Proceedings of the Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI 2024)
IEEE Xplore Part Number : CFP24VG0-ART ; ISBN : 979-8-3315-4066-1

Optimizing Stroke Prediction with Ensemble Learning: A Comparative Study of AdaBoost, SVM, and KNN Models

Ankita Sharma
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab
India
ankita.3921@chitkara.edu.in

Sonam Mittal
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab
India
sonam.mittal@chitkara.edu.in

***Abstract*— Stroke is the main cause of long-term disability and death worldwide; it is a terrible medical condition caused by disrupted blood flow to the brain. This work intends to predict stroke occurrence using lifestyle, clinical, and demographic factors. Using Machine Learning (ML) methods including AdaBoost, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). An ensemble model called a voting classifier is used to increase model accuracy. The dataset consists in information including gender, age, hypertension, heart disease, average glucose level, BMI, smoking status, and habitation type. The overall accuracy of the Voting Ensemble algorithm is 90%. higher as compared to the other ML algorithms. Early stroke detection is essential since fast response considerably reduces the degree of the condition and long-term consequences. This work emphasizes the potential of ML in stroke prediction so supporting activities for preventative healthcare and personalized treatment strategies for stroke prevention.**

Keywords: Artificial Intelligence, Machine Learning, AdaBoost, SVM, KNN, Voting Classifier, Brain Stroke, and Ensemble Learning.

I. INTRODUCTION

Another name for a brain attack, which results from a halted blood flow in the brain is brain stroke. Sometimes a brain stroke that is, the damage or death of a portion of the brain—causes abrupt death. There are two kinds of strokes: hemorrhagic and ischemic ones. Usually from the buildup of a blood clot or plaque deposit, an ischemic stroke is a cerebrovascular accident resulting from an impediment or narrowing of an artery supplying blood to the brain. More than 85% of all strokes are of this kind; hemorrhagic stroke results from weak blood vessels rupturing and blood seeping into the brain causing damage. Out of all the strokes, about 15% are hemorrhagic ones. Those with high blood pressure, high blood sugar, and cholesterol issues are more prone to have brain stroke. The regional director general of Southeast Asia, Poonam Khetrapal Singh, claims that low-middle-income countries have about 11 million cases of brain stroke annually, which causes 4 million deaths annually from which 30% of patients are seriously disabled and 70% of survivors recover.

In the field of healthcare, artificial intelligence (AI) is indispensable as, by using its several approaches to the particular condition, these algorithms enable the early stage disease prediction. Under this technique, supervised learning, artificial intelligence (AI) includes a subpart called ML with several approaches or algorithms for the prediction of brain stroke.

A. AdaBoost: Designed to combine weak classifiers into a strong classifier to improve their performance, Adaptive Boost—sometimes known as AdaBoos, is a complex ensemble learning method. It uses an iterative technique to teach a series of weak models—often decision mistakes. Initially, all training samples have the same weights; but, the algorithm modifies these weights following each classifier's training: raising them for mistakenly categorized samples to stress difficult situations and lowering them for precisely identified ones. Every weak classifier has a weight based on its accuracy; the weighted group of all the classifiers generates the final forecast from which those performing better have more impact. Since AdaBoost is well-known for reducing variance and bias, it is a good tool for increasing classification accuracy. If not precisely tweaked, it could overfit even considering noisy data and outliers.

B. SVM: Strong and adaptable supervised learning approach Applied for both classification and regression Support Vector Machine (SVM). SVM is really based on the search for a hyperplane in the feature space that best splits several classes. Expressed in two-dimensional space, this hyperplane is a line splitting the data into two classes with the largest margin between their nearest points—also known as support vectors. This margin is absolutely important since it improves the generalizing capacity of the model toward unprocessed data. Thanks to the use of kernel functions, SVM is especially successful in high-dimensional spaces and with datasets nonlinearly separable. These kernels translate the input data into higher-dimensional environments where a linear separation could be feasible: polyn, radial basis function (RBF), sigmoid, Though it can be computationally demanding, particularly with complicated kernels, SVM is also well-known for its resilience and efficiency in managing vast amounts. It performs especially in circumstances when the margin of separation is clear and can even function in noisy environments. SVM has strengths, however, to reach the best performance it needs careful parameter and kernel choice adjustment.

