Unveiling the Innate Potential of Ensemble Techniques in Advanced Brain Stroke Classification

Preet Singh
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
s.preet@chitkara.edu.in

Taniya Hasija
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
taniya@chitkara.edu.in

KR Ramkumar
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
k.ramkumar@chitkara.edu.in

2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE) | 979-8-3503-7700-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSPCRE62303.2024.10675075

***Abstract*— The emerging scope of new technology in the classification and early prediction of various diseases is increasing. Artificial Intelligence (AI) and Machine Learning (ML) techniques have automated data analysis, to recognize hidden patterns, and trend identification in large datasets, which leads to the development of new diagnostic and early prediction methods. In this research article, the brain stroke classification analysis is done using the ML-supervised algorithms: XGBoost, CatBoost and LGBoost. To enhance the performance of the classification analysis the Weighted Voting Ensemble Classifier is used. This ensemble technique improves the prediction performance by combining the accuracy of these models, by assigning the weights. The Weighted Voting Ensemble significantly increases the accuracy by up to 96.7% as compared to the other classifiers. This results in the utilization of a best-prediction model for the categorical analysis of brain stroke. The enhanced performance can anticipate the early prediction of the disease and take the necessary precautionary measures to lead a sustainable lifestyle.**

Keywords*— *Brain Stroke, Supervised Machine Learning, Healthy and Sustainable Lifestyle, Weighted Voting Ensemble Algorithm, SMOTE Analysis.

I. Introduction

Brain Stroke is a condition under which the supply of oxygenated blood becomes less to that part of the brain and it starts to malfunction. It happens when a brain region cannot get the required oxygen supply because of a blockage in the artery supply [1]. This blockage can be caused by a clot or ruptured blood vessels. The symptoms of the stroke vary from speaking disability to mobility issues and even paralysis. This occurs based on which part of the brain is affected. When the brain is deprived of oxygen for an extended period, the cells die, leading to permanent damage to the affected areas. Thus, a part of the brain is damaged, and the surrounding cells lose their vitality and stop functioning properly, which can lead to malfunction of the affected area of the body. Stroke is categorised into three main types: Ischemic stroke, haemorrhagic stroke and Transient ischemic attack [2]. The term "ischemic stroke" is a condition in which arteries contract and get clogged with fat deposits in the walls. In ischemic stroke, arteries may get clogged in either the n-sector (neck) or the m-sector (brain). The second form of stroke is a haemorrhagic stroke, which is caused by bleeding into and around the brain [2,3]. The next stroke type called Transient Ischemic Attack, is an ischemic condition that may last from a few minutes to 24 hours. Early diagnosis must be required like CT scans and MRI images of the brain. But nowadays Artificial Intelligence (AI) and Machine Learning (ML) algorithms [5-7] make it easy for the early prediction of these diseases in patients with certain common conditions. In this paper, the classification analysis of brain stroke is done using XGBoost, CatBoost and LGBoost. To enhance the accuracy and effectiveness of brain stroke categorization, we propose the implementation of the Weighted Voting Ensemble Technique. In this paper, Section II elaborates on the machine-learning approaches along with a literature review in Section III. Implementation methodology and results are given in Section IV and Section V. The results are concluded in section VI.

II. Machine Learning

ML is a fundamental building block of AI that facilitates the handling of huge datasets and tracing the patterns in the dataset for detailed analysis and forecasting the future values. ML is further divided into two parts that are supervised and unsupervised learning. The Supervised Learning machine will understand the patterns by calibrating the labels and predicting the output. It is further classified into classification and regression. The regression tasks involve the prediction of the dependent variables based on the independent variables. The relationship between these variables is analyzed using regression analysis like linear regression, and multi-linear regression [9]. The classification algorithm is based on the categorization of labelled data. The data is set to be categorized for the prediction. All the datasets of similar categories are to be classified under the same class and later the prediction of discrete categories is done [10]. The Gradient Boosting algorithm (GBA) is an ML-based classification algorithm. GBA is an approach under which the predictions of various weak models are combined to create a strong model for the prediction analysis. Weak Model training is the first step involved in this algorithm. The algorithm utilizes the ensemble technique in which the predictions of the several weak models are combined and the resultant output has less incorporated error, considerably improving the accuracy of the model. A decision tree is used to predict the categorization task on training data. This first model makes certain predictions that may not be perfect. So, the difference between the actual values and predicted values is calculated which is known as residuals [8]. In Gradient Boosting Algorithm, a new model is trained using the residuals from the previous model. This process is repeated until a strong model is obtained, which can make highly accurate predictions [9-10].

