JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Cardiovascular Disease Using Machine Learning Techniques

1Harichselvam C, 2Dr. Jerald F, 3Dr. T.V. Ananthan, 4Harishwar J, 5Maheswari A
1Student, Dept. of CSE, Dr. M.G.R Educational and Research Institute, Chennai, India
2Associate Professor, Dept. of ECE, Dr. M.G.R Educational and Research Institute, Chennai, India
3Professor, Dept. of CSE, Dr. M.G.R Educational and Research Institute, Chennai, India
4Student, Dept. of CSE, Dr. M.G.R Educational and Research Institute, Chennai, India
5Associate Professor, Dept. of CSE, Dr. M.G.R Educational and Research Institute, Chennai, India

Abstract :

Cardiovascular Diseases (CVDs) are the leading global cause of mortality, driving the need for innovative early prediction methods. This study utilizes Machine Learning (ML) models, including Random Forest, XGBoost, and Logistic Regression, to predict CVD using a dataset of 70,000 patient records and 11 features. Ensemble learning via a Voting Classifier improved accuracy. Techniques like SMOTE addressed class imbalance, and feature scaling ensured performance. The Voting Classifier achieved the highest accuracy (74%), with XGBoost performing similarly. These results demonstrate ML's potential in offering accurate, non-invasive tools for assessing cardiovascular risk.

Index Terms – Cardiovascular Disease Prediction, Machine Learning, Ensemble Learning, XGBoost, Random Forest, SMOTE

INTRODUCTION

Cardiovascular diseases (CVDs) pose a significant global health challenge, causing millions of deaths annually and placing a burden on healthcare systems. Early detection and prevention of CVD can improve patient outcomes and reduce healthcare costs. Traditional diagnostic approaches often involve invasive procedures, are resource-intensive, and may be limited in predictive accuracy.

Advancements in machine learning (ML) offer transformative potential for healthcare, enabling the analysis of complex datasets to identify patterns and make predictions. ML techniques can provide faster and more accurate diagnoses, aiding clinicians in risk assessment and treatment planning. However, challenges such as class imbalance in medical datasets, model interpretability, and integration into clinical workflows persist.

This study aims to develop an ML-based predictive framework for cardiovascular disease using a publicly available dataset. By leveraging algorithms like Random Forest, XGBoost, and Logistic Regression, combined with ensemble learning, this research seeks to enhance prediction accuracy and address dataset challenges such as class imbalance.

Methodology

Dataset Description

The dataset, sourced from Kaggle, contains 70,000 patient records with 11 features and a binary target variable (cardio), indicating the presence or absence of CVD. The features include:

- Objective: Age, Height, Weight, Gender
- Examination: Systolic and diastolic blood pressure, Cholesterol, Glucose
- Subjective: Smoking, Alcohol intake, Physical activity

Data Preprocessing

- Cleaning: The id column was removed as it was non-informative.
- Transformation: Age was converted from days to years for better interpretability.
- Encoding: Categorical variables were one-hot encoded.
- Scaling: Features were standardized using the StandardScaler.
- Handling Class Imbalance: SMOTE was applied to generate synthetic samples for the minority class.

Model Development

Random Forest: A tree-based ensemble method with 100 estimators was used to assess feature importance and prediction accuracy.

As shown in Fig. 1 and Fig. 2, the confusion matrix and ROC curve respectively illustrate the model's ability to identify both classes effectively.

XGBoost: A gradient boosting algorithm with GPU acceleration was employed for efficiency and improved performance.

As shown in Fig. 3 and Fig. 4, XGBoost's confusion matrix and ROC curve demonstrate its strong performance.

Logistic Regression: A linear model served as a baseline for comparison.

As seen in Fig. 5 and Fig. 6, Logistic Regression also provides reliable predictions as depicted in its confusion matrix and ROC curve.

Voting Classifier: An ensemble approach combining the strengths of the above models through soft voting.

As shown in Fig. 7 and Fig. 8, the ensemble model achieves the best balance between precision and recall, visualized through its confusion matrix and ROC curve.

Evaluation Metrics

Accuracy :

Definition:

The ratio of correctly predicted instances to the total instances.

