Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.778151

EFFECTIVE PREDICTION FOR ISCHEMIC BRAIN STROKE USING ENSEMBLE BOOSTING ALGORITHMS

C. Tamilselvi
Research scholar, Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamilnadu, India

S. Ramamoorthy
Professor, Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamilnadu, India

V. N. Rajavarman
Professor, Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamilnadu, India

ABSTRACT:

The explosive increase of medical data on a worldwide basis is unsustainable since it can only be handled by BigData analytics. BigData Analytics can deal with rapidly exploding medical data by focusing on fundamental characteristics such as volume, velocity, variety, veracity, and value. Ischemic Stroke is a medical condition in which when the blood arteries supplying blood to the brain is clogged, causing brain damage. According to the World Health Organization (WHO), stroke is the leading cause of death and disability globally. Early recognition of the multiple strokes warning indicators can reduce the severity of the stroke.The primary goal of this research is to use Ensemble Boosting Algorithms to efficiently predict the likelihood of a brain stroke happening at an early stage.In order to evaluate the algorithm's effectiveness, an effective dataset for stroke prediction was obtained from the Kaggle website. Several Ensemble Boosting Algorithms including AdaBoost, Gradient Boosting Machine, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine and CatBoost were successfully used in this study. Among the above algorithms Extreme Gradient Boosting (XGBoost) give a high accuracy rate of 98.2% and provides a fast-training speed, LightGBM second the XGBoost with an accuracy rate of 95.7% but fastest of all the algorithms. This work has the potential to substantially boost the medical system's ability to prevent and alleviate the damage caused by ischemic strokes.

Keywords— *Ensemble Boosting Algorithms, XGBoost, AdaBoost, Gradient Boost Machine, Light GBM, CatBoost*

1. INTRODUCTION

Predictive analytics is a type of BigData analytics that makes predictions about potential occurrences using historical data, statistical modelling, data mining tools, and machine learning & Ensemble Boosting algorithms. Predictive analytics uses BigData comprehending to furnish intelligence about the future.The use of a Big Data technique may

be helpful in the identification of relevant indicators for stroke diagnosis and risk assessment [1]. Big data–enabled research can
be distinguished from other types of research based onthe attributes described in the 5Vs framework, including:volume (the size or number of records), variety (heterogeneityor diversity of type, structure and setting of data), velocity (rapid generation and reporting of data),veracity (data quality and reliability), and variability(variations between different data sources and datasets)[2, 3]. In stroke research, big data are valuable, low-cost resources for making evidence-based decisions toguide practice and policy on the prevention of adverseoutcomes following stroke [2].

Ischemic stroke, a neurological condition caused by insufficient blood supply to the brain, hasemerged as a rising public health concern with serious effects, including pain and fatality, in today's society. Ischemic stroke occurs when an arterial blockage reduces or affects blood supply to brain cells, killing the cells and leading to die in minutes. Hemorrhagic stroke, on the other hand, occurs when the blood vessels in the brain are severely damaged as a result of hypertension, high cholesterol, and other risk factors [4, 5]. Many Big data predictive analytics and machine learning (ML) models have been developed to predict the probabilities of a stroke occurring in the brain. Strokes are caused by a variety of risk factors, including medical conditions such as high blood pressure, heart disease, diabetes, high cholesterol, and atrial fibrillation, as well as harmful habits such as smoking, obesity, eating unhealthy foods, and not exercising [6]. Stroke is one of several serious medical conditions that can be prevented if diagnosed early and can be successfully treated if predicted in the early stages. Machine learning is crucial for the diagnosis and forecasting of diseases in the field of health care. Machine learning techniques are currently used to predict the incidence of stroke [7, 8].

The main objective of this study is to demonstrate the application of machine learning ensemble boosting techniques to the prediction of brain stroke.The key contribution of this study is the application of different Ensemble Boosting algorithms (XGBoost, AdaBoost, Gradient Boost Machine, Light GBM, CatBoost) to a dataset that is freely accessible from the Kaggle website, facilitating for comparison and identification of the optimum approach for prediction stroke.To identify which ensemble algorithm predicts the dataset most accurately, a variety of performance metrics (accuracy, precision, recall, F-1 score and MCC) are obtained and compared amongst the algorithms.

2. RELATED WORK

Vrdoljak et al. [9] used four machine learning classifiers—XGBoost, Random Forest, Logistic Regression, and Univariate Logistic Regression—to predict breast cancer lymph node metastases in patients who were suitable for neo adjuvant therapy. XGBoost performed best in this study, with a mean AUC of 0.762 (95% CI: 0.726-0.794).

