The Seybold Report

ISSN 1533-9211

DOI 10.17605/OSF.IO/B7H8R

ISCHEMIC STROKE PREDICTIVE ANALYTICS USING FUSION GRADIENT BOOSTING DECISION TREE ALGORITHM COMBINING LIGHTGBM AND XGBOOST IN BIGDATA

C. TAMILSELVI

Research scholar, Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamilnadu, India. Email: tamilselvi.cse@drmgrdu.ac.in

Dr. RAMAMOORTHY. S

Professor, Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamilnadu, India. Email: ramamoorthy.s@drmgrdu.ac.in

Dr. RAJAVARMAN. V. N

Professor, Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamilnadu, India. Email: rajavarman.vn@drmgrdu.ac.in

Abstract

Health care is a wide area where data plays a major role. Scope of Big data analytics is extremely high and important for the prediction and prevention of various diseases. The Big Data analytics is effectively applied since we have extremely diversified data form from different sources. In this paper we have proposed a fusion prediction model for Ischemic Stroke based on the Light GBM algorithm and the XGBoost algorithm with GBDT. The prediction model is built using the user's physical examination data and major biomedical indicator like blood pressure, BMI, Heart diseases, Avg glucose level, and Cigarette smoking are used as the auxiliary judgment criteria. The LightGBM model's predicted results and the XGBoost model's expected results are then both fed into the GBDT model. Finally, the fusion of the two prediction results is obtained. This fusion prediction model results are compared with various other machine learning classifiers. The experimental result proves that the fusion predictive model is more accurate than the other classifiers.

Keywords: Big Data Analytics, Light GBM, XG Boost, GBDT

1. INTRODUCTION

Stroke is a medical condition in which the blood vessels in the brain rupture, causing brain damage. Symptoms may appear if the brain's flow of blood and other nutrients is disrupted. Stroke is the leading cause of death and disability worldwide, according to the World Health Organization (WHO) almost fifteen million people suffer from stroke globally every year, out of which one person dying every four to five minutes [1]. Early awareness of the numerous stroke warning symptoms can assist to lessen the severity of the stroke. The two most common forms of strokes are ischemic and hemorrhagic. Ischemic stroke occurs when a blockage reduces or disrupts blood supply to brain cells, destroying the cells in minutes and resulting in death. Hemorrhagic stroke, on the other hand, happens when weak blood arteries are extensively damaged as a result of hypertension, excessive cholesterol, and other risk factors [2]. To forecast the likelihood of a stroke happening in the brain, many Big data predictive analytics and machine learning (ML) models have been developed. Strokes are caused by a

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/B7H8R

number of risk factors, including medical issues such as high blood pressure, heart disease, diabetes, high cholesterol, and atrial fibrillation, as well as bad habits such as smoking, obesity, unhealthy meals, and a lack of physical activity [3].

Globally, the exponential growth of medical data is unsustainable since it can only be handled through BigData analytics. Along with volume, velocity of the data at which it is created is accelerated at the great level. BigData Analytics as primary characteristics such as volume, velocity, variety, veracity and value of data can deal with the rapidly growing medical data. Predictive analytics is a form of BigData analytics that uses historical data, statistical modeling, data mining techniques, and machine learning to create predictions about future happenings. Predictive analytics provides the intelligence about the future using the insight of BigData.

The early prediction of brain stroke impact in order to address risk factors is regarded as a lifesaving matter. Because of considerable progress in predicting different diseases, Big Data Analytics, machine learning and AI approaches can be used to predict the chance of a stroke developing. Various categorization methods have been employed with reasonable results to predict strokes [4] [5] [6]. Because of its accuracy in predicting various diseases, the ensemble approach is widely employed in medical applications [7] [8] [9]. These methods combine the predictions of different classification models to boost overall performance. The classification process in stacking is done in two stages: the first is training several base models on the full dataset, and the second is training a meta-learner classifier on the first layer's prediction outcomes to deliver the final prediction [10].

