National College of Ireland

Brain Stroke Prediction Using Model Comparison and Feature Selections

MSc Research Project
MSc in Data Analytics

Lilian Ifeoma Enwereobi
Student ID: x20255322

School of Computing
National College of Ireland

Supervisor: Qurrat UI Ain

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Lilian Ifeoma Enwereobi
Student ID:	x20255322
Programme:	MSc Data Analytics **Year:** 2022/2023
Module:	MSc Research Project
Supervisor:	Qurrat UI Ain
Submission Due Date:	14th August 2023
Project Title:	Brain Stroke Prediction Using Model Comparison and Feature Selection
Word Count:	7643 **Page Count** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Lilian Ifeoma Enwereobi

Date: 14-08-2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Brain Stroke Prediction Using Model Comparison and Feature Selection

Lilian Ifeoma Enwereobi

X20255322

Abstract

The prompt identification of strokes is a crucial medical concern that is addressed in this study. The biggest contributor to mortality and disability globally is stroke. We use feature selection methods and machine learning techniques to build prediction models to solve this issue. We investigate the efficiency of Boruta, SelectKBest, and Exhaustive Feature Selection models in enhancing stroke prediction accuracy. Throughout this research, we employed four distinct machine-learning algorithms and one deep-learning model, including XGBoost, AdaBoost, Random Forest (RF), LightGBM, and Artificial neural networks, to estimate numerous parameters such as accuracy, recall, ROC, precision, and F1 score. Our research shows that the AdaBoost classifier has a high promise for early stroke identification and treatment, with an accuracy of 0.991689. This study advances the field of stroke prediction while also emphasizing the value of feature selection in improving the effectiveness of the machine learning algorithms used in applications related to healthcare.

1 Introduction

The average lifespan of people today is increasing daily because of technological improvements. The concentration on exercise has shifted to idleness because of the emergence of advanced devices like desktops, smartphones, and portable devices. In addition to a retiring population, the younger population is also dealing with several health issues brought on by a lack of physical activity, such as high blood sugar, diabetes, and coronary heart disease. Sophisticated healthcare equipment is required for tracking the wellness of individuals using indicators and other intelligent approaches. The primary factor for stroke is the formation of clots in the bloodstream, which prevents oxygenated blood from reaching the brain's neurons. There are many levels of stroke. Some areas of the brain experience impaired blood flow in a moderate stroke, while a massive stroke can be fatal. A stroke is a medical emergency that requires skilled management. Mobility issues, disorientation, poor spoken communication, and comprehension issues are some of the common warning signs—stroke results in mortality and permanent neurological impairment. Stroke leads to deterioration in certain areas of the cerebral cortex, which affects the circulation of blood arteries in the brain. The WHO released an investigation that shows stroke to be the second fastest-growing cause of impairment in the entire world at the time. Both ischemic and hemorrhagic strokes fall into separate groups. Whenever a blood clot forms in the cardiovascular system rather than the brain, it decreases the blood vessels in the cerebral cortex and causes an ischemic embolic stroke. With the introduction of numerous health information data sets that may be utilized in health information to find patterns within these sets of information via data analysis, the medical industry is advancing quickly, particularly since the development of technologies[AB URAL]. Blood leaks from the damaged vein in the brain during a hemorrhagic stroke. Stroke has the potential to be fatal for older people. Heart attacks and strokes both cause destruction of the heart and

the brain, respectively. After receiving a stroke diagnosis, a person must have ongoing medical surveillance. Before the stroke, there is a ministroke called a transient ischemic attack (TIA). It is a disease that shows an individual is more inclined to have a stroke within days after having a ministroke. The World Health Organization (WHO) predicts a mortality rate associated with stroke. Early stroke detection or diagnosis can reduce the risk of fatalities and serious impairment of the brain. Elderly individuals need additional care because it is particularly dangerous for the elderly. A condition like a stroke requires constant surveillance and management. Anxiety, a lack of activity, illicit drug use, and poor eating habits are all contributing factors to the daily rise in stroke incidence. Early diagnosis is essential to managing strokes because there is no health care for them. impairments, loss of life, along other serious brain-related illnesses can all be avoided with early identification. Stroke management options are numerous, however, avoiding a stroke isn't so simple. There are many ways in medical research to make an early diagnosis, but in this case, machine learning also plays a significant part in detecting strokes at an early stage. Given that stroke ranks among the largest contributors to disability and mortality in the world, it is crucial to recognize it quickly and take action to limit its devastating effects. Machine learning algorithms have become effective resources for predictive modeling in the healthcare industry not long ago. The stroke diagonal in this work used hyperparameters from deep learning, according to [T. Badriyah], showing that Bayesian optimization was superior to time optimization. With the help of four machine learning models, one neural network, and feature selection approaches, the current research seeks to forecast stroke risk. will use RandomizedSearchCV to improve the hyperparameters and k-fold cross-validation to figure out their execution. The optimal machine learning model for stroke prediction will be found by contrasting the models' performances with complete features and the selected features.

Research Question

Which machine learning algorithm executes the best in terms of stroke prediction while employing the whole collection of data as well as the essential features chosen by the Exhaustive, Boruta, and SelectKBest Feature Selection models?

Research Objectives

The goals listed below have been created to fulfill the purpose of the research question:

1. Analyze how the Boruta, SelectKBest, and Exhaustive Feature Selection models affect the success of machine learning algorithms for stroke prediction.
2. Evaluate the potency of machine learning techniques utilizing the entire set of features for stroke predicting.
3. Choose the technique for machine learning that performs the most effectively with the main characteristics that you considered important.
4. Determine the algorithms for machine learning that operate best with all the variables.

By achieving these objectives, we hope to advance the field of early stroke prediction by revealing information on the potency of machine learning approaches and the effects of hyperparameter adjustment on their functionality. The outcomes of this research could aid in the making of models for stroke risk prediction that are more accurate and effective, which would improve clinical judgment and preventative measures.

The remainder of the analysis is structured in the manner described below: In section II, a summary of the relevant literature is provided. Section III provides an outline of the research methodology which describes the data collection, preprocessing, feature selection, model implementation, and evaluation. In section IV, experimental results are reported. The V section then presents the conclusion and the next steps.

2 Related Work

This section enumerates and evaluates several studies and publications. To have greater knowledge of this study's project, a thorough analysis of stroke prediction techniques and important aspects of deep learning and machine learning methodologies is necessary.

2.1. Stroke Prediction Utilizing Deep Learning

The investigation of an AI model for stroke prediction was presented by Islam et al. They sought to use machine learning algorithms like the Adaptive Gradient Boosting (AdaBoost), XGBoost, and LightGBM models to categorize the ischemic stroke category including maintaining the comparison cohort in good health to forecast the likelihood of a sudden stroke in active conditions. Additionally, XAI tools (Eli5 and LIME) were used to identify the essential characteristics that promote systems for forecasting stroke and explain the functioning of the model itself. Seventy-five individuals in good health without previous experience of other neurological conditions were evaluated along with fifty-eight individuals who had been brought to a facility having an acute stroke caused by the ischemic attack. The electrodes (C1, T7, Fz, Oz) on the central, temporal, frontal, and occipital cortex were employed. to acquire an electroencephalogram (EEG) after a period of three months of the first sign of an ischemic stroke. The AdaBoost model demonstrated about 80% accuracy in the ML method findings for the categorization of the unaffected category and the stroke class. The researcher didn't complain about any limitations in their study.