Rado et al. [10] developed an ensemble model and compared the results of the homogeneous ensemble methods Random Forest (Bagging), Adaptive Boosting, and Stacking. They compared the model's performance against standalone classifiers using accuracy, Mean Squared Error (MSE), precision, and F-measure. Their results suggest that ensemble classifiers outperform standalone classifiers in terms of accuracy. The stacking classifier is the most accurate, with an accuracy of 87.58%.

Rezazadeh A et al. [11] applied ultrasound images to create an explainable machine-learning way for breast cancer diagnosis. For this experiment, they used a probabilistic ensemble of decision tree classifiers. The model's most important texture features were found by measuring feature importance with SHAP values. According to this experiment, the LightGBM approach with 500 iterations performs best, with 0.91 accuracy, 0.94 precision, 0.93 recall, and 0.93 F1 score.

Fernandez-Lozano, C. et al and Ntaios, G. et al[12,13] states that Machine learning (ML) algorithms are currently widely used in medical research, and numerous mortality prediction models for ischemic stroke have been developed. It allows researchers to create accurate models. When there is high collinearity in the data, using ML techniques can be advantageous.

Luca et al. [14] used proven ML prediction models to integrate approaches such as feature selection and ensemble learning to diagnose medical illnesses.

Kumar et al. [15] developed an ensemble method for automatically detecting brain disorders. It is a learning-based method that employs many underlying baseline methods in the ensemble to improve accuracy.

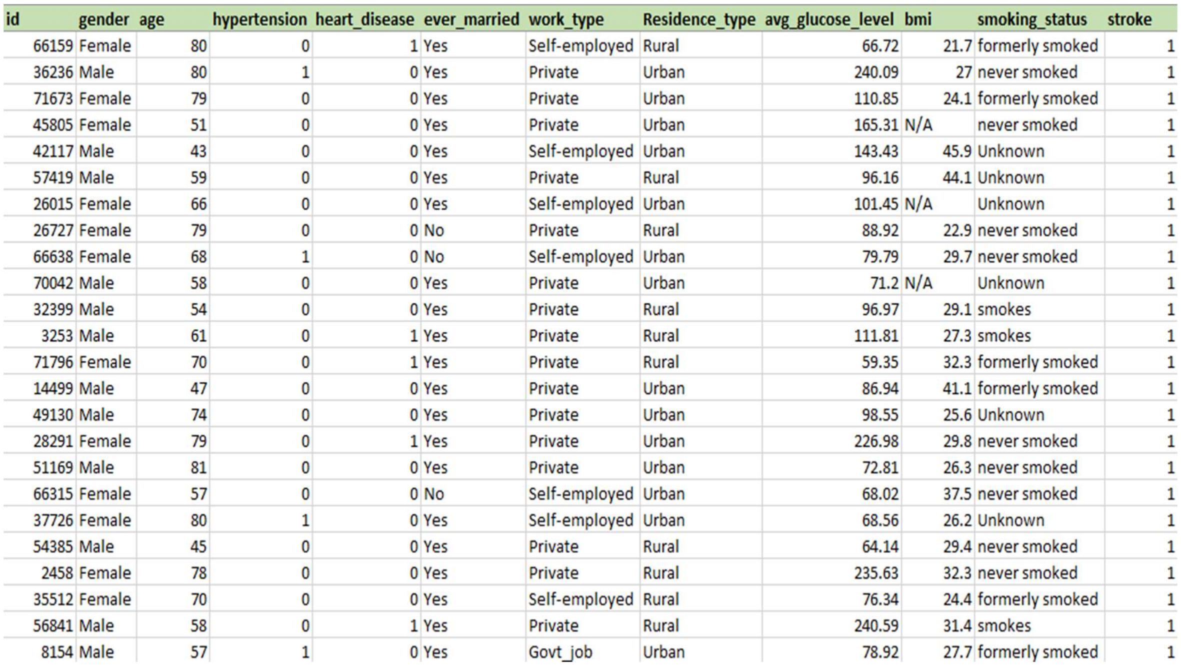
Agnes et al. [16] concentrated on using ensemble learning for brain stroke detection and clinical diagnosis.

3. METHODOLOGY

This section has been split into four parts: dataset description, data preprocessing, Ensemble Boosting algorithms, and implementation procedure. Each subsection is detailed in detail below.