The main goal of this paper is to propose a fusion model with GBDT which combine LightGBM and XGBoost algorithms to predict effective brain stroke with BigData

2. RELATED WORK

Ali et al. [11] enhanced their stroke prediction model using distributed machine learning algorithms and Apache Spark, a well-known big data platform. A Decision Tree, Support Vector Machine, Random Forest, and Logistic Regression classifiers were used to build the prediction model. They employed a healthcare stroke dataset in their research. They used accuracy, precision, recall, and the f1-measure to assess the model's performance. Random Forest has the highest accuracy of 90% among all classifiers.

Sailasya and Kumari [12] used six machine learning classifiers to train their stroke prediction model: logistic regression, decision tree, random forest, K-nearest neighbour, support vector machine, and Nave Bayes. They used a dataset comprising stroke risk variables. They also created an HTML page as a user interface to collect the user's stroke parameter values and offer him with the prediction result. They used the F1 score, accuracy, precision, and recall to evaluate overall performance. The results show that the Nave Bayes classifier achieved the best accuracy of 82% when compared to the other classifiers utilized. The achieved accuracy is insufficient to forecast such a serious medical condition.

Nwosu et al. [13] used multiple machine-learning classifiers to create a prediction model for brain strokes. The prediction model was created by combining three classifiers: a neural

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/B7H8R

network (multi-layer perception), a decision tree, and a random forest. Using the neural network classifier, they reached 75% accuracy. The primary goal of any medical prediction model is to improve accuracy, yet their outcomes in this study are insufficient to be trusted.

Mahesh and Srikanth [14] intended to utilize decision trees, naive Bayes, and artificial neural network classification algorithms for machine learning to create a stroke prediction model. Their research focuses on the effects of modifiable and non-modifiable risk factors for stroke. High blood pressure, smoking, and other risk factors are included in the data collection. The AUC (area under the curve) and ROC (receiver operating characteristics) are used to assess the overall performance of prediction models. The better the prognosis, the higher the AUC result. Their findings suggest that the three algorithms provide adequate prediction accuracy. The user interface for providing stroke risk alerts was a web application. The AUC_ROC score cannot be used to evaluate a prediction model on its own.

Monteiro et al. [15] developed a predictive model for stroke functional diagnosis. A total of 541 patient's data were included in the study. The prediction model was built using popular algorithms such as logistic regression, decision trees, support vector machines, random forests, and XGBoost. The final performance of the models was evaluated using the area under the curve (AUC), which was greater than 90%.

Rado et al. [16] 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%.

Jeena and Kumar [17] designed a stroke prediction model that predicts the chance of having a stroke based on a variety of risk conditions. Age of the model, atrial fibrillation, gait problems sight impairment, and so on. Support vector machine classification was used to develop predictive models using various kernel functions such as linear, quadratic, RBF, and polynomial. The linear kernel function was the most accurate, with 91% accuracy. The key disadvantage here is that the database size is insufficient to make the prediction findings more credible, and it only has 350 cases. However, the unbalance in the stroke dataset was not taken into account in this study, resulting in erroneous conclusions.

3. PROPOSED WORK

1.1 Method

XGBoost Algorithm

The XGBoost algorithm [18] is on if the renowned gradient boosting Techniques with improved performance and speed. XGBoost is classified as a boosting strategy in Ensemble Learning. Ensemble learning combines different models into a collection of predictors to improve prediction accuracy.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/B7H8R

In the boosting methodology the defects created by prior models are attempted to be repaired by subsequent models by adding weights to the models. It is a finest implementation of the GBDT algorithm [19]. XGBoost is a more advanced version of gradient boosting which includes regularization factors that helps in minimizing overfitting. The XGBoost algorithm supports parallel processing with efficient memory management for BigData. In addition to the above features, they can handle missing values, takes care of outliers and auto tree- pruning. XGboost has been recognized to be the most efficient Scalable Tree Boosting Method.

