EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

Harnessing the Power of Ensemble Machine Learning for the Heart Stroke Classification

Purnima Pal[1*], Manju Nandal[2], Srishti Dikshit[3], Aarushi Thusu[4], Harsh Vikram Singh[5]

[1]Research Scholar, Kamla Nehru Institute of Technology, Sultanpur
[2]Assistant Professor, Noida Institute of Engineering and Technology, Greater Noida
[3]Research Scholar, Dr. C.V. Raman University, Bihar
[4]Assistant Professor, Noida Institute of Engineering and Technology, Greater Noida
[5]Professor, Kamla Nehru Institute of Technology, Sultanpur

Abstract

A heart stroke, also known as a myocardial infarction or heart attack, is a critical medical condition that arises when there is an obstruction in the coronary arteries that provide blood to the heart muscles. This blockage results in a diminished flow of blood and oxygen to a specific area of the heart. This abrupt interruption initiates a gradual sequence of heart muscle damage, which can lead to varying degrees of functional impairment. The severity of these impairments is primarily determined by the precise location of the heart muscle affected. Therefore, it is of utmost importance to identify the warning signs and symptoms of a stroke as soon as possible. This is the objective of this paper is to early recognition and prompt action can significantly improve the chances of a healthy and fulfilling life following a stroke. In this research work, the Stroke dataset is pre-processed and on pre-processed dataset machine learning and ensemble machine learning techniques were employed to develop and assess several models aimed at creating a stable framework for predicting the enduring stroke risk. And various matrices like accuracy, F1 score, ROC, precision, and recall are calculated. Among all models, AdaBoost model demonstrated exceptional performance validated through multiple metrics, including Precision, AUC, recall, accuracy, and F1-measure. The results underscored superiority of the AdaBoost classification method, achieving an impressive Accuracy of 99%. AdaBoost model may serve as a stable framework for predicting enduring stroke risk, emphasizing its potential utility in clinical settings for identifying individuals at higher risk of experiencing a stroke.

Keywords: Ensemble Machine Learning, Machine Learning, Performance Metrics, Stroke Prediction

Received on 19 October 2023, accepted on 11 December 2023, published on 15 December 2023

doi: 10.4108/eetpht.9.4617

1. Introduction

Heart strokes, commonly referred to as myocardial infarctions or heart attacks, are a critical and prevalent health concern. These events constitute a sudden, often life-threatening condition that necessitates our understanding and vigilance. The heart, a vital organ that sustains our existence, takes centre stage in a heart stroke. Any disruption in its rhythmic functioning can yield grave consequences. This disruption arises from an interruption in the heart's essential blood supply, an indispensable factor for its optimal performance [1]. Strokes are a serious medical emergency that needs quick medical help [2]. Timely identification and prompt intervention are essential in preventing additional harm from heart complications. The occurrence of fatalities from heart attacks is on the rise in contemporary society. Several behaviours, including alcohol consumption, smoking, and

[*]Corresponding author. Email: purnima22pal@gmail.com

high cholesterol levels, are contributing factors to heart-related issues [2]. In accordance with World Stroke Organization, approximately 13 million individuals experience a stroke annually, leading to around 5.5 million fatalities [3]. This condition stands as the main provider to both global mortality and disability, exerting a profound impact across various facets of life. Stroke doesn't solely affect the individual experiencing it but also extends its repercussions to the patient's family, social circle, family, and professional life. Moreover, in contrast to usual misconceptions, strokes may happen in individuals of any gender, age, or physical condition [4].

As technology continues to progress within the medical arena, the capability to forecast the onset of a stroke has become attainable through Machine Learning application. The algorithms inherent to Machine Learning prove invaluable in delivering precise predictions and accurate analyses. For predicting heart strokes, Machine Learning [5–7]and Ensemble machine learning [8–10] methods are broadly uses which could potentially improve the classification of heart diseases.

structure is as follows: First, Heart Stroke is introduced in Section 1. Section 2 provides an overview of pertinent research, while Section 3 elaborates on the dataset employed in our study. Our proposed methodology is outlined in Section 4, followed by the presentation and discussion of results in Section 5. Lastly, Section 6 concludes the paper and explain future research.

