Harnessing Ensemble Models for Cardiovascular Stroke Forecasting

Mr. P P N G Phani Kumar[1], P. Meghana[2], G. Tejasri[3], B. Dilliswari[4], Mohammad Abdul Basit[5]
[1] *Associate Professor, Dept. Of Computer Science and Engineering*
[2-5] *B. Tech Student, Dept. Of CSE (Data Science)*
[1-5] *Raghu Engineering College, Visakhapatnam*

***Abstract*—cardiovascular diseases execute the highest number of deaths internationally and disregard economic levels or geographical boundaries. The identification of cardiovascular conditions at an early stage alongside quick medical care substantially decreases disease prevalence while improving medical results. A complete Cardiovascular Disease Prediction System will be developed through machine learning algorithms and ensemble techniques according to this project's objective. The prediction system uses Logistic Regression along with K- Nearest Neighbors, Random Forest, Decision Tree and XG-Boost and Ensemble Learning to assess multiple cardiac health signals which help determine cardiovascular event risks. This predictive system incorporates multiple models to reach high reliability and accuracy thus enabling healthcare professionals to take proactive healthcare measures. The innovative solution provides medical staff with productive information to boost diagnostic accuracy alongside preventive healthcare initiatives.**

***Index Terms*—Ensemble Learning, Machine Learning, Voting Classifier, Logistic Regression, K-Nearest Neighbors, Random Forest, Decision Tree, XG-Boost.**

I. INTRODUCTION

Stroke presents itself as an extensive worldwide health problem among cardiovascular diseases because it poses substantial danger to patients who could potentially die from it. Medical staff can offer prompt lifesaving medical treatment to those who are at risk when they receive early identification. Machine learning particularly serves healthcare purposes by helping medical staff detect diseases without difficulty. The analysis creates a stroke prediction model through ensemble learning methods to boost the accuracy in forecasting outcomes. The employed analysis depends on three main models comprising Logistic Regression with Decision Tree and Random Forest algorithms and K- Nearest Neighbors (KNN). Precision and balanced results emerge from a Voting Classifier which uses multiple models to combine their output predictions for reliable predictions. Ensemble methods help the predictive model cut down errors while providing better generalization results to multiple patient data points. Research data from Kaggle presents health metrics that include patient age, cholesterol, glucose measurements and hypertension status and smoking habits statistics. The framework runs with high operational speed while maintaining expansion capabilities together with straightforward usability for assisting healthcare professionals to evaluate stroke risk at initial stages. Research effectiveness depends on first measuring the single model effectiveness before choosing the best method from an ensemble approach. The investigation focuses on stroke prediction system development through machine learning models with an added objective to optimize ensemble learning accuracy by evaluating models through F1-score metrics which measure precision and recall.

The project uses multiple ML techniques to help earliest possible stroke identification which drives better preventative healthcare solutions.

1. CARDIOVASCULAR DISEASE

A group of medical problems called cardiovascular diseases (CVDs) affects the heart alongside blood vessels through six conditions including coronary artery disease

1

(CAD), stroke, heart failure, arrhythmias and hypertension. According to the World Health Organization (WHO) cardiovascular diseases rank as the second biggest fatal group that kills 17.9 million people throughout the globe annually. The medical community can prevent multiple severe diseases when patients receive early detection and introduce lifestyle changes combined with appropriate medical attention.

1.1 Types of Cardiovascular Diseases

Several cardiovascular diseases exist that target

distinct areas of the circulatory system including heart diseases characterized by narrowing of arteries.

a. Coronary Artery Disease (CAD)

Atherosclerosis leads to coronary artery disease because the plaque buildup blocks the blood supply to oxygen-rich heart tissue through the coronary arteries which leads heart attacks and heart failure.

b. Stroke

Brain damage together with death becomes possible when a stroke cuts off blood flow to the brain. The human body experiences two distinct kinds of stroke. A blood clot blocking brain supply arteries creates Ischemic Stroke. Bleeding within brain vessels during a hemorrhagic stroke result in neurological destruction because blood vessels break open.

c. Heart Failure

Acute heart failure develops when the heart muscles become weak and fail to circulate blood correctly thus causing fluid to expand throughout the lungs and other body tissues.

d. Hypertension

Hypertension also known as High blood pressure persists over an extended period in the body when someone has hypertension. The absence of symptoms from hypertension increases the danger of heart attacks and strokes and kidney disease development.

e. Peripheral Artery Disease (PAD)

PAD develops when arterial blockages decrease blood movement particularly toward the feet and lower legs. The condition produces pain alongside numbness which makes walking into a troublesome act. The presence of PAD signals the spread of atherosclerosis which raises individual risk for heart attack and stroke.

