International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Machine Learning based Brain Stroke Prediction using Light Gradient Boosting Machine Algorithm

[1]Jarapala Parvathi, [2]Dr. Saikiran Ellambotla

Submitted: 16/03/2024 **Revised**: 26/04/2024 **Accepted**: 02/05/2024

Abstract: Timely detection and proactive measures to prevent stroke are of highest priority due to the effective likelihood of extreme disabilities or destructive effects associated with this disease. Stroke diseases can be separated into two categories those are ischemic stroke and Hemorrhagic stroke, and they should be minimized by emergency treatment such as thrombolytic or coagulant administration by type. The Early detection of the multiple stroke warning symptoms can facilitate the stroke's harshness. The main purpose of this study is to predict the possibility of a brain stroke happening at an early stage using machine learning algorithms. To gauge the effectiveness of the algorithm, a reliable dataset for stroke prediction was taken from the Kaggle website. Various classification models, including, K-Nearest Neighbor (K-NN), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) were successfully utilized in this analysis for classification studies. The performance of the methodology is evaluated using Precision, Recall, and F1-Score evaluation metrics. With experimental results, we can show that the proposed LightGBM classifier has 99% classification accuracy, which was the highest (among the machine learning classifiers).

Keywords*: Brain stroke prediction, machine learning, Logistic Regression, K-Nearest Neighbor, Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine*.

I. Introduction

Strokes, often called strokes or cerebrovascular accidents (CVAs), are caused by a blockage of blood in a part of the brain, resulting in brain cell damage or loss of life. Blood vessels that supply the brain can also become blocked (ischemic stroke) or bleed (hemorrhagic stroke). A stroke will lead to severe consequences with cognitive and physical impairments and will require immediate hospital treatment. Treatment and rehabilitation options vary depending on the type and amount of stroke. Stroke is a surprising medical condition that involves spark treatment to tame its side effects. The World Health Organization (WHO) published its 2019 Causes of Death Report in December 2020. The report found that 55% of all estimated deaths in 2019 (or about 55.4 million people) were caused by the top 10 causes of harm. According to Life in the United States [1], the number of strokes is clearly excessive: a stroke occurs every 40 seconds, affecting approximately 800,000 people and disabling one person each year. Major stroke is also the sixth leading cause of death [2].

Additionally, recent studies have shown that COVID-19 is associated with stroke and that people are more likely to die from stroke [3]. According to Kummer and colleagues, the mortality rate among COVID-19 patients with a history of stroke is significantly higher than among those without a history of stroke. A stroke can be diagnosed by imaging tests, including computed tomography (CT), magnetic resonance imaging (MRI), CT angiography (CTA), magnetic resonance angiography (MRA), blood tests, ECG, and transcranial Doppler (TCD).) Included. The most popular ultrasound methods for diagnosing stroke are CT and MRI, but these involve exposure to radiation or possible allergic reactions to similar chemicals. The risks of these strategies are that they are time-consuming to administer, costly to test, and challenging to conduct real-time research

1Research Scholar, Department of Computer Science &Engineering, Chaitanya Deemed to be University, Warangal. parvathi.cse@gmail.com
2Assistant Professor, Department of Computer Science& Engineering Chaitanya Deemed to be University, Warangal. Kiran.09528@gmail.com

at an early stage. Access to CT and MRI scans may be limited in sensitive health settings, particularly in resource-constrained areas or emergencies. The availability of these imaging modalities can be a challenge and delay obtaining timely scans for stroke prediction and diagnosis. Recent research has attempted to predict stroke complications using statistics or units, taking into account elements of positive probability and suggesting strategies to overcome these limitations. Insights were also gained.

Previous studies on stroke have targeted predictors of heart attack in various subjects. There are not many studies on cerebral palsy. The primary purpose of this article is to explain how device dominance algorithms, augmentation strategies, and artificial neural networks (ANNs) can be used to predict when cerebral palsy will occur. Will At the heart of this research is a software program that compares and describes qualitative methods for predicting stroke onset by using several algorithms on a freely accessible dataset (from the Kaggle website). In this description, neural networks and machine recognition algorithms are used as type algorithms to predict the lifestyle of stroke patients through the distribution of relevant functions. The principal component analysis (PCA) method reduces the dimensionality. After lowering the metrics, we reserved the most critical tasks for hit prediction. Accuracy, precision, omission, f-1 score and AUC curve (under the roc curve) of various general performance measures of beauty fashion are evaluated and compared with each other. Which one makes the most accurate prediction on the data set? The experimental results of the proposed strategies were then compared with existing studies to demonstrate the novelty of the images.