Calculations:

Accuracy=(TruePositives+True Negatives)/Total Instances

It provides an overall measure of the model's performance but may not be reliable if the dataset is imbalanced.

Precision, Recall, and F1-score :

Precision:

The proportion of true positive predictions among all positive predictions made by the model.

Precision=True Positives / (True Positives + False Positives)

Recall (Sensitivity):

The proportion of actual positive cases correctly identified by the model.

Recall= True Positives /(True Positives+False Negatives)

F1-score:

The harmonic mean of Precision and Recall, balancing the trade-off between the two.

F1-score=2×(Precision+Recall)/(Precision×Recall)

Confusion Matrix :

A tabular summary of model predictions vs. actual labels, showing:

1. True Positives (TP): Correct positive predictions.
2. True Negatives (TN): Correct negative predictions.
3. False Positives (FP): Incorrect positive predictions.
4. False Negatives (FN): Incorrect negative predictions.

It is useful for visualizing model performance and identifying misclassification patterns.

ROC-AUC Curve :

ROC (Receiver Operating Characteristic):

Plots the true positive rate (Recall) against the false positive rate (FPR) at various threshold settings.

FPR=False Positives/(False Positives + True Negatives)

AUC (Area Under Curve):

A single metric summarizing the ROC curve. Higher AUC values indicate better model performance in distinguishing between classes.

Results

The Voting Classifier achieved the highest accuracy (74%), demonstrating the benefits of ensemble learning. XGBoost performed similarly, showcasing its effectiveness in handling complex data. Logistic Regression provided competitive baseline performance.

Random Forest: Accuracy of 72%, demonstrating strong performance in identifying both classes.

XGBoost: Achieved the highest standalone accuracy of 74%, benefiting from advanced boosting techniques.

Logistic Regression: Accuracy of 72%, offering a competitive baseline.

Voting Classifier: Outperformed individual models with an accuracy of 74%, highlighting the advantage of ensemble learning.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	72%	0.72	0.72	0.72
XGBoost	74%	0.74	0.74	0.74
Logistic Regression	72%	0.73	0.72	0.72
Voting Classifier	74%	0.74	0.74	0.74

Discussion

The study highlights the potential of ML techniques in predicting CVD, with ensemble learning yielding the best performance. XGBoost's comparable accuracy demonstrates its robustness for medical datasets. While the results align with existing studies, limitations include reliance on a single dataset and lack of external validation. Future directions include:

- Expanding datasets
- Enhancing model interpretability
- Validating with real-world clinical data

Conclusion

This research highlights the potential of machine learning in early CVD detection, with ensemble models demonstrating superior performance. By integrating diverse algorithms and addressing data imbalances, the study establishes a strong foundation for further advancements in predictive healthcare. Future efforts should focus on enhancing model interpretability, expanding feature sets, and deploying user-friendly tools for clinical use.

Acknowledgments

The author expresses gratitude to the Dr. MGR Educational and Research Institute, Chennai, for their support and resources in conducting this research. Additionally, the availability of open-access datasets on platforms like Kaggle was invaluable in enabling this work.