1) Extreme Gradient Boosting (XGBoost): The technique combines a gradient-boosted algorithm and decision trees. In this machine-learning approach, the weights assigned to the independent variables decide the decision tree. A categorical analysis is carried out by combining the prediction values of

979-8-3503-7700-2/24/$31.00 ©2024 IEEE

many decision trees. The XGBoost algorithm [11] is employed to examine large datasets. The tree grouping method employs a binary decision-making approach, where trees are categorized as either yes or no. This procedure may sometimes get intricate in many scenarios. In order to mitigate these issues, the XGBoost algorithm incorporates tree rectification to enhance the efficiency of the decision-making process. The memory utilization of this device is minimal, hence enhancing its performance. The training model employs the bagging technique, followed by the creation of decision trees. The resulting outputs are then integrated [11-12].

2) Categorical Boosting (CatBoost): The Categorical Boosting method is a supervised machine learning strategy that utilizes diverse categorical data for the purpose of categorization. CATBoost is built around the Gradient boosting technique pioneered by Yandex. The CATBoost [11] has three primary processes. The initial stage entails transforming the data into a binarized format. The next stage is transforming the category data into numerical representation. The third step includes the implementation of features via the decision tree method. The development of trees relies on a decision-making process that efficiently calculates the end-node leaf using the best possible criteria, ensuring that no category is overlooked in the decision-making process. CATBoost concurrently generates the training and testing datasets to enable effective working of the algorithm in creating randomized categorization under the decision tree. It also entails determining the number of iterations required to train the dataset by allowing the number of epochs to be adjusted based on a yes or no condition [11-13].

3) Light Gradient Boosting (LGBoost): This is a machine learning ensemble strategy that involves integrating the forecasting accuracy of many base models. The model's precision improved dramatically. This technique assigns weights based on the accuracy of the foundational models and the cumulative sum of their accuracies.The aggregate method improves the accuracy of forecasts by combining the forecasts of multiple models using the weights generated by the base models [14]. The Light Gradient Boosting method is a simple machine learning technique that uses decision trees to enhance the model's decision capability.It achieves this by acquiring knowledge of initial decision trees. Furthermore, there is the benefit of reduced memory usage for data loading. The decision procedure requires a significant amount of time to input the data, which is beneficial considering the issues associated with other gradient-based robustness approaches that need to iterate over the complete data numerous times. 12-13] in the table. The decision tree is extended using a leaf-by-leaf tree growth chart, resulting in an efficient and effective decision-making process.

4) Weighted Voting Ensemble Algorithm (WVE): This is a machine learning ensemble strategy that involves integrating the forecasting accuracy of many base models. The model's precision improved dramatically. This technique assigns weights based on the accuracy of the foundational models and the cumulative sum of their accuracies. The ensemble approach improves the accuracy of predictions by integrating the predictions of many models using weights generated by the base models [14].

III. Relevant Literature

The study done by various authors using different ML algorithms to predict brain stroke is given in Table I. Many attributes used for the prediction analysis have been collected from the hospitals and clinics.