In this study, we perform stroke prediction tasks using the entire dataset obtained from the Kaggle Stroke Prediction Dataset. Our goal is to develop a model that outperforms existing strategies in terms of accuracy and robustness. To achieve this, we propose an ensemble method with an improved LiteGBM algorithm.

II.Literature Survey

Previous studies have addressed several issues related to stroke prediction. Several studies use device specialization (ML) techniques to achieve this. At this stage, significant contributions to the investigation have been identified. In today's interpretation, an Artificial Neural Networks (ANN) technique trained with a Multilayer Perceptron (MLP) algorithm is used to predict the mortality of stroke patients, and the results are 8.7%. Corrected.

The overarching goal of this review is to predict the likelihood of developing cerebral palsy at an early stage using depth information and machine learning strategies. A reliable dataset for hit prediction was obtained from the Kaggle website to measure the effectiveness of the algorithm. Different modes like Extreme Gradient Boosting (XGBoost), Ada Boost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K Neighbours, SVM - Linear Kernel, Naive Bayes and Deep Neural Networks (3-Layer and 4-Layer) was modified to be effectively used for classification tasks in this research. Random forest classifier has 99% species accuracy. This is the highest price (of many devices tested in the class). A 3-layer deep neural ensemble (four-layer ANN) produced 92.39% better accuracy than the 3-layer ANN approach using the selected capabilities as input.

Shehzada Mushtaq et al. [2023] This paper presented an improved stroke detection algorithm for predicting stroke events. To develop a predictive algorithm, they used a dataset of essential parameters that could be charged for cerebral palsy, including age, body mass index (BMI), gender, cardiovascular disease, smoking reputation and Others. The dataset was modified for the first arrangement. They were dealing with missing values, binomial abilities, and balancing datasets. They used various classification algorithms, including Naïve Bayes, Logistic Regression, XgBoost, Selection Shrubs, AdaBoost, K-Nearest Neighbor, Random Forests, Voting Classifiers, and others, to develop their prediction model. Various metrics such as accuracy, F1 score, observability and precision were used to evaluate the models.

Santosh E et al. [2023] The brain is the most complex organ in the human body. Stroke is a permanent disability that occurs globally and is a significant cause of loss of life. A stroke occurs when blood flow to the brain is reduced and stops the brain from functioning. There are two leading causes of stroke: a blocked artery (ischemic stroke) or a ruptured or ruptured artery (hemorrhagic stroke). Prediction of early stroke is extra powerful when miles are helpful early on. People's lifestyle choices often lead to stroke, especially those that include diabetes, heart disease, weight problems,

diabetes and high blood pressure - the modern scenario. CNN, Densen, and VGG16, as well as various machine information acquisition algorithms (ML), are used in this observation. This working setup uses one of the following algorithms that can expect hits and provide new facts with precision.

Sairam Vasaet al. [2023] These studies aim to develop highly efficient models using Machine Learning Algorithms (MLAs), namely Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest Classifier (RFC) and Support Vector Machine. (SVM), Naive Bayes Classifier (NBC), KNN Classifier (KNN) and XGBoost Classifier (XGB). Implemented the above algorithm with Hyperparameter with GridSearchCV (CV=5) on the given dataset. A given data set tends to be more balanced. During the training of the models, some problems were encountered, including a terrible, record-breaking zero value in the data and a risky version of the real one to improve fashion and data performance. The numbers need to be balanced. These are the use of SMOTE-related reality sampling methods. Among the seven models, XGB is the most desirable version, with a typical accuracy value of 96.34%.