2. Related Work

In [11], the researchers utilized five machine learning methods to predict strokes using the CVS (Cardiovascular Health Study) dataset. Their findings revealed that the most effective strategy involved the fusion of Decision Tree algorithm with C4.5. In [12], the researchers conducted stroke prediction by using enhanced RF (Random Forest) model. They applied the algorithm to assess risk measures associated with strokes. According to the authors, this method demonstrated superior performance when compared to existing algorithms. In their study [13], the researchers gathered data using EEG (Electroencephalography) and transferred the bio-signal data to a server. Subsequently, they conducted data preprocessing and applied various Deep Learning models to foresee the likelihood of stroke. Among these methods, the Gated Recurrent Unit Network (GRU) exhibited the highest performance with an accuracy of 95.6%, followed closely by BILSTM with an accuracy of 91%. Authors in [12] proposed a stroke prediction approach that incorporated machine learning classification and deep learning algorithms. They sourced their stroke dataset from Kaggle and found that machine learning methods, particularly Random Forest, outperformed deep neural network techniques with an outstanding accuracy of 99%. For their study, the researcher [14] acquired stroke data from Sugam Hospital located in Kumbakonam, Tamil Nadu, India. They classified the types of strokes using various data mining and machine learning methods. Among these methods, Artificial neural network (ANN) trained using the SGD (stochastic gradient descent) approach outperformed other techniques, achieving strong classification results with an accuracy exceeding 95 percent. In contrast, Support Vector Machine (SVM) and ensemble methods provided an accuracy of 91 percent. In the research [15] employed the Kaggle dataset and recommended using a diverse set of machine learning techniques, that includes Decision trees, RF, logistic regression, KNN, NB, and SVM for stroke prediction. Surprisingly, the Naive Bayes method outperformed other algorithms, achieving an accuracy of 82%. Authors [16] conducted a study aimed at early detection of ischemic stroke using algorithmic methods. They focused on developing a CNN-based approach for automating the identification of primary ischemic strokes. They gathered 256 images for training and testing the CNN model. Employing data augmentation techniques to enhance their image data by eliminating unrealistic stroke-related areas, their CNN method achieved an impressive accuracy of 90%. A study conducted by [17] to predict stroke risk levels, a variety of models such as voting, Bayesian network, logistic regression, decision tree, NB, neural network, boosting, and random forest were used. The final output of the analysis revealed that the Random Forest model outperforms with the highest precision at 97.33% and the highest recall at 99.94%.text.

3. Dataset Description

The data utilized for this study was obtained from a Kaggle dataset credited to Fedesoriano [18]. This dataset comprises a total of 5110 entries, each representing a patient and includes 12 attributes. Remarkably, only 249 attributes out of the 5110 candidates in the dataset were recorded as having experienced a stroke, resulting in a substantial data imbalance. The attributes encompass categorical information such as smoking state, work type, gender, marital status and residence type. Additionally, there are numerical attributes including age, body mass index and average glucose level. Given the aim to predict the stroke using various learning models, the target attribute in this dataset is stroke, denoted as 0 for no stroke and 1 for stroke cases. Table 1 represent the attributes of the stroke dataset.

Table 1. Dataset Description

S.No.	Attribute	Values
1	Patient ID	[1, 43400]
2	Gender	Female, Male

3	Age	[0.08, 82]
4	Married	Yes, No
5	Work type	Employed, Private
6	Avg-glucose	[55-291]
7	Smoking status	Smoked, Never, formerly
8	Hypertension	Yes, No
9	BMI	[10.1, 97.6]
10	Residence type	Rural, Urban
11	Heart disease	Yes, No
12	Stroke	Yes, No

Figure 1 shows the correlation between the target variable i.e., stroke and other variables. Analysis of the heatmap reveals a strong correlation between hypertension, age, heart disease, and average glucose level with the occurrence of stroke in the dataset.