C. KNN: Simple but effective supervised learning tool K-Nearest Neighbors (KNN) is applied for classification and regression applications. The fundamental concept of KNN is to predict based on the closest feature space data

2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) | 979-8-3315-4066-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOICI62503.2024.10695950

points. KNN assigns the majority class among those neighbors for classification using the 'k' nearest neighbors of a query location.

By averaging the values of the 'k' nearest neighbors, regression projects the output value. The choice of 'k' is quite crucial; a small value of 'k' may render the algorithm sensitive to data noise, whilst a large 'k' can generate excessively smooth forecasts and loss of detail. KNN is versatile and fit to numerous data forms since it is non-parametric, that is, it does not assume any underlying data distribution. Nevertheless, its computational complexity increases with dataset growth since it has to determine the distance between the query point and every other point. Usually applied are Euclidean, Manhattan, or Minkowski distance calculations. When the decision boundary is complex and not easily stated by a simple model even if it is basic, KNN can be helpful for employment.

D. Voting Classifier: A voting classifier is an ensemble learning method meant to pool the forecasts of several base models so improve general classification performance. The basic idea of providing a more strong and accurate result is aggregating the forecasts from various algorithms. Every individual model—or " Vote"—in a voting classifier produces a prediction; the voting process chooses the final class label. Usually, voting consists in two forms: majority and weighted voting. Usually speaking, the class label with the most votes from the individual classifiers makes the last prediction. More exactly, under weighted voting every vote of every classifier is weighted in line with its performance, therefore affecting the final choice. Combining numerous models with different strengths and weaknesses—such as neural networks, support vector machines, and decision trees—helps voting classifiers to be very successful. This variance helps to lower the boundaries of every one model and improves generalizing capacity. Voting classifiers are easy to develop and applied to generate strong predictive models by aggregating the talents of numerous algorithms to get improved performance and dependability in predictions.

II. Relevant Literature

This research identifies the ML technology for predicting brain stroke, such as KNN, CART (Classification and Regression Trees), SVM, and Logistic regression. According to the author, they surveyed the data set from different hospitals in Bangladesh. Approximately 100 patients' data is collected from which 80% are men and the rest is women. 70% of the data is split into training sets and 30% into testing sets. After applying the models to the data set, the KNN model performed better with 97% accuracy, and the author also calculated the confusion matrix with its different parameters such as precision, recall, f1-score, and support with different values of 0.97, 0.97, 0.97, 60 [1].

The author employs diverse ML techniques to forecast brain stroke. The different algorithms are XGBoost, Light Gradient Boosting Machine, Random Forest, Naive Bayes Decision Tree, LR, K-NN, Ada Boost, SVM - Linear Kernel, and deep neural networks like ANN (3-layer and 4-layer ANN) are utilized as the classification work. The data set is collected from the Kaggle site, which consists of 12 columns and 5110 rows having different parameters like BMI, heart disease, name, gender, hypertension, married or unmarried, etc. The Random Forest gave the highest accuracy of 99%. The accuracy of the three-layer deep neural network (4-layer ANN) is 92.39% greater than that of the three-layer ANN approach using the chosen features as input [2].

In this research work author utilized the 1-D Convolutional Neural Network (CNN) model for the prediction of brain stroke. The author also discusses two different functions Random Over Sampler and Standard Scaler to control the imbalanced data and for the removal of outliers. For decaying the learning rate from 0.001 to 0.000001 Adam optimizer is used by which the accuracy of the model performed better with 98% accuracy. The data is taken from the Kaggle site which contains 4981 data in the form of rows and columns with nine features such as ever_married, BMI, disease, age, gender, hypertension, residence, etc [3].

In this paper, the author researches the prediction of tissue outcome and assessment of treatment effect in Acute Ischemic Stroke (AIS) using the Deep Learning (DL) model. The author did the hypothesis in the CNN model with 〖CNN〗_deep for the enhancement of the accuracy and changes in the frequency of the models. For treating acute magnetic resonance imaging, the data set consists of 222 patients, of which 91 are women. The shallow CNN model is compared with 〖CNN〗_deep to evaluate the performance of the data. 〖CNN〗_deep gives better performance than CNN. Both models' results are (AUC=0.88 ± 0.12) than the generalized linear model (AUC=0.78±0.12; P=0.005) [4].

The author explained the ML approaches for the prediction of AIS using brain MRI-based biomarkers. In this article, NLP-based ML models are used on the data set of MRI text data. The data set consists of a total of 1840 subjects out of which 645 patients (35.1%) gave poor outcomes 3 months after the stroke onset. Random forest works as the best classifier and gave the best results of (0.782 of AUROC). This paper represents the considerable advances in forecasting poor outcomes using DL models such as CNNs and LSTMs, notably through document-level NLP techniques. The findings highlight the potential of NLP-based DL algorithms as useful tools for improving clinical decision-making using unstructured MRI text data [5].