Mr.M.Thirunavukkarasu et al.[2023] The purpose of these studies was to explore how they could improve the subjects. They used Kaggle's Stroke Disorders document. Patients can benefit from pre-processed statistics. Ischemic stroke and stroke are types of hemorrhagic stroke. Individuals were classified using systems study techniques. In this study, the gadget evaluation technique was used in seven examples. Logistic Regression, Support Vector Machine (SVM), Random Forest, Cat Boost, Multilayer Perceptron (MLP), Naive Bayes, and K-nearest Neighbors. Therefore, our results show that Cat Boost offers higher accuracy with marginal values and f1-Score. ' Forget that.

Latharani T R et al. [2023]A stroke is a medical emergency due to the possibility of death or lifelong disability. Ischemic stroke is manageable. However, this treatment should be started within hours of the onset of stroke signs and symptoms. If a stroke is suspected, the patient, family, or witnesses should seek emergency medical attention immediately. A transient ischemic attack (TIA or mini-stroke) is an acute ischemic stroke in which signs and symptoms disappear on their own. This condition also requires rapid diagnosis to reduce the chances of a life-threatening stroke. If all symptoms resolve within 24 hours, it is, by definition, a stroke, not a TIA. According to the World Health Organization (WHO), stroke is the second leading cause of global death. Their machine learning model uses data units to predict survival based on factors such as sex, age, comorbidities, and smoking records to judge a victim's likelihood of stroke.

A.Srinivas et al.[2023] The proposed model was a pairing machine that combines the predictions of several individual classifiers to derive insights from hard and fast policies, including random forests, extremely random trees, and histogram-based pure gradient boosting. Each classifier returns a probability estimate for each class, and the final estimate is based entirely on the joint value of these probabilities. The weight assigned to each classifier can be based entirely on the overall performance of the validation set or can be uniformly generated. The proposed proposal voting model performed better in terms of final prediction accuracy and robustness than the rating model. Difficulty tracking stroke type can help improve resource utilization and reduce healthcare costs.

Prasad Gahiwad et al. [2023]This paper used a convolutional neural network for stroke detection in CT experimental images. After training the model and testing it on the CT benchmark dataset of 2551 snapshots, we achieved a peak accuracy of 90%. For stroke analysis and treatment, mental CT control images need to be subjected to digital quantitative assessment. An essential tool for damage detection is provided by deep neural networks that have a world-class ability to achieve real-world expertise.

Abdur Nur Tusheret al. [2022] There are many causes of cerebral palsy, such as abnormal blood flow to the brain. Diagnosing cerebral palsy or cerebral palsy is now very important for the scientific medical doctor but the real challenge is to identify the diseases successfully. In most cases, doctors look at the final result and decide whether it is too unbelievable or not. In this research, they developed a device that can diagnose cerebral palsy for the first time. Some elegant algorithms like Logistic Regression, Category, Regression Tree, K-Nearest Neighbor and Support Vector Machine are used to train this device version. The KNN rule set achieves the highest accuracy at 97%. The recommended gadget can be reliable, automatic and time-saving.

Bonna Akter et al. [2022]This paper proposed a recipe that preserves the method for obtaining

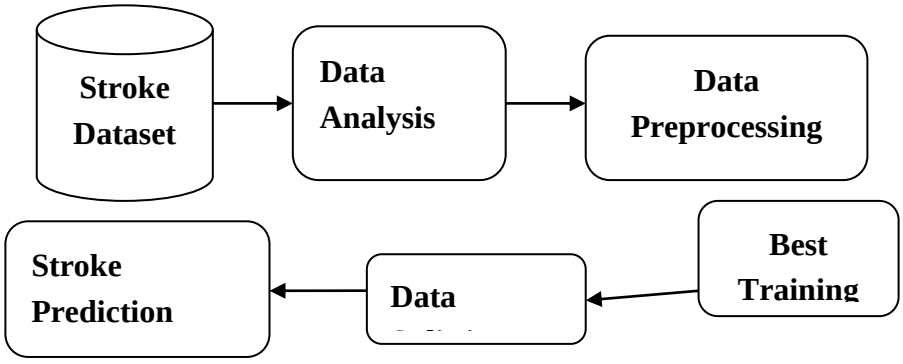
accurate predictions of cerebral palsy. Effective fact-gathering, pre-fact processing and record transformation techniques are applied to provide reliable data for the success of our proposed model. A "contour dataset" was used to collect the versions. Standardization techniques are used to standardize statistics. Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT) classifiers are used in training and testing strategies. Record the overall performance, accuracy, precision (SEN), error rate, false positive rate (FPR), false low rate (FNR), root mean square error, and log of each classifier.