TABLE I. Literature Review Summary

Year / Reference	Name of the Author	Name of the Model	Summary	Accuracy
2020 / [8]	Yu et al.	Decision Tree algorithm (DTA)	From the National Institute of Health Stroke Scale, the authors obtained the dataset. For training and testing, an additional 25% and 75% of the data are separated. The model's accuracy is determined to be 91.11%.	91.11%
2020 / [9]	Govindarajan et al.	Artificial Neural Network (ANN)	The authors conducted the study of 507 patients of the Tamil Nadu hospital and applied the ANN algorithm by using the stochastic gradient descent algorithm and this model gives an accuracy of 95%.	95%
2017 / [10]	Singh et al.	Back Propagation Neural Network (BPNN), DTA	In this research article, the authors used the dataset from the Cardiovascular Health Study and applied the ML algorithm DTA and BPNN. By comparing and contrasting the accuracy of these models, the better model is the BPNN with an accuracy of 96%.	96%
2017 / [11]	Chin et al.	Convolutional Neural Network (CNN) using the DL approach	The authors used various techniques to improve CT scan images of patients' brains. CT images were preprocessed before feeding them to the model for training and testing. The model achieved 95% training accuracy and 92.6% testing accuracy.	95%
2023 / [12]	Rahman et al.	LR, SVM, KNN, ANN (3,4 Layers), XGBoost, RandomForest	This research article used a Kaggle dataset to analyse stroke prediction via classification. The RandomForest classifier had the highest accuracy rate of 99% among the ML models. The best deep learning model was the 4-layered ANN with an accuracy rate of 92%.	92%

2022 / [13]	Tusher et al.	LR, KNN, SVM	The Authors in this article perform the early-prediction analysis of brain stroke using the classification models. A model which outperforms others is KNN with an accuracy of 96.45%.	96.45%
2022 / [15]	Kaur et al.	LSTM, FFNN	In this research article, the authors used the results of EEG for the early-prediction analysis of brain stroke. The ML approach used by researchers is LSTM and FFNN with the accuracy of 87% and 83% respectively.	87%
2016 / [16]	Adam et al.	KNN and Decision Tree	The authors used the KNN and Decision tree for the early forecast of ischemic brain Stroke disease for the prevention and early diagnosis. The dataset is collected from the hospital database of Sudan. The dataset used is then split into 70-30%. The choice tree model outperforms the other model as the former's accuracy is higher than the latter's.	93%

IV. Methodology

A. Dataset

The dataset for the categorical analysis is collected from the Kaggle, Open Database Repository [17]. The various parameters are taken into account for this analysis. The exploratory data analysis shows some insightful patterns in the dataset. As depicted in fig. 1, statistical data shows older patients have a high probability of Brain Stroke. The next factor which influences the chances of brain stroke is hypertension. Fig. 2 shows that patients with hypertension have a high chance of Brain Stroke. The statistical analysis of the datasets shows in Fig 3 that the patients with a chance of heart disease are also prone to brain stroke. This shows a strong correlation between the likelihood of brain stroke occurrence and the individuals who are at higher risk of developing heart disease.