This article talks about brain stroke prediction or hemorrhagic transformation in arterial ischemic stroke (AIS) patients. Used ML approaches DT and MLR (Multivariate LR) for the prediction of brain stroke. The data set and other support for the work are provided by the Beijing Hospital Clinical Research 121 Program. The ROC curve (AUC) is 81.7%. both models confirmed the lower collateral status and highest PLR (platelet to lymphocyte ratios) that the risk of hemorrhagic transformation is increasing in AIS patients [6].

This research paper works on the concept of ML with its various algorithms XGBoost, AdaBoost, LR, DT, KNN, SGD, Gaussian Classifier, QDA, Multilayer Perceptron, and Gradient Boosting Classifier. this paper also used a weighted voting classifier by using ensemble learning, after results this model works better than other algorithms. The data set is from a Bangladesh medical clinic that has 5110 patient's information with different parameters age hypertension, gender, etc. The weighted voting classifier gave the best accuracy of 97% and proves that this classifier works better on the data set of brain stroke [7].

In this study, ML techniques are used to identify and predict strokes using medical data. This work is restricted in its ability to forecast risk variables for distinct types of strokes. To solve

this restriction, a Stroke Prediction (SPN) algorithm is developed that employs the enhanced random forest to analyze the degrees of risk received from strokes. When compared to previous models, the (SPR) model utilizing ML enhanced prediction accuracy to 96.97%. The data set used in this study includes 4,799 participants, with 3,123 males and 1,676 females [8].

III. METHODOLOGY

The approach of the paper is discussed in this part together with several ML techniques applied for the brain stroke classification. It is projected and classified using AdaBoost, SVM, KNN models, and an ensemble with the Weighted Voting classifier to get higher accuracy. They are using the method known as ensemble learning where several weak models create a strong model for the classification of brain stroke.

A. *Dataset*

The data for this study consists of many characteristics of patient health and lifestyle choices that can affect stroke risk. The data set consists of 4982 rows with 12 parameters, such Primary variables are gender (male, female, other), age, and hypertension. It also covers heart disease, ever-married status (no, yes), and work type (children, govt job, never worked, private, self-employed). The residence type falls into either urban or rural. Furthermore included in the dataset are average glucose level, BMI, and smoking status (previously smoked, never smoked, smokes, Unknown). Stress is the goal variable. This extensive dataset facilitates the study of several elements related to stroke incidence and helps to clarify how diverse variables influence stroke risk, therefore guiding focused preventive and intervention plans.

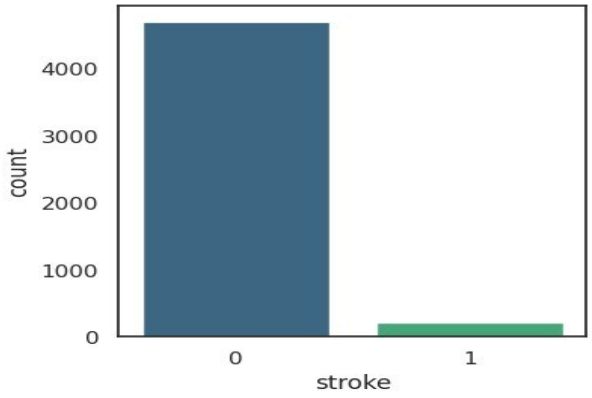

Fig 1 Sample Data Set

B. *Implementation Environment*

This paper works on two classifiers for the classification of breast cancer. For the implementation of the HP VICTUS 7000 series, Ryzen 5 processor. Data set is taken from Kaggle, coding part is implemented using Pycharm community 2023.2.3. Different modules are used in Python code Pandas as pd, seaborn as SNS, Numpy as np, Matplotlip as plt, XGBoost, AdaBoost and confusion matrix, etc.

C. *Proposed Architecture*

The methodology of the proposed architecture is given in this section as shown in figure 2.