Tahia Tazin et al. [2021] This study used a fixed set of physical parameters and machine mastering algorithms, including logistic regression (LR), decision tree (DT) classifier, random forest (RF) classifier, and voting for 4 correct reliability estimates. Classification. The Random Forest classifier produced an excellent-looking rule for training models with about 96% accuracy for this assignment. The dataset used to develop the technique becomes an open-access entry in the Strike Prediction dataset. The accuracy percentage of the models used in this research is relatively high compared to previous studies. This indicates that the models used in it are more reliable. A comparison of different models has confirmed its robustness and can be estimated by reading the diagram.

III.Proposed Methodology

This phase represents the proposed methodology stages, such as the experimental dataset, data analysis, balance of data with SMOTE, and training mode. The proposed work is represented in Fig. 1 below. As shown in the figure.1, first gathers the Stroke-related data from the Kaggle Dataset, then executes the data analysis using bi-variate, categorical distribution, and pairwise analysis. The next phase is data preprocessing using the SMOTE algorithm to balance the stroke dataset. The gathered dataset is divided into training and testing data in the ratio of 80:20%. The ensemble model is designed for stroke prediction using machine learning algorithms such as Logistic Regression (LR), K-Nearest Neighbor (K-NN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), proposed Light Gradient Boosting Machine (LightGBM). After machine learning models are used, the best training and testing results are obtained.

Proposed system architecture



A) Data Analysis

After informative training, the data set is tested to decide whether data cleaning is necessary or not. This step also includes an analysis of the appropriate size capabilities. The identification of outliers and anomalies in records allows the selective determination of hypotheses, which is the main reason for statistical evaluation. Records are assessed, considered, and updated effectively without having to make assumptions about first-order records and incremental fashion. It includes both numeric and specific types of variables by exploring the dataset. The absolute capacity of the record set is "five", and the numerical capacity is "7", with a 4% missing cost in the characteristic BMI. Because variables require entering numbers, studies should code superficial characteristics before moving on to versions. This study used secondary tests for numerical attributes, distributions of specific abilities, and some binary tests to observe all abilities. Some search strategies rely on record classification and use target

evaluation. In two forms of analysis, this research analyzes how the target country is reduced by functions that best match the target's characteristics. Age, mean glucose level and BMI are used for 2 types of diagnosis. People between the ages of sixty-five and eighty-five are most likely to have a stroke. BMI no longer discriminated stroke patterns and had a 4% missing value. As a result, we are dropping this feature. A discrete hazard distribution, called the express distribution, describes the probability that the charge of the random variable corresponds to one of K training where each class has a corresponding hazard. In the distribution of express variables, gender, first married, image type, house shape, smoking prevalence, high blood pressure and coronary heart disease abilities are observed, which in these cases affect the effects of stroke and are used to determine the results. Both can be researched. This approach analyzes a greater diversity of hit (stroke = 0) instances in each feature than the other subtypes and cannot speak of different learning based on 'gender'. A dual grand measure of household type and smoking reputation can be accomplished with a dyadic analysis. Observation showed that both urban and rural residents did not affect stroke incidence, but this could not be noted. In a binary analysis by job type and gender, it was observed that the top '2' stroke cases occurred among the youth, and girls and older women working in government jobs were more likely to have a stroke. An analysis of the type of practice and prevalence of smoking suggests that the majority of stroke victims are recruited into the 'personal' sphere.