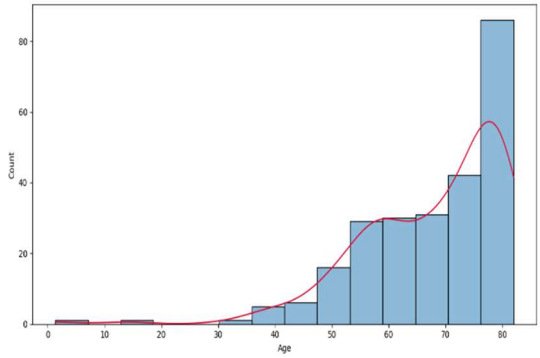

Fig. 1. Role of Age in Brain Stroke

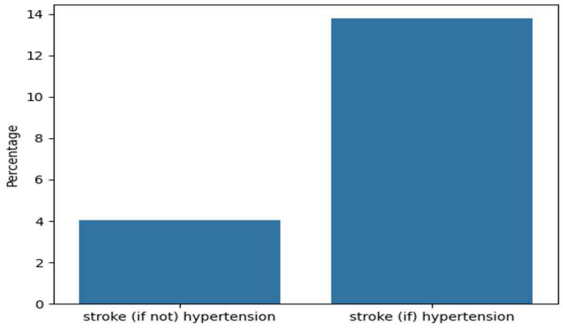

Fig. 2. Brain Stroke and Hypertension

B. Classification Approach

ML supervised learning technique involves the classification based on the gradient-based approach used for brain stroke categorization. The labelled values linked to the patient with stroke and no stroke are class 1 and class 0 respectively. The model in classification attempts to predict the appropriate label of a specified dataset through supervised classification techniques. Before using the classification-based algorithms to forecast newly discovered data, it must be thoroughly trained by training data and assessed using test data. The classifiers are further divided into two parts Lazy Learners and Eager Learners based on the ability to learn the pattern in the dataset.

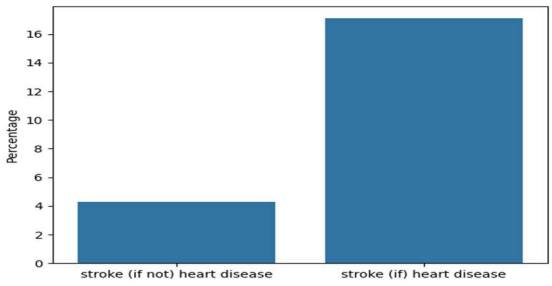

Fig. 3. Heart Disease and Brain Stroke

C. Performance-Evaluation Metrics

1) Confusion Matrix (ConfM): The ConfM is a tool used in evaluating the performance of an ML model. It consists of True-Positive (TP), False-Negative (FN), False-Positive (FP) and True-Negative (TN).

2) Precision: As defined by equation 3, precision is defined as the total percentage of correct values predicted by the supervised ML-classification model.

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

3) Recall: Equation 4 defines the computation of recall as a performance evaluation criterion that validates the model training of data points.

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

4) F1-Score: The use of equation 5 gives us the harmonic mean values of precision and recall.

$$F1 - score = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad (5)$$

5) Accuracy: As defined by equation 6, accuracy is a performance evaluation criterion that evaluates the model's true prediction divided by the total number of predictions made by the model.

$$Accuracy = \frac{TP}{TP+TN+FP+FN} \quad (6)$$

The model which performs best is considered for the categorical analysis to predict the brain stroke according to the class defined.

D. Proposed Architecture

For the Brain Stroke Classification analysis, Data Preprocessing is the foremost step. In this step, the process of cleaning and integrating the dataset happens. For analysis and interpretation, data preprocessing is necessary. This step ensures that the data is accurate, complete and consistent for the categorical analysis. In data preprocessing, the outlier detection and the Skewed Data Detection analysis are also performed. Outliers are values that are odd ones, these values are not related to other dataset column values. The outlier detection helps in making the data suitable for analysis. The skewed data analysis is used to make the distribution of the dataset normal throughout the analysis. The skewness test helps to check whether the data is symmetrical or not. This makes the dataset suitable for the classification analysis. The correlation analysis examines the relationship between each value in the column with every other parameter value. It determines the strength and direction of the variables' relationship. The next step involves the Synthetic Minority Oversampling Technique (SMOTE) Analysis. The SMOTE is used to detect the imbalanced labels in the dataset. The SMOTE technique helps in the generation of the synthetic data values for the minority class. This algorithm works under the principle of synthetic oversampling of the data to overcome the problem of random sampling. Overfitting is preferred instead of underfitting because underfitting might delete some useful information and labels from the dataset. The dataset is split into training and testing data with an 80-20 ratio. The classification algorithms used for brain stroke classification are XGBoost, CatBoost and LGBM. The classification report and the confusion matrix are generated for the analysis purpose. The ensemble technique is used to further increase the accuracy of the classification task. Ensemble learning involves combining multiple algorithms to obtain better performance and accuracy. The weighted voting ensemble combines the three classification models and predicts the values using testing data. The computation of the difference between the predicted values and the testing values is assigned to the weights of the algorithm. The model's weights play a crucial role in the optimization of its performance and overall effectiveness in categorizing the values. The performance evaluation is done using the confusion matrix for the computation of accuracy, precision and recall. The Steps involved in the Stroke Classification are shown in Fig 4.