LightGBM Algorithm

LightGBM mainlycomprises two algorithms, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

The Gradient-based One-Side Sampling (GOSS) algorithm ignores most samples with small gradients and calculates the information gain using only the remaining samples. Despite the fact that the GBDT method has no data weight, according to the concept of information gain [20], each data instance has a different gradient. The information gain is more impacted by the large gradient instance. As a result, while sampling, we should keep the sample with a large gradient. The threshold can be preset or samples with small gradients can be removed at random. Researchers have indicated that this method's measurement findings are more accurate than random sampling results at the same sample rate, especially when the range of information gain is large.

In Exclusive Feature Bundling (EFB) various features are almost mutually exclusive in the sparse feature space; for example, many features rarely use non-zero values at the same time. Usually, it is in the application, despite the fact that the number of features is higher; nevertheless, because the feature space is so limited, can you develop a lossless way to lower the effective features? Many characteristics are almost mutually exclusive, especially in the limited feature area (for example, many features are not non-zero at the same time, like a hot spot). We can combine features that are mutually exclusive into a single feature. Then, using the greedy technique, we reduce the bundling problem to a graph coloring problem and obtain an approximate solution.

Fusion Model Using GBDT Algorithm

Friedman et al. introduced the GBDT method in 2001 [20], which is a boosting algorithm. When building the model, the GBDT algorithm uses the previously established negative gradient direction of the model loss function, and then iteratively improves the model's accuracy.

Gradient Boosting Decision Tree (GBDT)

GBDT, a composite algorithm, is composed of a number of linear sub model combinations. It uses the regression tree as the sub model and iterates based on the iteration principle then adds sub models one at a time to reduce loss function.

GBDT can be expressed as equation (1)

ISSN 1533-9211

DOI 10.17605/OSF.IO/B7H8R

$$F_i(X) = \sum_{i-1}^{k} f_i(x|\theta_i) \quad (1)$$

Where $f_i(x|\theta_i)$ is the is the regression tree sub model that was introduced in the i-th iteration, θ_i is the parameter sub model, k is the number of sub models, x is the data sample.

θ_i Is gained by optimising the loss function by the equation (2)

$$\theta_i = \min L[F_{i-1}(x) + f_i(x|\theta_i)] \quad (2)$$

Where *L* is the loss function used for prediction

GBDT includes a pre-sorting technique for feature selection and splitting, making it time and memory consuming and unsuitable for huge processing. LightGBM, on the other hand, replaces GBDT's pre-sorting technique and layer growth strategy with histogram algorithms and leaf growth strategies, significantly boosting the algorithm's speed and efficiency [21].

1.2 Modeling Process

The proposed prediction fusion model was developed using a Kaggle stroke prediction dataset [22], as shown in Fig. 1. As indicated in Table I, 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.

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
48405	Male	80	0	1	Yes	Private	Urban	68.53	24.2	smokes	1
36706	Female	76	0	0	Yes	Self-employed	Urban	106.41	N/A	formerly smoked	1
41069	Female	45	0	0	Yes	Private	Rural	224.1	56.6	never smoked	1
71639	Female	68	0	0	No	Govt_job	Urban	82.1	27.1	Unknown	1
53401	Male	71	1	1	No	Govt_job	Rural	216.94	30.9	never smoked	1
60744	Male	61	1	0	Yes	Self-employed	Rural	76.11	27.3	smokes	1
7547	Male	74	0	0	Yes	Private	Urban	72.96	31.3	smokes	1
31720	Female	38	0	0	No	Self-employed	Urban	82.28	24	formerly smoked	1
5563	Female	77	0	0	Yes	Private	Urban	105.22	31	never smoked	1
68798	Female	58	0	0	Yes	Private	Rural	59.86	28	formerly smoked	1
72918	Female	53	1	0	Yes	Private	Urban	62.55	30.3	Unknown	1
13491	Male	80	0	0	Yes	Private	Rural	259.63	31.7	smokes	1