3.1Dataset description

The dataset used in this study was obtained from the Kaggle data repository [17], The overall number of participants was 5110, with 2115 males, 2994 females and 1 other.As indicated in Table 1, the attributes include ID, gender, age, hypertension, heart disease, ever married, work type, home type, average glucose level, BMI, and smoking status. Stroke is the target column.The number '0' denotes the absence of any stroke risk, while the number '1' denotes the possibility of stroke risk.



id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
66159	Female	80	0	1	Yes	Self-employed	Rural	66.72	21.7	formerly smoked	1
36236	Male	80	1	0	Yes	Private	Urban	240.09	27	never smoked	1
71673	Female	79	0	0	Yes	Private	Urban	110.85	24.1	formerly smoked	1
45805	Female	51	0	0	Yes	Private	Urban	165.31	N/A	never smoked	1
42117	Male	43	0	0	Yes	Self-employed	Urban	143.43	45.9	Unknown	1
57419	Male	59	0	0	Yes	Private	Rural	96.16	44.1	Unknown	1
26015	Female	66	0	0	Yes	Self-employed	Urban	101.45	N/A	Unknown	1
26727	Female	79	0	0	No	Private	Rural	88.92	22.9	never smoked	1
66638	Female	68	1	0	No	Self-employed	Urban	79.79	29.7	never smoked	1
70042	Male	58	0	0	Yes	Private	Urban	71.2	N/A	Unknown	1
32399	Male	54	0	0	Yes	Private	Rural	96.97	29.1	smokes	1
3253	Male	61	0	1	Yes	Private	Rural	111.81	27.3	smokes	1
71796	Female	70	0	1	Yes	Private	Rural	59.35	32.3	formerly smoked	1
14499	Male	47	0	0	Yes	Private	Urban	86.94	41.1	formerly smoked	1
49130	Male	74	0	0	Yes	Private	Urban	98.55	25.6	Unknown	1
28291	Female	79	0	1	Yes	Private	Urban	226.98	29.8	never smoked	1
51169	Male	81	0	0	Yes	Private	Urban	72.81	26.3	never smoked	1
66315	Female	57	0	0	No	Self-employed	Urban	68.02	37.5	never smoked	1
37726	Female	80	1	0	Yes	Self-employed	Urban	68.56	26.2	Unknown	1
54385	Male	45	0	0	Yes	Private	Rural	64.14	29.4	never smoked	1
2458	Female	78	0	0	Yes	Private	Rural	235.63	32.3	never smoked	1
35512	Female	70	0	0	Yes	Self-employed	Rural	76.34	24.4	formerly smoked	1
56841	Male	58	0	1	Yes	Private	Rural	240.59	31.4	smokes	1
8154	Male	57	1	0	Yes	Govt_job	Urban	78.92	27.7	formerly smoked	1

Fig. 1. Stroke prediction dataset.

Table1. Dataset Attributes and Their Description

Attribute Number	**Attribute Name**	**Type (Possible Values)**	**Description**
1	id	Numeric	A unique code for the patient
2	gender	String(Male, Female)	Refers to the gender of the patient
3	age	Numeric	Define the age of the patient
4	hypertension	Numeric (0, 1)	Denotes whether the patient suffering from hypertension or not
5	heart_disease	Numeric (0, 1)	Denotes whether the patient is suffering from any heart disease or not
6	ever_married	Nominal (Yes, No)	Denotes if the patient is married or not
7	work_type	String (private, self-employed, govt_job, children)	Denotes to the work type of the patient
8	Residence_type	Nominal (Urban, Rural)	Denotes to the type of the patient's residence

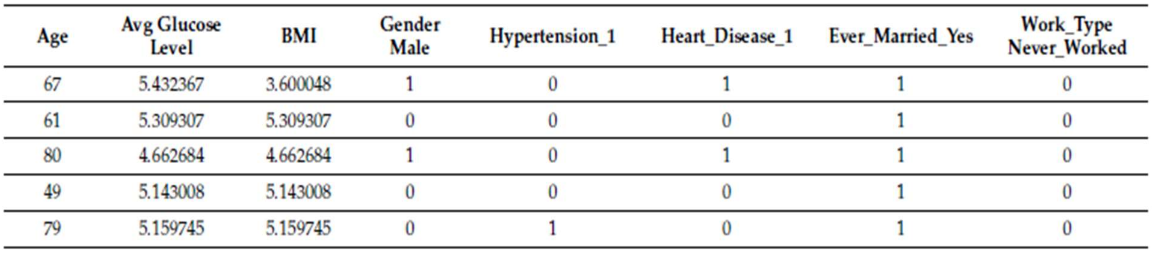
9	avg_glucose_level	Floating point number (55.12 to 271.74)	Denotes to the patient's level of blood sugar
10	bmi	Floating point number (14 to 48.9)	Denotes to the patient's body mass index
11	smoking_status	String (formerly smoked, never smoked, smokes, unknown)	Denotes whether the patient smokes or not
12	stroke	Numeric (0, 1)	Denotes whether the patient had a stroke or not

3.2 Data Preprocessing

To attain better accuracy, the data pre-processing technique is used to balance the dataset, it includes replacing the missing values, outliers' elimination, over-sampling, one-hot encoding, and normalization.