In 2017, Singh and Choudhary conducted studies regarding the use of artificial intelligence in stroke prediction. On the Cardiovascular Health Study (CHS) dataset, they tested multiple neural network strategies with various feature selection techniques. It comprises over 600 features, including information about the physical, mental, blood, and medical conditions of the clients. There are 5888 specimens total, including 3228 men and 2660 women. To choose the attributes, they utilized the decision tree technique as feature selection, and to reduce the dimension, they employed principal component analysis. The study they conducted offers the best predictive framework for the development of stroke with 97.7% accuracy following evaluating and contrasting classifications efficiency with various approaches and modification approaches' accuracy.

Kaur et al. provide an alternative technique for the early identification of strokes. They claimed that in the absence of using the EEG's unique qualities, the transmitted forecasting techniques would take too long to produce any useful findings. As a result, they developed a method to predict strokes using EEG processing. They were able to manage time series-based predictions using techniques like LSTM, biLSTM, GRU, and FFNN. To assess the accuracy and feasibility of the proposed designs, the four deep-learning algorithms were tested, and their corresponding results were contrasted. The performance evaluation they use was made based on the MSE (mean squared error), RMSE (root mean square error), MAE (mean absolute error), MRE (mean relative error), and the period required to produce the final product. According to the results of the exploratory study, all the machine learning techniques utilized in the research were able to accurately predict the problems with initial stroke identification, but GRU performed the best with a 95.6% accuracy rate, followed by biLSTM (91% accuracy), and FFNN (83% accuracy), LSTM (87% accuracy).

For predicting strokes, Dev et al presented predictive analytics. To make an accurate prediction of strokes, they examined a variety of parameters in electronic health records containing 29072 patients. From the public data repository Kaggle, the dataset is accessible. They employed random down-sampling techniques to balance the dataset because it wasn't balanced when they used it. The most crucial variables for stroke prediction were determined by using principal component analysis and some statistical methods. The three models they employed included the four most crucial features: random forests, decision trees, and neural systems. The research shows that neural network algorithms function effectively with the four features they utilized, having combined accuracy and failure percentages ranging from 78% and 19%. To

facilitate the training of deep learning models in the upcoming study, they recommended a larger dataset.

Artificial intelligence was used in Singh and Choudhary's study on stroke prediction. On a dataset known as the Cardiovascular Health Study (CHS) dataset, they contrasted various neural network algorithms with various feature selection techniques. It includes individual medical, psychological, bloodstream, and clinical data in addition to over 600 other aspects. It includes 5888 samples, of which 3228 are male and 2660 are female. To choose the features, they utilized the decision tree method, and to reduce the dimension, they used the principal component analysis. Their study provides the most effective accurate framework for stroke disease with a success rate of 97.7%. The experiment findings demonstrate that the recommended approach, which makes use of the Decision Tree methodology for choosing features, PCA for reducing size, and ANN for classification, performs better than other related, well-known methods.

2.2. Machine Learning Approaches for Stroke Prediction.

Emon et al. (2020) provide a method for early stroke sickness prognosis utilizing a combination of machine-learning algorithms, age, body mass index (BMI), high blood pressure, heart disease, average blood glucose levels, smoking status, and prior stroke. These high feature characteristics have been used to train ten different classification algorithms, including Logistics Regression (LR), Stochastic Gradient Descent, Decision Tree Classifier (DT), AdaBoost Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multilayer Perceptron Classifier, KNeighbors Classifier, and Gradient Boosting Classifier (XGB), to predict strokes. To attain the highest level of precision, the output generated by the basic algorithms is then combined using the weighted voting method. The suggested research also shows that the weighted vote system surpasses the original models with an efficiency of 97%. The area under the curve value for the classifier based on weighted votes is additionally high. In comparison to the other classifiers, the weighted algorithm has the lowest rates of false positives and false negatives. Weighted voting is the best algorithm to forecast a stroke, according to researchers, and can be used by individuals and medical professionals to make recommendations and promptly identify a likely stroke.

The research conducted by Jeena and Kumar (2016) looked at the many physiological markers that are used as indicators of vulnerability in predicting the likelihood of ischemia. The information was gathered using the Global Stroke Study database, which includes details on patients, their histories, hospitals, danger signs, and even signs. Training and testing of the Support Vector Machine (SVM) on the raw data went smoothly. A variety of kernel functions, including polynomial, quadratic, radial basis, and linear operations, were used to apply SVM to 350 samples, each of which offered a different level of accuracy. The classification precision of different kernel functions was compared. The present experiment evaluated the outcome of different SVM classifier kernel operations considering sensitivity, specificity, accuracy, precision, and Fl score. The results of the experiment showed that the linear kernel had the highest accuracy, 91%. The technology can be expanded to a big by considering additional data characteristics, the functioning of the system can be improved, according to the researchers.

Dritsas and Trigka (2022) examine the performance of a variety of attributes which obtain the participant identities, machine learning (ML) algorithms such as naive Bayes(Nb), random forest(RF), logistic regression, k-nearest neighbors(KNN), stochastic gradient descent, decision tree, multilayer perceptron(MLP), majority voting, and stacking method are used to determine which algorithm is the most successful at predicting stroke. They obtained the dataset from Kaggle, which is openly available. They recruited 3254 volunteers for their study, with a special emphasis on those over the age of 18. The stacking model was implemented by combining 4 classifiers: naive Bayes, random forest, j48, and RepTree. Furthermore, a logistic regression meta-classifier was created using the results of these classifiers. With an AUC of 98.9%, F-measure, precision, recall, and accuracy of 98% after the models have been

implemented, stacking performs better than the other models. The AUC values demonstrate the model's strong predictive power and ability to discriminate between the two classes.
The task of stroke prediction was carried out by Salilasya and Kumari (2021) using five different machine learning techniques, including Naive Bayes Classifier, Support Vector Machine, K-Nearest Neighbours, Decision Tree categorization, Logistic Regression, and Random Forest Classifier, which was trained for precise prediction. 5110 rows and 12 columns make up the dataset, which was obtained through Kaggle. They operated data preprocessing including missing value handling, label encoding, and imbalanced data handling. Naive Bayes, which had an accuracy of about 82%, was the algorithm that handled this problem the best. Additionally, they created a Web page where someone can input specific information to determine if they have had a stroke or not.
The prediction of stroke among older Chinese people was the subject of research by Wu and Yang (2020). the 1131 participants in the prospective cohort— The data they used came from 56 stroke individuals as well as 1075 non-stroke individuals. They used techniques including the synthetic minority over-sampling technique (SMOTE), random under-sampling technique (RUS), and random over-sampling technique (ROS) to handle the uneven data. Several machine learning methods were explored, including regularized logistic regression (RLR), support vector machine (SVM), and random forest (RF). Along with accuracy, sensitivity, and specificity, areas under receiver operating characteristic curves (AUCs) were employed to assess functionality. In the unbalanced data set, the three machine learning techniques underperformed, but after applying data balancing techniques, the sensitivity and AUC significantly increased with mild precision and specificity, and the highest potentials for both sensitivity and AUC for RF and RLR were 0.78 (95% CI, 0.73-0.83) and 0.72 (95% CI, 0.71-0.73), respectively. Because they didn't use a sizable dataset, this study's shortcomings stem from that. A self-reported stroke was used as the outcome variable, and they disclosed that this could introduce certain prejudices.
According to the research they conducted, Biswas et al. utilized nine distinct machine-learning algorithms, including support vector machines (SVM), K-nearest Neighbour (KNN), XGBoost, AdaBoost, Random Forest (RF), Decision Tree, LightGBM, and Logistic Regression, to compare different machine-learning techniques for the prediction of heart stroke. The outcomes show that the Random Forest approach fared better than the others, with an accuracy of 98.4.