1. ***Figures and Tables***

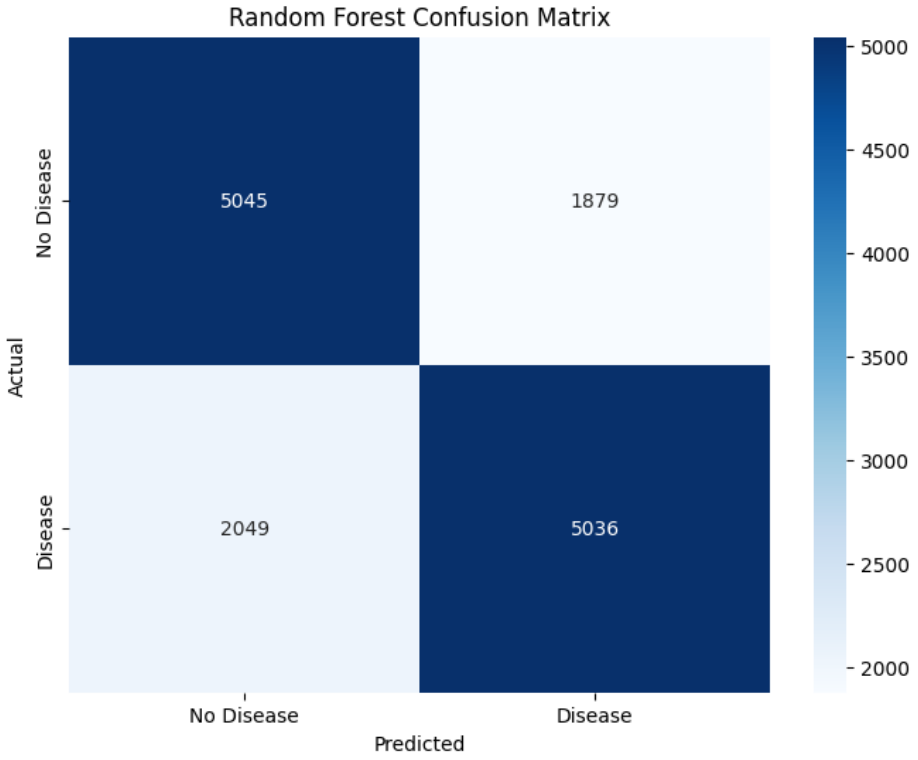


Fig. 1

R