1. Data collecting and compilation of a brain stroke patient corpus inclusion of relevant traits (age, gender, medical history, symptoms, diagnostic).
2. Data cleaning, addressing missing values or imputation, data cleansing, and integration help to produce uniformity and consistency of data.
3. Getting Data Ready separating data into testing and training sets
4. Model Development, AdaBoost, SVM, and KNN method Implementation
5. Individual model training on a dataset Creation of ensemble models by combining weighted voting-based individual model predictions
6. Model Evaluation: assessing an ensemble model using a testing dataset
7. Confusion matrix, accuracy, precision, recall, and F1-score based assessment
8. Possible Improvements and Research of Other ML Techniques Hyperparameter tuning in search for best performance.

The selection of an algorithm depends on performance criteria, problem complexity, and features of the dataset. One should experiment with several algorithms. Careful data preparation, feature engineering, hyperparameter tuning, and ensemble methods used together will produce high accuracy. Including more pertinent characteristics, investigating cutting-edge technologies, and cross-valuation will help to improve model performance. Comparative analysis and repeatability depend on a thorough knowledge of a dataset including size, features, distribution, and imbalance. Iteratively refining model architecture, hyperparameters, and feature engineering grounded on performance measures can help to maximize stroke prediction. To enhance the brain stroke classification the weighted voting ensemble technique is employed for the categorization combining the accuracies of the multiple algorithms: SVM, KNN, and AdaBoost.

D. *Evaluating Parameters*

Different parameters are used for the performance evaluation of models, like Precision, Recall F1-Score, and Accuracy.

- *Confusion Matrix:* It is a tool used for the performance evaluation of the model, with different parameters. Like True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The confusion matrix evaluates the Accuracy, Precision, Recall, and F1-Score of the model.

- *Precision:* The ratio of True Positives to all Positives is called precision. That would be the proportion of patients that we accurately diagnose with breast cancer out of all those who genuinely have it, according to our issue statement.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

- *Recall:* Recall is a metric that indicates how well our model finds True Positives. Recall therefore indicates the number of people that we accurately diagnosed as having breast cancer out of all those who genuinely have heart disease.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

- *F1-Score:* The harmonic mean of a classification model's precision and recall is known as the F1 score.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

- *Accuracy:* It is defined as the overall performance performed by the model or how often an ML model correctly predicts the outcome is called accuracy.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

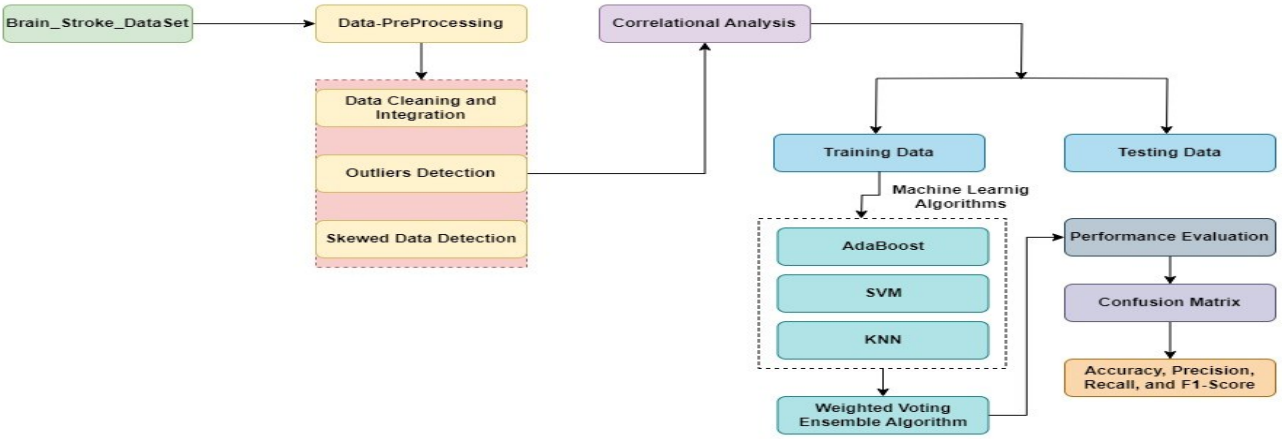

Fig 2 Proposed Architecture of Brain Stroke Using Ensemble Techniques

IV. RESULT

In this paper, ML approaches are used to predict brain stroke, and AdaBoost, SVM, KNN, and Voting Classifier models are utilized. The voting classifier performs better with the highest accuracy of 90%.

