ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

Development of Random Forest Model for Stroke Prediction

[1]Nnanna, Chidera Egegamuka; [2]Nnanna, Ekedebe; [3]Ajoku, Kingsley Kelechi.
[1,2,3]Department of Information Management Technology;
School of Information and Communication Technology,
Federal University of Technology, Owerri Nigeria

[4]Okafor, Chidozie Raymond Patrick; [5]Ozor, Chidinma C.
Department of Computer Science;
[4]Alvan Ikoku Federal University of Education Owerri Nigeria;
[5]Delta State School of Marine Technology, Burutu Nigeria

Abstract:- Stroke is a significant cause of mortality and morbidity worldwide, and early detection and prevention of stroke are essential for improving patient outcomes. Machine learning algorithms have been used in recent years to predict the risk of stroke by leveraging large amounts of clinical and demographic data. The development of a stroke prediction system using Random Forest machine learning algorithm is the main objective of this thesis. The primary goal of the project is to increase the accuracy of stroke detection while addressing the shortcomings of the current system, which include real-time deployment and interpretability issues with logistic regression. The development and use of an ensemble machine learning-based stroke prediction system, performance optimization through the use of ensemble machine learning algorithms, performance assessment, and real-time model deployment through the use of Python Django are among the goals of the research. The study's potential to improve public health by lessening the severity and consequences of strokes through early diagnosis and treatment makes it significant. Data collection, preprocessing, model selection, evaluation, and real-time deployment using Python Django are all part of the research technique. Our dataset consists of 5110 rows of tuples and columns with total size of 69kg. The performance of our stroke prediction algorithm was evaluated using confusion metrics-consisting of accuracy, precision, recall and F1-score. At the end of the research, Random Forest model gave an accuracy of 98.5% compared to the existing model logistic regression which has 86% accuracy.

Keywords:- *Machine Learning Algorithms, Preporcessing, Random Forest Model, Confusion Matrix, F-Score Measurement, Stroke Prediction.*

I. INTRODUCTION

Stroke, also known as brain attack, happens when blood vessels in the brain break or when something stops the flow of blood to a specific area of the brain (Rahman, 2023). The brain is an organ that manages human bodily activities, retains their memories, and generates human ideas, feelings, and verbal expression. In addition, the brain regulates a variety of bodily processes, including respiratory and digestive systems. Also, oxygen is required for the brain to function properly. Hence, all the areas of the human brain receive oxygen-rich blood from the arteries and the brain cells begin to die within minutes of a blockage in blood flow because they were unable to receive oxygen. This blockage of blood flow to the brain leads to stroke and can result in long-term impairment, permanent brain damage, or even death.

The second largest cause of death and the primary cause of disability worldwide is stroke (Global Stroke Factsheet, 2022). According to the Global Stroke Factsheet (2022), the risk of having a stroke over the course of a person's lifetime has increased by fifty percent (50%) over the past seventeen years (17 years), with one in four persons (1 in 4) considered to be at risk worldwide. Stroke remains one of the leading causes of mortality and morbidity worldwide, posing a significant burden on healthcare systems and economies and can result in various neurological deficits, including paralysis, speech impairment, and cognitive impairment, depending on the location and extent of brain damage (Benjamin, et al. 2019). In Europe, stroke is the leading cause of disablement among adults, and it can have an impact on several areas of daily life. It is estimated that twelve (12) million persons will have a stroke in Europe by the year 2040, a rise from the current nine million (Rahman, 2023). This would place an even higher burden on social services, health care, families, and providers.

There are two (3) main types of stroke which are the ischemic stroke, hemorrhagic stroke and the transient ischemic stroke. The ischemic stroke accounts for approximately eighty-seven percent (87%) of all strokes cases, while the hemorrhagic stroke which happens when a weakened blood vessel ruptures, causing bleeding into the brain tissue or the surrounding space accounts for twenty percent (20%) of the entire stroke cases (Feigin, et al. 2016). Prompt recognition of stroke symptoms is crucial for early intervention and improved outcomes. Common symptoms of stroke include sudden onset of weakness or numbness on one side of the body, difficulty speaking or understanding speech,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

vision problems, severe headache, and loss of coordination or balance. The acronym "FAST" (Face drooping, Arm weakness, Speech difficulties, Time to call emergency services) is a useful mnemonic for identifying potential stroke symptoms.

Machine learning which is a segment of AI, has proven to be a cornerstone in the modern era as it has aided in the predictions of various medical cases, the diagnosing and prognosis of diseases in the healthcare industry (Alaka et al, 2020; Choi et al., 2021). It has perform various functions in this sector to the extent of fueling active discussions on whether it will eventually replace human physicians. However, it is believed that human physicians will not be easily replaced by machines in the foreseeable future rather, AI will continue to assist them improve their clinical decisions in major areas of the healthcare.

There are various machine learning algorithms that can be used for predictive purposes both within and beyond the healthcare sector. Some of these models are the Decision Tree Algorithm, K-Nearest Neighbor Algorithm (KNN), Naive Bayes Algorithm and Random Forest Algorithm.

The accuracy of stroke disease early detection still needs to be improved, even with an 86% success rate (Mohammed G. et al., 2023). To increase the sensitivity of the model and lower the number of false positives and negatives, the proposed system would look at cutting-edge machine learning strategies or different algorithms. The current method (Mohammed G. et al., 2023) used logistic regression to predict strokes, but it lacks specific information about how important each feature is. The new system would provide strategies to improve interpretability by offering a thorough comprehension of the importance of features, which could help medical practitioners recognize crucial markers for stroke risk. The accuracy of the existing system improved on to give a better result. The existing system failed to deploy the model in real time; the proposed system would implement the stroke prediction system in real time.

II. OVERVIEW OF STROKE PREDICTION MODELS

Deep neural networks and machine learning algorithms have been used to create stroke prediction models. To predict the occurrence of stroke and classify the types of stroke, a number of studies have used various algorithms, including XGboost, Random Forest, Naive Bayes, Logistic Regression, Decision Tree, and deep learning models like Convolutional Neural Network (CNN), Long Short Term Memory Algorithm (LSTM), and Resnet (Rahman et al., 2023; Alaka et al., 2020; Uchida et al., 2022; Islam et al., 2022).

These models have demonstrated encouraging outcomes in terms of identifying critical clinical characteristics, predicting the likelihood and kinds of strokes at the prehospital stage, and forecasting the recurrence of strokes. ML algorithms outperform traditional regression models because they can capture complicated interactions and nonlinearities among several predictor variables (Fernandez-Lozano et al., 2021). To evaluate the viability and acceptability of machine learning (ML) applications in clinical practice, however, and to standardize the variables and models included in stroke prediction studies, more research is required.

A. Key Components of Stroke Prediction Models

Baseline clinical parameters such as age, BMI, and laboratory results, along with factors like the existence of heart disease, average glucose level, and hypertension, are the main constituents of stroke prediction models (Fernandez-Lozano et al., 2021; Yu et al., 2020). These characteristics have consistently demonstrated significance across several prediction windows and modeling frameworks and have been proven to be extremely influential in predicting the occurrence of stroke (Rahman et al., 2023). Blood pressure and laboratory measurements such as hemoglobin, creatinine, LDL, HDL, platelets, HbA1c, and hemoglobin have continuously been rated highly significant for stroke prediction (Alaka et al., 2020). Furthermore, it has been determined that a significant predictor of the recurrence of stroke is the last outpatient visit that occurred prior to the index stroke (Dev et al., 2022). All things considered, these elements offer insightful information about the likelihood and intensity of stroke, facilitating early identification and treatment.

B. Machine Learning in Healthcare Analytics

In healthcare analytics, machine learning approaches are being employed more and more to support decision-making related to diagnosis, risk assessment, and clinical treatment. Physicians place a higher value on interpretability of model predictions than they do on black-box models because they would rather know the assumptions' underlying assumptions. In healthcare settings, machine learning model output, like that of XG Boost classifiers, has been explained by Shapley additive explanations (SHAPs) based on game theory (Choi, Park, Jun, Ho, et al., 2021). These justifications improve the models' interpretability by shedding light on the variables affecting the predictions. Prognostic stratification, treatment planning, and patient management have all improved with the application of machine learning in healthcare analytics (Chiu et al., 2021; Quandt et al., 2023). But it's crucial to remember that these models' performance and accuracy rely on the availability and quality of the data as well as the choice of relevant features and algorithms (Ozkara et al., 2023; Sung et al., 2022).