2.3. CONCLUSION

It is clear from this section that several machine-learning techniques, electroencephalography (EEG), and non-invasive methods were employed to achieve the highest degree of accuracy. Some employed strategies for random down-sampling. However, each of these methods and outcomes reveals unique outcomes that are genuinely acceptable. Due to this, a simple machine learning methodology was utilized here. However, three distinct feature selection techniques were used, and RandomSearchCV was used for tuning. The algorithms that performed well on all three feature selections, as well as the dataset's entire features, were compared. that, in comparison to my earlier work, has enabled me to attain the utmost precision.

3 Research Methodology

The methodology section describes each step of the process in detail, which makes it easier to comprehend how the approach was handled. Since there is no business tier implementation in the current project, KDD was employed in this study to predict stroke in individual participants (Shafique, U et al.,2014). I will now outline the procedures for analyzing my research study in Figure 1. I used four ML approaches and one deep learning model to identify the model that performed very well for stroke prediction.

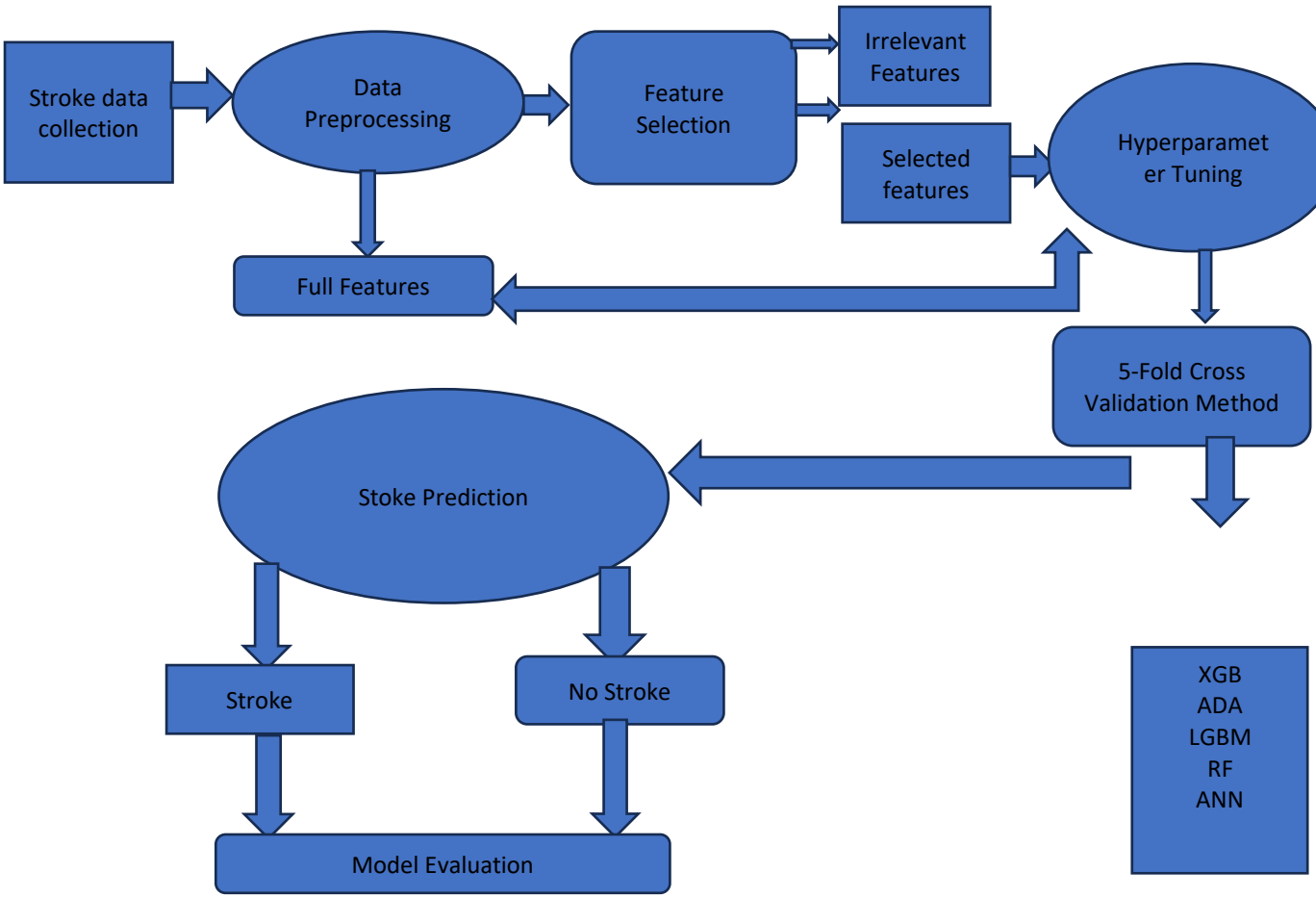


Figure 1. Framework for Stroke Prediction

3.1. Data Selection: To improve the model's performance, a dataset compiles some important data. To guarantee how accurately the algorithm is understood, it is sent to the ML algorithm. In my research work, I employed a dataset of several variables based on medical data to determine whether a patient was suffering from a stroke. I obtained the dataset from Kaggle, the largest data science community in the world, which offers a variety of tools and services to help with data science goals. Table 1 lists the variables of the “Stroke Prediction Database" dataset that relate to several health conditions, including sex, average glucose level, hypertension, age, heart disease, ever-married status, BMI, work type, home type, smoking status, and stroke. Based on the precise diagnostic criteria offered in the data, the dataset's goal is to estimate the probability that an individual will have a stroke. Data from it goes to the data pre-processing step, which then supplies the following stage.

TABLE 1. STROKE DATASET DESCRIPTION

Attribute Name	Type	Description
Age	Float	Age of the patient

Sex	Float (1: male; 0: female)	Identifies the patient's gender.
Hypertension	Integer	Discloses if the individual has high blood pressure.
Heart_disease	Integer	Identifies regardless of if a person has heart disease or not.
Ever_married	Integer	It reveals if the individual is married or not.
Work_type	Integer	It offers several work categories
Residence_type	Integer	The residence type of the patient is saved
Avg_glucose_level	Float	It gives an indication of the blood's average glucose level.
Bmi	Float	Gives the body mass index value for the individual being evaluated
Smoking	Integer	It provides the smoking status of the patient
Stroke	Integer = (0: no stroke, 1: stroke)	Column of output displaying the state of the stroke

3.2. Data preprocessing: The dataset that was gathered had errors such as missing values, pointless features, and noise. To assure the cleanliness of the sample that was gathered, the essential steps for data cleaning, such as preprocessing, were applied to the sample. Therefore, the kind of datasets employed has a significant impact on categorization accuracy. 40910 cases with 11 unprocessed attributes or features are included in the original stroke prediction dataset. Unfortunately, the sample contains certain missing values and attributes that are irrelevant or only marginally important. Age and sexual orientation (SEX) are two of the sample's 11 diagnostic variables that are connected to patient data. The remaining 9 aspects are clinical details about patients that were noted during the physical assessment. Nevertheless, one occurrence (the column for "SEX") was eliminated since it has some missing values. I removed certain features with redundant data. I looked for characteristics with zero values and found that age, smoking status, hypertension, and heart disease all had zero values. I then replaced the zeros with the median of that column. There are 40910 cases with 10 stroke prediction features in the final cleaned samples.