Figure 1: Stroke Prediction Dataset

Table 1. Dataset Attributes and Their Description

Attribute Number	Attribute Name	Description
1	id	A unique code for the patient
2	gender	Refers to the gender of the patient
3	age	Define the age of the patient
4	hypertension	Denotes whether the patient suffering from hypertension or not
5	heart_disease	Denotes whether the patient is suffering from any heart disease or not
6	ever_married	Denotes if the patient is married or not
7	work_type	Denotes to the work type of the patient
8	Residence_type	Denotes to the type of the patient's residence
9	avg_glucose_level	Denotes to the patient's level of blood sugar
10	bmi	Denotes to the patient's body mass index
11	smoking_status	Denotes whether the patient smokes or not
12	stroke	Denotes whether the patient had a stroke or not

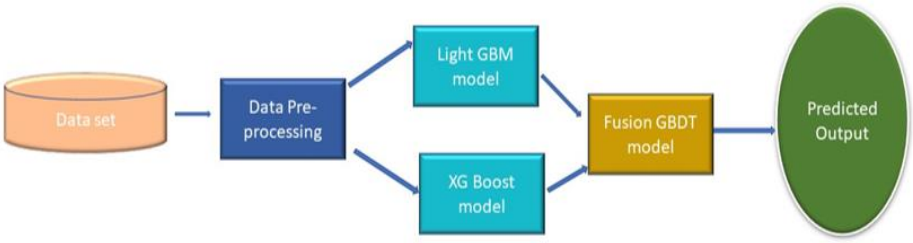
The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/B7H8R

Using predictive analytics on the above-mentioned processed dataset, we predict the findings as the target variable to assess the likelihood of users developing stroke. We train and develop the XGBoost and LightGBM models using the XGBoost and LightGBM algorithms, and then use the GBDT algorithm to fuse the outcomes of the two models to produce the final prediction results.

The Initial step of Data Preprocessing includes, normalizing the data, filling the missing values, eliminating the outliers, duplicates and sampling. After data preprocessing, we train the model using the XGBoost and LightGBM algorithms, respectively. The hyperparameter values are calculated using the cross-validation approach [23]. Hyperparameter tuning is the method of identifying the best hyperparameters for the classifier. It trials numerous combinations of the parameter values and identify the optimal values that gives maximum accuracy of the prediction model.

One of the most effective machine-learning techniques is stacking. It is a popular ensemble technique because it enhances model performance and solves challenging issues. A meta-model is used to aggregate predictions from multiple models. Stacking divides the dataset into two sets, the first being the training set and the second being the test set. This training set is separated into two parts: a training set provided by heterogeneous base learners to generate the first-level models and a validation set used by the models to provide level-one predictions, which are used as additional features for the second-level meta-learner. This meta-learner is trained using the newly acquired training data, which comprises of the first-level predictions, and makes the final prediction using the test set. The main goal is to build a meta-model that is trained on first-level results. This phase contributes to a more accurate final prediction. Fig. 2 shows how The LightGBM model's predicted results and the XGBoost model's expected results are then both fed into the Fusion GBDT model. Finally, the final stroke predictions are obtained.

Figure 2: Stroke prediction model using Fusion GBDT



K-Fold Cross-Validation

During the process of developing the fusion model, the dataset is divided into K collections of similar size using k-fold cross-validation, where K is an integer number. Folds are the name given to these collections. Making the number of iterations equal to the number of folds. Every iteration considers training with k-1 folds and validating with k-folds, changing the training

ISSN 1533-9211

DOI 10.17605/OSF.IO/B7H8R

and validating folds. Accuracy is calculated at every iteration and the average is calculated at the end.