The application of machine learning technologies now enables expanded stroke predictive assessments through large-scale healthcare data evaluations for identifying important risk indicators which help diagnose cardiovascular diseases. Traditional statistical methods fail to detect hidden patterns when it comes to multiple non-linear risk elements but ML algorithms succeed both at data pattern recognition and accurate prediction generation.

Prognosis for strokes is surging and improving because doctors have better tools involving cooler stats like machine learning thanks to more use of electronic health notes and devices that people wear. The implementation of these models allows healthcare professionals to implement early interventions while providing unique treatment strategies and making better medical choices to enhance stroke patient results and lower cardiovascular disease impact.

II. LITERATURE REVIEW

Researchers have extensively studied how machine learning (ML) algorithms predict cardiovascular diseases in various medical studies. Several ML prediction methods get widely adopted for medical applications. Multiple tests have evaluated these algorithms to assess their ability for both detecting and forecasting cardiovascular diseases.

Research by Kamutam Vinay et al. [1] demonstrated that Random Forest delivered the highest accuracy result of 99.17% compared to KNN and Decision Tree models thus becoming the top performing classifier based on their assessment.

Chua et al.'s[2] research on heart disease prediction involved a comparison between Naïve Bayes and SVM and showed KNN as the best performer with outstanding precision and recall values essential to medical diagnosis tasks.

According to Ambrish G et al.'s[3] research methods investigation of Logistic Regression reached 87.10% in accuracy evaluation. This research established that Logistic Regression shows excellent effectiveness for binary classifications yet falls behind other state-of-the-art ensemble learning approaches.

Kellen Sumwiza et al. [4] conducted research where Random Forest exhibited superior performance over KNN and SVM based on both accuracy levels and precision and recall rates.

The research by K. Hashi et al. [5] showed ensemble methods' significance through model heart disease prediction optimization using hyperparameter tuning. The accuracy results from 85.25% up to 91.80% strengthen the case for performing parameter

adjustments on ML algorithms to enhance predictive capabilities.

This research develops a Voting Classifier based on ensemble learning principles that combines advantages from Logistic Regression and Decision Tree with Random Forest and XGBoost and KNN. These results demonstrate how combining several algorithms through ensemble learning achieves a cardiovascular stroke prediction accuracy between 92.32% and 93.52%.

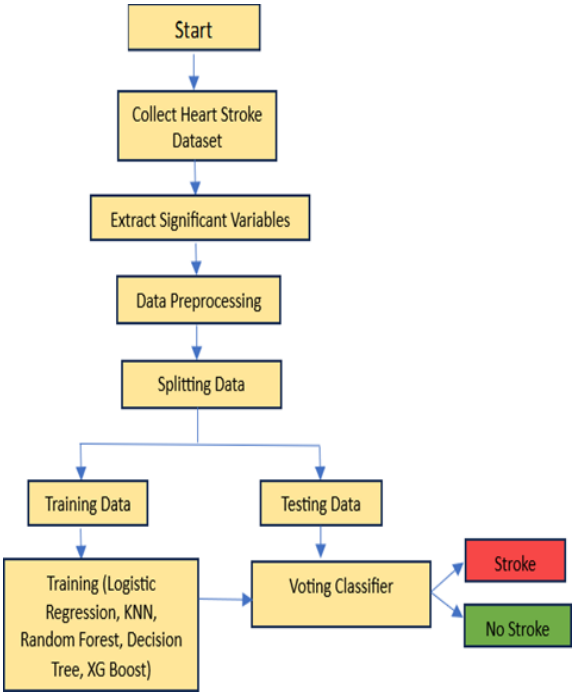
2. PROPOSED SYSTEM

The System introduces ensemble-based machine learning to handle existing solution failures. The system implements predictive accuracy improvement alongside robustness using five models including Logistic Regression and K-Nearest Neighbors and Random Forest and XGBoost and Decision Tree. The predictive system collects two individual model output results which combine their complementary capabilities to reduce errors and achieve enhanced generalization capabilities. The system utilizes real-time analysis of wide cardiac indicators to notify healthcare staff about high-risk patients which enables immediate intervention. The system provides preliminary capabilities to recognize heart disease symptoms together with evaluating patient risk levels.