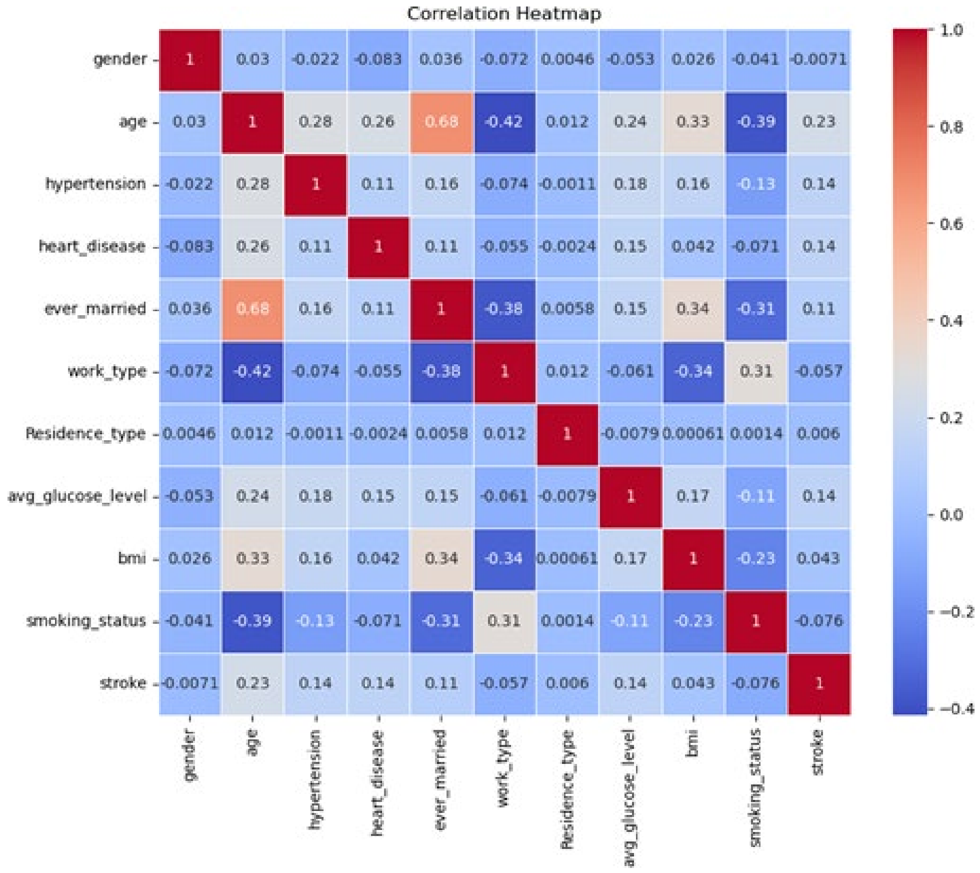


Figure 1: Heatmap of stroke Dataset

4. Proposed Methodology

The proposed diagnosis system (figure 2) comprises five phases such as data collection, data preprocessing, classification using machine and Ensemble Machine learning algorithms, evaluation of performance metrics to measure the effectiveness of the various algorithms and finally, selecting the best model for Stroke classification based on a comparative study of performance metrics. The data is split into two parts 80% for testing and 20% for training purposes.

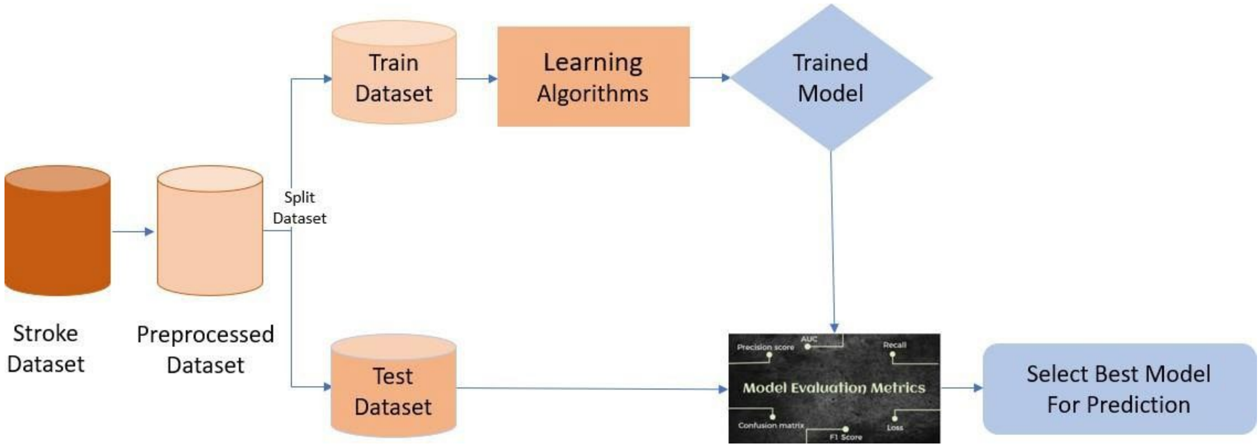


Figure 2: Proposed Methodology

4.1. Dataset Preprocessing:

The dataset contains 5110 instances and 12 attributes in which 201 instances have missing values and rows containing missing values have been eliminated. For preprocessing the dataset (figure 2), One Hot encoding and Normalization was employed.