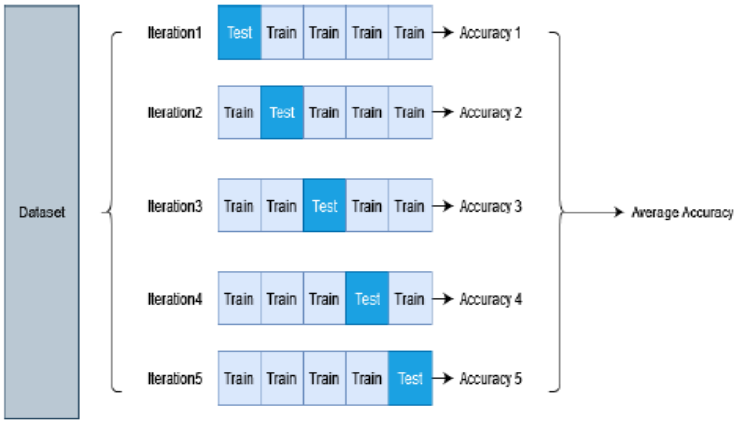


Figure 3: Sample Cross-Validation with k=5

4. RESULT AND DISCUSSION

The proposed fusion model uses various evaluation parameters, such as accuracy, precision, recall, f1-score, and MCC (Matthews Correlation Coefficient) [24]. The MCC value was considered for the classifiers in this paper since it is an effective metric for binary classification and unbalanced datasets, such as the used stroke dataset. It computes the correlation between the observed and expected values. If the correlation score is higher, the prediction is more accurate. It took into account all of the confusion matrix values. When the MCC value is near one, it indicates that the model accurately predicted both the actual and projected values.

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

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

$$3.\,\mathrm{RECALL} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

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

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

In the above equations, TP, TN, FP, and FN stand for True Positives, True Negatives, False Positives, and False Negatives, respectively. True positives and true negatives represent accurate predictions of whether or not an individual has stroke. However, the number of incorrect predictions made by the proposed model is determined by the number of false positives and false negatives. The classification methods were compared using the accuracy,

ISSN 1533-9211

DOI 10.17605/OSF.IO/B7H8R

precision, recall, f1 score, and MCC measures after developing the prediction model, as shown in Table II and Fig. 4:

Table 2: Comparison between the Base Classifiers and the Proposed Fusion Model

	Accuracy	Precision	Recall	F1 Score	MCC
SVM	0.90	0.91	0.93	0.94	0.85
Naïve Bayes	0.74	0.75	0.8	0.78	0.5
Logistic Regression	0.74	0.74	0.8	0.75	0.54
Random Forest	0.95	0.93	0.94	0.94	0.92
KNN	0.95	0.91	0.96	0.95	0.89
LightGBM	0.97	0.92	0.98	0.98	0.93
XGBoost	0.96	0.95	0.98	0.97	0.94
Fusion model	0.98	0.97	0.99	0.99	0.95

Figure 4: Comparison the Fusion Model with Other Classifiers

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/B7H8R

From table II, the fusion model scored the highest accuracy compared to the other standalone classifiers, with an accuracy of nearly 98%, illustrating the efficiency of the ensemble approaches. It also had an excellent MCC value of 95%, indicating that the fusion model provides accurate predictions because it accurately predicts both the actual and predicted values.

To predict the brain stroke with Bigdata the fusion model was proposed which had proved to achieve higher accuracy with comparison of various standalone classifiers. The graphical comparison of the predicted fusion model and various machine learning classifiers are shown in the below fig. 4

5. CONCLUSION AND FUTURE ENHANCEMENT

This paper demonstrated the fusion model for predicting brain stroke with the Big data. To enhance the prediction the new fusion model which used GBDT which in turn combines the results of LightGBM and XGBoost algorithms. According to the experimental results, utilizing a fusion model can greatly boost prediction accuracy, as it obtained approximately 98% and provided the highest MCC measure of approximately 95%, ensuring that the prediction is accurate. With the above finding this model can predict whether or not someone will have a brain stroke.

This study's future scope will involve using other combinations of the base model classifiers in the stacking model and to increase the size of the data randomly. It may also include using additional effective attributes to develop the prediction model.