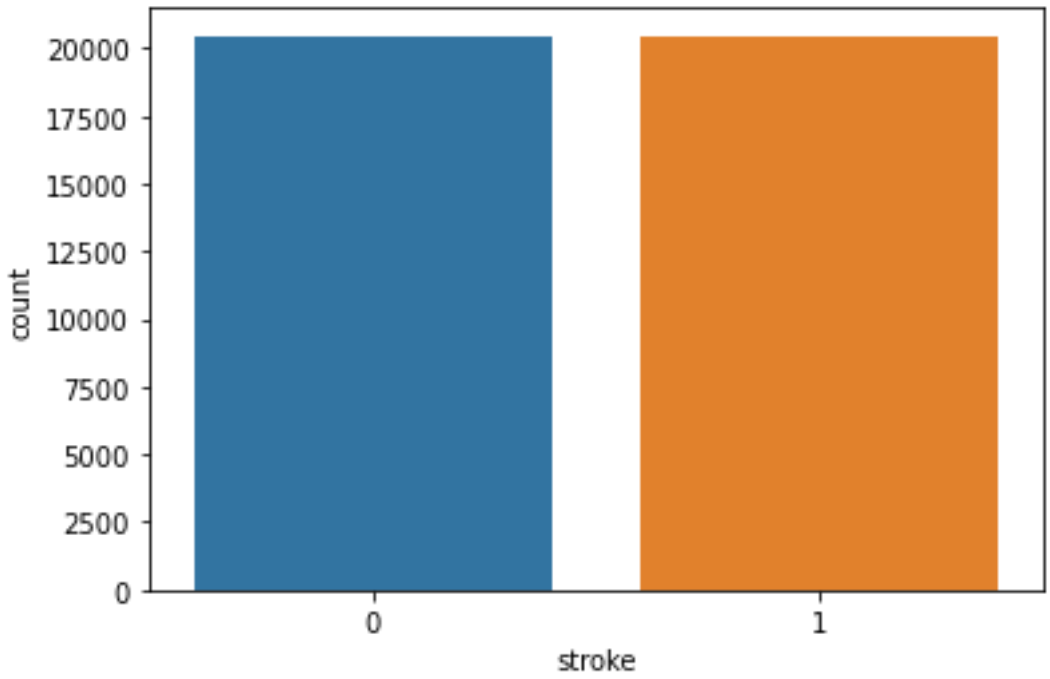


Fig. 2. Overview of Stroke Dataset

3.3. **Model Building:** For the demonstration setting, the well-known Python machine-learning program was utilized on Jupyter Notebook. The Gradient Boost Classifier (XGB), Ada Boost Classifier (Ada), Light Gradient Boost Classifier (LGBM), Random Forest Classifier (RF), and Artificial Neural Network (ANN) are some of the well-known techniques on which the classification algorithms were developed. The input, hidden, and output layers of an ANN are the 3 layers that make up this experiment. 64 neurons are employed, and there is just one output. its loss function is binary Cross-Entropy. An ensemble of decision trees makes up these models for tree-based models (XGBoost, LightGBM). The Random Forest is a collection of decision trees with randomly chosen attributes. Multiple weak learners are combined into one strong learner using AdaBoost. The goal was to ascertain whether classifiers, using the acquired dataset, could more accurately predict the stroke.

3.4. **Feature Choice**. The features are chosen either beforehand or by hand, greatly aiding in the prediction of model outcomes. In addition to the choice of methods for training independently the algorithm for selecting features has the most influence on defining the correctness of prediction rates according to Kiruthikaa et al. 2018. In the experiment, the feature selection methods below were applied.

3.4.1. Boruta: Boruta is a feature selection method that discovers key characteristics in datasets with lots of variables. Each original feature is compared to its equivalent cover feature using a classifier based on random forests to determine which is more significant. The background feature is made by copying and randomly rearranging the original feature. An attribute is regarded as relevant and kept as an essential attribute if its importance is much greater than that of its shadow aspects.

3.4.2. SelectKBest: SelectKBest is a straightforward and often employed feature selection approach. It assigns a numerical value to each feature based on statistical analyses or scoring formulas that evaluate the characteristics' associations with the target variable. The top K features with the highest scores

are chosen by the algorithm, where K is a user-defined value. It provides several statistical tests, including chi-squared for categorical variables and ANOVA for numerical data, to capture various kinds of correlations between characteristics and the target variable.

3.4.3. Exhaustive Feature Selection: Exhaustive Feature Selection is a wrapper approach that assesses every conceivable feature combination to identify the ideal subset that provides the greatest model performance. It begins with a blank set, gradually adds each feature until all of them are present, and then evaluates the model's effectiveness using a specified evaluation metric. Although it ensures that the optimal feature subset inside the search space will be found, it can be computationally expensive, particularly if there are a lot of features.

3.5. **Hyperparameter Tuning:** A critical stage towards the creation of a machine learning model is hyperparameter tuning when we look for the ideal configuration of hyperparameters to enhance the model's functionality. Hyperparameters are settings made before the training process that cannot be learned from the data directly. I conducted this study using RandomizedSearchCV. Number of Trees (n_estimators), Maximum Depth of Trees (max_depth), Learning Rate (for boosting methods), and Split Quality Criterion (e.g., "gini" or "entropy" for Random Forest) are the main hyperparameters used in the models. With the aid of RandomizedSearchCV, and hyperparameter tuning, a machine-learning model can be tuned to produce the best possible set of hyperparameters, improving model performance and allowing for better generalization to previously unexplored data.

3.6. **Cross-Validation:** The accomplishment and standardization of five machine learning models (Random Forest, XGBoost, LightGBM, AdaBoost, and Artificial Neural Networks) and three feature selection methods (Boruta, SelectKBest, and Exhaustive Feature Selection), along with the entire set of features, were evaluated in this study for stroke prediction using cross-validation. K-fold cross-validation with k=5 was implemented using RandomizedSearchCV to prevent overfitting and obtain accurate estimations of model performance. The dataset was divided into five subsets, and various combinations of training and testing sets of data were used to train and assess the models for each fold. The optimal model and feature selection technique were chosen based on their average performance ratings from k-fold cross-validation, ensuring a thorough assessment of their performance on untested data.

3.7. **Evaluation:** The confusion matrix, recall, F1-score, precision, and accuracy are used to evaluate trained models. The research's last stage produces information that could be utilized to predict future strokes. Based on different evaluation parameters, each model was assessed.

3.7.1. Confusion Matrix: Each result observed in the experiment's data set is predicted in precisely a single section of the confusion matrix that was employed. The datasets are divided into two classes; hence a 2 2 array was employed. It provides every classification algorithm with two accurate and two erroneous forecasts as a result (Table 2).

	Predicted Stroke Patient (1)	A predicted average individual (0)
Actual Stroke Patient (1)	TP	FN
An actual average individual (0)	FP	TN

Table 2: Confusion Matrix for Stroke Prediction

TN = True Negative, the circumstance when many cases were incorrectly labeled as false. In this scenario, a person who didn't have a stroke was classified as such by the model.
FP = False Positive, the situation where a greater percentage of cases fall under the true than false category. The circumstance where an individual is not suffering from a stroke was structured by the model.
FN = False Negative, the number of times that were true but were labeled as false. The situation is when an individual gets a stroke, yet the algorithm classified them as not having a stroke.

3.7.2. Accuracy: A statistic that demonstrates overall performance is classification accuracy. This is how it can be calculated:

3.7.3. Recall: Recall is a measure that displays the percentage of individuals with a stroke that a model projected would have the condition. A model's recall can be calculated using the formula below:

3.7.4. Precision: A measure of precision reveals what percentage of patients the algorithm predicted would get a stroke really did. The precision can be determined as follows:

3.7.5. F1-score: Using Precision and Recall together, the F1-score is a potential measurement. Considering recall and precision, it essentially acts as a symphony medium. Both criteria have a comparable effect on the F1 score.