The Missing values are identified and were predicted using the mean values and are replaced. Calculating the top and lower boundaries of the feature allowed the outliers to beremoved. The interquartile range was determined, as well as the first and third quartiles, for each feature. In one-hot encoding, k binary features with just 0 or 1 potential values replace the category feature with k possible values and k > 2. The hot feature, one of these k features, is exactly equal to 1, hence the term one-hot encoding. If the categorical feature only has two possible values, 0 or 1 is substituted. Table 2 depicts the dataset used to develop the one-hot encoding approach.

Table 2. The stroke prediction dataset after applying the one-hot encoding.



Age	Avg Glucose Level	BMI	Gender Male	Hypertension_1	Heart_Disease_1	Ever_Married_Yes	Work_Type Never_Worked
67	5.432367	3.600048	1	0	1	1	0
61	5.309307	5.309307	0	0	0	1	0
80	4.662684	4.662684	1	0	1	1	0
49	5.143008	5.143008	0	0	0	1	0
79	5.159745	5.159745	0	1	0	1	0

We implemented feature scaling using Z-score normalization [18], which allows features with extremely dissimilar ranges of values to have similar ranges of values.Because the Stroke Prediction dataset's training set included 3901 rows for the no-stroke class and 187 rows for the stroke class, it was an imbalanced set due to a large skew in the distribution of the two classes.Many ML algorithms can be influenced by this bias in the training set, causing the conventional error metrics like accuracy not work very well when the ratio of positive to negative samples is significantly skewed and far from 50:50. As a result, class balancing via oversampling should be applied to the training set. To balance the samples of the two classes, we used the SMOTE to adjust the imbalanced participant distribution between two-stroke and non-stroke in the training set.

3.3 Ensemble Boosting algorithms

3.3.1 AdaBoost

AdaBoost, or adaptive boosting, is one of the most fundamental boosting algorithms. Decision trees are commonly used in modeling. Multiple sequential models are constructed, each of which corrects the errors of the previous model. AdaBoost provides weights to erroneously predicted observations, and the succeeding model works to predict these values properly [19].

At first, all observations in the dataset are assigned equal weights. A model is constructed from a subset of data. Predictions are produced on the entire dataset using this model. The predicted and actual values are compared to calculate the errors. Higher weights are assigned to data points that were erroneously predicted in the next model. The error value can be used to compute weights. For example, the greater the error, the greater the weight assigned to the observation. This approach is repeated until the error function remains constant or the maximum number of estimators is reached.

3.3.2 Gradient Boosting Machine

Gradient Boosting, also known as GBM, is another ensemble machine learning approach that may be used to solve both regression and classification problems [20]. GBM employs the boosting strategy, which involves combining several weak learners to generate a strong learner. As a base learner, regression trees are utilized, and each succeeding tree in the series is built on the errors calculated by the preceding tree.

For the overview of the GBM algorithm, we shall use a basic example. To forecast the age of a group of people.For all observations in the dataset, the mean age is considered to be the expected value. The errors are estimated using the mean prediction and the actual age values. The errors calculated above are used as the target variable in a tree model. Our goal is to discover the best split to minimize the error. This model's predictions are combined with prediction 1 and is the new prediction. This anticipated value and the actual value are used to calculate new errors. Steps are continued until the maximum number of iterations is reached (or the error function remains constant).

3.3.3 XGBoost

XGBoost is scalable, quick, and the most extensively used ML method for creating decision tree ensembles because it can handle big datasets and attain cutting-edge efficiency in many classification and regression tasks. It runs swiftly, the open-source implementations are simple to use. It is an ensemble learning technique that provides a stronger prediction by integrating the predictions of numerous weak models [20]. In XGBoost, we will evaluate the trained decision trees as well as the samples that we are still working with. When we create the next decision tree, we will pay more attention to the examples where we are failing. As a result, rather than evaluating everything, we focus more on the subset of training samples that continues to perform well. As a result of this, a new decision tree was created. Regularization is integrated into XGBoost to prevent overfitting.