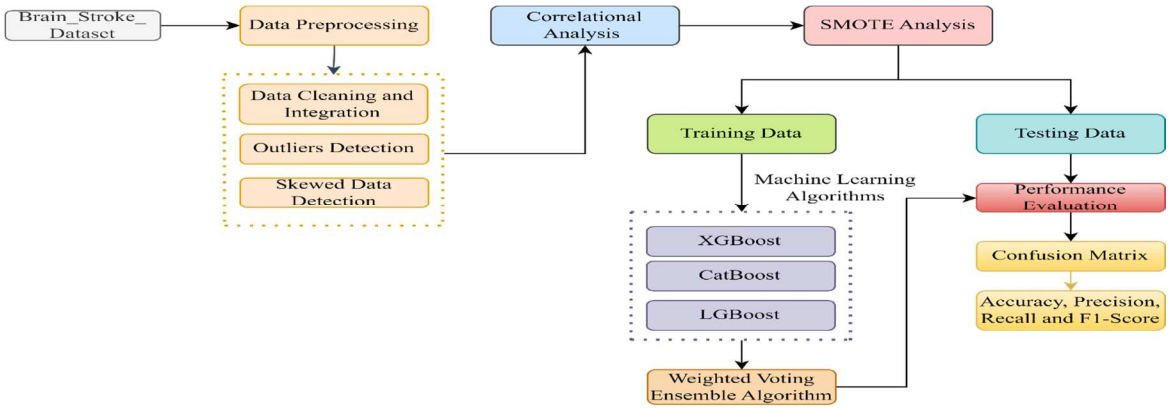

Fig. 4. Proposed Architecture for Stroke Classification

V. Results and Analysis

A. Results

In this section of the article, we discuss the results of three different classification algorithms: XGBoost, CatBoost and LGBoost. For classification purposes, class 0 represents patients with no brain stroke records and class 1 represents patients with a brain stroke history. The performance results of the XGBoost algorithm are given in Table II. The heatmap of the confusion matrix for the XGBoost is shown in Fig. 5.

The inference drawn from fig. 5 shows that the true positives determined by the confusion matrix are 925 for class 0 and the value for class 1 is 896. The 'false alarms' as the model categorizes as true but it is false labelled. The values as false positives are 40 and the false negatives as 33. The confusion matrix of the CatBoost classifier is shown in the fig 6. Based on the information presented in Fig. 6, the confusion matrix of the CatBoost, it can be inferred that there are 932 true positives for class 0 and 888 true positives for class 1. The model has also identified some false positives, which are instances where it categorizes something as true even though it is false. Specifically, there are 48 false positives, as well as 26 false negatives.

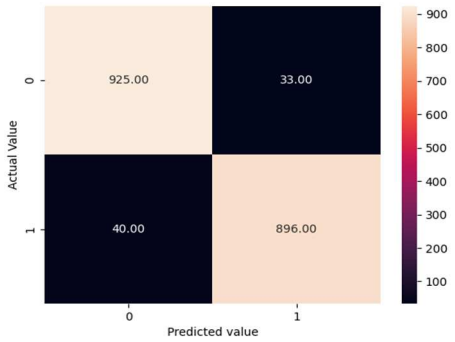

Fig. 5. Confusion Matrix of XGBoost

The performance evaluation of the LGBoost classifier as determined by the confusion matrix is shown in fig 7. The

values determined by the confusion matrix give the true positive value of 926 for class 0 and 889 for class 1. The false alarm as determined by the algorithm is given as 47 in the case of class 0 and 32 in the case of class 1. The confusion matrix heatmap for the Weighted Voting Ensemble algorithm is shown in Fig. 8. The inference from Fig. 8 shows that the TP value is 951, the FN value is 20, the FP value is 41, and the TN value is 882.