andom Forest Confusion Matrix

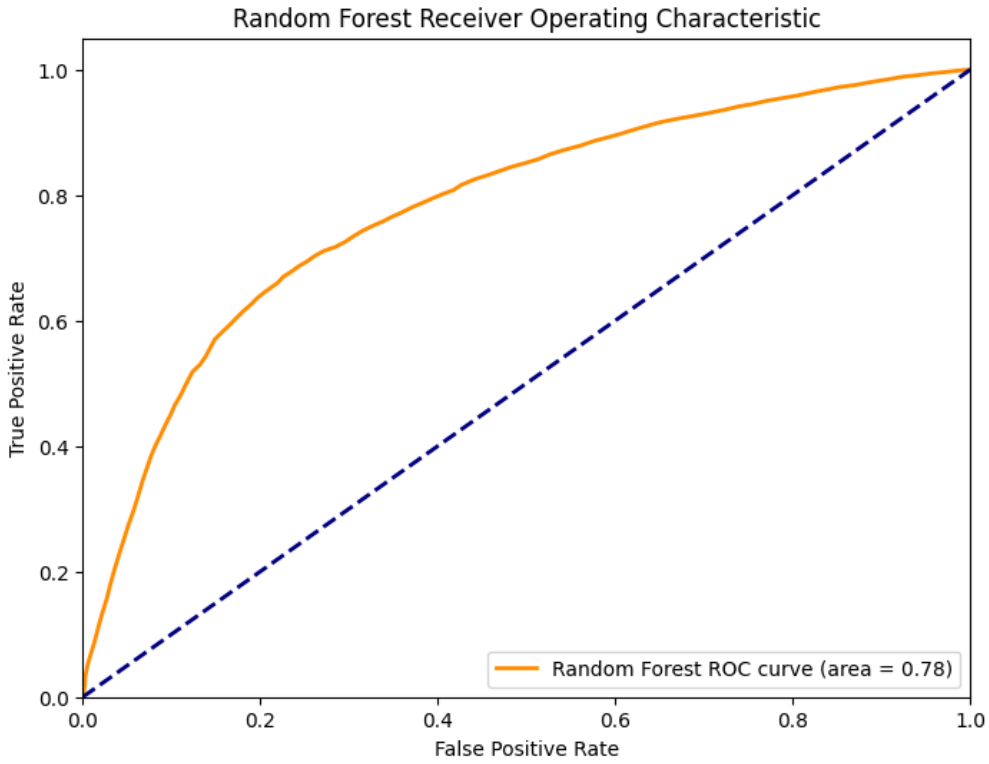


Fig. 2

R

andom Forest ROC Curve