3.3.4 LightGBM

It is a gradient-boosting machine-learning approach. Light GBM has numerous advantages like parallel training, sparse optimization, early termination, multiple loss functions, regularization, and bagging. The way the trees are built is a significant difference between the two. The LightGBM does not build a tree row by row, level by level. Instead, it divides into trees and chooses the leaf with the greatest reduction. Furthermore, LightGBM does not use the sorted-based decision tree learning technique, which seeks the optimal split point based on sorted feature values. LightGBM performs a much-enhanced histogram-based decision tree learning calculation, which has a considerable influence on both productivity and memory utilization. The LightGBM method employs two specific approaches known as Exclusive Feature Bundling (EFB) and Gradient-Based One-Side Sampling (GOSS) to operate faster while maintaining high precision.Light GBM beats all the other algorithms when the dataset is extremely large [21].

3.3.5 CatBoost

Handling categorical variables is a time-consuming procedure, especially when you have a large number of variables. When your categorical variables contain too many labels (i.e. they are highly cardinal), applying one-hot-encoding on them exponentially increases the dimensionality and makes working with the dataset extremely challenging. CatBoost can deal with categorical variables automatically and does not require considerable data preprocessing like other machine learning techniques.

3.4 Implementation Procedure

This study builds a prediction model using Ensemble Boosting Algorithms to predict stroke. The Kaggle repository data sets were evaluated to discover the best and most probable association between them. Fig 2. depicts the working procedureas (1) pre-processing of the Stroke Prediction dataset, (2) splitting the dataset,(3) applying the SMOTE on the training set, (4) tuning the five Ensemble Boosting techniquessuch as Adaptive Boosting (AdaBoost), Gradient Boosting, XG Boost, Light GBM, and CatBoost, (6) utilizing the measured metrics to evaluate the of prediction of stroke using various evaluating parameterssuch as accuracy score, precision score, recall score, F1-measure, and MCC.