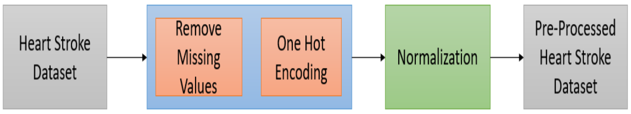


Figure 3: Data Pre-processing

One-hot encoding is a necessary step when dealing with categorical variables in the dataset, such as gender or smoking status, which cannot be directly used as input for machine learning algorithms[20]. One-hot encoding converts categorical variables into a binary (0 or 1) format, creating new binary columns for each category within the original variable. This transformation allows machine learning models to work with categorical data effectively. It ensures that categorical data doesn't introduce bias into the predictive models.

Normalization is a crucial step in preparing numerical features within the dataset. It scales the values of different features to a standardized range, ensuring that no single feature dominates the learning process. In this work we have utilized the Min-Max scaling for normalization [21].

4.2. Machine Learning Models:

KNN (K-nearest neighbor) algorithm is a classification method that relies on measuring distances between data points. When it comes to classifying a new, unseen sample "s" with N attributes, the algorithm selects the K closest training samples based on the Euclidean Distance[22] (as defined by Equation 1). In this research paper, a value of K = 5 has been chosen for consideration.

$$\text{Euclidean Distance}(s, s_i) \;=\; \sqrt{\textstyle\sum_{j=1}^{N}(s_j - s_{ij})^2} \qquad (1)$$

Support Vector Machine (SVM) is a technique that measures a hyperplane representing the optimal separation boundary between two classes. This hyperplane aims to maximize the distances between itself and the closest positive and negative examples, with each example being represented as a vector within a high-dimensional space. A feature vector is constructed using the specified set of features[23]. In this research paper, hyperparameter tuning is employed to optimize the model's performance on unseen data, and cross-validation is conducted by partitioning the dataset into five subsets.

Logistic Regression (LR) stands out as a widely used supervised machine learning model for the prediction of heart disease. LR is tailored for binary classification tasks, where the dependent variable takes on discrete or binary categorical values, such as 1 or 0[24]. LR employs the sigmoid function as its cost function, which maps predicted real values to probabilistic values within the range of 0 to 1. This characteristic facilitates the interpretation of the model's output as probabilities.

$$f(x) = \frac{1}{1+e^{-x}} \qquad (2)$$

Regularization[25] is a method used to mitigate overfitting and enhance the generalization ability of a model by introducing a penalty term to the loss function. In this work L2 regularization was implemented.

4.3. Ensemble Machine Learning Models:

Random Forest stands as a classification algorithm grounded in the principles of decision trees[26]. True to its name, this algorithm assembles a forest comprising numerous individual trees. It falls within the ensemble algorithm category, a group of techniques that harness multiple algorithms to formulate predictions. Random Forest generates a collection of decision trees through the utilization of random subsets from the training dataset. This procedure is iteratively executed with various random subsets, and the final decision is determined by a majority consensus among these trees. One of the distinctive strengths of the Random Forest algorithm lies in its effective handling of missing data, yet it is susceptible to overfitting. To mitigate this overfitting risk, practitioners employ appropriate parameter tuning techniques.

Adaboost is an algorithm that employs the boosting technique to enhance the performance of less potent learners[27]. Initially, a classifier is trained on the original dataset. Afterward, multiple iterations involve training additional copies of the classifier, each with the aim of rectifying errors made by its predecessor. This process generates diverse subsets of the dataset by assigning varying weights to data elements. Instances that are incorrectly classified are accorded greater weights, increasing their likelihood of inclusion in subsequent subsets. This iterative procedure continues, resulting in the training of several models sequentially. These feeble classifiers are then amalgamated using a cost function to create a robust classifier. The final prediction is influenced by the accuracy of each classifier, with higher accuracy classifiers carrying more weight. Notably, the AdaBoost algorithm provides the flexibility to specify the weak classifier to be boosted as a parameter.

To evaluate the performance of a model the following metrics are examined[28]:

Accuracy: The accuracy metric of the model is used to define its performance across all classes. Accuracy helps when all classes are equally important. It can be calculated as the ration of the total number of predictions to the number of predictions that were accurate.