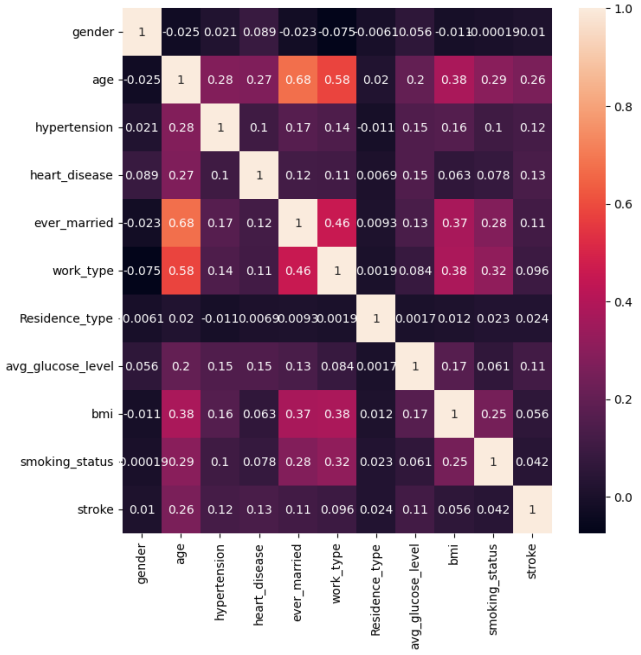

Fig 3 Correlation Heatmap

Figure 3 shows a visual correlation heat map. Brighter hues suggest closer links between variables. Age seems to be favorably linked with variables including heart disease, hypertension, and ever married. With BMI, smoking status indicates a somewhat positive connection. Fascinatingly, home type appears to have very little influence on other variables.

A. *AdaBoost:*

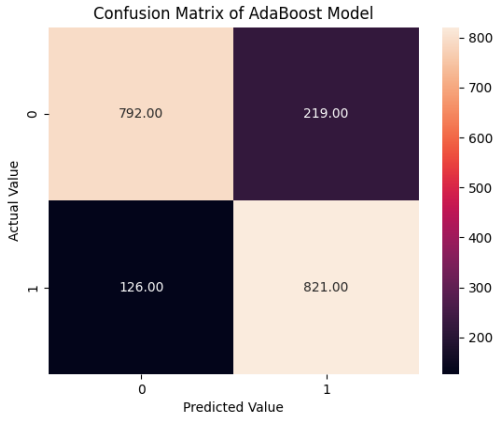

Fig 4 Heatmap of AdaBoost Model

Figure 4 shows the heatmap of the AdaBoost Model with the different evaluation parameters like TP, TN, FP, and FN. The true prediction made by the model for class 0 is 792 and for class 1 is 821. The false predictions made by the algorithm is 345.

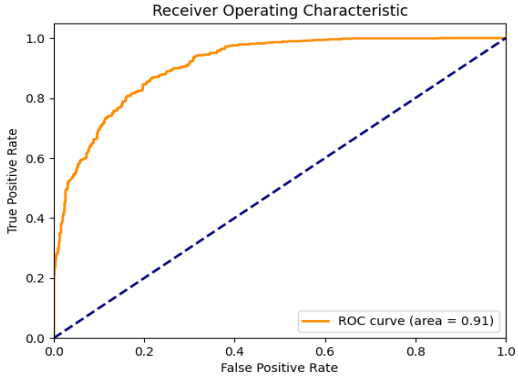

Fig 5 ROC Curve of AdaBoost Model

Fig 5 depicts the ROC Curve of the AdaBoost model in which the x-axis shows a false positive rate and the y-aix shows a true positive rate. The AUC value for the AdaBoost algorithm is 0.91.

B. *SVM:*

The confusion matrix visualization for the SVM model is given in Fig 6. The inference from the Fig. 7 shows that the TP value for Class 0 is 805. The algorithm predicted 858 instances as TN. The false predictions made by the algorithm are 295. ROC curve for the SVM algorithm is depicted in Fig. 6 and the value of AUC for the SVM algorithm is 0.92.