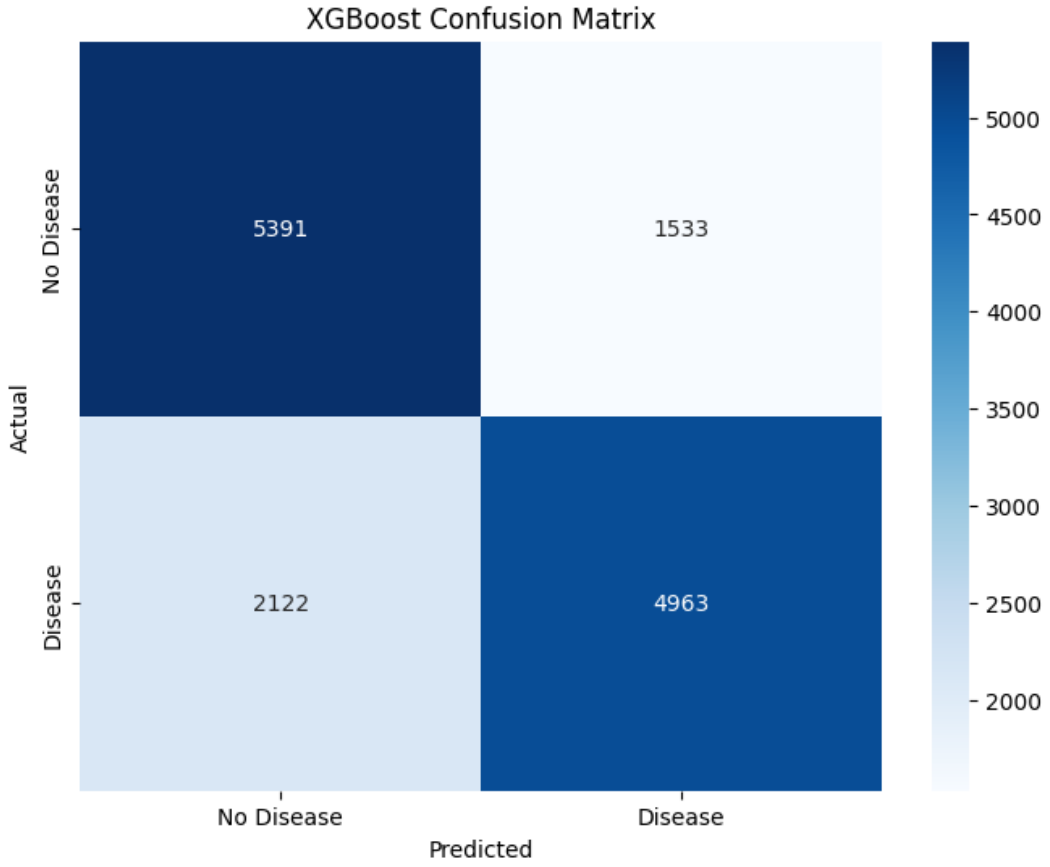
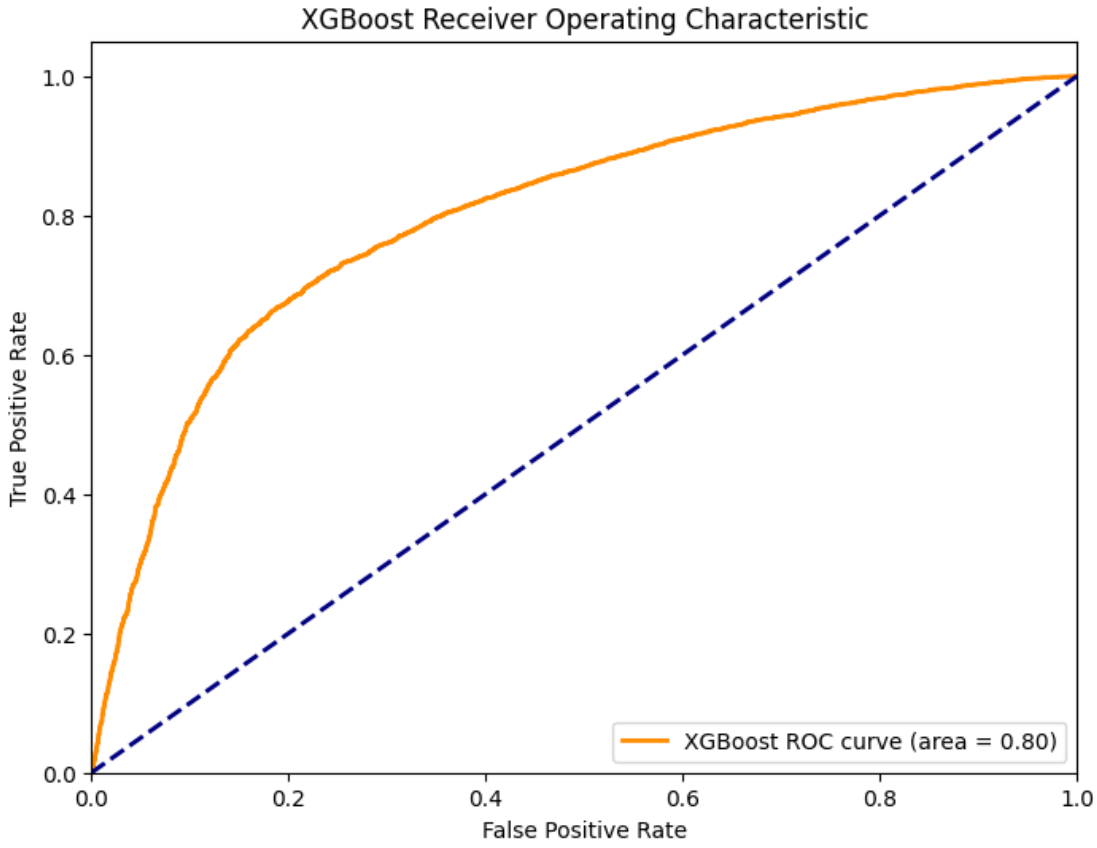