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \qquad (3)$$

Recall: The recall measures how well the model can classify Positive samples. The recall is calculated as the ratio of Positive samples that were correctly classify as Positive to all Positive samples. The more positive samples that are identified, the recall value is larger.

$$Recall = \frac{t_p}{t_p + + f_p} \qquad (4)$$

Precision: The precision is obtained as the ratio of Positive samples that were classified correctly to all the samples that were classified as Positive (either incorrectly or correctly). Precision measures how well the model categorizes a sample as positive.

$$Precision = \frac{t_p}{t_p + + f_n} \qquad (3)$$

F1-score: The F1 score is precisely calculated as the harmonic average between recall and precision.

$$F1\ score = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad (4)$$

Where: t_p is correctly predicted, f_p is incorrectly predicted instances, t_n is negatively predicted instances and f_n is the negatively predicted instances.

5. Result and Discussion

In this research, we assessed various machine learning models for predicting strokes. The evaluated models encompassed k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Random Forest, and AdaBoost. We provide an overview of essential evaluation metrics, such as precision, accuracy, F1-score and recall. Table 2 displays the performance metrics resulting from this analysis.

The accuracy metric gauges the overall accuracy or correctness of the model's predictions. Among the models, AdaBoost outperforms with the highest accuracy at 99%, closely followed by Random Forest with an accuracy of 98.43%. These models demonstrate excellent predictive capabilities in identifying stroke cases. Precision indicates the ability of a model to make accurate positive predictions.

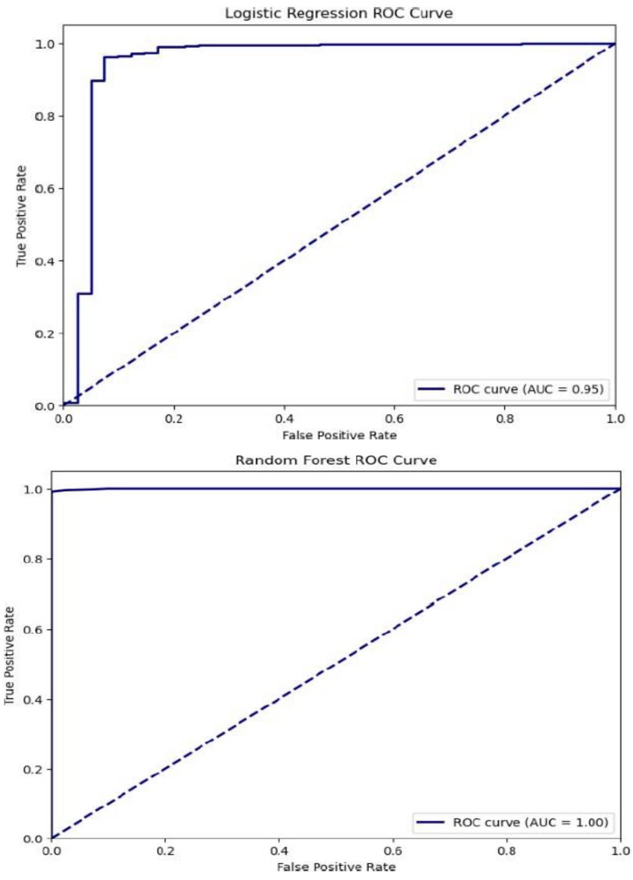
Logistic Regression, SVM, Random Forest, KNN, and AdaBoost all exhibit high precision values, with AdaBoost and Random Forest achieving perfect precision scores of 1. This suggests that these models are highly reliable in correctly identifying individuals at risk of stroke without generating many false positive predictions.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9659	0.9681	0.995	0.981
SVM	0.9584	0.9604	0.995	0.977
Random Forest	0.9843	0.997	0.995	0.996
KNN	0.956	0.958	0.995	0.976
AdaBoost	0.99	1	1	1

Table 2: Performance Metrics of the Models

Across all models, there is consistently high recall, with values of 0.995 or higher. This signifies that the models are effective in capturing most stroke cases within the dataset. All models, except Logistic Regression, achieved F1-Scores of 0.976 or higher, indicating strong overall performance.

The AUC metric evaluates a model's capacity to differentiate between negative and positive classes. An AUC value of 1 signifies flawless accuracy, while a value of 0 indicates incorrect test performance.