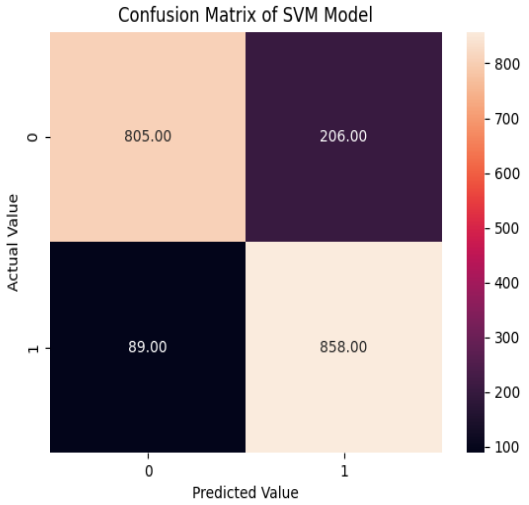

Fig 6 SVM Confusion Matrix Visualisation

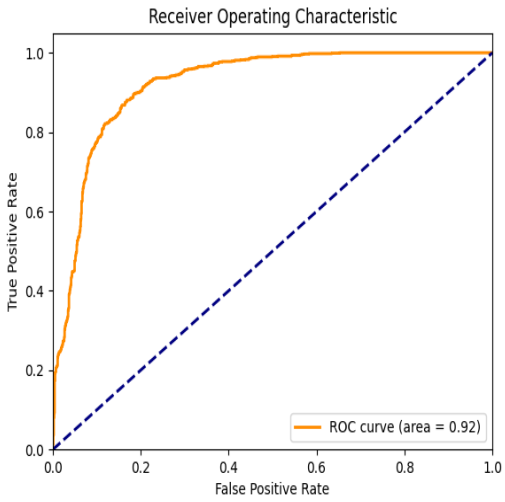

Fig 7 ROC Curve for SVM Model

C. *KNN:*

The Confusion Matrix heatmap for the KNN algorithm is depicted in Fig. 8. The model for stroke classification predicted 807 as TP values for class 0. The true prediction for class 1 made by the algorithm is 928. The false prediction done as a misclassification by the algorithm is 223. The ROC curve is shown in Fig. 9 in which the AUC value for the plot is given by 0.95.

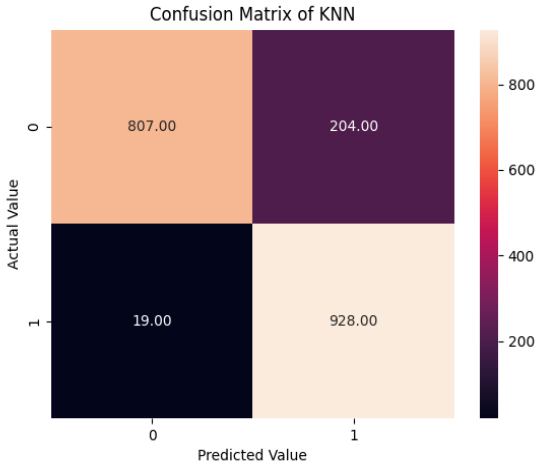

Fig 8 Confusion Matrix Heatmap for KNN

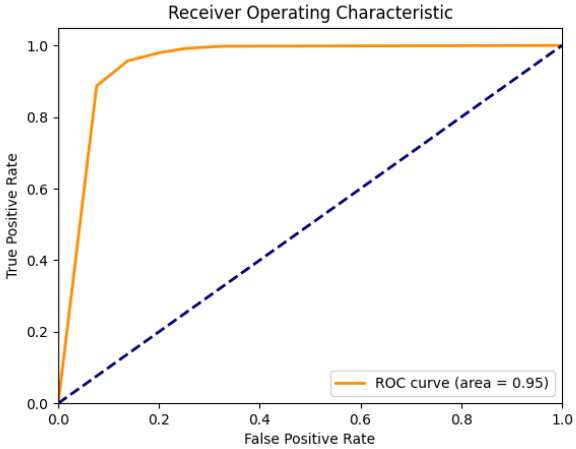

Fig 9 ROC curve of KNN

D. *Voting Classifier:*

The Voting Classifier employs the ensembling of the AdaBoost, KNN, and SVM models combined to make the classification of the stroke visualized in the form of a confusion matrix in Fig. 10. The True Prediction made by the ensemble model is 847 for class 0 and 923 for Class 1. The misclassification made by the algorithm is 188. The value of AUc for the Ensemble Voting Classifier is 0.96 as shown in Fig 11.