Fig. 3

XGBoost Confusion Matrix



F

ig. 4

XGBoost ROC Curve

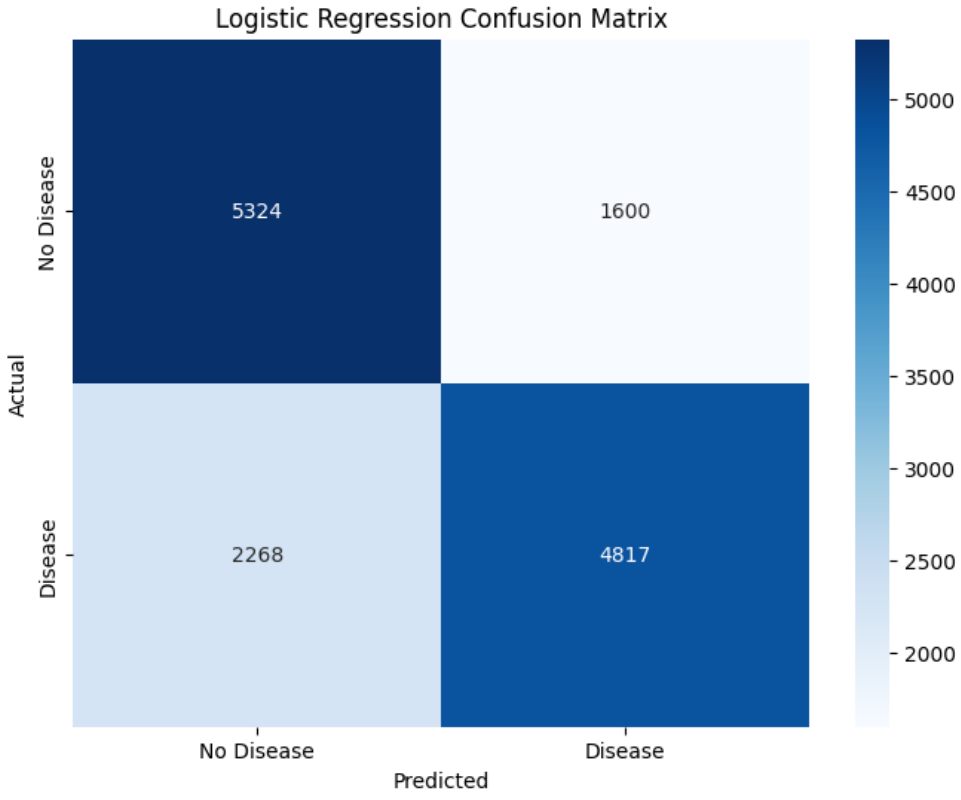


Fig. 5

Logistic Regression Confusion Matrix

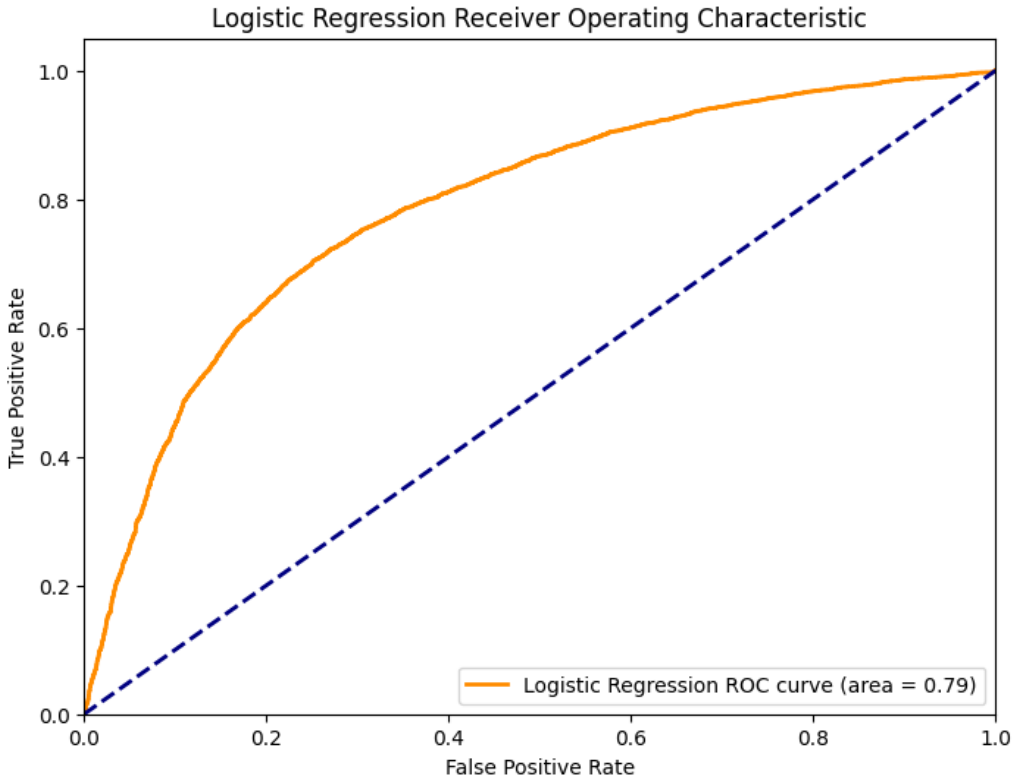


Fig. 6

L

ogistic Regression ROC Curve

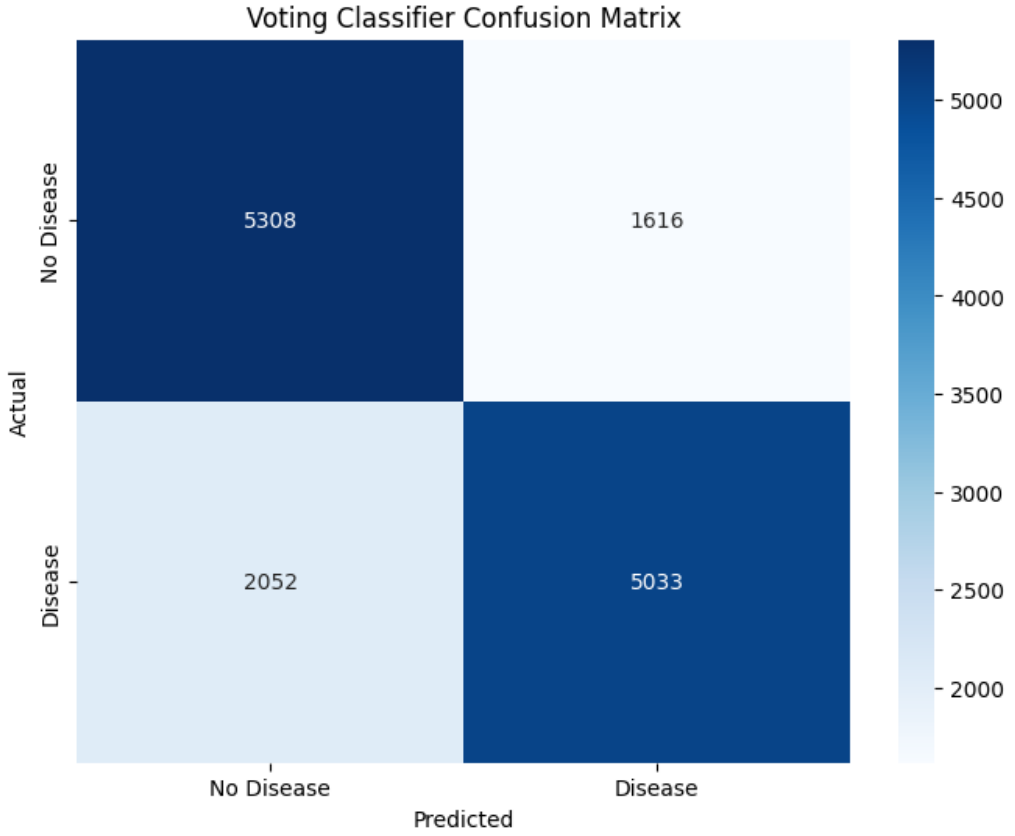


Fig. 7

Voting Classifier Confusion Matrix

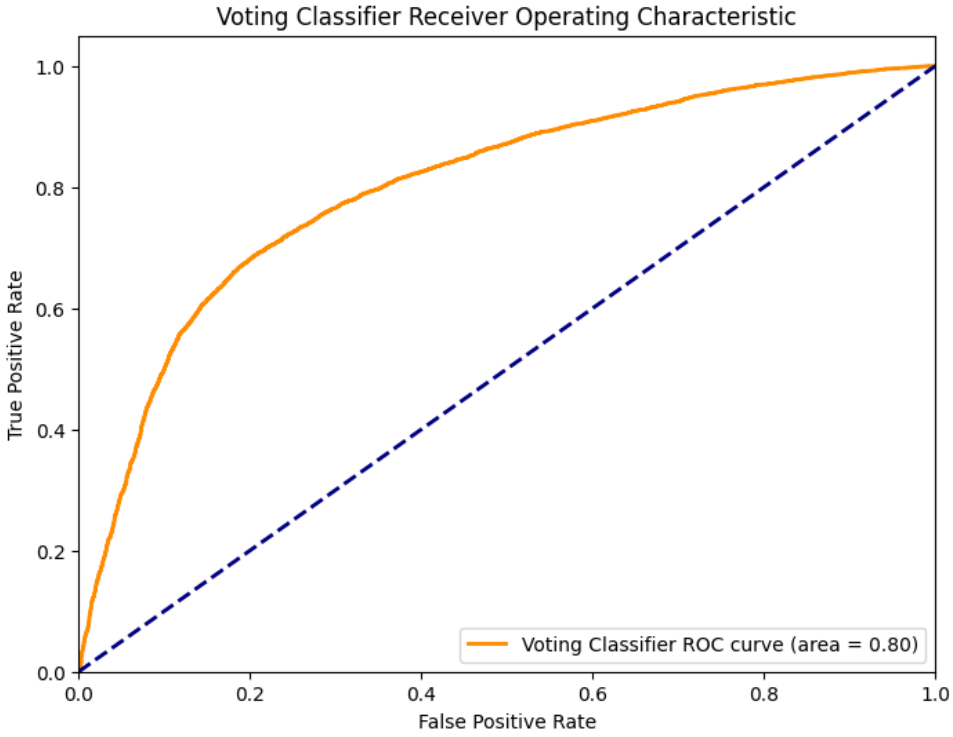


Fig. 8

Voting Classifier ROC Curve

References

Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001;29(5):1189-1232.

Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321-357.

Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.

Pedregosa F, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-2830.

He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009;21(9):1263-1284.

Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. Morgan Kaufmann; 2016.