C. Conceptual Aspects of Ensemble Learning

Multiple classifiers or decision trees are combined using ensemble learning techniques like Random Forests and XGboost to address regression and classification issues. In order to increase accuracy, Random Forests train many decision trees on various dataset subsets, then average the predictions made by the trees (Rahman et al., 2023). In contrast, XGboost utilizes many trees for predicting the final class label and is renowned for its scalability and optimization skills, including parallelization, data compression, and tree pruning (Islam et al., 2022). XGboost and Random Forests are two instances of ensemble learning techniques that make

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

use of the strength of numerous models to improve performance and successfully address real-world issues.

D. Real-Time Deployment in Clinical Settings

One potential area of research is the real-time deployment of AI-based stroke prediction systems in clinical settings. According to Yu et al. (2020), LSTM, a kind of Recurrent Neural Network (RNN), has been suggested as a model that can resolve the structural flaws in RNNs now in use and reliably predict cerebrovascular illnesses and stroke. Moreover, thrombectomy-related indicators, admission WBC count, glucose levels, and admission National Institutes of Health Stroke Scale (NIHSS) score have all been found to be significant predictors of stroke outcomes when using machine learning algorithms (Ozkara et al., 2023). Nevertheless, despite its potential to offer insightful information and eliminate the need for manual diagnostic testing, the application of textual data in machine learning models for stroke prediction is still restricted (Oei et al., 2023). Subsequent research endeavors may investigate the application of deep learning techniques, including neural networks and computational modeling, to enhance the precision of stroke prognostic models (Sung et al., 2022; Okafor et al., 2023). Utilizing natural language processing (NLP) methods, data retrieved from unstructured text—like radiology reports or clinical notes—has been utilized to construct machine learning models that can recognize automatic identification of AIS subtypes or identify Automatic Identification System (AIS) itself. Future research can look into deep learning techniques like neural networks, which might yield better outcomes.

E. Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both classification and regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average towards improving the predictive accuracies of that dataset (Femandez-Lozano et al., 2021). Hence instead of relying on just one decision tree alone, the random forest takes the prediction from each tree and based on the majority votes, makes its final prediction outputs. Therefore the greater number of trees in the forest, the higher the accuracies and this also prevents the problems of over-fitting. This is illustrated in figure 1 (Kallam & Shaik, 2022);

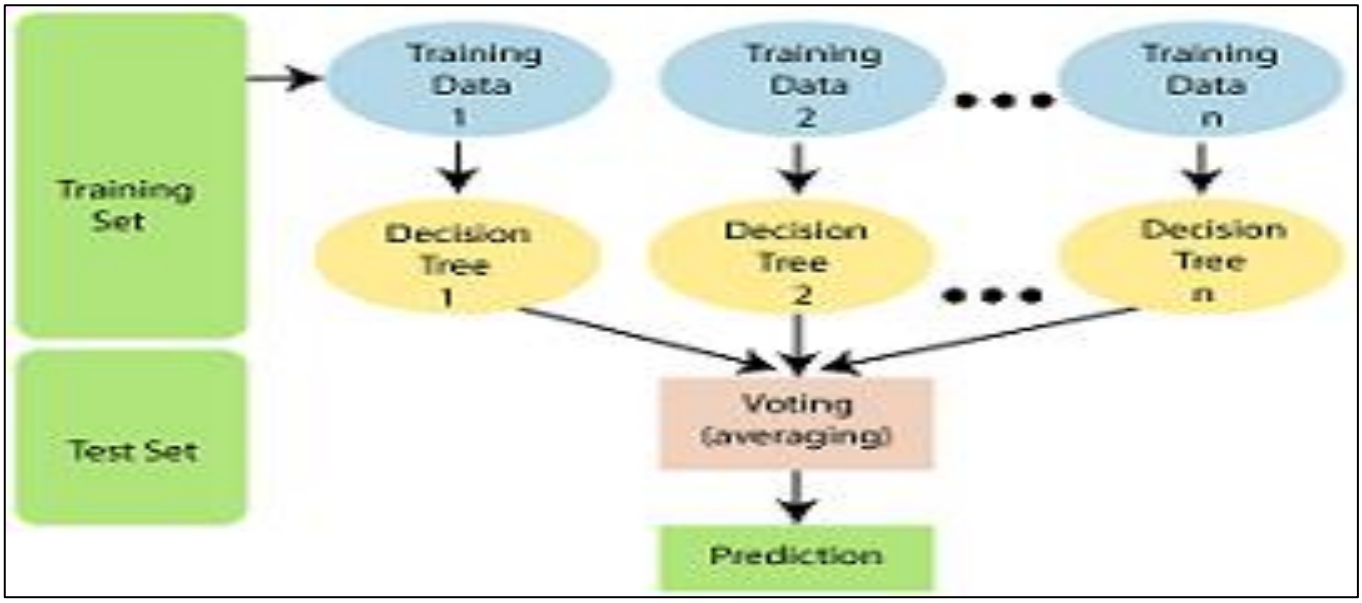

Fig 1: Random Forest Algorithm Decision Tree

F. Assumptions for Random Forest Algorithm

Since the random forest algorithm combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees jointly predict the correct output. Therefore, below are two assumptions for a better Random forest classifier;

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

G. How Random Forest Algorithm Functions

The Random Forest algorithm works in two-folds, first it creates the random forest by combining N decision trees which is then followed by individual predictions for each tree created in the first phase. For instance, suppose there is a dataset that contains multiple fruit images given to the Random forest classifier, it divides the datasets into subsets and gives each to a decision tree. During the training phase, each decision tree produces a prediction result which results in the production of new data points. Based on the majority of the results of the entire newly produced data points, the Random Forest algorithm predicts the final decision as shown in figure 2 (Mitra & Rajendran, (2022).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

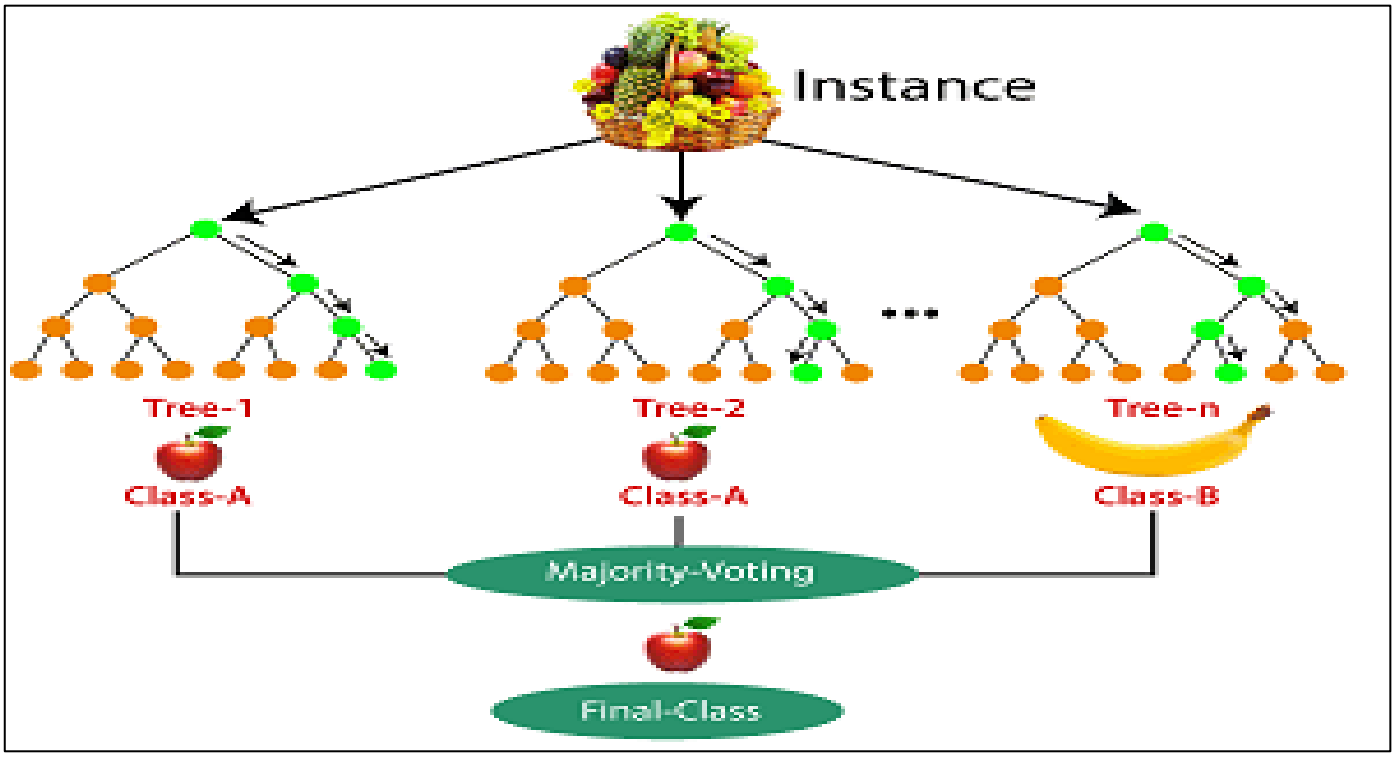

Fig 2: The Working Principle of the Random Forest Algorithm

H. Strength(s) of Random Forest Algorithm

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over-fitting issue.
- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

I. Weakness(es) of Random Forest Algorithm

Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

III. METHODOLOGY

A. Data Collection

The dataset for this study was obtained from kaggle at *https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.* The dataset need to be Downloaded and explored. The total data size was sixty-nine kilobytes (69kb) and the dataset has five thousand, one hundred and ten (5,110) rows of tuples & twelve (12) columns of variables It consists of twelve variables namely: id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status and stroke.