B) Data Preprocessing

Data processing involves standardizing some information and breaking it down into numerical values. The pre-processing part of the information is essential because it allows us to obtain more accurate capabilities and increases the overall performance of the version. In this article, advanced scalar techniques are used to encode discrete facts.

a) Encoding Categorical Variables

Categorical variables are complex for many machine learning algorithms. It is essential to convert some variables into numeric records. This is an important step in the robust functioning of the implemented algorithm. The way variables are encoded affects how smoothly different algorithms work. One or more tags can be added to a feature's dataset in word or numeric order. This interprets facts less complex for humans but more understandable for computer systems. Therefore, in this statement, coding is used to make these labels understandable with the help of a computer system. Some of the coding techniques include one-hot coding, hash coding and others. Let's take a look at this tag, which is an encoding method used to encode certain variables.

b) Label Encoder

Label encoding allows numeric labels to be encoded in ML models. This is an essential step in pre-processing the facts for supervised learning techniques. A label encoder assigns specific values to each label to distinguish each unusual label value within the dataset. Tags can be used when other preferences dictate it. This technique usually substitutes numbers from 0 to N - 1 for each charge in a transparent column. A label encoder assigns a zero-to-1 ratio to each instant function in this paper.

c) Data Scalar

Standardization is an optimization strategy commonly used to standardize facts. A normalized scalar is a crucial device that is typically used as a preprocessing piece to normalize the operating range of an input dataset before capturing the instrument's variational information. Scales a feature in-unit variance after subtracting the implied value for normalization. Since the fashionable scalar predicts that the facts are normally distributed for each interval, it is superimposed on everything such that the distribution effect is zero and has a fashionable deviation (SD) of 1[5]. These studies take advantage of traditional scalar techniques to normalize the functional diversity of the bat input dataset.

C) Data Splitting

Building models comes after pre-processing the information and dealing with imbalances. The re-sampling dataset is divided into training and control, with a percentage of 80% in the training part and 20% in the testing information to increase the accuracy and efficiency of this method. After the segmentation process, different classifiers are used to build the model. The classification techniques used in this function are LR, KNN, RF, XGBoost and LightGBM.

Balance Dataset with SMOTE

Real-world information units typically have a high proportion of "normal" instructions and a low

proportion of "strange" instructions. If the schooling distribution of the classifier is heterogeneous across the data set, unbalanced statistics are taken into account. Specifically, two strategies are used to balance the dataset: the first is oversampling applied to minority elites, and the second is under sampling applied to the general population. This method offers high overall classification performance. SMOTE is a definitive solution to the type imbalance problem and has achieved robust results in many fields. The finite training set combines synthetic statistics with the SMOTE technique to create a balanced and robust set of facts. The significant imbalance problem is associated with inequality between the categories of the statistical set used to construct the forecast approach. This is a huge challenge now reserved for clinical records. Class imbalance problems are addressed by manipulating data and algorithms or by improving model performance when the training set is disproportionately small, which produces dangerous results. Class algorithms validate traditional generic classes with a more economical range of assumptions than optimal prediction types. The primary step of the technique is to oversample and under sample randomly large and small samples. This research dataset contains training: stroke and no stroke. Wish the stat set was more balanced. Therefore, we used the SMOTE technique to stabilize the facts. Before applying the SMOTE technique, the untrained model performed poorly with a sufficient number of instances of each grace. Using the SMOTE approach works perfectly with proper system training to understand the version hierarchy.

Model Evaluations

A) Classification Models

A classification model aims to emerge from some entries in the training set. Types and complexity labels will await updated statistics. This study used various gadget mastering algorithms, such as LR, KNN, RF, XGBoost, and LGBM. Extensions to these algorithms are given below.

a) Logistic Regression (LR)

LR is a supervised learning algorithm that define the probability that a target's significance will emerge. These objective cost values may be more stable. They recommend the most effective effects possible. The parameter is binary, and to make it easier to identify, records are represented as 1 (for a stroke) or zero (for no stroke).

b) K-Nearest Neighbour(K-NN)

KNN is the least complex algorithm that uses supervised learning. It is used to create similar facts by combining and storing unique information. As new records are made, the class can recognize the use of a method that is appropriate for its properties. Although Miles is used in styles for type and regression, it is also used to solve class-related problems. It keeps the data at any stage of training and reflects any changes by reclassifying the data as if they were just given. The KNN approach provides the best overall performance when optimally selecting K faces.

c) Random Forest (RF)

Researchers can use machine learning to solve regression and type problems. Random forests are a popular class of artificial intelligence developed through supervised data collection. Decision trees are constructed on different subsets of the provided dataset to improve the prediction accuracy. This species is only sometimes entirely dependent on a single tree. Specific types of wood were created to find the same effect. The most accessible, freshest effects are selected from these forests over a long period.

