appl 82, 31759–31773 (2023). https://doi.org/10.1007/s11042-023-14817-z
[11] Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. Processes, 11(4), 1210. https://doi.org/10.3390/pr11041210

[12] Mohammed Nasir Uddin, Rajib Kumar Halder, An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach, Informatics in Medicine Unlocked, Volume 24, 2021, 100584, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100584.

[13] Almulihi, A., Saleh, H., Hussien, A. M., Mostafa, S., El-Sappagh, S., Alnowaiser, K., Ali, A. A., & Refaat Hassan, M. (2022). Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction. Diagnostics, 12(12), 3215. https://doi.org/10.3390/diagnostics12123215

[14] Mehmood, A., Iqbal, M., Mehmood, Z. et al. Prediction of Heart Disease Using Deep Convolutional Neural Networks. Arab J Sci Eng 46, 3409–3422 (2021). https://doi.org/10.1007/s13369-020-05105-1

[15] El-Sofany, H., Bouallegue, B. & El-Latif, Y.M.A. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Sci Rep 14, 23277 (2024). https://doi.org/10.1038/s41598-024-74656-2

[16] Rahman, A.U., Alsenani, Y., Zafar, A. et al. Enhancing heart disease prediction using a self-attention-based transformer model. Sci Rep 14, 514 (2024). https://doi.org/10.1038/s41598-024-51184-7

[17] Gayathri R, Sangeetha S. K. B, Sandeep Kumar Mathivanan, Hariharan Rajadurai, Benjula Anbu Malar MB, Saurav Mallik, Hong Qin, Enhancing heart disease prediction with reinforcement learning and data augmentation, Systems and Soft Computing, Volume 6, 2024, 200129, ISSN 2772-9419, https://doi.org/10.1016/j.sasc.2024.200129.

[18] Chaithra, C.S., Siddesha, S., Aradhya, V.N.M., Niranjan, S.K. (2024). A review of machine learning techniques used in the prediction of heart disease. Revue d'Intelligence Artificielle, Vol. 38, No. 1, pp. 201-212. https://doi.org/10.18280/ria.380120

[19] CS, C., & Niranjan, S. K. (2024). Ten Years of Research on Cardiac Imaging in Diagnosing Heart Diseases. Grenze International Journal of Engineering & Technology (GIJET), 10.

[20] UCI Machine Learning Repository. (n.d.-b). https://archive.ics.uci.edu/dataset/45/heart+disease