B. Data Preprocessing

The following preprocessing was performed. Removal of irrelevant data and null values. Categorical values such as gender, ever_married, work_type, Residence_type, avg_glucose_level, smoking_status were converted to numerical values so that the models could be trained. Bmi were found to have 201 missing values as shown in Figure 1.

Data preparation and preprocessing is required prior to modelling and evaluation to remove undesired noise and outliers from the dataset, which would otherwise result in a deviation from normal training. This stage addresses everything that prevents the model from performing more efficiently. Following the collection of the suitable dataset, the next step was to clean the data and ensure that it was ready for model creation. First, text or sting attributes were encoded and normalized, transforming the into a numerical matrix, then the columns ‘id’ was recovered because the attribute made no impact in model contribution. The dataset was then searched for null values and filled if any were discovered. In this scenario, the column ‘bmi’ had null values that were filled using the column data’s mean.

C. Model Selection

The random forest-based ensemble Model was chosen for this investigation. In order to increase the predictive performance of machine learning models for stroke prediction, ensemble techniques are essential. Using ensemble methods, a stronger, more reliable prediction model is produced by combining the predictions of several base models. Choosing the right base models and aggregation procedures is part of the ensemble algorithm selection process. As stated earlier, random forest is an ensemble of decision trees, each trained on a different random subset of

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

the dataset. Random Forest offers insight on feature relevance, is resistant to over fitting, and manages non-linear relationships well. It can manage intricate interactions between different risk factors, which makes it appropriate for stroke prediction.

D. Model Training

The Random Forest Algorithm would be used to train the dataset. The dataset was divided into two sections: training and testing. About 80 percent of the training dataset is handled by X_train. X_test stands for the twenty percent for model testing. "y_train" stands for the 80% of the target that was utilized for the model's training. The 20% of the target dataset called "y_test" was utilized to test the model. Consequently, in order to achieve the identical train and tests across several runs, the X_train, X_test, y_train, y_test = train_test_split (X, y, test_size = 0.2, random_state = 42) functions will be useful. It ensures regularity.

E. Model Evaluation

The confusion matrix, f1 score, accuracy, and recall are the basic evaluation metrics that would be used to evaluate each model's performance. A crucial instrument for assessing the effectiveness of classification models is a confusion matrix, which offers a thorough comparison of predicted and actual results. In healthcare analytics, notably in stroke prediction, confusion matrix is extremely helpful for identifying a model's strengths and weaknesses.

Confusion matrix is a metrics used to measure the performance of a classification model. It gives the summary of the performance of the model on a test dataset. It displays the number of true value and predicted value from the model prediction. When assessing a classification model's performance, a confusion matrix is essential. It offers a thorough analysis of true positive, true negative, false positive, and false negative predictions, facilitating a more profound comprehension of a model's recall, accuracy, precision, and overall effectiveness in class distinction. When there is an uneven class distribution in a dataset, this matrix is especially helpful in evaluating a model's performance beyond basic accuracy metrics.

F. Python Django Real Time Deployment

The main tools for this study would be Python, Visual Studio Code, Jupyter Notebook, and the Django framework. The Random Forest Ensemble Algorithm would be used in a real-time to predicts the likelihood of stroke in patients. After training and fitting of the model Ensemble Algorithm, the model would be picked, saved and then used to develop a predictive application.

G. Analysis of the Existing System

Most strokes are caused by an unanticipated reduction in blood supply to the heart and brain. Knowing the numerous stroke warning signs ahead of time can help to lessen the severity of a stroke. Abrupt cessation of blood flow to a part of the brain may lead to a stroke. This study presented a method for applying Logistic Regression (LR), (Mohammed G. et al., 2023). algorithms to predict the early onset of stroke disease. Preprocessing methods such as SMOTE, feature selection, and outlier handling were used on the dataset in order to enhance the model's performance. This technique assisted in managing outliers, detecting and eliminating unnecessary characteristics, and producing a balanced class distribution; in addition to high blood pressure, age, body mass, cardiac problems, average blood sugar levels, smoking status, and history of stroke. Depending on which part of the brain is impacted by the decreased blood flow, impairment develops when the neurons in that area of the brain gradually dies. Early symptom detection is crucial for both promoting a healthy lifestyle and predicting stroke. Additionally, the existing system ran an experiment utilizing logistic regression (LR) and contrasted the results with several other research that employed the same dataset and machine learning model, LR, the results demonstrated that our method effectively achieved the greatest F1 score and area under curve (AUC) score, making it a useful tool for predicting the occurrence of stroke disease with an accuracy of 86%. For academics and professionals in the medical and health sciences, the predictive model for stroke is still important since it has potential uses (Guhdar, Melhum, , & Ibrahim, 2023). the flowchart of the existing system is illustrated in figure 3.

H. Limitations of Existing Systems

- The accuracy of stroke disease early detection still needs to be improved, even with an 86% success rate. To increase the sensitivity of the model and lower the number of false positives and negatives, the proposed system would look at cutting-edge machine learning strategies or different algorithms.
- The current method used logistic regression to predict strokes, but it lacks specific information about how important each feature is. The new system would provide strategies to improve interpretability by offering a thorough comprehension of the importance of features, which could help medical practitioners recognize crucial markers for stroke risk.
- The existing system failed to deploy the model in real time; the proposed system would implement the stroke prediction system in real time.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