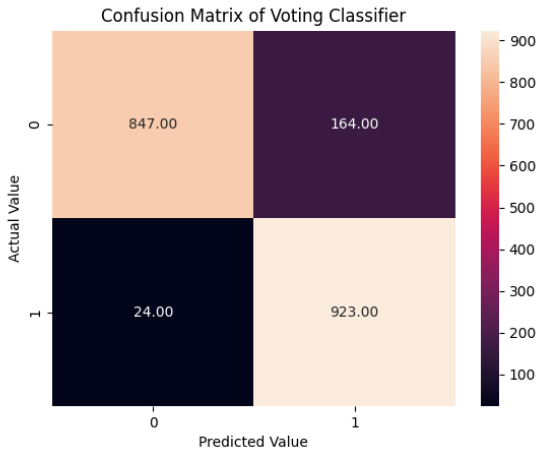

Fig 10 Confusion Matrix Heatmap for the Voting Classifier

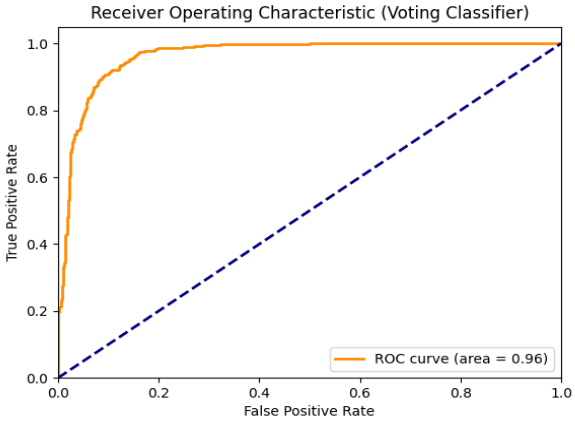

Fig 11 ROC curve for the Voting Classifier

E. *Comaparative Evaluation*

This article discusses the comparison between ML Models or classifiers, AdaBoost, SVM, KNN, and Voting Classifier with Lasso CV on XGBoost Classifiers. It analyses different parameters of the confusion matrix like precision, Recall, F1-Score, and Accuracy. The precision, Recall, F1-Score, and accuracy of AdaBoost, SVM, and KNN models are 0.86, 0.78, 0.82, 82% and 0.90,0.80,0.85, 85%, and 0.98, 0.80,0.88, 89% respectively. By applying the ensembling technique on the voting classifier, it achieves the highest accuracy of 90%, with precision, recall, and f1-score of 0.97, 0.84, and 0.90 respectively as shown in Table 1.

Table 1

Parameter	AdaBoost	SVM	KNN	Voting Classifier
Precision	0.86	0.90	0.98	0.97
Recall	0.78	0.80	0.80	0.84
F1-Score	0.82	0.85	0.88	0.90
Accuracy	82%	85%	89%	90%

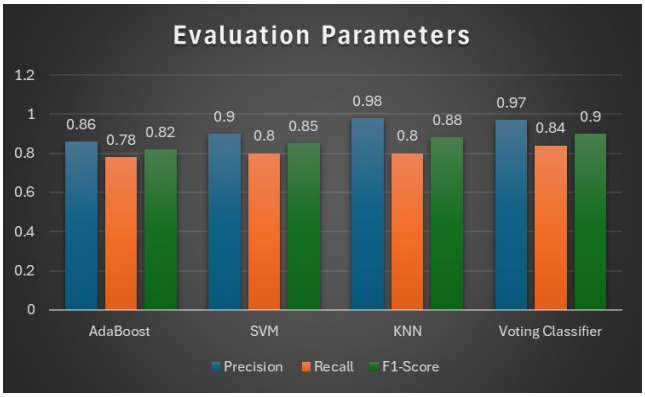

Fig 12 Performance Metrics of Models

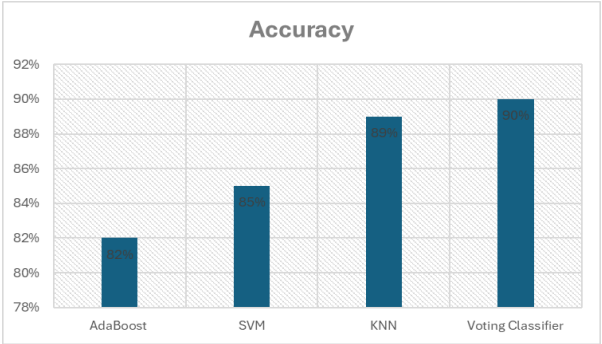

Fig 13 Accuracy of the Models

VI Conclusion

This work investigates several ML techniques to use patient data to predict stroke events. Using AdaBoost, SVM, and KNN showed the efficiency of each model in stroke estimate; each model had unique strengths. With an accuracy of 90%, integrating these models using a voting classifier produced the best performance. This united approach raised predicted accuracy and underlined the efficiency of merging several ML methods to handle challenging healthcare problems. The results highlight the need to use cutting-edge ML techniques for early stroke classification, which assists in quick interpolation and reduces the long-term impacts of stroke. Future research could add more characteristics and improve model parameters to promote individualized healthcare approaches and increase prediction capacity.