d) XGBoost

XGBoost is a notable implementation of the gradient-boosting machine learning algorithm. It uses an active ensemble approach that combines valid prediction models (usually a selection tree) to create a robust prediction version. XGBoost aims to reduce feature loss by adding new wood to the community. The final prediction is obtained by summing the predictions about each tree in the version. XGBoost contains regularization expressions to manage overfitting and offers flexibility in defining custom loss properties. In XGBoost, we use L1 and L2 orchestration techniques to change initialization complexity and reduce overfitting. L1 regularization or Lasso regularization supports the separation of feature weights by accurately adding weight values to the loss feature.

e) Light Gradient Boosting Machine (LightGBM)

A lightweight Gradient Boosting Machine (LGBM) is one of all other gradient boosting methods and combines learning methods. The main reason for LGBM is to offer large slopes that increase the rate of return. LGBM makes the shrub upward using the sensitive leaf style. For splitting, the branch that minimizes the loss is chosen. A leaf that minimizes

damage is roughly chosen to split and prepare the tree. LGBM detected perfect bin references using a histogram-based aggregation approach. Gradual preference-based one-way sampling (GOSS) techniques are used to judge the importance of information samples in improving LGBM training. The main reason for this is to ignore recording efforts with small gradients and handle with extra-large gradients. The basic assumption is that the error in the mean-slope statistic may be less and more informative. GOSS allows discarding a small amount of the dataset and using the entire dataset to classify the information obtained by determining the proximity of suitable discriminants. However, this amount can change the distribution of the file and introduce inadequacies for models with large slopes. GOSS solves this challenge by running all trials with large slopes and periodically deciding on facts with small slopes. Since the sample is for data with large gradients, GOSS increases the weight of samples with small gradients when calculating the registration gain.

LGBM operates the Exclusive Feature Bundling process to manage dataset sparsity. Since it aggregates independently entire features in a closely lossless way, the number of features is decreased while the most informative ones are observed.

Algorithm1:

1: **Input**: Stroke healthcare datasetD_s
2: **Output**: Model Performance
3: Data Analysis (DA)
4: Bi − variate Analysis
5: categorical Variables Distribution
6: Pairwise Analysis
7: Data Preprocessing (D_p)
8: x, y {Dataset D_s}
9: $x_{train}, x_{test}, y_{train}, y_{test}$ {data split into train and test set}
10: SM = smote () {smote formula}
11: Ensemble {Create Ensemble Models}
12: RF ← RandomForest ()
13: XGB ← XGBoost ()
14: LGBM ← LGBmodel ()
15: Em
← Accuracy, Precision, Recall, F1
− Measure {Evaluation metrics}
16: Return ← Best Training Results
17: Ensemble {Recreate Model for test results}
18: **Return** ← Result

Algorithm 1 describes the prevalent working of the suggested ensemble model. The ensemble model is constructed using three machine learning algorithms (RF, XGB, LBGM). The stroke dataset is taken as input, and the output is the implementation of the proposed ensemble learning model. Data analysis is utilized for the visualization of all the features. Bi-variate, definite variables distribution, and pairwise analysis methods are used. Data is distributed into training and tests with sizes of 0.8 and 0.2 and pre-processed using SMOTE (SM) to balance the data. In this Evaluation, metrics of accuracy, precision, recall, and F1-score are used to calculate the proposed methodology performance. The best results are obtained on the training dataset. Then, the model and outcomes on the test dataset are recreated.

IV.Experiments

The performance of the ensemble model and individual models in stroke prediction is assessed using the area under the receiver operating characteristic curve (AUC) metric. In Table 1, we provide the performance evaluation received by each model, emphasizing the significant progress accomplished.