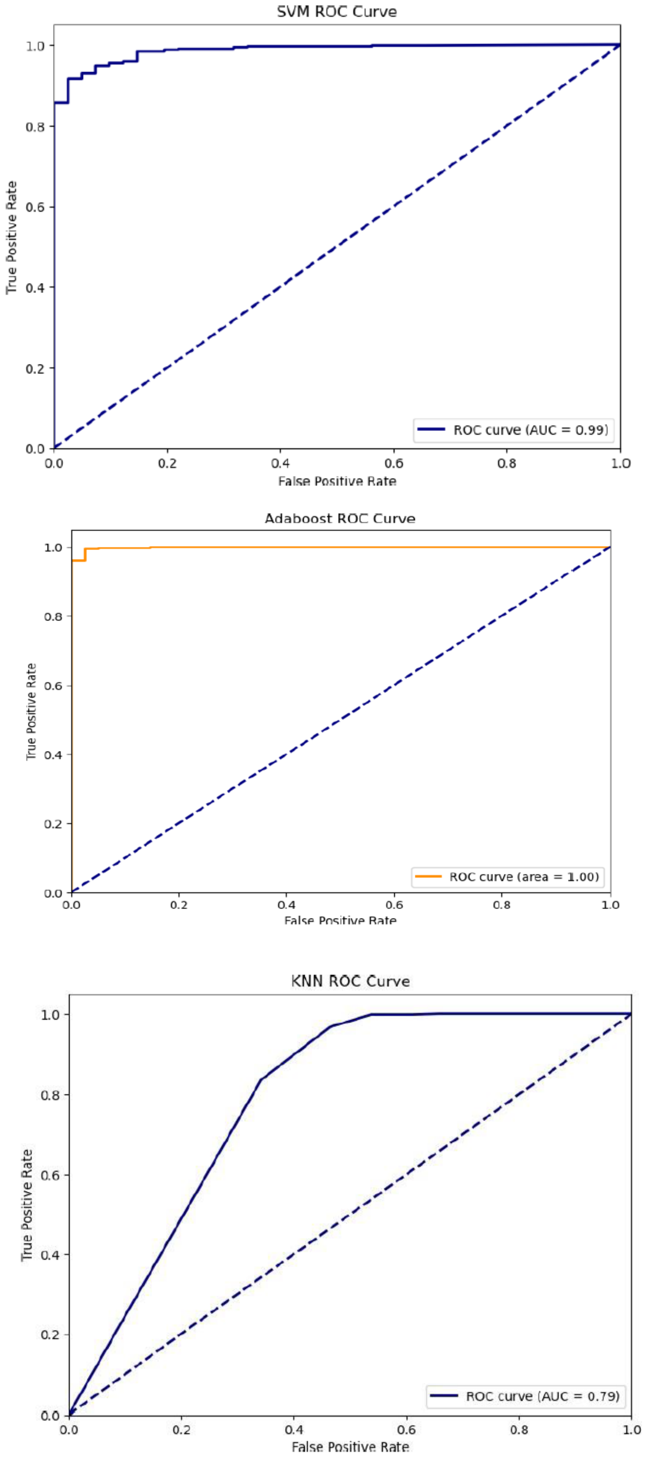


Figure 3: ROC Curve of the Learning Models

In Figure 3, the ROC curve is presented, and notably, among all the models, the ensemble learning model AdaBoost attains the highest AUC Area Under the Curve (AUC) with a perfect score of 100%. This outstanding performance underscores AdaBoost's exceptional predictive capability of Stroke Prediction.

5.1. Comparison with existing work:

Table 3 provides comparison with existing work.

Author and year	Methodology	Result	Proposed Methodology
Das et. Al,2023[3]	Random Forest	Accuracy-98.4%	Random Forest (Accuracy)-98.43%
Rahman et al., 2023 [30]	ANN	Accuracy-92.39%	KNN (Accuracy)-95.6%
Sharma et al., 2022 [31]	Random Forest	Accuracy-98.94%	Adaboost (Accuracy)-99%
Rana et al., 2021[32]	ANN	ROC-0.84	Adaboost (ROC)-100%

6. Conclusion

A stroke is a serious medical condition that requires immediate treatment to prevent it from worsening further. In this research, we have investigated the durability of diverse ML (machine learning) and Ensemble machine learning models for prediction of heart strokes. Our findings highlight the promise of the Adaboost model as a valuable tool for classifying stroke occurrences. This suggests that Adaboost could play a crucial role in enhancing the accuracy of stroke prediction, ultimately aiding healthcare professionals in making timely and informed decisions.

However, our research represents just one step on the path to improving stroke prediction. Future studies can delve deeper into understanding feature importance through advanced techniques, such as deep neural networks. Furthermore, sensitivity analysis can be employed to identify specific physiological indicators that could serve as targeted control points for stroke prevention.