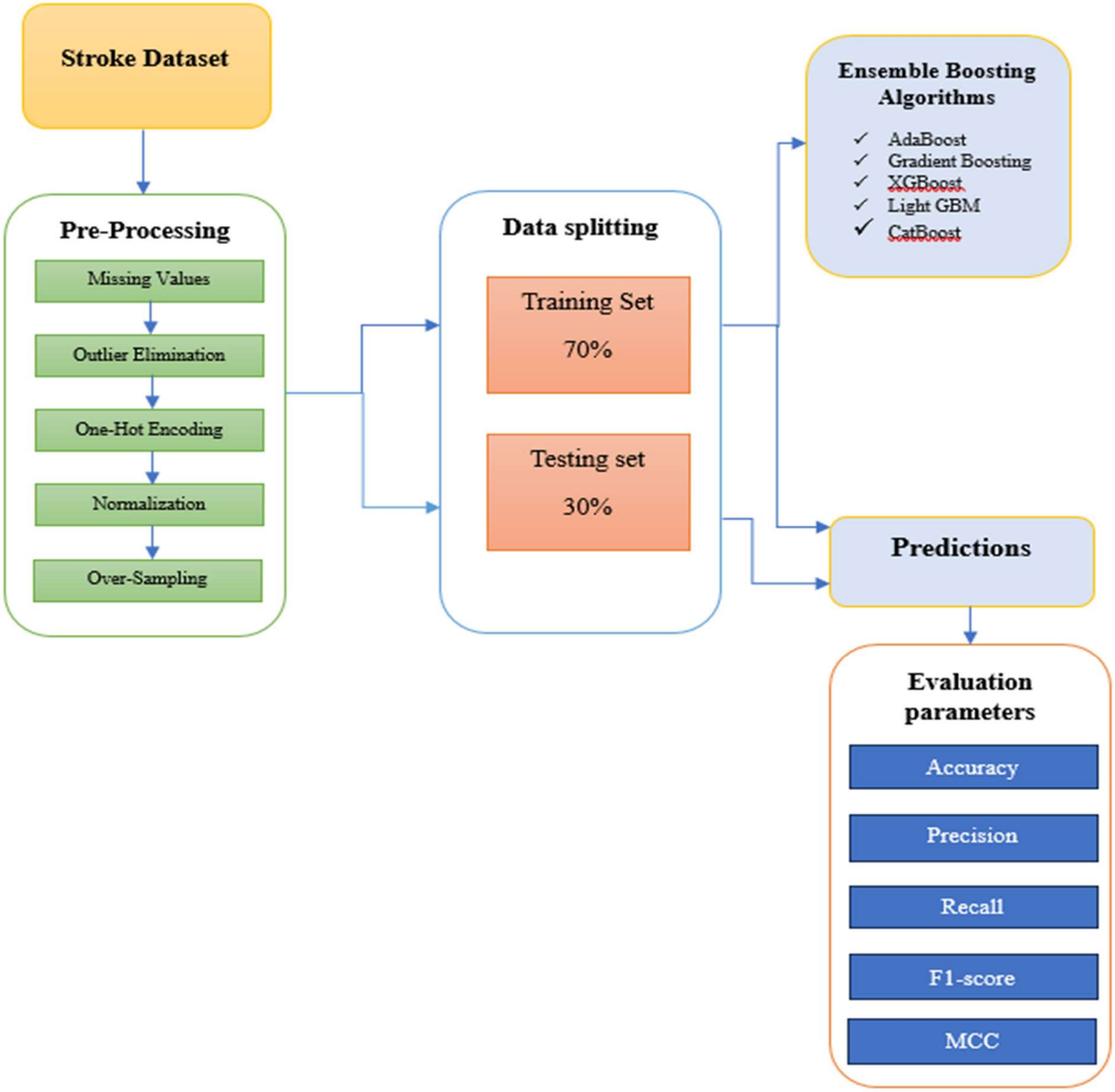


Fig. 2. Work flow of Stroke prediction using Ensemble Boosting Algorithms.

We pre-processed the dataset at the start of our experiment by replacing missing values, removing outliers, using one-hot encoding, and normalizing the data. The stroke dataset was then divided into a training set of 4088 rows (3901 no-stroke and 187 strokes) (70%) and a test set of 1022 rows (30%).To train and validate the models in this study, we used a 10-fold cross-validation technique. We employed classification techniques to train the model after using the 10-fold cross-validation technique

4. RESULT AND DISCUSSION

The Ensemble Boosting algorithms are used to predict stroke uses various evaluation parameters, such as accuracy, precision, recall, f1-score, and MCC (Matthews Correlation Coefficient) [22].

$$1. ACCURACY = \frac{TP}{TP + TN + FP + FN}$$

$$2. PRECISION = \frac{TP}{TP+FP}$$

$$3. RECALL = \frac{TP}{TP+}$$

$$4. F1 - \ SCORE = \frac{2(RECALL*PRECISION)}{RECALL+PRECISIO}$$

$$5. MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

True Positives, True Negatives, False Positives, and False Negatives are denoted by TP, TN, FP, and FN in the above equations. True positives and true negatives are accurate forecasts of whether or not a person will have a stroke. However, Number of inaccurate predictions is defined by the number of false positives and false negatives.

To train and validate the models in this study, we used a 10-fold cross-validation method. We employed Ensemble Boosting Algorithms (AdaBoost, GBM, XGBM, LightGBM, CatBoost) to train the model after using the 10-fold cross-validation method. Table 3 shows the performance results derived from the various Ensemble Boosting Algorithms (AdaBoost, GBM, XGBM, LightGBM, CatBoost) utilized to train the model.

Table 3. Performance of Ensemble Boosting Algorithm for Predicting Stroke

Ensemble Boosting Algorithms	Accuracy	Precision	Recall	F1 Score	MCC
Adaboost	0.862	0.876	0.888	0.869	0.862
Gradient Boosting	0.923	0.912	0.936	0.920	0.924
LightGBM	0.957	0.933	0.965	0.955	0.905
XGBoost	0.982	0.953	0.974	0.964	0.953
CatBoost	0.944	0.925	0.947	0.936	0.887

From the above table 3 we have evidence that AdaBoost Ensemble Algorithm as achieved 86.2%, 87.6%, 88.8%, 86.9%, and 86.2% for accuracy, precision, recall, f1-score, and MCC respectively. The Gradient Boosting has achieved 92.3%, 91.2%, 93.6%, 92.0% and 92.4% for accuracy, precision, recall, f1-score, and MCC respectively. The LightGBM has achieved 95.7%, 93.3%, 96.5%, 95.5%, and 90.5% for accuracy, precision, recall, f1-score, and MCC respectively. The XGBoost has achieved 98.2%, 95.3%, 97.4%, 96.4, and 95.3 for accuracy, precision, recall, f1-score, and MCC respectively. The CatBoost has achieved 94.4%, 92.5%, 94.7%, 93.6% and 88.7% for accuracy, precision, recall, f1-score, and MCC respectively. It so evident that XG Boost Ensemble algorithm has the highest Accuracy rate of 98.2% in predicting Ischemic stroke efficiently with BigData and LightGBM algorithm stands next with the accuracy rate of 95.7%