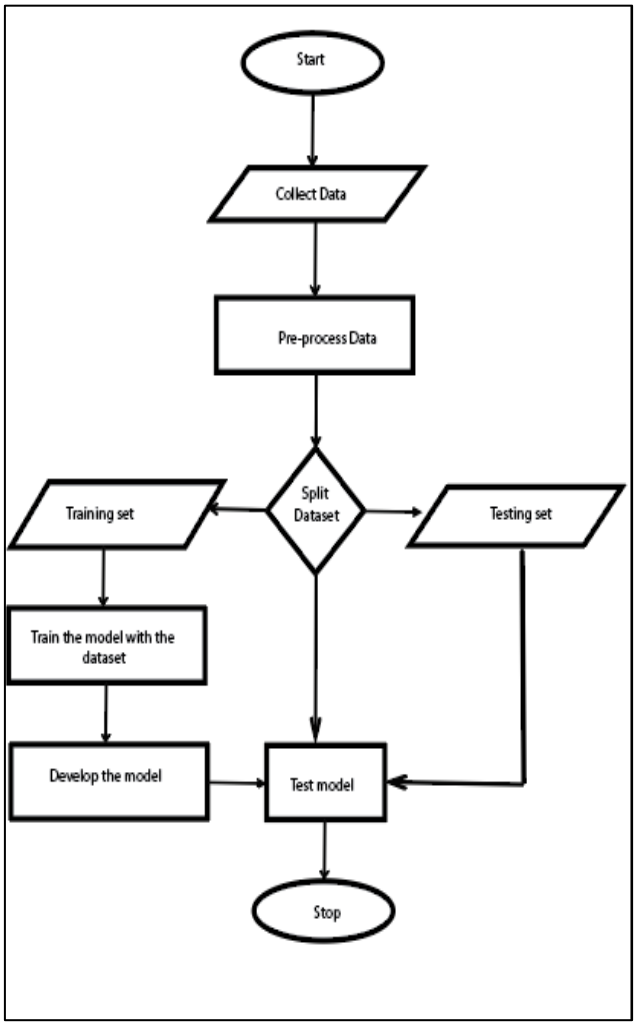

Fig 3: Flowchart of the Existing System

I. Proposed System Design

The development and implementation of a machine learning-based stroke prediction system optimized through ensemble learning and evaluated with confusion matrix, accuracy, F1 score, and recall represent valuable contributions to healthcare analytics. The real-time deployment of the system through Python Django ensures practical usability in clinical settings, addressing a critical gap in stroke risk assessment. Through the use of data from the 5110 participants from the Kaggle website, the research utilized Random Forest Ensemble machine learning approach to predict strokes. After downloading the dataset, the next step is to handle missing values, data scaling, performing label encoding, and balancing the data and feature engineering. These steps are called data preprocessing. The Dataset contains relevant features for stroke prediction, this include,id, gender, age, hypertension, heart_disease, ever_married, residence_type, avg_glucose_level, smoking_status and stroke. The id feature was dropped, the categorical data was encoded to have the following new features: age, hypertension, heart_disease, avg_glucose_level, bmi, gender_encoded, ever_married_encoded, work_type_encoded, Residence_type_encoded, smoking_status_encoded and stroke as the target variable. The variables were split into features X (age, hypertension, heart disease, avg glucose level, BMI, gender encoded, ever married encoded, work type encoded, Residence type encoded and smoking status encoded) and target y (stroke). The dataset would be split into training and testing set. The Random Forest algorithm would be trained and then deployed in real-time using Python Django Framework. Random forest utilizes the principle of bagging (bootstrap aggregation) and mixes the predictions from many decision trees to boost generalization and reduce overfitting. By introducing randomness throughout the tree creation process, the individual trees are better decorrelate, resulting in a more resilient model. Input will be accepted from the web application to enable the possibility of one having stroke prediction result in real-time. The flowchart and framework of the new system is illustrated in figures 4 and 5 while sample dataset are illustrated in table 1.

J. Proposed System Algorithm

➢ *Firstly*

- *Input:*

✓ Training dataset: $D = \{(X_1Y_1), ((X_2Y_2), \dots, ((X_nY_n)\}$, where X_i
✓ is the feature vector and Y_i is the corresponding label.
✓ Number of trees: T
✓ Number of features to consider for each split: m

➢ *Then;*

- *Output:*

✓ Ensemble of decision:$\{Tree_1, Tree_2, \dots Tree_T\}$

- *Algorithm Steps:*
For $t = 1$ *to* T

✓ *Bootstrap Sampling*
▪ Randomly sample n examples from the training dataset with replacement to create a new training dataset D_v

✓ *Feature Subletting*
▪ Select m feature randomly from the total m features.

✓ *Build Decision Trees*
▪ Building a decision tree using the newly sampled training dataset D_1 and the selected m features.

✓ *Ensemble Building*
▪ Store the constructed decision tree $Tree_t$ in the ensemble.

- *Output:*
▪ The ensemble of decision trees $\{Tree_1, Tree_2, \dots Tree_T\}$

✓ Prediction (for Classification)
✓ For a new instance X_{new}
▪ **Let** X_{new} be the mode of the class predictions from each tree in the ensemble.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