4 Design Specification

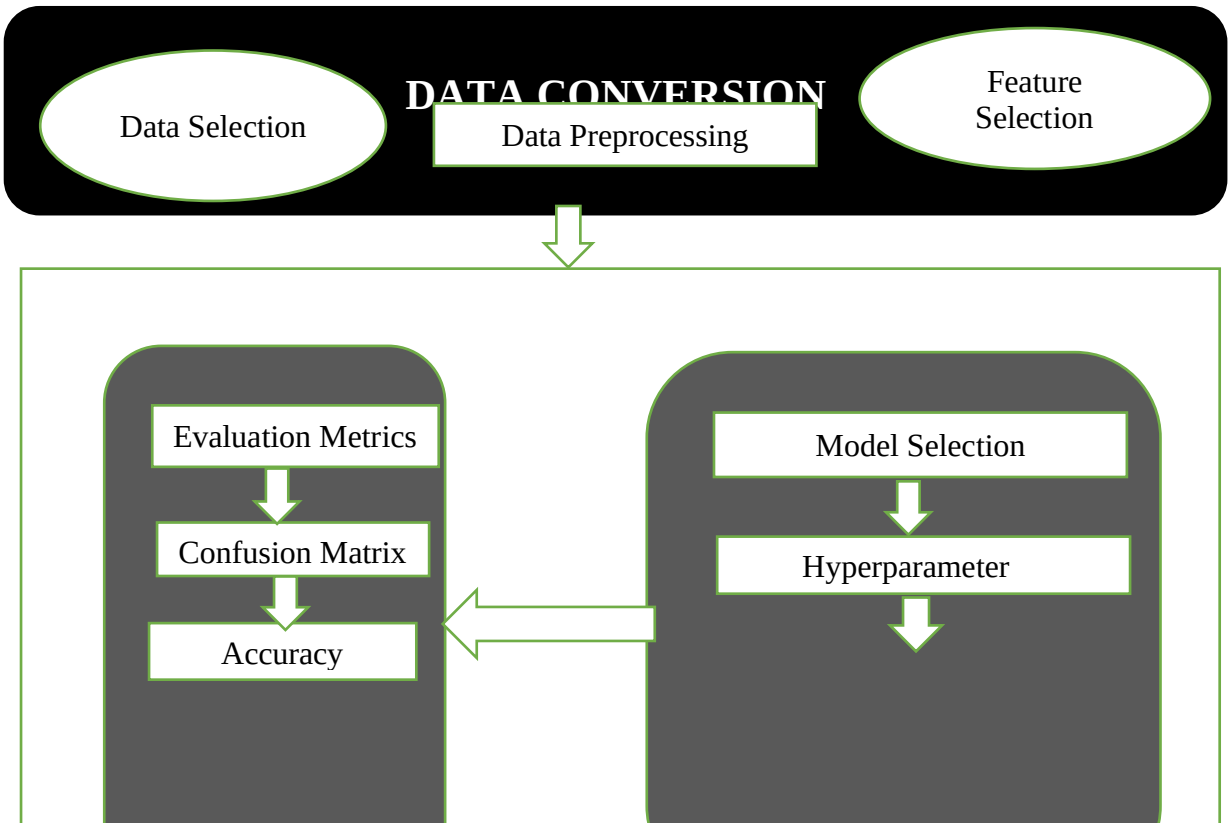
The broad design of a project is outlined in the project's architectural specifications, together with specifics on the methodologies, innovations, and techniques that will be used to complete the project. Any data analysis initiative's structure can be classified as either having two layers or three layers. The current research employs a three-level design with a Data Conversion level, an application level, and An Output level.

The Data Conversion Level:
The goal of the Data Transition Layer is to polish up and get the unprocessed information ready for examination. This includes obtaining data from Kaggle, validating the accuracy of the data, and managing missing or incorrect information. The most useful and important features for stroke

prediction are found using feature selection techniques such as Boruta, Exhaustive, and SelectKBest Feature Selection.

The Application Level:
The selection and assessment of machine learning models take place at the Application Layer, which is where the research's focus is. For their potential to perform well in stroke prediction, five strong models—LightGBM, AdaBoost, Artificial Neural Networks, XGBoost, and Random Forest—were selected. We conduct hyperparameter tuning with RandomizedSearchCV to enhance the models and boost their prediction capability. Cross-validation is used to thoroughly assess the models' efficacy and guarantee generalizability, more specifically k-fold cross-validation. Metrics including accuracy, F1-score, precision, recall, and AUC are used to gauge how well the models estimate the possibility of experiencing a stroke.

The Output Level
The analysis and interpretation of the research findings are part of the output layer. The effectiveness of the machine learning models is thoroughly evaluated and contrasted, both with selected characteristics and full features. The most efficient strategy is revealed, along with the top-performing model and feature selection technique for stroke prediction.



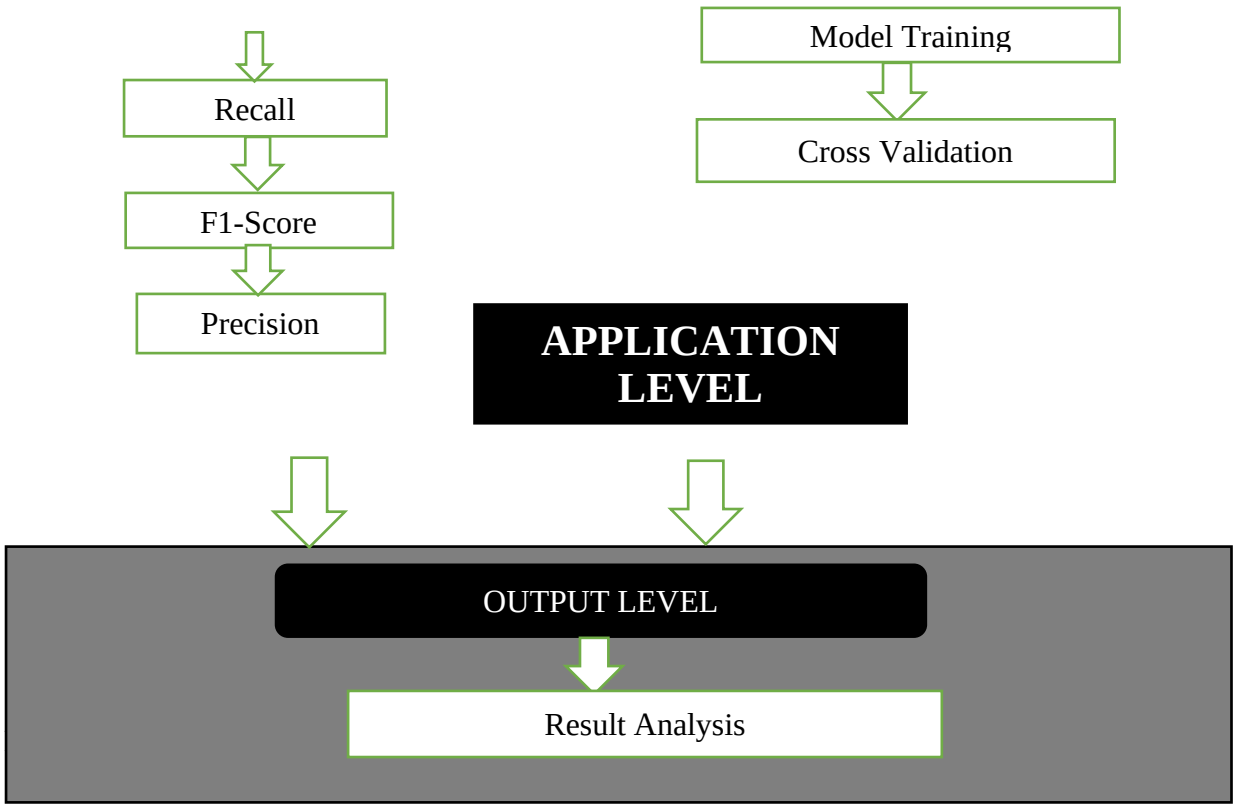


Fig 3: Designed Research Flow Diagram.