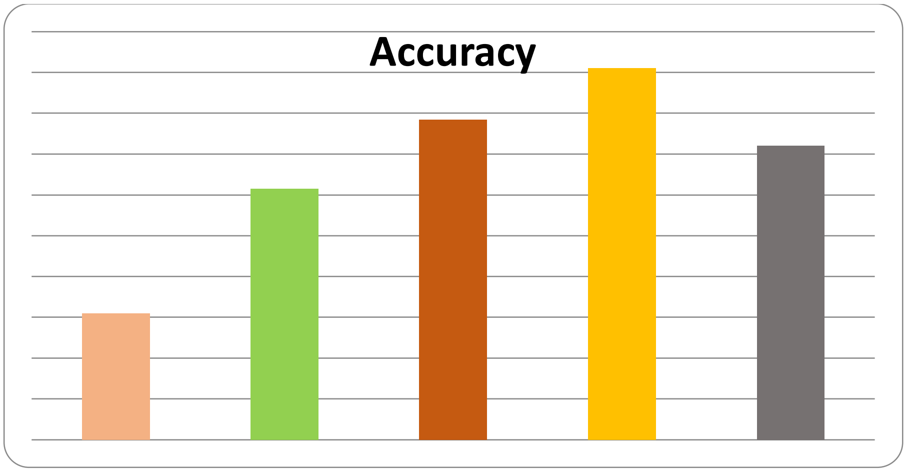


Fig 3. Accuracy rate achieved by Ensemble Boosting Algorithms

Fig 4 shows a comparison of several performance characteristics such as precision, recall, and F-Measure obtained after training different models employed in this study.

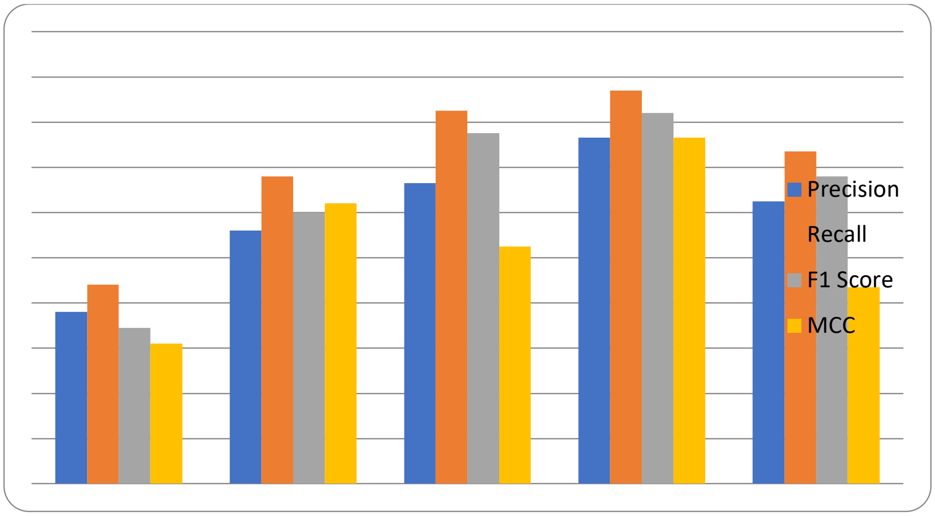


Fig 4. Other Performance Evaluation metrics achieved by Ensemble Boosting Algorithms

The main drawback of this study is that it relied on publicly available datasets. The dataset contains few attributes therefore it may not accurately predict the outcome. If real data from the hospital or medical institution with the patient's full health profile is available, a more accurate model can be developed using XGBoost ensemble algorithm for better predictions for Ischemic Stroke with BigData.

5. CONCLUSION AND FUTURE ENHANCEMENT

Our study's goal is to predict early ischemic stroke. Stroke is one of the leading causes of mortality and is becoming more common by the day. Several stroke risk factors are to blame for various types of strokes. BigData predictive analytics has proved to be very efficient in handling exploding medical data currently. The design of an Ensemble Boosting Algorithms can aid in the early detection of stroke and minimize its severe effects. These performance indicators of different Ensemble Boosting algorithms(AdaBoost, GBM, XGBM, LightGBM, CatBoost) used in this study show the effective prediction of stroke. We began with preprocessing, handling missing values, eliminating outliers, using one-hot encoding, and normalizing the features with different ranges of values. After Splitting the dataset we trained with Ensemble Boosting Algorithms and achieved high accuracy, with XGBoost performing the best at 98.2%, followed by LightGBM at 95.7%, CatBoost at 94.4%., Gradient Boosting at 92.3% and AdaBoost at 86.2%

In the future, we can extend the research by applying multiple deep learning classification techniques and designing a framework with Ensemble Boosting Algorithm to achieve efficient Ischemic Stroke Prediction from live data of the Patients from hospitals.