Key features of proposed system:

- The system achieves higher accuracy levels in addition to increased robustness through model integration which decreases predictive errors.
- The Voting Classifier in Ensemble learning helps reduce both model-based biases and estimation variances which results in improved decision outcomes.
- The analysis tool detects the most vital health elements which affect stroke risk levels.
- The system shows ability to manage extensive data as well as process new data points for human health development.
- The interface of the system was specifically designed to enhance user convenience when healthcare professionals interact with it.
- Precise predictions help medical professionals carry out early diagnoses which decreases the probability of developing critical cardiovascular diseases.

III. FLOWCHART



The workflow diagram presents the stroke prediction sequence that begins by collecting data which is followed by feature extraction and preprocessing steps. Several machine learning models including Logistic Regression, KNN, Random Forest, Decision Tree, and XGBoost receive trained data from the division of the dataset into training and testing parts. A Voting Classifier operates on testing data to establish final predictions therefore labeling individuals either as having a Stroke condition (Red) or without a Stroke condition (Green) as indicated by model results.

3. METHODOLOGY

A cardiovascular stroke prediction system functions through the combination of various machine learning models as well as an ensemble Voting Classifier. This approach follows multiple stages to achieve its methodology.

3.1. Dataset and Preprocessing

Kaggle's Cardiovascular Disease Dataset provides the source for the data which contains features including age and cholesterol levels and smoking status and glucose levels and hypertension along with 12 attributes among 5110 records.

An explanation of features with brief descriptions:

1. ID: All patients in the dataset receive their own unique ID to serve as identifier.

2. Gender: Specifies the gender of the individual as Male, Female.
3. Age**:** Represents age of the patient.
4. Hypertension**:** A binary variable where 0 indicates the absence of hypertension while 1 signifies the presence of hypertension.
5. Heart Disease**:** Indicates whether the patient has been diagnosed with heart disease. A value of 0 means the patient does not have heart disease, while 1 confirms the presence of heart disease.
6. Married Status: Participants falling into unmarried status receive a "No" label while married participants get marked as "Yes."
7. Employment Type: Specifies the nature of the patient's job. Values include: Children, Government job, never worked, Private, Self-employed.
8. Residence Type: The variable Residence Type determines 0 for rural residence and sets 1 for patients living in urban areas.
9. Average Glucose Level: Represents the mean glucose level in the patient's blood.
10. BMI: A numerical value of Body Mass Index (BMI) serves as a height-to-weight measurement for adult body weight assessment.
11. Smoking Status: The smoking status field confirms if patients were smokers. The smoking status questionnaire contains three options which are Formerly smoked, never smoked and Smokes.
12. Stroke: 1 indicates detection of stroke and 0 indicates detection of no stroke.

The process of data preprocessing consists of treating missing data points and applying encoding methods to categories and normalization techniques to numerical values for better model results.

The initial Id feature gets eliminated because it provides no essential value to the analysis. The procedure for BMI contains null values while the median values fill in the blanks. The gender variable contains multiple values but the data only keeps male and female categories. The training process of models requires numerical variables which leads to the implementation of Feature Encoding techniques on gender, martial status, employment type, residence type, smoking status variables. The data allocation step follows the completion of feature encoding operations.

It becomes essential to deal with the imbalanced dataset because the unbalanced samples might result in prediction biases from the models. SMOTE effectively produces new synthetic samples from minority class data to make the dataset more evenly balanced. The dataset received balance treatment through SMOTE application.

3.2. Model Selection and Training

Five Machine Learning algorithms serve to predict cardiovascular stroke prediction: Logistic Regression, Decision Tree, Random Forest, XGBoost and K-Nearest Neighbor (KNN).

The applied models including Logistic Regression and Decision Tree and Random Forest and KNN and XGBoost operate under scikit-learn and XGBoost software libraries for training. The models implement fit() method on the training data X_train, y_train to develop patterns from available information while receiving appropriate parameter settings. The models generate their predictions through the predict() method for both training and testing data once training has finished. The model assessment includes accuracy score evaluation along with cross-validation tests and classification metric measurement through precision, recall, and F1-score. A confusion matrix represents the prediction performance output which enables users to evaluate how well models classify the dataset information.