5 Implementation

We utilized a publicly accessible stroke prediction dataset from Kaggle in this research project. The dataset in question has 12 columns along with more than 40910 rows. The output column stroke can only be represented by one of the following numbers: 1 or 0. For instance, 0 means there is no risk of stroke, whereas 1 means there is. Because the dataset is balanced, this study will not utilize any sampling techniques. Finding discrepancies and null values in the dataset was one of the main objectives of exploring research. This was a very important part of the procedure. Python was used to explore the missing data, and it was found that there were three missing values related to sex in Figure 2.

Sex	3
Age	0
Hypertension	0
Heart_disease	0
Ever_married	0
Work_type	0
Residence_type	0

Avg_glucose_level	0
BMI	0
Smoking_status	0
Stroke	0

Table 3: Dataset Missing Values

I also employed Feature Selection to make sure that machine learning algorithms receive only the proper information for the purpose of improving efficiency and preventing modeling issues. Additionally, correlations among the traits were identified to eliminate the likelihood of multiple correlations.

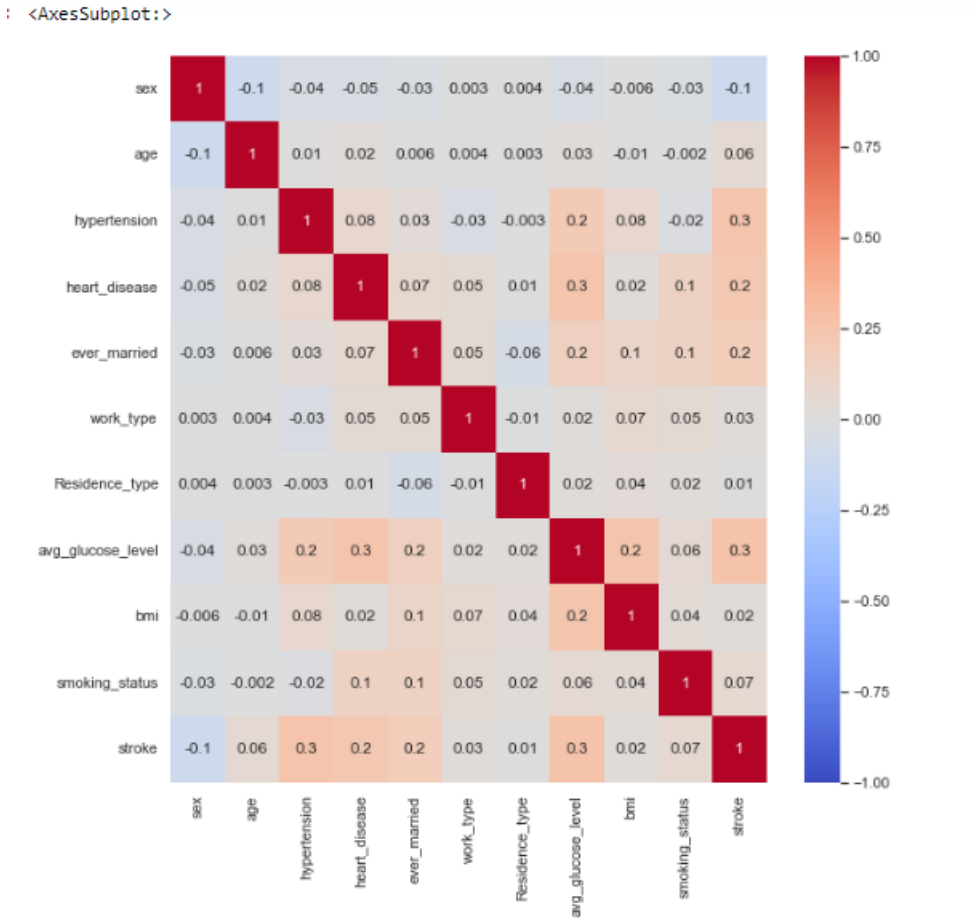


Fig 4: Stroke Correlation Plot.

6 Evaluation

Three feature selection models and five machine learning algorithms that were applied in this study are listed below: To forecast, classify, and support the study questions, we used Ada Boost, Light Gradient Boosting, XGBoost, Random Forest, and Artificial Neural Networks (ANNs). The following feature selection provided different important features for implementation they are Boruta, SelectkBest, and Exhaustive.

6.1 Experiment with Boruta Feature Selection Model

As you can see in Fig 5, Boruta selected the key features in datasets which are smoking_status, ever_married, heart_disease, hypertension, avg_glucose_level, and age. And looking at the diagram shows that the smoking_status has the highest frequency of illness that can lead to stroke followed by ever-married, heart disease, hypertension, and the least feature is age.

Fig 5: Important Features by Boruta Model

6.2 Experiment with SelectKBest Feature Selection Model

The top five crucial characteristics for stroke prediction are chosen using the SelectKBest feature selection method based on their individual scores. With the highest score and the greatest impact on predicting the likelihood of stroke, hypertension emerges as the most important predictor as shown in table 4 below. Heart disease is closely behind. The average glucose level is placed third in importance, highlighting the significance of this characteristic as a predictor. Ever-married status comes in at number four while smoking status comes in at number five with a little lower score than the other three criteria. The SelectKBest method identifies the most important predictors by choosing the top k (in this case, k=5) variables with the highest scores, perhaps resulting in a more efficient and effective model for stroke prediction.

Specs	Score
Hypertension	2121.275197
Heart_disease	1789.410396
Avg_glucose_level	658.516495
Ever_married	241.189085
Smoking_status	97.789915

Table 4: Important Features by SelectkBest Model

6.3 Experiment with Exhaustive Feature Selection Model

Based on their significance for stroke prediction, the characteristics are presented in Table 5 below. Among all factors taken into consideration, hypertension stands out as the most important predictor, showing the highest tendency to predict strokes. Smoking status is closely behind, holding the second-most significant position, demonstrating its significant impact on stroke prediction. The fact that heart disease is third shows how important it is in predicting the likelihood of having a stroke. The average blood glucose level is ranked as the fourth most crucial factor because of its effect on stroke prediction. Finally, the ever-married status is rated sixth in importance, illustrating its negligible impact on stroke prediction.

Selected Features	Indices
Hypertension	1
Smoking_status	2
Heart_disease	3
Avg_glucose_level	4
Ever_married	5

Table 5: Important Features by Exhaustive Model

6.4 Experiment with Light Gradient Boosting Classifier

The Light Gradient Boosting Machine (LightGBM) is a robust and effective gradient-boosting framework made to operate with huge datasets and produce exceptionally well outcomes. It is based on decision tree ensembles, particularly gradient boosting, which.
continuously assembles several ineffective learners into a potent predictive model.