Through these ongoing endeavors, we aim to push the boundaries of stroke prediction, enhance the accuracy of models, and mature our understanding of this critical healthcare challenge. Ultimately, these advancements can translate into more effective early interventions, potentially saving lives and improving the quality of care for individuals at risk of stroke. Our research underscores the importance of continued exploration in this field, with the goal of reducing the devastating impact of strokes on individuals and healthcare systems worldwide.

References

[1] Gorelick, P.B., Scuteri, A., Black, S.E., DeCarli, C., Greenberg, S.M., Iadecola, C., Launer, L.J., Laurent, S., Lopez, O.L., Nyenhuis, D., Petersen, R.C., Schneider, J.A., Tzourio, C., Arnett, D.K., Bennett, D.A., Chui, H.C., Higashida, R.T., Lindquist, R., Nilsson, P.M., Roman, G.C., Sellke, F.W., Seshadri, S.: Vascular Contributions to Cognitive Impairment and Dementia: A Statement for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 42, 2672–2713 (2011). https://doi.org/10.1161/STR.0b013e3182299496.

[2] Das, M.C., Liza, F.T., Pandit, P.P., Tabassum, F., Mamun, M.A., Bhattacharjee, S., Kashem, M.S.B.: A comparative study of machine learning approaches for heart stroke prediction. In: 2023 International Conference on Smart Applications, Communications and Networking (SmartNets). pp. 1–6. IEEE, Istanbul, Turkiye (2023). https://doi.org/10.1109/SmartNets58706.2023.10216049.

[3] Learn about Stroke: . [(accessed on 25 May 2022)]. Available online: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke.

[4] European Stroke Initiative Executive Committee, EUSI Writing Committee, Olsen, T.S., Langhorne, P., Diener, H.C., Hennerici, M., Ferro, J., Sivenius, J., Wahlgren, N.G., Bath, P.: European Stroke Initiative Recommendations for Stroke Management-update 2003. Cerebrovasc Dis. 16, 311–337 (2003). https://doi.org/10.1159/000072554.

[5] Emon, M.U., Keya, M.S., Meghla, T.I., Rahman, Md.M., Mamun, M.S.A., Kaiser, M.S.: Performance Analysis of Machine Learning Approaches in Stroke Prediction. In: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 1464–1469. IEEE, Coimbatore, India (2020). https://doi.org/10.1109/ICECA49313.2020.9297525.

[6] Dev, S., Wang, H., Nwosu, C.S., Jain, N., Veeravalli, B., John, D.: A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics. 2, 100032 (2022). https://doi.org/10.1016/j.health.2022.100032.

[7] Uttam, A.K.: Analysis of Uneven Stroke Prediction Dataset using Machine Learning. In: 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). pp. 1209–1213. IEEE, Madurai, India (2022). https://doi.org/10.1109/ICICCS53718.2022.9788309.

[8] Khosla, A., Cao, Y., Lin, C.C.-Y., Chiu, H.-K., Hu, J., Lee, H.: An integrated machine learning approach to stroke prediction. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 183–192. ACM, Washington DC USA (2010). https://doi.org/10.1145/1835804.1835830.

[9] Paikaray, D., Mehta, A.K.: An Extensive Approach Towards Heart Stroke Prediction Using Machine Learning with Ensemble Classifier. In: Dua, M., Jain, A.K., Yadav, A., Kumar, N., and Siarry, P. (eds.) Proceedings of the International Conference on Paradigms of Communication, Computing and Data Sciences. pp. 767–777. Springer Singapore, Singapore (2022). https://doi.org/10.1007/978-981-16-5747-4_66.

[10] Kumar, K.L., Reddy, B.E.: Heart Disease Detection System Using Gradient Boosting Technique. In: 2021 International Conference on Computing Sciences (ICCS).

pp. 228–233. IEEE, Phagwara, India (2021). https://doi.org/10.1109/ICCS54944.2021.00052.

[11] Singh, M.S., Choudhary, P., Thongam, K.: A Comparative Analysis for Various Stroke Prediction Techniques. In: Nain, N., Vipparthi, S.K., and Raman, B. (eds.) Computer Vision and Image Processing. pp. 98–106. Springer Singapore, Singapore (2020). https://doi.org/10.1007/978-981-15-4018-9_9.

[12] Bandi, V., Bhattacharyya, D., Midhunchakkravarthy, D.: Prediction of Stroke Severity Using Machine Learning. RIA. 34, 753–761 (2020). https://doi.org/10.18280/ria.340609.