3.3. Ensemble Learning Approach

Ensemble Learning: It is one of the machine learning techniques, used to increase predictive performance and robustness of a model by combing multiple models.

Voting Classifier: It is a ensemble technique, it combines the predictions from different models use soft or hard voting to make a final decision.

The model implements Voting Classifier to ensemble multiple classifiers including Logistic Regression, Decision Tree, Random Forest, XGBoost and KNN for accuracy enhancement. StandardScaler standardization ensures all features receive uniform scaling before KNN application because this technique works best with distance-based models. The Voting Classifier functions using "soft" voting which selects predictions from the model with the greatest overall probability average because this technique generates predictions that are more reliable and balanced. Performance evaluation of the ensemble model occurs on the test set following its training on scaled training data (X_train_scaled, y_train) through accuracy score along with classification report and confusion matrix

analysis. The classification report provides important performance measurements which consist of precision, recall and F1-score for each category.

3.4. Model Evaluation

The Performance of the model is evaluated based on the performance metrics like precision, recall, accuracy, f1-score.

Precision: focuses on how many of the cases that the model predicted as positive (disease cases) were actually correct.

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

Accuracy: measure signifies the number of successful predictions by the model from all its predictions made. It gives an overall idea of the model's correctness.

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

Recall: It tells us how many of the real disease cases were correctly identified.

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

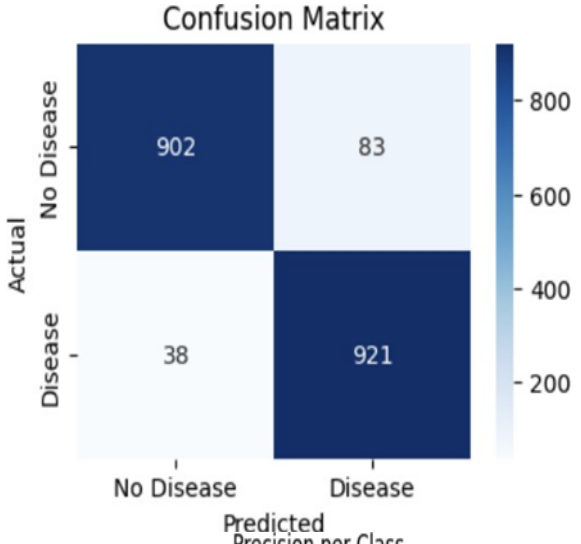
F1-score: For imbalanced datasets you can use F1-score as the harmonic mean of precision and recall to achieve better performance assessment.

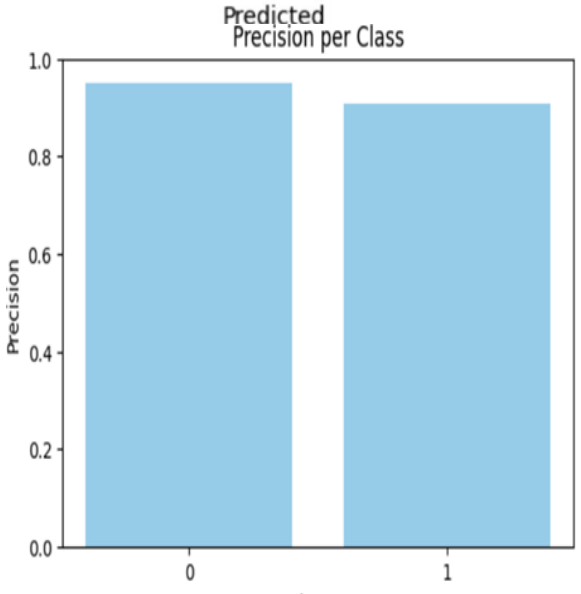
$$F1\text{-}score = \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}}$$

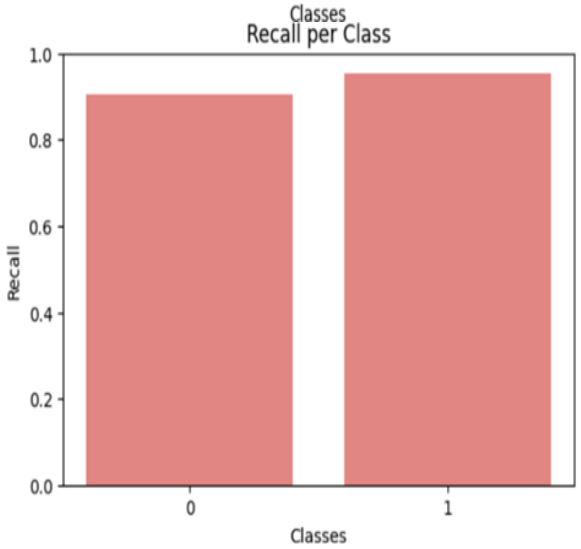
Confusion matrix: is a table shows the true positives, true negatives, false positives and false negatives. This method shows the number of successful predictions among all forecasted values.