References

[1] M. Toğaçar, B. Ergen, and Z. Cömert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Med. Hypotheses*, vol. 135, p. 109503, 2020, doi: 10.1016/j.mehy.2019.109503.

[2] S. Singh, S. Mittal, and S. Singh, "Analysis and Forecasting of COVID-19 Pandemic Using ARIMA Model," *ACCESS 2023 - 2023 3rd Int. Conf. Adv. Comput. Commun. Embed. Secur. Syst.*, no. May 2021, pp. 143–148, 2023, doi: 10.1109/ACCESS57397.2023.10199278.

[3] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, *Artificial intelligence to deep learning: machine intelligence approach for drug discovery*, vol. 25, no. 3. Springer International Publishing, 2021. doi: 10.1007/s11030-021-10217-3.

[4] S. Singh and S. Mittal, "Pandemic Outbreak Prediction using Optimization-based Machine Learning Model.," in *3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 2023, pp. 154–159.

[5] C. Ravi and N. Khare, "AN ADABOOST OPTIMIZED CCFIS BASED CLASSIFICATION MODEL FOR BREAST CANCER DETECTION," 2017.

[6] "No Title," [Online]. Available: https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/

[7] Rahmanul Hoque, Suman Das, Mahmudul Hoque, and Mahmudul Hoque, "Breast Cancer Classification using XGBoost," *World J. Adv. Res. Rev.*, vol. 21, no. 2, pp. 1985–1994, Feb. 2024, doi: 10.30574/wjarr.2024.21.2.0625.

[8] "No Title," [Online]. Available: https://www.geeksforgeeks.org/xgboost/

[9] A. Derangula, S. R. Edara, and P. K. Karri, "Feature selection of breast cancer data using gradient boosting techniques of machine learning," *Eur. J. Mol. Clin. Med.*, vol. 7, no. 2, pp. 3488–3504, 2020, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096914275&partnerID=40&md5=476b9182339725809c258a0b63a14a48

[10] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, "Developing A Web based System for Breast Cancer Prediction using XGboost Classifier." [Online]. Available: www.ijert.org

[11] E. Sugiharti, R. Arifudin, D. T. Wiyanti, and A. B. Susilo, "Convolutional neural Network-XGBoost for accuracy enhancement of breast cancer detection," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jun. 2021. doi: 10.1088/1742-6596/1918/4/042016.

[12] P. Liu, B. Fu, S. X. Yang, L. Deng, X. Zhong, and H. Zheng, "Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 1, pp. 148–160, Jan. 2021, doi: 10.1109/TBME.2020.2993278.

[13] *Das Kapital und Vorarbeiten. Bd. 9.[2], Capital, a critical analysis of capitalist production. London 1887 : Apparat*.

[14] I. Shokripoor Bahman Bigloo *et al.*, "A Parallel Genetic Algorithm Based Method for Feature Subset Selection in Intrusion Detection Systems Presentation of an Efficient Automatic Short Answer Grading Model Based on Combination of Pseudo Relevance Feedback and Semantic Relatedness Measures Pages 17-30 Pages 31-52 Image Encryption by Using Combination of DNA Sequence and Lattice Map Pages 61-74 Designing a Trust-Based Recommender System in Social Rating Networks," 2019.

[15] K. Varshini, R. K. Sethuramamoorthy, V. Kumar, S. A. Shree, and S. Deivarani, "Breast cancer prediction using machine learning techniques," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 6 Special Issue, pp. 2026–2032, 2020, doi: 10.5120/ijca2022922490.

[16] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3240–3247, 2009, doi: 10.1016/j.eswa.2008.01.009.

[17] O. F. Ereken and C. Tarhan, "Breast Cancer Detection using Convolutional Neural Networks," *ISMSIT 2022 - 6th Int. Symp. Multidiscip. Stud. Innov. Technol. Proc.*, vol. 027, no. 2018, pp. 597–601, 2022, doi: 10.1109/ISMSIT56059.2022.9932694.

[18] S. Guan and M. Loew, "Breast cancer detection using transfer learning in convolutional neural networks," *Proc. - Appl. Imag. Pattern Recognit. Work.*, vol. 2017-Octob, no. January, 2017, doi: 10.1109/AIPR.2017.8457948.