References

1. Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis, and Mohammad Monirujjaman Khan , Stroke Disease Detection and Prediction Using Robust Learning Approaches, Hindawi Journal of Healthcare Engineering Volume 2021, Article ID 7633381, 12 pages https://doi.org/10.1155/2021/7633381
2. CDC, “About stroke,” Centers for Disease Control and Prevention, 06-May-2022. [Online]. Available: https://www.cdc.gov/stroke/about.htm.
3. "Stroke”, nhs.uk,2022. [Online]. Available: HTTPs://www.nhs.uk/conditions/stroke/.
4. A. Roy, A. Kumar, K. Singh, and D. Shashank, “Stroke Prediction using Decision Trees in Artificial Intelligence. Stroke Prediction Using Decision Trees in Artificial Intelligence,” IJARIIT, vol. 4, pp. 1636–1642, 2018.
5. B. Khalid and N. Abdelwahab, “A model for predicting Ischemic stroke using Data Mining algorithms,” IJISET, vol. 2, no. 11, 2015.
6. M. Rajora, M. Rathod, and N. S. Naik, “Stroke prediction using machine learning in a distributed environment,” in Distributed Computing and Internet Technology, Cham: Springer International Publishing, 2021, pp. 238–252.
7. I. D. Mienye, Y. Sun, and Z. Wang, “An improved ensemble learning approach for the prediction of heart disease risk,” Inform. Med. Unlocked, vol. 20, no. 100402, p. 100402, 2020.
8. K. Shilpa and T. Adilakshmi, “Applying ensemble techniques of machine learning to predict heart disease,”

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/B7H8R

in Proceedings of the International Conference on Cognitive and Intelligent Computing, Singapore: Springer Nature Singapore, 2022, pp. 775–783.

9. Z. Asghari Varzaneh, M. Shanbehzadeh, and H. Kazemi-Arpanahi, “Prediction of successful aging using ensemble machine learning algorithms,” BMC Med. Inform. Decis. Mak., vol. 22, no. 1, p. 258, 2022.
10. J. Brownlee, “Stacking ensemble machine learning with python,” Machinelearningmastery.com, 09-Apr-2020. [Online]. Available:
11. Ali, Abdelmgeid A., “Stroke Prediction using Distributed Machine Learning Based on Apache Spark,” Stroke 28(15), pp. 89-97, 2019.
12. G. Sailasya and G. L. A. Kumari, “Analyzing the performance of stroke prediction using ML classification algorithms,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, 2021.
13. C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John, “Predicting Stroke from Electronic Health Records,” Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2019, pp. 5704–5707, Jul. 2019, DOI: https://doi.org/10.1109/EMBC.2019.8857234.
14. M. Kunder Akash and S. Srikanth, “Prediction of Stroke Using Machine Learning,” ResearchGate, 2020.
15. Monteiro, M., Fonseca, A. C., Freitas, A. T., Pinho e Melo, T., Francisco, A. P., Ferro, J. M. and Oliveira, A. L., "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients", 2018.
16. 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.
17. R. S. Jeena and S. Kumar, “Stroke prediction using SVM,” IEEE Xplore, Dec. 01, 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7988020.
18. T Chen, C Guestrin (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
19. FRIEDMAN J H (2001). Greedy function approximation: a gradient boosting machine．The Annals of Statistics，29(5): 1189-1232．
20. Kent J T (1983). Information gain and a general measure of correlation[J]. Biometrika, 70(1): 163-173.
21. Lulu Liang, Wei Hu, Yiwei Zhang, Kun Ma, Yujia Gu, Bei Tian, Hongqiang Li. An algorithm with LightGBM + SVM fusion model for the assessment of dynamic security region. E3S Web of Conferences **256**, 02022 (2021). https://doi.org/10.1051/e3sconf/202125602022
22. Fedesoriano, “Stroke prediction dataset,” Kaggle, 26-Jan-2021. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.
23. C.W. Hsu, C.J. Lin (2002). A comprasion of methods for multiclass support vector machines. IEEE transactions on Neural Networks, 13(2): 415–425.
24. 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.