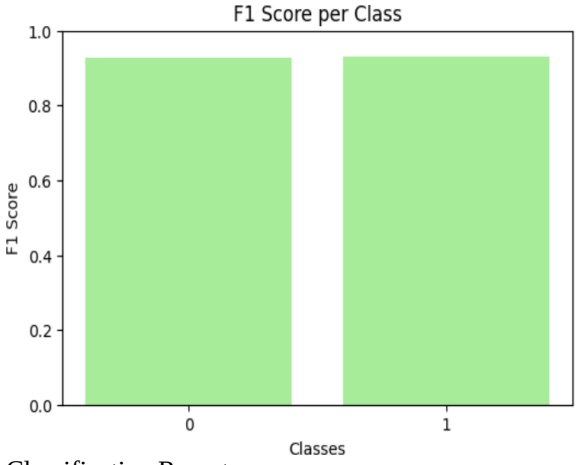
IV. RESULTS AND DISCUSSIONS

In the proposed model, Ensemble model achieved an accuracy of 92% to 93%, precision of 91%, recall of 95% and f1- score of 93%. From these results, we recommend using the ensemble method since it utilizes the predictive accuracy of the models, which in turn improves the classification model. The high recall value indicates proper stroke detection capability that decreases the number of patients wrongly diagnosed with false negatives. The model demonstrates its usefulness in time- based cardiovascular disease prediction through its high F1-score which strikes a proper balance between recall and precision.









Classification Report

	precision	recall	F1-score	support
0	0.96	0.92	0.94	985
1	0.92	0.96	0.94	959
Accuracy	-	-	0.94	1944
Macro avg	0.94	0.94	0.94	1944
Weighted avg	0.94	0.94	0.94	1944

V. CONCLUSION

The project team established a Cardiovascular Stroke Forecasting System through Ensemble Learning methods for better predictive ability. Our merged machine learning models which include Regression Logic and Nearest K-Neighbors together with Random Forest along with Decision Trees and XG-Boost and Voting Classifier led to better results than single systems. Ensemble methods achieve improved accuracy of 92% to 93%, precision of 91%, recall of 95% and f1- score of 93%.

The research demonstrates how artificial intelligence tools detect strokes at an early stage thus enabling medical staff to identify patients with high-risk potential. Design improvements for this model should consist of deep learning technologies as well as real-time patient tracking and larger available clinical data to improve predictive accuracy and reliability.

REFERENCES

[1] Kamutam Vinay, Marneni Yashwant, Prashanth Mulla - Heart Stroke Prediction using Machine Learning, in 2023 ReasearchGate.

[2] S. Chua, V. Sia, P.N.E. Nohuddin - Comparing Machine Learning Models for Heart Disease Prediction, in 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET).

[3] Bah Ibrahima, Xue Yu - KNN Algorithm used for Heart Attack Detection, in 2021 FES Journal of Engineering Science.

[4] Ambrish G, Bharathi Ganesh, Dhanraj - Logistic Regression technique for Prediction of Cardiovascular Disease, in 2022 KeAi Chinese Roots Global Impact in Global Transitions Proceedings.

[5] B.P. Deepak Kumar, Sagar Yellaram, Sumanth Kothamasu – Heart Stroke Prediction Using Machine Learning, in 2021 International Journal of Creative Research Thoughts (IJCRT).

[6] Archana Singh, Rakesh Kumar – Heart Disease Prediction using Machine Learning Algorithms in 2022 International Journal of Advanced Research in Science, Communication and Technology (IJARSCT).

[7] Emil Agbemade - Predicting Heart Disease using Tree-based Model, Data Science and Data Mining, University of Central Florida, 2023.

[8] Anjali Sharma, Cheena Dhingra, Hina Bansal - Implementation of ML Algorithms for Cardiovascular Disease Prediction, in 2024 Emergent Converging Technologies and Biomedical Systems.