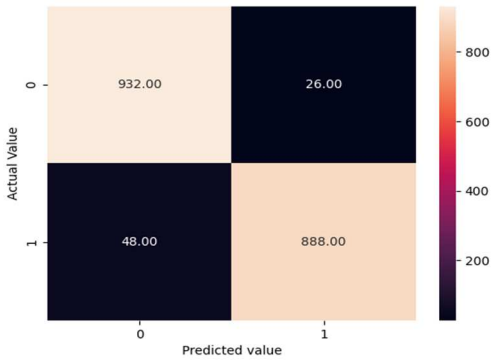

Fig. 6. Confusion Matrix of CatBoost

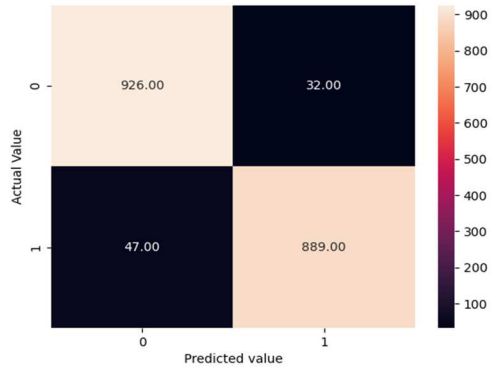

Fig. 7. Confusion Matrix of LGBoost

B. *Comparative Analysis*

The comparative analysis of various machine learning (ML) models is conducted based on their accuracy, precision, recall, and f1-score. The graphical depiction of the accuracy of these models is presented in Figure 9 for different classification algorithms, which helps in understanding the performance of the models across various classes.

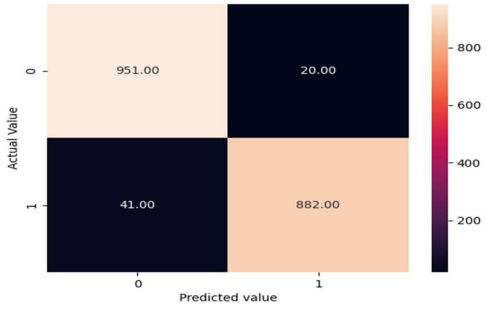

Fig. 8. Confusion Matrix of Ensemble Algorithm

Here class 0 represents the data with patients who have no stroke and class 1 represents the data of patients with stroke. Table II depicts the performance evaluation metric values of class 0 for the different classifiers used for the analysis.

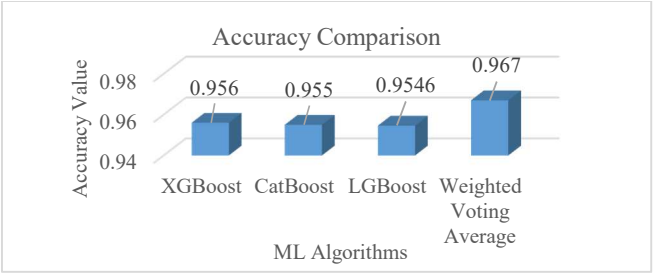

Fig. 9. Accuracy of the Different ML models

The values indicate that the precision, recall and F1-score of the Weighted Voting Ensemble algorithm is higher as compared to the other ML models. In Table III, the values of precision, recall and f1-score of the different classifiers used for class 1 are given.