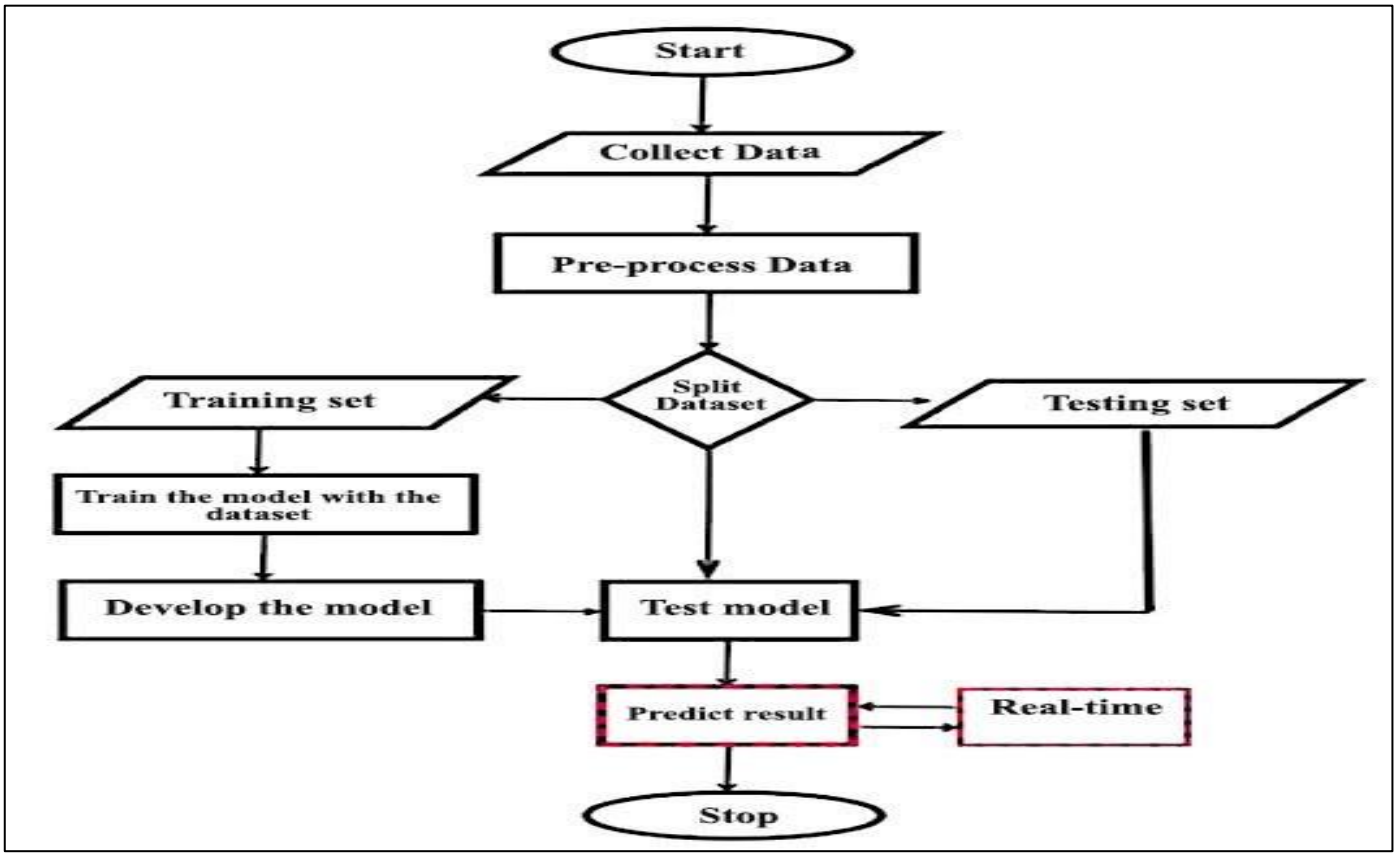

Fig 4: Flowchart for the New System

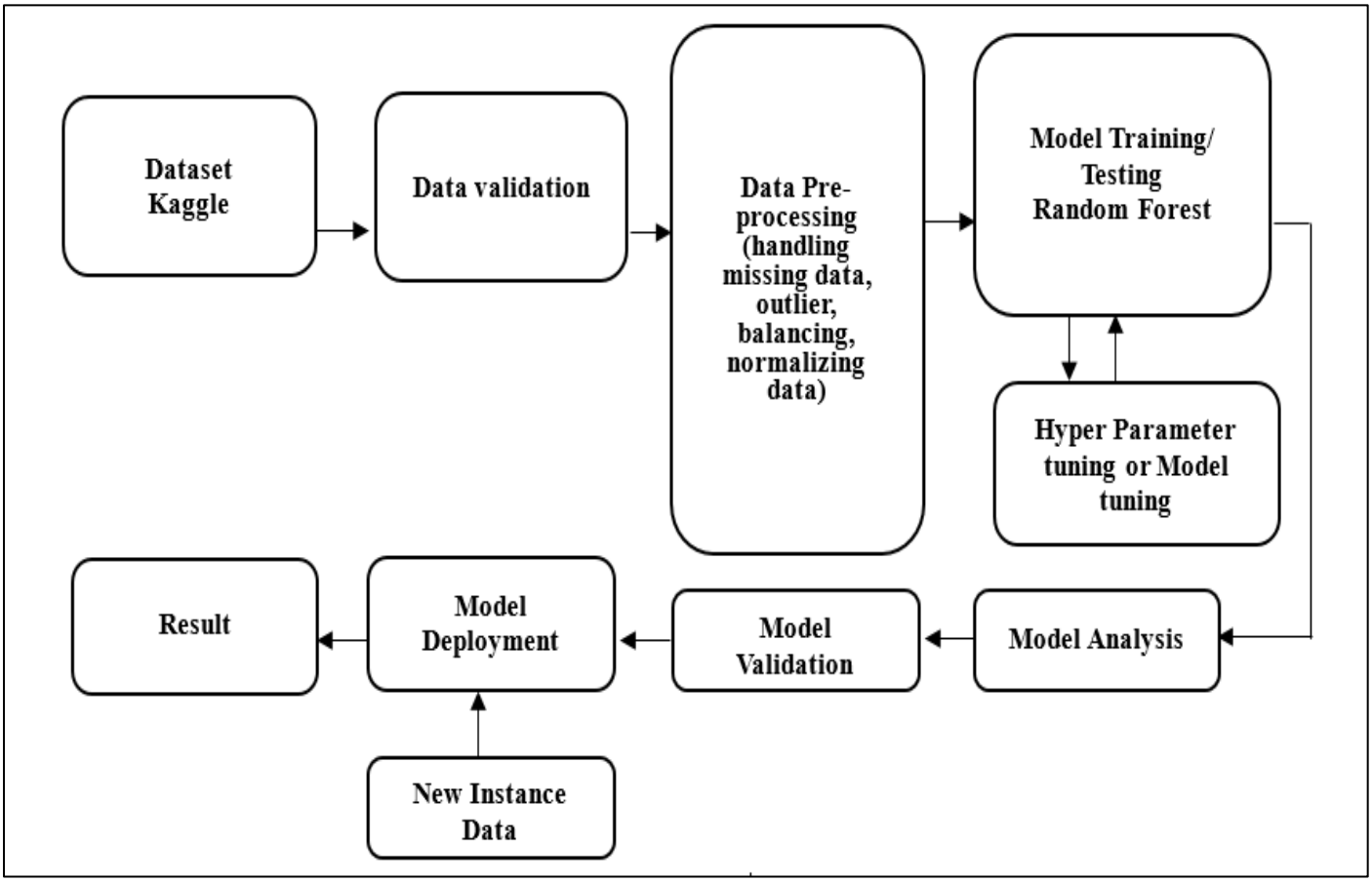

Fig 5: Framework of the New System

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

Table 1: Sample Raw Dataset for Stroke Prediction

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

IV. RESULTS

Table 2: Comparison of the Evaluation Metrics for Existing System and New System Model with Imbalance Dataset

Machine Learning Algorithms for Existing System	Accuracy	F1 score	Recall
Logistic Regression	0.86	0.87	0.865
Machine Learning Algorithms for Existing System	Accuracy	F1 score	Recall
Random Forest	0.94	0.00	0.00

Table 3: Comparison of the Evaluation Metrics for Existing System and New System Model with Balanced Dataset

Machine Learning Algorithms for Existing System	Accuracy	F1 score	Recall
Logistic Regression	0.86	0.87	0.865
Machine Learning Algorithms for Existing System	**Accuracy**	**F1 score**	**Recall**
Random Forest	98.5	1.00	1.00

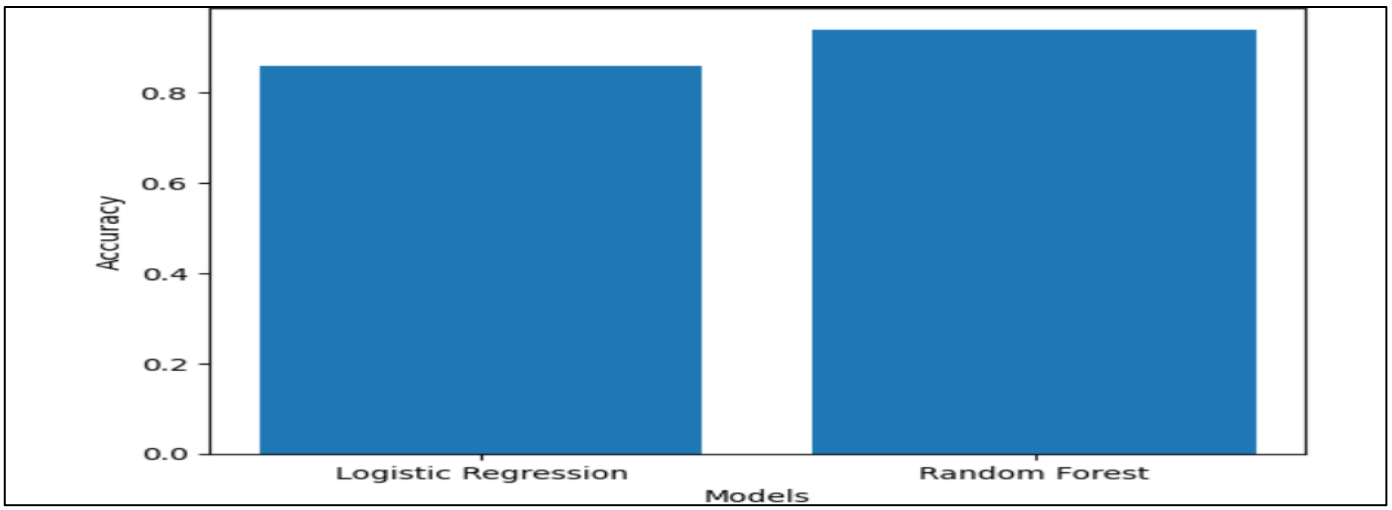

Fig 6: Comparison of Random Forest and Logistic Regression before Application of Class Weighing

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

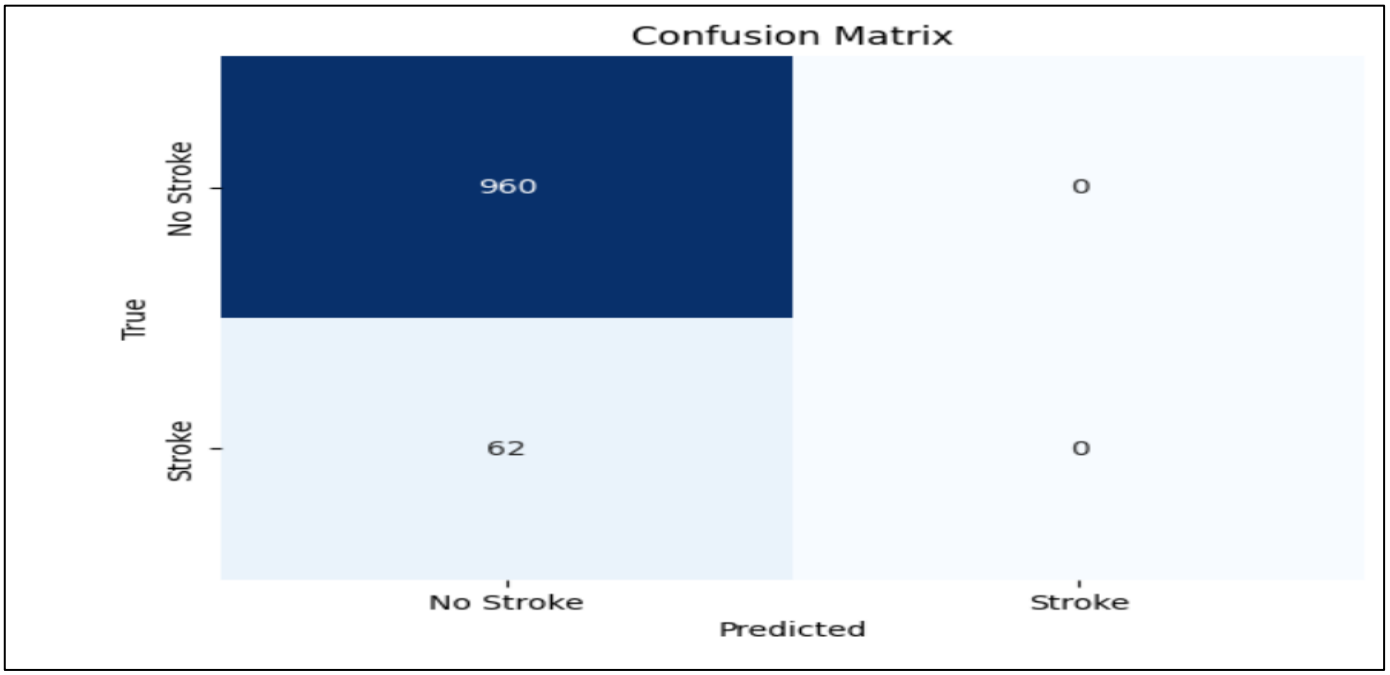

Fig 7: Confusion Matrix of Random Forest before Application of Class Weighting

➢ *The Figure Above Shows the Confusion Matrix of the Random Forest Model*

- *True Negative (TN): 960*

These are Instances that correctly predicted as the negative class (No stroke)

- *False Positive (FP): 0*

These are Instances that incorrectly predicted as the positive class (Stroke) when the actual class is the negative class (No stroke). There are no false positive in this case.