[13] Kaur, M., Sakhare, S.R., Wanjale, K., Akter, F.: Early Stroke Prediction Methods for Prevention of Strokes. Behavioural Neurology. 2022, 1–9 (2022). https://doi.org/10.1155/2022/7725597.

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

[15] Sailasya, G., Kumari, G.L.A.: Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. IJACSA. 12, (2021). https://doi.org/10.14569/IJACSA.2021.0120662.

[16] Chin, C.-L., Lin, B.-J., Wu, G.-R., Weng, T.-C., Yang, C.-S., Su, R.-C., Pan, Y.-J.: An automated early ischemic stroke detection system using CNN deep learning algorithm. In: 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST). pp. 368–372. IEEE, Taichung (2017). https://doi.org/10.1109/ICAwST.2017.8256481.

[17] Li, X., Bian, D., Yu, J., Li, M., Zhao, D.: Using machine learning models to improve stroke risk level classification methods of China national stroke screening. BMC Med Inform Decis Mak. 19, 261 (2019). https://doi.org/10.1186/s12911-019-0998-2.

[18] Stroke Prediction Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[19] Al-Zubaidi, H., Dweik, M., Al-Mousa, A.: Stroke Prediction Using Machine Learning Classification Methods. In: 2022 International Arab Conference on Information Technology (ACIT). pp. 1–8. IEEE, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.1109/ACIT57182.2022.10022050.

[20] Singh, D., Singh, B.: Feature wise normalization: An effective way of normalizing data. Pattern Recognition. 122, 108307 (2022). https://doi.org/10.1016/j.patcog.2021.108307.

[21] Pawlovsky, A.P.: An ensemble based on distances for a kNN method for heart disease diagnosis. In: 2018 International Conference on Electronics, Information, and Communication (ICEIC). pp. 1–4. IEEE, Honolulu, HI, USA (2018). https://doi.org/10.23919/ELINFOCOM.2018.8330570.

[22] Çınar, A., Tuncer, S.A.: Classification of normal sinus rhythm, abnormal arrhythmia and congestive heart failure ECG signals using LSTM and hybrid CNN-SVM deep neural networks. Computer Methods in Biomechanics and Biomedical Engineering. 24, 203–214 (2021). https://doi.org/10.1080/10255842.2020.1821192.

[23] Majumder, A.B., Gupta, S., Singh, D.: An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour. J. Phys.: Conf. Ser. 2286, 012017 (2022). https://doi.org/10.1088/1742-6596/2286/1/012017.

[24] Yang, Z., Liang, Y., Zhang, H., Chai, H., Zhang, B., Peng, C.: Robust Sparse Logistic Regression With the L_{q} ($0 < \text{q} < 1$) Regularization for Feature Selection Using Gene Expression Data. IEEE Access. 6, 68586–68595 (2018). https://doi.org/10.1109/ACCESS.2018.2880198.

[25] Babu, G.H., Jayasree, G., Ashika, C., Ahalya, V., Niroopa, K.A.: Heart Disease Prediction System Using Random Forest Technique. IJRASET. 11, 1133–1141 (2023). https://doi.org/10.22214/ijraset.2023.48764.

[26] Li, R., Shen, S., Chen, G., Xie, T., Ji, S., Zhou, B., Wang, Z.: Multilevel Risk Prediction of Cardiovascular Disease based on Adaboost+RF Ensemble Learning. IOP Conf. Ser.: Mater. Sci. Eng. 533, 012050 (2019). https://doi.org/10.1088/1757-899X/533/1/012050.

[27] Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 21, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7.

[28] Mishra, I., Mohapatra, S.: An enhanced approach for analyzing the performance of heart stroke prediction with machine learning techniques. Int. j. inf. tecnol. 15, 3257–3270 (2023). https://doi.org/10.1007/s41870-023-01321-8.

[29] Sharma, C., Sharma, S., Kumar, M., Sodhi, A.: Early Stroke Prediction Using Machine Learning. In: 2022 International Conference on Decision Aid Sciences and Applications (DASA). pp. 890–894. IEEE, Chiangrai, Thailand (2022). https://doi.org/10.1109/DASA54658.2022.9765307.

[30] Rana, C., Chitre, N., Poyekar, B., Bide, P.: Stroke Prediction Using Smote-Tomek and Neural Network. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–5. IEEE, Kharagpur, India (2021). https://doi.org/10.1109/ICCCNT51525.2021.9579763.