REFERENCES

[1] Liu Y, Luo Y, Naidech AM. Big data in stroke: how to use big data to make the next management decision. Neurotherapeutics. (2023) 20:744– 57. doi: 10.1007/s13311-023-01358-4

[2] Ung D, Kim J, Thrift AG, Cadilhac DA, Andrew NE, Sundararajan V, et al. Promising use of big data to increase the efficiency and comprehensiveness of stroke outcomes research. Stroke. 2019; 50(5):1302–9. https:// doi. org/ 10. 1161/ STROK EAHA. 118.020372.

[3] Olaronke I, Oluwaseun O, editors. Big data in healthcare: prospects, challenges and resolutions. 2016 Future Technologies Conference (FTC); 2016 6–7 Dec. 2016.

[4] CDC, "About stroke," Centers for Disease Control and Prevention, 06-May-2022. [Online]. Available: https://www.cdc.gov/stroke/about.htm.

[5] Bustamante, A.; Penalba, A.; Orset, C.; Azurmendi, L.; Llombart, V.; Simats, A.; Pecharroman, E.; Ventura, O.; Ribó, M.; Vivien, D.;et al. Blood Biomarkers to Differentiate Ischemic and Hemorrhagic Strokes. Neurology 2021, 96, 1928–1939. [CrossRef] [PubMed]

[6] "Stroke", nhs.uk,2022. [Online].Available: HTTPs://www.nhs.uk/conditions/stroke/.

[7] Jeena RS, Kumar S. Stroke prediction using SVM, International Conference on Control. Instrumentation, Communication and Computational Technologies (ICCICCT), 2016: 600–602.

[8] Hanifa SM, Raja SK. Stroke risk prediction through non-linear support vector c lassification models. Int. J. Adv. Res. Comput. Sci., 2010; 1(3).

[9] Vrdoljak J, Boban Z, Baríc D, et al. Applying explainable machine learning models for detection of breast cancer lymph node metastasis in patients eligible for neo adjuvant treatment. Cancers (Basel). 2023;15(3):634. doi:10.3390/cancers15030634

[10] O. Rado, M. Al Fanah, and E. Taktek, "Ensemble of Multiple Classification Algorithms to Predict Stroke Dataset," Advances in Intelligent Systems and

Computing, vol. 998, pp. 93–98, 2019, DOI: https://doi.org/10.1007/978-3-030-22868-2_7.

[11] Rezazadeh A, Jafarian Y, Kord A. Explainable ensemble machine learning for breast cancer diagnosis based on ultrasound image texture features. Forecasting. 2022;4(1):262-274. doi:10.3390/forecast4010015

[12] Fernandez-Lozano, C. et al. Random forest-based prediction of stroke outcome. Sci. Rep. 11, 1–12 (2021).

[13] Ntaios, G. et al. Machine-learning-derived model for the stratification of cardiovascular risk in patients with ischemic stroke. J. Stroke Cerebrovasc. Dis. 30, 106018 (2021).

[14] Brunese, Luca; Mercaldo, Francesco; Reginelli, Alfonso; Santone, Antonella (2019). An Ensemble Learning Approach for Brain Cancer Detection exploiting Radiomic Features. Computer Methods and Programs in Biomedicine, 105134, p1-45.

[15]. Patra, Amit Kumar; Ray, Ratula; Abdullah, AzianAzamimi; Dash, Satya Ranjan (2019). Prediction of Parkinsonâ€™s disease using Ensemble Machine Learning classification from acoustic analysis. Journal of Physics: Conference Series, 1372(), 012041, p1-8.

[16]. Gyorfi, Agnes; Kovacs, Levente; Szilagyi, Laszlo . IEEE International Conference on Systems, Man and Cybernetics (SMC) - Bari, Italy (2019.10.6-2019.10.9)] 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) - Brain Tumor Detection and Segmentation from Magnetic Resonance Image Data Using Ensemble Learning Methods. ,IEEE 2019,p909–914.

[17] Fedesoriano, "Stroke prediction dataset," Kaggle, 26-Jan-2021. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

[18] Tan, P.; Steinbach, M.; Karpatne, A.; Kumar, V. Introduction to Data Mining; Computers; Pearson: New York, NY, USA, 2018; pp.1–864. Available online: https://www-users.cse.umn.edu/~kumar001/dmbook/index.php (accessed on 4 February 2023).

[19] D.-C. Feng et al., "Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach."

[20] Sailasya, G.; Kumari, G. Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. Int. J. Adv. Comput.Sci. Appl. 2021, 12, 539–545.

[21] Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree, https://github.com/Microsoft/LightGBM

[22] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, 2020.