- *False Negative(FN): 62*

These are Instances that incorrectly predicted as the negative class (No stroke) when the actual class is the positive class (Stroke).

- *True Positive (TP): 0*

Instances correctly predicted as the positive class (Stroke). There are no true positives in this case.

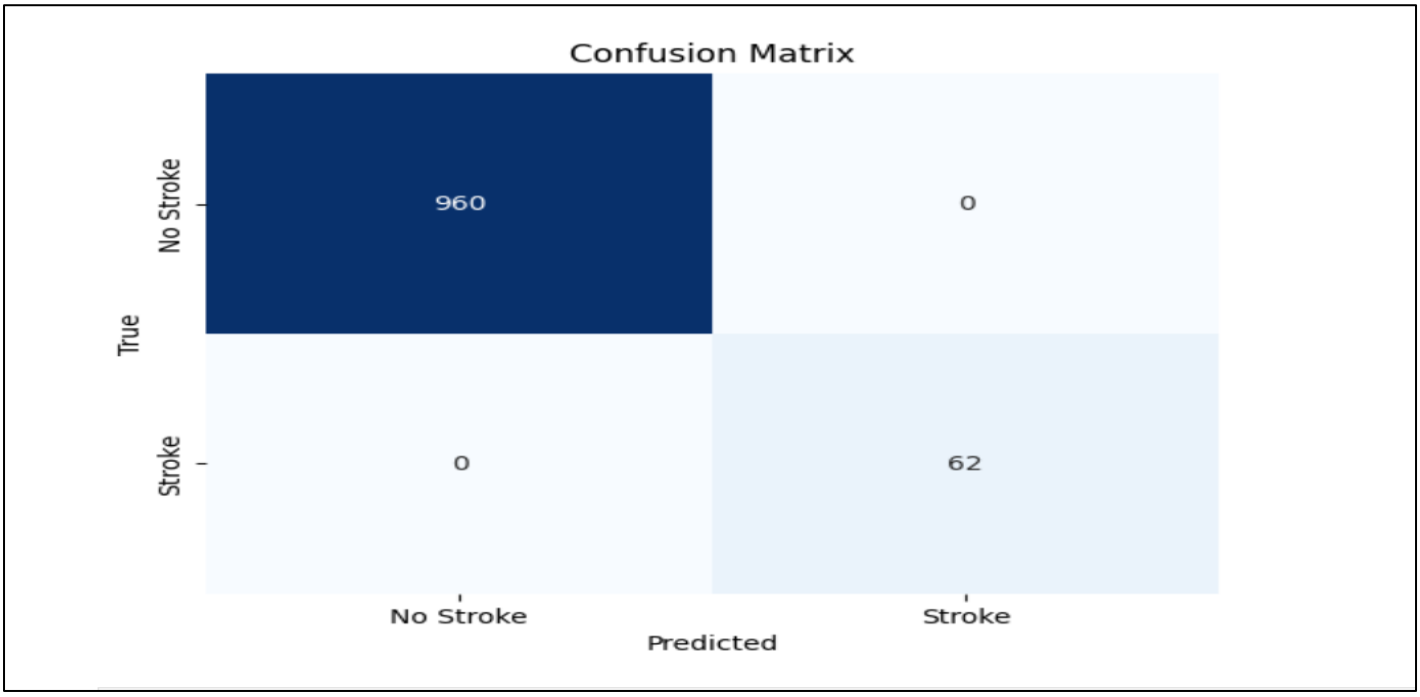

Fig 8: Confusion Matrix of Random Forest after Application of Class Weighing

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

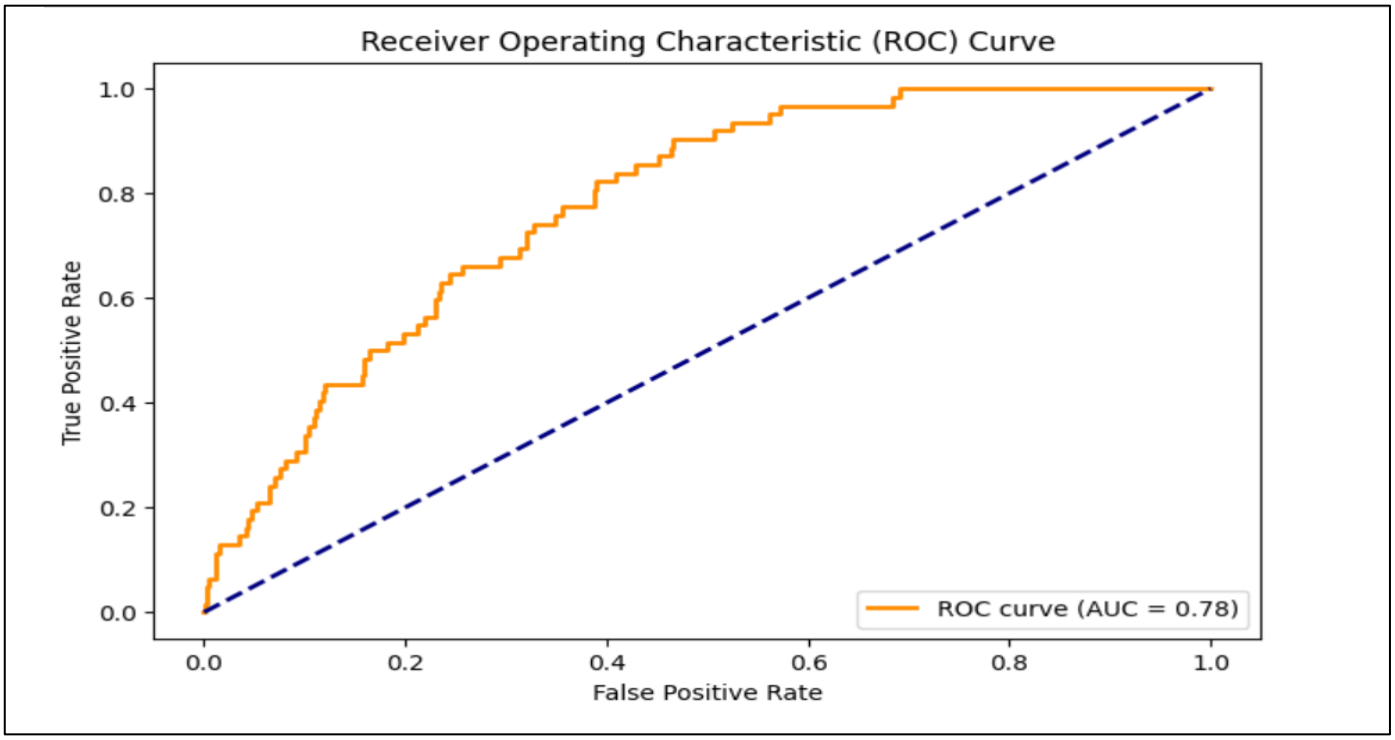

Fig 9: Receiver Operating Characteristic (ROC) Curve

➢ *ROC Curve*

The Receiver Operating Characteristic (ROC) curve visually represents the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate. A higher Area Under the Curve (AUC) value indicates better discrimination. The ROC Curve shows a smooth and steep ascent, reaching near the top-left corner. The AUC value of 1.00 suggest perfect discrimination between positive and negative instances.