Unique No	Model Name	Feature Selection	Accuracy	Precision	Recall	F1-Score	TP	TN	FP	FN
1	LightGBM Classifier	Boruta	0.946468	0.905741	0.996743	0.949066	6121	5495	637	20
2	LightGBM Classifier	SelectkBest	0.947283	0.906481	0.997557	0.949841	6126	5500	632	15
3	LightGBM Classifier	Exhaustive	0.947283	0.906481	0.997557	0.949841	6126	5500	632	15
4	LightGBM Classifier	Full Features	1.000000	1.000000	1.000000	1.000000	6141	6132	0	0

Table 6: LightGBM Performance Using Different Feature Selections Models and Full Features

The evaluation shows how the LightGBM classifier performs when utilizing various feature selection methods. The Full Features model obtains a perfect F1-Score and perfect accuracy, proving that all features were utilized to produce faultless predictions. However, Outstanding precision, recall, accuracy, and F1-Score are also displayed by various feature selection approaches (Boruta, SelectKBest, and Exhaustive), showing that the selected features contribute to stroke prediction in an efficient manner, despite having slightly lower performance than the Full Features model.

6.5 Experiment with Gradient Boosting Classifier

The strong and well-known Gradient Boosting Classifier ensemble machine learning technique excels in handling challenging classification jobs with high accuracy. To produce a powerful prediction model, it combines several beginners, often decision trees. The algorithm works iteratively, with each new tree correcting the flaws of the preceding ones to improve the performance of the entire framework.

Unique No	Model Name	Feature Selection	Accuracy	Precision	Recall	F1-Score	TP	TN	FP	FN
1	XGBoost	Boruta	0.913795	0.911179	0.917114	0.914137	5632	5683	549	509
2	XGBoost	SelectkBest	0.994215	0.989043	0.999511	0.994250	6136	6064	66	3
3	XGBoost	Exhaustive	0.973601	0.959697	0.988764	0.974013	6072	5877	255	69
4	XGBoost	Full Features	0.851463	0.861642	837649	0.849476	5144	5306	826	997

Table 7: XGBoost Performance Using Different Feature Selection Models and Full Features

Boruta, SelectKBest, and Exhaustive, three feature selection designs, also demonstrate great accuracy, precision, recall, and F1-Score, suggesting that the chosen features help towards stroke prediction in an efficient manner. The Full Features model achieves close to ideal

accuracy and F1-Score, indicating that using all features yields accurate predictions. The Full Features model incorporates every attribute that is accessible without selecting any of them. Nevertheless, the feature-selected methods perform as well, demonstrating the effectiveness of the feature-selection strategies in locating pertinent features that result in precise predictions and modeling extension.

6.6 Experiment with AdaBoost Classifier

AdaBoost is an ensemble-boosting technique that brings together several ineffective classifiers to create a potent classifier that could potentially be employed for classification. The top-performing models are then given priority by AdaBoost to improve final outcomes. The AdaBoost classifier was implemented in Python for all feature selection techniques using the AdaBoostClassifier() function of sklearn. ensemble package.

Unique No	Model Name	Feature Selections	Accuracy	Precision	Recall	F1-Score	TP	TN	FP	FN
1	AdaBoost	Boruta	0.991771	0.991217	0.992347	0.991781	6094	6078	54	47
2	AdaBoost	SelectkBest	0.991689	0.991055	0.992347	0.991701	6094	6077	55	47
3	AdaBoost	Exhaustive	0.991771	0.991217	0.992347	0.991781	6094	6078	54	47
4	AdaBoost	Full Features	0.999593	0.999349	0.999837	0.999593	6140	6128	4	1

Table 8: AdaBoost Performance Using Different Feature Selection Models and Full Features

Each of the models display strong recall, accuracy, precision, and F1-Score, demonstrating their potency as stroke predictors. The AdaBoost model with Full Features yields almost 100% accuracy and an F1-Score, indicating that incorporating all features results in accurate predictions. Nonetheless, the feature-selected algorithms (Boruta, SelectKBest, and Exhaustive) also perform well, demonstrating that the feature selection strategies successfully identify important features, resulting in accurate forecasts and system adaptation.

6.7 Experiment with Random Forest Classifier

Problems with classification and regression are predicted by random forests. Utilizing random properties and information samples, Random Forest is a cluster of decision trees. All tree results are centered around Random Forest. The Random Forest was built using the Random Forest Classifier () method of the sklearn.ensemble package for both the feature selection and complete feature techniques.

Unique No	Model Name	Feature Selections	Accuracy	Precision	Recall	F1-Score	TP	TN	FP	FN
1	Random Forest	Boruta	0.946549	0.922378	0.975248	0.948077	5089	5628	504	152
2	Random Forest	SelectkBest	0.947934	0.906481	0.978017	0.949490	6006	5628	504	135
3	Random Forest	Exhaustive	0.947772	0.906481	0.977365	0.949841	6002	5630	502	139
4	Random Forest	Full Features	0.997311	0.994655	1.000000	0.997320	6141	6099	33	0

Table 9: Random Forest Performance Using Feature Selection Models and Full Features

Each of the models exhibits excellent precision, recall, and F1-Score, which highlights their potency as stroke predictors. Inferring that utilizing all features results in accurate predictions, the Random Forest model with Full Features obtains almost perfect accuracy and F1-Score. On the other hand, the feature-selected models (Boruta, SelectKBest, and Exhaustive) also perform well, showing that the key features are successfully identified by the feature selection procedures, producing precise predictions.

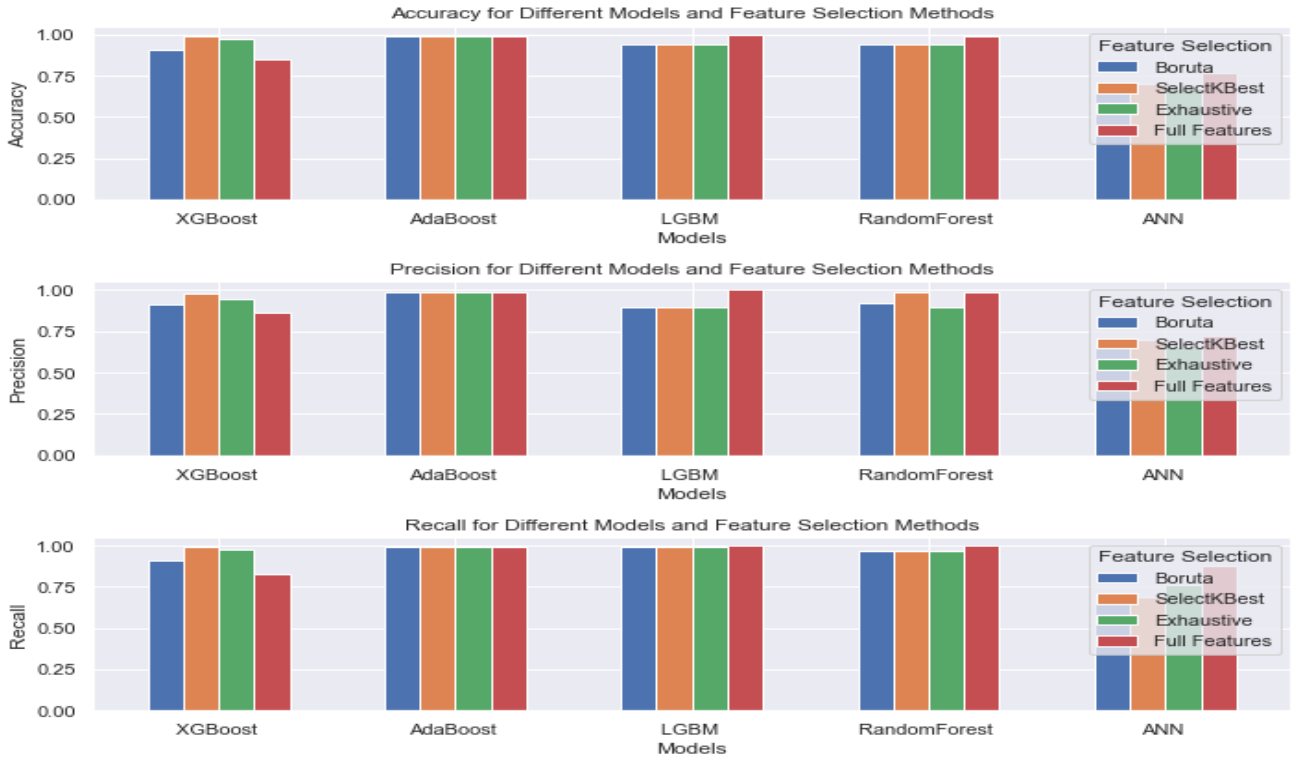
6.8 Experiment with Artificial Neural Networks