TABLE II. PERFORMANCE EVALUATION OF CLASS 0

Performance Evaluation Metric	XGBoost	CatBoost	LGBoost	Weighted Voting Ensemble Algorithm
Precision	0.95	0.9448	0.9451	0.9536
Recall	0.9781	0.9802	0.98	0.9844
F1-score	0.9686	0.9622	0.9648	0.9688

TABLE III. PERFORMANCE EVALUATION OF CLASS 1

Performance Evaluation Metric	XGBoost	CatBoost	LGBoost	Weighted Voting Ensemble Algorithm
Precision	0.977	0.9788	0.98	0.9834
Recall	0.9571	0.9411	0.9411	0.9844
F1-score	0.967	0.9596	0.9622	0.9668

The graph in Fig 10 and 12 shows the comparison of classification models for the precision and recall values. The value of the precision and recall for the WVE algorithm is 0.9536 and 0.9844 respectively for the class 0.

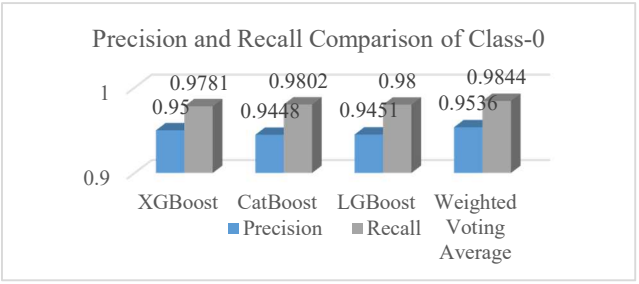

Fig. 10. Precision and Recall Comparison of Class-0

The analysis gives the results that the precision and recall for class 1 are 0.9834 and 0.9844 respectively. WVE Technique outperformed other ML classifiers in this article as shown in Figure 10 and Fig 12. Also the conclusion from Fig 11 and 13 shows that the f1-score of the WVE is more than as compared to the other three classifiers. The value of the f1-score for class 0 (No stroke) is 0.9688 and the value of f1-score for the class-1 (Stroke) is 0.9668.

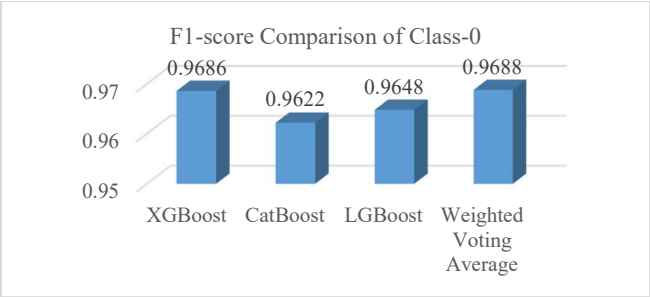

Fig. 11. F1-score Comparison of different ML models

The value of Accuracy is 96.7% for Weighted Voting Ensemble Technique. The precision of the KNN model is 0.9536 and the recall value is 0.9844. The harmonic mean calculated as the F1-score is 0.9688.