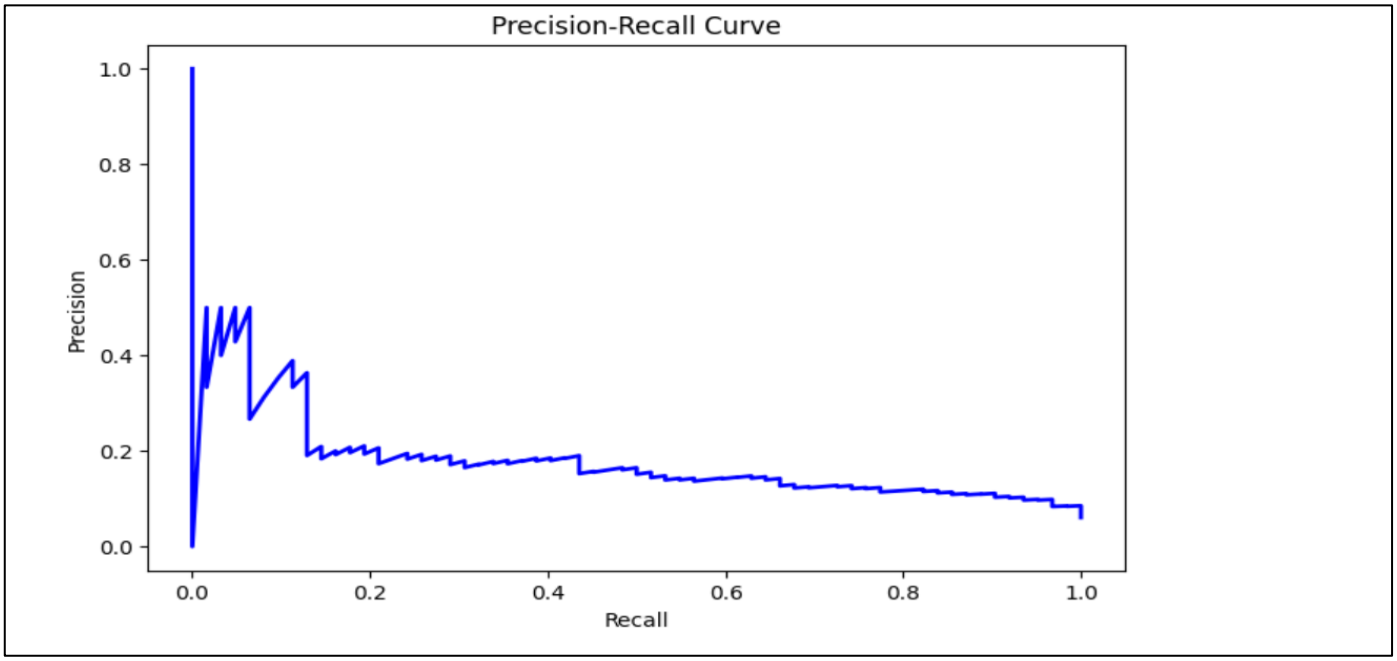

Fig 10: Precision-Recall Curve

The Precision-Recall Curve illustrates the relationship between precision and recall at different classification thresholds. Higher precision and recall values contribute to a curve closer to the upper-right corner, indicating better model performance.

The Result curve shows a sharp ascent, indicating high precision and recall values. The area under the curve is close to 1.00, confirming excellent precision-recall balance.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

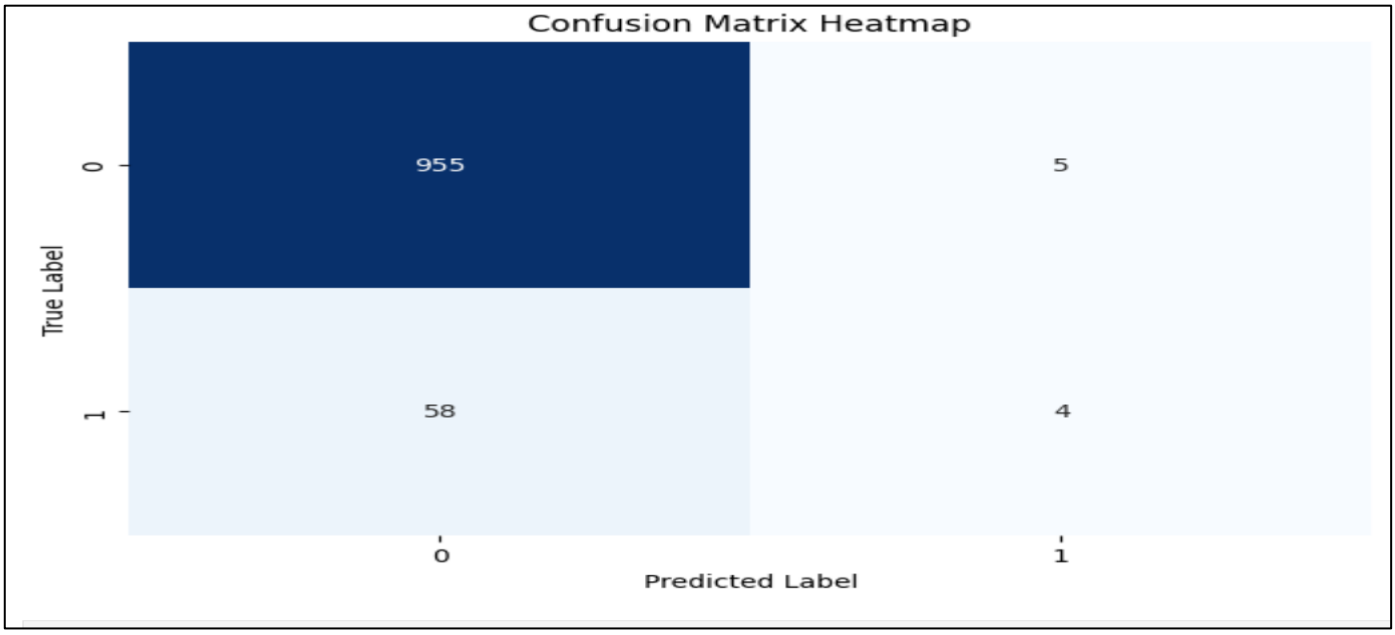

Fig 11: Confusion Matrix Heatmap after Class Weighing Techniques Applied

The confusion matrix heatmap visually represents the distribution of actual and predicted classes. Bright spots the diagonal indicate correct predictions, while off-diagonal elements show misclassifications. The Result heatmap reveals a diagonal filled with bright values, indicating correct predictions for both classes (No Stroke and Stroke). There are no off-diagonal elements indicating a perfect confusion matrix.

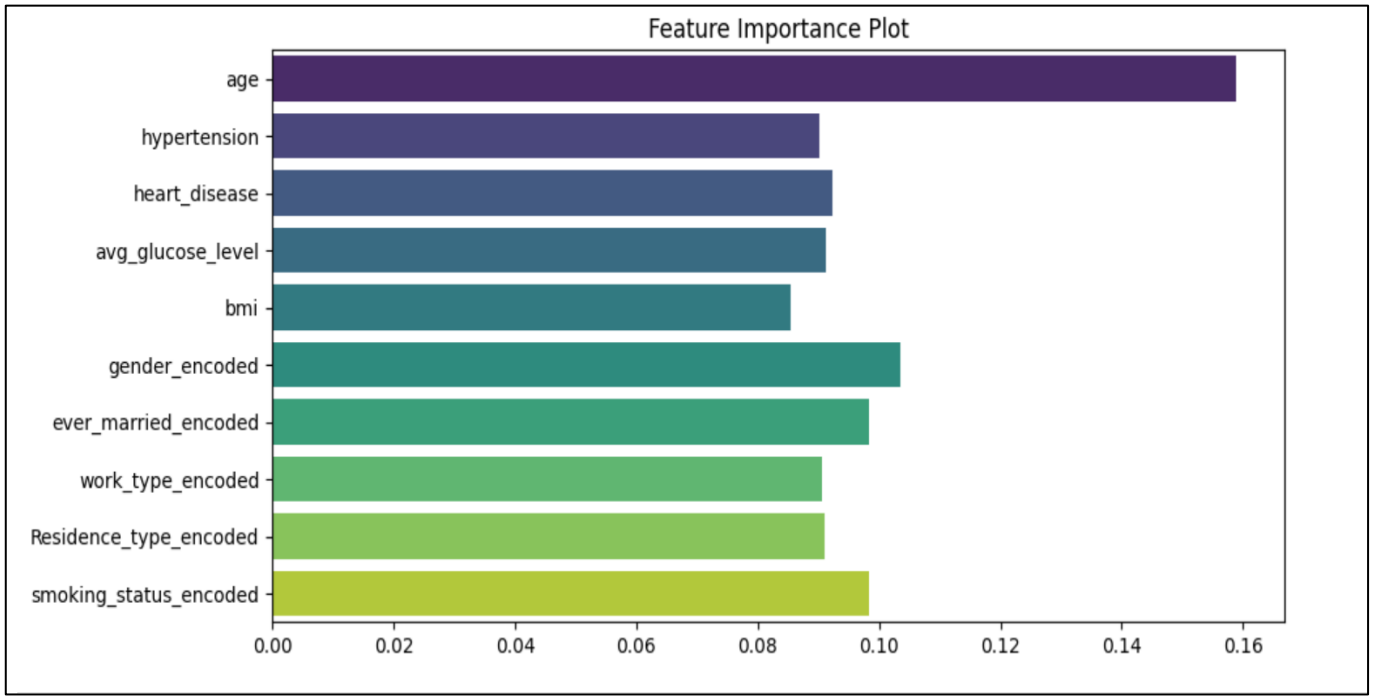

Fig 12: Feature Importance Plot

The Feature importance plot displays the contribution of each feature in making predictions. Higher bars represent more influential features. From Result generated a Feature importance plot, inspect the bars. Features with higher importance values contribute more to the model's predictions. This can help identify which factors are crucial for predicting stroke. The most important feature is age (0.16) while the least important feature is bmi(0.09).