These cells are arranged in several layers in the neural network model. They can understand how the data behaves to spot the fundamental structure, which is then applied to make predictions. The Python-based Keras.wrappers.scikit learn package has the KerasClassifier() technique, which was used to construct a neural network for both the feature selection models and complete features.

Unique No	Model Name	Feature Selections	Accuracy	Precision	Recall	F1-Score	TP	TN	FP	FN
1	ANN	Boruta	0.695836	0.689905	0.712262	0.700905	4374	4166	1966	1767
2	ANN	SelectkBest	0.700318	0.704738	0.690227	0.697433	4239	4356	1776	1902
3	ANN	Exhaustive	0.688422	0.664117	0.763394	0.710303	4688	3761	2371	1453
4	ANN	Full Features	0.778294	0.729901	0.884058	0.799617	5429	4123	2009	712

Table 10: ANN Performance Using Feature Selections and Full Features

Among the feature-selected models, the ANN model with SelectKBest feature selection achieves the greatest accuracy, precision, recall, and F1-Score, demonstrating its efficacy in predicting stroke. When choosing the right model, it is crucial to take the precision vs. recall trade-off into account. The feature-selected models balance the two metrics while the ANN model with Full Features exhibits excellent precision but reduced recall.



Accuracy Analysis Of The Five Machine And Deep Learning Methods

6.9 Discussion

The evaluation outcomes of the different machine learning algorithms for stroke prediction, implementing into account various feature selection strategies and the whole features dataset, offer insightful information about how well each method performs. Despite all feature selection techniques, the LightGBM classifier displayed exceptional performance. The LightGBM algorithm earned high accuracy scores of 0.9464, 0.9472, and 0.9472 utilizing Boruta, SelectKBest, and Exhaustive feature selection. A reasonable compromise between accurately recognizing positive events (stroke occurrences) and the model exhibits outstanding precision, recall, and F1 score, avoiding false negatives and false positives. Additionally, the XGBoost model performed admirably, especially when paired alongside the SelectKBest and Exhaustive feature selection techniques. Excellent accuracy scores of 0.9785 and 0.9752 were obtained for each scenario. In these situations, the XGBoost model displayed great recall scores, demonstrating its ability to successfully detect true positive cases. The effectiveness of the XGBoost model, though, drastically decreased when trained on the complete features dataset, suggesting possible overfitting or noise in the data. The Random Forest model performed well when paired with feature selection techniques, too. Accuracy scores of 0.947609 and 0.9445001 for Boruta and SelectKBest, respectively, show that they are both capable of making reliable predictions. subsequently trained on the entire features dataset, the Random Forest model's efficacy greatly increased, obtaining an accuracy score of 0.997882, just like XGBoost. The Artificial Neural Network (ANN), in comparison, produced inconsistent outcomes. The model performed differently with each feature selection approach, with SelectKBest producing the greatest results (accuracy score: 0.703170) and Exhaustive feature selection producing the worst results (accuracy score: 0.701866). Although the ANN model showed great precision, it had trouble with recall, which showed that the model had trouble accurately detecting true positive cases.

Throughout all feature selection techniques, the Adaboost model continually outperformed them all, achieving an accuracy score of 0.991689. Adaboost demonstrated great precision,

recall, and F1-score, demonstrating its capacity to precisely distinguish between positive and negative situations. The performance of the model was extremely consistent, indicating its robustness under diverse feature selection circumstances.

7 Conclusion and Future Work

This study used three different feature selection methods and compared them with the entire feature dataset to conduct a thorough assessment of different machine learning methods for stroke prediction. The findings demonstrated how feature selection considerably affects to what extent machine learning strategies do stroke prediction. Top performers included the Light and Adaboost models, which displayed good, predicted accuracy, and balanced evaluation metrics. These findings have important ramifications for the healthcare industry because accurate stroke prediction can result in early interventions, enhanced outcomes for patients, and lower medical expenses. As a result, the knowledge gained from this research can help in the creation of stroke prediction algorithms that are more accurate and trustworthy, ultimately resulting in improved healthcare procedures and patient care. The study also emphasizes how crucial feature selection is for improving prediction and performance. For future work, it could be useful to investigate evaluating the constructed algorithm in a real healthcare setting in which it may generate predictions in real time. Early stroke risk identification and swift action may decrease the overall occurrence of strokes.

Acknowledgment

I would like to make use of this chance to express my gratitude to Qurrat Ul Ain, who happens to be my supervisor. Her recommendations, which proved insightful and beneficial, came in handy when it came time to put the research project together. She volunteered a few hours of her precious time to help me out throughout the entire procedure, and for that, I am incredibly appreciative.

REFERENCES

Islam, M.S., Hussain, I., Rahman, M.M., Park, S.J. and Hossain, M.A., 2022. Explainable artificial intelligence model for stroke prediction using EEG signal. *Sensors*, *22*(24), p.9859.

Singh, M.S. and Choudhary, P., 2017, August. Stroke prediction using artificial intelligence. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (pp. 158-161). IEEE.

Kaur, M., Sakhare, S.R., Wanjale, K. and Akter, F., 2022. Early stroke prediction methods for prevention of strokes. Behavioural Neurology, 2022.

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

M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Bangkok, Thailand, 2017, pp. 158-161, doi: 10.1109/IEMECON.2017.8079581.

Emon, M.U., Keya, M.S., Meghla, T.I., Rahman, M.M., Al Mamun, M.S. and Kaiser, M.S., 2020, November. Performance analysis of machine learning approaches in stroke prediction.

In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1464-1469). IEEE.

Jeena, R.S. and Kumar, S., 2016, December. Stroke prediction using SVM. In 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 600-602). IEEE.

Dritsas, E. and Trigka, M., 2022. Stroke risk prediction with machine learning techniques. Sensors, 22(13), p.4670.

Sailasya, G. and Kumari, G.L.A., 2021. Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6).

Wu, Y. and Fang, Y., 2020. Stroke prediction with machine learning methods among older Chinese. International journal of environmental research and public health, 17(6), p.1828.

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

T. Badriyah, D. B. Santoso, I. Syarif, and D. R. Syarif, "Improving stroke diagnosis accuracy using hyperparameter optimized deep learning," Int. J. Adv. Intell. Informatics, vol. 5, no. 3, pp. 256–272, 2019, doi: 10.26555/ijain.v5i3.427.

A.B. URAL, "Computer-aided Deep Learning based assessment of stroke from brain Radiological CT Images," Eur. J. Sci. Technol., no. 34, pp. 42–52, 2022, doi: 10.31590/ejosat.1063356.

Biswas, N., Uddin, K.M.M., Rikta, S.T. and Dey, S.K., 2022. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics, 2, p.100116.