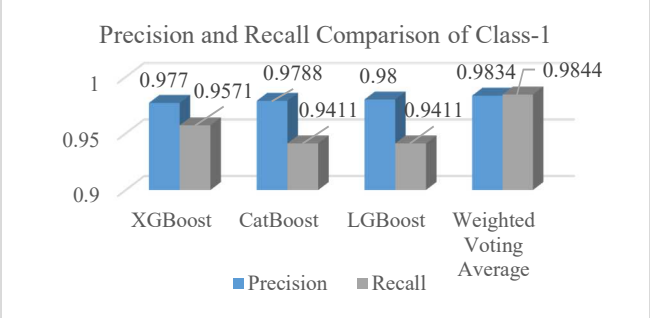

Fig. 12. Precision and Recall Comparison of Class-1

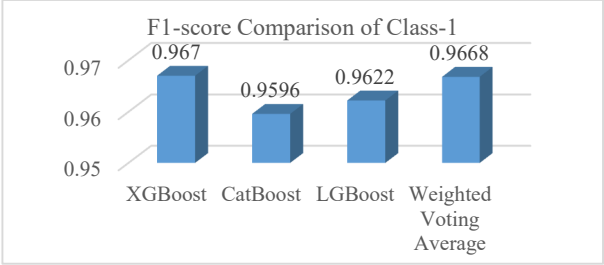

Fig. 13. F1- score Comparison of Class-1

VI. Conclusion

Brain Stroke is a life-threatening disease, that needs to be diagnosed at an early stage. The approach used in this article involves the use of ML classification algorithms: XGBoost, CatBoost and LGBoost. The Ensemble technique is also involved in the categorization analysis to combine the weights of all the ML algorithms and then utilize them to improve the functioning and performance of the model. The accuracy improves significantly with the use of a Weighted Voting Ensemble Technique. The accuracy of the WVE model is 96.7%. The ML algorithm results help in the prediction of the brain at an early stage and the needful treatment and precautions can be done to save the life. The upcoming task involves utilizing deep learning algorithms to classify brain CT scans accurately.

References

[1] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine Learning for Brain Stroke: A Review," *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 10, p. 105162, Oct. 2020, doi: https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162.

[2] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches," *Journal of Healthcare Engineering*, vol. 2021, p. e7633381, Nov. 2021, doi: https://doi.org/10.1155/2021/7633381.

[3] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," *European Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 23–30, Jan. 2023, doi: https://doi.org/10.24018/ejece.2023.7.1.483.

[4] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine Learning for Brain Stroke: A Review," *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 10, p. 105162, Oct. 2020, doi: https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162.

[5] G. Sharma, V. Anand, and S. Gupta, "Utilizing the Inception-ResNetV2 Pre-trained Model for Binary Classification of Leukemia Cells: An Advanced Approach to Hematological Diagnostics," *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2023, pp. 1-6, doi: 10.1109/GCAT59970.2023.10353360.

[6] Taniya, V. Bhardwaj and V. Kadyan, "Deep Neural Network Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit," *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2020, pp. 374-378, doi: 10.1109/ICCCA49541.2020.9250780.

[7] S. Singh, S. Mittal and S. Singh, "Analysis and Forecasting of COVID-19 Pandemic Using ARIMA Model," *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, Kalady, Ernakulam, India, 2023, pp. 143-148, doi: 10.1109/ACCESS57397.2023.10199278.

[8] Y. Yu et al., "Use of Deep Learning to Predict Final Ischemic Stroke Lesions From Initial Magnetic Resonance Imaging," JAMA Network Open, vol. 3, no. 3, p. e200772, Mar. 2020, doi: https://doi.org/10.1001/jamanetworkopen.2020.0772.

[9] Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H. *et al.* Classification of stroke disease using machine learning algorithms. *Neural Comput & Applic* **32**, 817–828 (2020). https://doi.org/10.1007/s00521-019-04041-y

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

[11] C. -L. Chin *et al.*, "An automated early ischemic stroke detection system using CNN deep learning algorithm," *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, Taichung, Taiwan, 2017, pp. 368-372, doi: 10.1109/ICAwST.2017.8256481.

[12] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," *European Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 23–30, Jan. 2023, doi: https://doi.org/10.24018/ejece.2023.7.1.483.

[13] A. N. Tusher, M. S. Sadik and M. T. Islam, "Early Brain Stroke Prediction Using Machine Learning," *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2022, pp. 1280-1284, doi: 10.1109/SMART55829.2022.10046889.

[14] A. Dogan and D. Birant, "A Weighted Majority Voting Ensemble Approach for Classification," 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 2019, pp. 1-6

[15] M. Kaur, S. R. Sakhare, K. Wanjale, and F. Akter, "Early Stroke Prediction Methods for Prevention of Strokes," *Behavioural Neurology*, vol. 2022, p. e7725597, Apr. 2022, doi: https://doi.org/10.1155/2022/7725597.

[16] S. Adam, A. Yousif, and M. Bashir, "Classification of Ischemic Stroke using Machine Learning Algorithms," *International Journal of Computer Applications*, vol. 149, no. 10, pp. 975–8887, 2016, Accessed: Mar. 10, 2024.

[17] Kaggle, "Datasets | Kaggle," *Kaggle.com*, 2019. https://www.kaggle.com/dataset