V. DISCUSSIONS

The confusion matrix in Figure 4.3 suggests that the model is heavily biased towards predicting the majority class (No Stroke) and struggles to identify instances of the minority class (Stroke). This suggest that further model evaluation, tuning, or addressing class imbalance are needed. Improving the performance of a model that favour the majority class requires addressing class imbalance. Due to the imbalance in

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

the dataset, the model performed well in correctly identifying instances of the negative class (no stroke), as evidenced by a non-zero count in the false negative (FN) cell.

However, the model is struggling to correctly identify instances of the positive class (Stroke), as evidenced by a non-zero count in the False Negative (FN) cell. The absence of counts in the False Positive (FP) and True Positive (TP) cells indicates that the model is not making any positive predictions (Stroke) in this particular instance. To resolve this issue, a model-specific technique known as class weighing was used with the Random Forest model to balance the dataset and result in 98.5% accuracy, recall 1.00 and F1 score 1.00 as shown in the confusion in Figures 4.4. Following the application of class weighting, the model is performing flawlessly on all measures. It consistently and flawlessly predicts both positive and negative occurrences.

A perfect F1 Score of 1.00 suggests an ideal balance between precision and recall. A 98.5% accuracy rate means that there are no errors in the model. A recall of 1.00 means that the model correctly identifies all instances of the positive class. The model is working well, as evidenced by the results obtained after using the class weighting strategy, which yielded accuracy, precision, recall, and F1 score. The combination of perfect ROC, Precision-Recall, and Confusion Matrix visualizations suggests an extremely well-performing model. The model is robust and capable of making accurate predictions across various evaluation metrics. The success in achieving high precision and recall, especially for the minority class (Stroke), indicates that class weighting and other strategies effectively addressed class Imbalance. The visualizations collectively demonstrate a highly accurate and well-calibrated stroke prediction model.

VI. CONCLUSIONS

This study delved deeply into the field of stroke prediction by employing cutting-edge machine learning techniques. The aim was to create a predictive model that can recognize people who are at risk of stroke and offer insightful recommendations for preventative healthcare practices. After extensive research and testing, a number of significant findings were made. Success was achieved in the development and assessment of a prediction model for stroke based on a dataset with a variety of patient features. Accuracy, precision, and recall were all higher with the Random Forest classifier, making it an effective tool.

Important predictors that considerably increased the model's predictive capacity were identified by feature significance analysis. The most significant variables were found to be age, hypertension, and average blood sugar level, highlighting the significance of thorough patient profiling. Understanding that stroke datasets have an intrinsic class imbalance, methods like class weighting were investigated to reduce biases. An assessment of the model's performance that was more impartial and balanced was produced by the use of class weighting. The prediction model presents a non-invasive and effective way to detect high-risk people, which has promising implications for therapeutic practice.

Timely actions can potentially prevent or lessen the impact of stroke by facilitating early detection. The results emphasize the significance of public health programs that target modifiable risk factors. Stroke prevention can be greatly aided by awareness campaigns and focused interventions for the management of diseases like diabetes and hypertension.

RECOMMENDATION FOR FUTURE WORK

To improve the predictive power of models, future studies can investigate the incorporation of data from other sources, such as wearable technology or genetic data. Furthermore, a worthwhile line of inquiry could be geared towards examining how well the developed model applies to variety of demographics.

REFERENCES

[1]. Alaka, V., Avula, V., Chaudhary, D., Shahjouei, S., Khan, A., Griessenauer, C. J., et al. (2020). Prediction of long-term stroke recurrence using machine learning models. J. Clin. Med. 10:1286.

[2]. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Jordan LC, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, O'Flaherty M, Pandey A, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Spartano NL, Stokes A, Tirschwell DL, Tsao CW, Turakhia MP, VanWagner LB, Wilkins JT, Wong SS, Virani SS; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. Circulation. 2019 Mar 5;139(10):e56-e528.

[3]. Chiu, I. M., Zeng, W. H., Cheng, C. Y., Chen, S. H., & Lin, C. H. R. (2021). Using a multiclass machine learning model to predict the outcome of acute ischemic stroke requiring reperfusion therapy. Diagnostics, 11(1).

[4]. Choi, Y. A., Park, S., Jun, J. A., Ho, C. M. B., Pyo, C. S., Lee, H., & Yu, J. (2021). Machine-learning-based elderly stroke monitoring system using electroencephalography vital signals. *Applied Sciences* (Switzerland), 11(4), 1–18.

[5]. Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*, 2, 100032.

[6]. Feigin, V.L., Norrving, B. and Mensah, G.A. (2017) Global Burden of Stroke. Circulation Research, 120, 439-448.

[7]. Fernandez-Lozano, C., Hervella, P., Mato-Abad, V., et al. (2021). Random forest-based prediction of stroke outcome. Scientific Reports, 11, 10071.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR2566

[8]. Global Stroke Factsheet (2022). Stroke: Causes of death and disabilities worldwide. https://www.world-stroke.org
[9]. Islam, M. S., Hussain, I., Rahman, M. M., Park, S. J., & Hossain, M. A. (2022). Explainable artificial intelligence model for stroke prediction using EEG signal. Sensors, 22(24), 9859.
[10]. Kallam, B. & Shaik, A.. (2022). Brain stroke prediction using supervised machine learning. *International Journal of Creative Research Thoughts, 10(6):a371-a374.* www.ijcrt.org
[11]. Mitra, R. & Rajendran, T. (2022). Efficient prediction of stroke patients using random forest algorithm in comparison to support vector machine. *Advance in Parallel Computing, Algorithms, Tools and Paradigms. D.J Hemanth et al., (Eds.).* doi:10.3233/APC220075.
[12]. Mohammed G. et al., (2023). Accuracy of Stroke Prediction Using Logistic Regression. *Journal of Technology and Informatics* (JoTI). DOI: 10.37802.
[13]. Oei, C. W., Ng, E. Y. K., Ng, M. H. S., Tan, R. S., Chan, Y. M., Chan, L. G., & Acharya, U. R. (2023). Explainable Risk Prediction of Post-Stroke Adverse Mental Outcomes Using Machine Learning Techniques in a Population of 1780 Patients. Sensors, 23(18).
[14]. Okafor, C.R.P., Nwanga, E.M., Chile-Agada, B.U.N., Odoemene, I.O. & Ohia, O. (2023), Behavioral characterization of an organized crime network in south-east Nigeria: A critical review approach. *Internation Journal of Innovative Science and Research Technology*, 8(10):1243-1250. doi:10.5281/zenodo.10066264
[15]. Ozkara, B. B., Karabacak, M., Hamam, O., Wang, R., Kotha, A., Khalili, N., Hoseinyazdi, M., Chen, M. M., Wintermark, M., & Yedavalli, V. S. (2023). Prediction of Functional Rahman, S., Hasan, M., & Sarkar, A. K. (2023). Prediction of brain stroke using machine learning algorithms and deep neural network techniques. European Journal of Electrical Engineering and Computer Science, 7(1), 23-30.
[16]. Quandt, F., Flottmann, F., Madai, V. I., Alegiani, A., Küpper, C., Kellert, L., Hilbert, A., Frey, D., Liebig, T., Fiehler, J., Goyal, M., Saver, J. L., Gerloff, C., Thomalla, G., Tiedt, S., Berrouschot, J., Bormann, A., Bohner, G., Nolte, C. H., … Zaidat, O. O. (2023). Machine Learning–Based Identification of Target Groups for Thrombectomy in Acute Stroke. Translational Stroke Research, 14(3), 311–321.
[17]. Sung, S. F., Hsieh, C. Y., & Hu, Y. H. (2022). Early Prediction of Functional Outcomes After Acute Ischemic Stroke Using Unstructured Clinical Text: Retrospective Cohort Study. JMIR Medical Informatics, 10(2).
[18]. Uchida, K., Kouno, J., Yoshimura, S., Kinjo, N., Sakakibara, F., Araki, H., & Morimoto, T. (2022). Development of Machine Learning Models to Predict Probabilities and Types of Stroke at Prehospital Stage: the Japan Urgent Stroke Triage Score Using Machine Learning (JUST-ML). Translational Stroke Research, 13(3), 370–381.
[19]. Yu, J., Park, S., Kwon, S.-H., Ho, C. M. B., Pyo, C.-S., & Lee, H. (2020). AI-based Stroke Disease Prediction System Using Real-Time Electromyography Signals. Applied Sciences, 10, 6791.