DATASET

The analysis was carried out using the stroke prediction dataset available on the Kaggle website. In this dataset, there were 12 columns and 5110 rows. The description of the dataset is given in below Table I. The output column's stroke has a value of either 1 or 0. A stroke risk was found when the value 1 was detected, but a stroke risk was not found when the value 0 was shown. In this dataset, there is a greater chance of a 0 in the output column (stroke) than a 1 there. The stroke column alone has 249 rows with a value of 1, whereas 4861 rows have a value of 0. In the output column before preprocessing, Fig. 1 shows the Visualizing Count of classes along with the number. Preprocessing (oversampling) of the data is used to balance the data and improve accuracy. After Oversampling, the value of stroke with 1 has risen to the level of 4861 while the value of non-stroke data with 0 remains the same as before (i.e., 4861)

Evaluation metrics

The proposed diagnostic method is expected in general performance situations using a well-known matrix that includes accuracy, precision, recall, and F1-score. These metrics are calculated using.

$True\ Positive\ (TP)$,$True\ Negative\ (TN)$, $False\ Positive\ (FP)$,and$False\ Negative\ (FN)$ parameters. When TP is likely to cause most stroked in a brain stroke patient, FPs will likely find that the rate at which brain stroke is detected is .

the rate at which a healthy person is found. TN hopes to reveal that a person with brain stroke is healthy. Therefore, FN is prescribed when a healthy man or woman has brain stroke

Fig.2 Performance comparison between machine learning algorithms

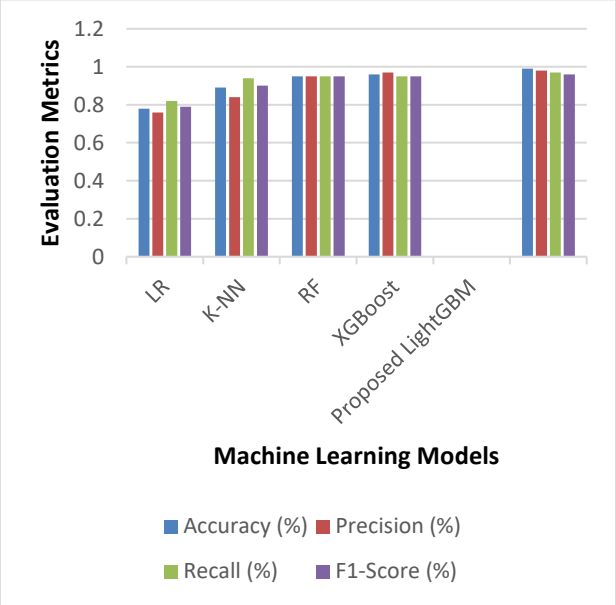


Table.1 Performance evaluation of Machine learning algorithms

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score(%)**
LR	0.78	0.76	0.82	0.79
K-NN	0.89	0.84	0.96	0.90
RF	0.95	0.95	0.95	0.95
XGBoost	096	0.97	0.95	0.95
Proposed LightGBM	0.99	0.98	0.97	0.96

The total effectiveness of ML algorithms on the stroke dataset is shown in Table1. The LR algorithm achieved 0.78% accuracy, 0.76% precision, 0.82% recall, and 0.79% f1-score. K-NN algorithm obtained 0.89% accuracy, 0.84% precision, 0.96% recall, and 0.90% f1-score. DT obtained 0.91% accuracy, precision, recall, and f 1-score. RF achieved 0.95% accuracy, 0.95% precision, 0.95% recall, and 0.95% f1-score. The XGBoost algorithm gained 0.96% accuracy, a precision of 0.98%, are call of 0.95%, and an f1-score of 0.95%. The proposed LGBM algorithm performed an accuracy of 0.93%, precision of 0.96%, recall of 0.90%, and 0.93% f1-score.

Confusion matrix

We have imported the Confusion matrix library for image Classification, which is the procedure of splitting a given set of data into lessons. In machine learning (ML), we prepare the problem, aggregate and smooth the data, load some vital feature variables, flag the version, estimate its performance, and improve it with some shipping properties.

A much higher way to estimate the performance of a classifier is to look at the confusion matrix. The well-known idea is to count the instances of class A in which complexity times are classified as Class B.

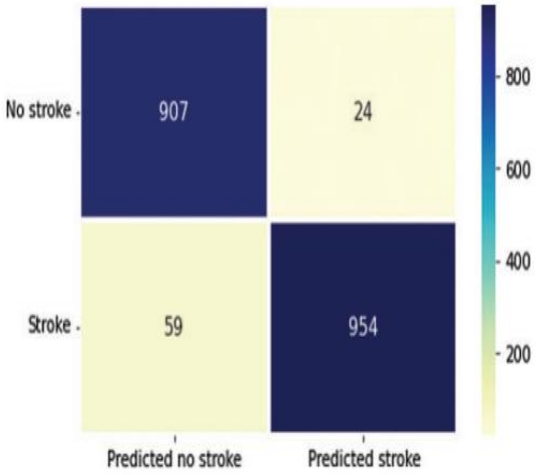


Fig.3 Confusion matrix of proposed model

Fig.3 graphically illustrates the confusion matrix (CM) of the suggested method. It determines the stroke in two classes: stroke and no stroke. Because it has more continuous, better true positive and negative results and fewer false positive and negative values, the proposed technique performs better.

V.Conclusion

In this paper, we proposed an integrative paradigm for improved stroke prediction by synergizing the machine learning-based LightGBM algorithm. Our analysis creates valuable assistance to the progress of machine learning in the area of stroke prediction, delivering helpful knowledge for healthcare specialists in identifying individuals at higher risk of stroke and allowing timely interventions and personalized care. The experiments conducted on the brain stroke dataset gathered from a publicly available online Kaggle data set and the proposed machine learning-based LightGBM algorithm performance are evaluated using accuracy, precision, recall, and f1-score and compared with various existing machine learning algorithms of LR, K-NN, RF, and XGBoost. The LGBM algorithm achieved an accuracy of 0.93%, precision of 0.96%, recall of 0.90%, and 0.93% f1-score. With the experimental results, we can show that the proposed algorithm achieves better performance compared with the previous models.

References

[1] Senjuti Rahman, and Ajay Krishno Sarkar, 2023, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques", pp.23-30.

[2] Shehzada Mushtaq and Kamaljit Singh Saini, 2023, "Machine Learning for Brain Stroke Prediction", IEEE.

[3] Santosh E, Hruthik Gowda MP, 2023, "Brain Stroke and Its Stages Prediction Using Deep Learning", IRJMETS, pp. 2575-2580.

[4] Sairam Vasa, PremKumar Borugadda, 2023, "A Machine Learning Model to Predict a Diagnosis of Brain Stroke", ICDT.

[5] A. Srinivas, Joseph Prakash Mosiganti, 2023, "A brain stroke detection model using soft voting based ensemble machine learning classifier", Volume 29, October 2023, p.100871.

[6] Prasad Gahiwad and Sachet Karnakar, 2023, "Brain Stroke Detection Using CNN Algorithm",DOI: 10.1109/I2CT57861.2023.10126125 .

[7] Abdur Nur Tusher andMd. Tariqul Islam, "Early Brain Stroke Prediction Using Machine Learning", IEEE

[8] Bonna Akter, Sadia Sazzad, 2022, "A Machine Learning Approach to Detect the Brain Stroke Disease", IEEE, **DOI:** 10.1109/ICSSIT53264.2022.9716345.

[9] Tahia Tazin Md Nur Alam, Mohammad Sajibul Bari, "Stroke Disease Detection and Prediction Using Robust Learning Approaches", Hindawi, Journal of Healthcare Engineering, pp.1-12.

[10] U. R. Acharyaand D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," Computer Methods and Programs in Biomedicine, vol. 161, pp. 103–113, 2018.

[11] Y. H. Kwon, S. B. Shin, "Electroencephalogra phy based fusion two-dimensional (2D)-convolution neural networks (CNN) model for emotion recognition system,", vol. 18, no. 5, p. 1383, 2018.

[12] B. Kimand C. J. Winstein, “Corticospinal tract microstructure predicts distal arm motor improvements in chronic stroke,” Journal of Neu rologic Physical Therapy, vol. 45, no. 4, pp. 273–281, 2021.

[13] H. A. Adhi, and M. Rezal, “Automatic detection of ischemic stroke based on scaling exponent electro encephalogram using extreme learning machine,” Journal of Physics: Conference Series, vol. 820, pp. 12005–12013, 2017.

[14] E. C. Djamal, and D. Djajasasmita, “Identification of post-stroke EEG signal using wavelet and convolutional neural networks,” Bulletin of Electrical Engineering and Informatics, vol. 9, no. 5, pp. 1890–1898, 2020