

Boosting stroke prediction with ensemble learning on imbalanced healthcare data

Outmane Labaybi¹, Mohamed Bennani Taj¹, Khalid El Fahssi², Said El Garouani²,
Mohamed Lamrini¹, Mohamed El Far¹

¹Laboratory LPAIS, Department of Computer Science, Faculty of Science Dhar El Mahraz, University Sidi Mohamed Ben Abdellah, Fez, Morocco

²Laboratory LISAC, Department of Computer Science, Faculty of Science Dhar El Mahraz, University Sidi Mohamed Ben Abdellah, Fez, Morocco

Article Info

Article history:

Received Mar 25, 2024

Revised Nov 6, 2024

Accepted Nov 11, 2024

Keywords:

Artificial intelligent

Ensemble learning

Imbalanced dataset

Random forest

Stroke prediction

ABSTRACT

Detecting strokes at the early day is crucial for preventing health issues and potentially saving lives. Predicting strokes accurately can be challenging, especially when working with unbalanced healthcare datasets. In this article, we suggest a thorough method combining machine learning (ML) algorithms and ensemble learning techniques to improve the accuracy of predicting strokes. Our approach includes using preprocessing methods for tackling imbalanced data, feature engineering for extracting key information, and utilizing different ML algorithms such as random forests (RF), decision trees (DT), and gradient boosting (GBoost) classifiers. Through the utilization of ensemble learning, we amalgamate the advantages of various models in order to generate stronger and more reliable predictions. By conducting thorough tests and assessments on a variety of datasets, we demonstrate the efficacy of our approach in addressing the imbalanced stroke datasets and greatly enhances prediction accuracy. We conducted comprehensive testing and validation to ensure the reliability and applicability of our method, improving the accuracy of stroke prediction and supporting healthcare planning and resource allocation strategies.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mohamed Bennani Taj

Laboratory LPAIS, Department of Computer Science, Faculty of Science Dhar El Mahraz

University Sidi Mohamed Ben Abdellah

Fez, Morocco

Email: bennani.taj@gmail.com

1. INTRODUCTION

According the world health organization (WHO), about 15 million people have strokes [1] every year all over the world [2], [3]. WHO defines stroke as a brain-related illness that leads to the dysfunction of the brain. There are two types of strokes, hemorrhagic stroke (when a blood vessel breaks and causes bleeding in the brain) and ischemic stroke (when a blood vessel gets blocked) [4], [5]. This dysfunction is mainly a result of vessel problems, and it lasts for longer than 24 hours. It's important to know the type of stroke because treatments depend on it. Detecting stroke early is crucial for better treatment results [6].

It's a critical medical condition, that requires accurate and timely prediction to facilitate preventive measures and improve patient outcomes. Traditional models often face challenges in handling complex patterns and relationships within healthcare datasets [7]. To address these challenges, we explore a scientific approach that leverages the synergy of ensemble learning, hyperparameter tuning and machine learning (ML) algorithms.

ML, it's a part of artificial intelligence, has revolutionized healthcare by providing tools to analyze vast datasets, detect patterns, and make predictions. In the context of stroke, ML algorithms offer the potential to refine risk prediction models, contributing to early diagnosis and preventive strategies.

An increasing number of researches have investigated the application of ML models in stroke prediction in the last decade. Maheshwari *et al.* [8] provides a study of various risk factors to understand the probability of stroke. It used a regression-based approach to identify the relation between a factor and its corresponding impact on stroke.

Exploring a Kaggle dataset, Sailasya and Kumari [9] delved into stroke prediction using various ML algorithms, including logistic regression (LR), k-nearest neighbour (KNN), random forest (RF), support vector machine (SVM), Naïve Bayes (NB), and decision tree (DT) algorithms. To address imbalanced data, an undersampling method was employed. The findings revealed that NB exhibited the highest performance, boasting an overall accuracy of 82%. In comparison, KNN and SVM both achieved an 80% accuracy, while LR yielded a slightly lower accuracy of 78%.

Nwosu *et al.* [10] harnessed electronic health records and utilized a dataset provided by McKinsey and company, encompassing 11 distinct attributes such as body mass index, heart disease, marital status, age, average blood glucose, and smoking status. Within this dataset, 548 patients had experienced a stroke, while 28524 patients had not encountered any prior strokes. Due to the dataset's imbalance, 1000 downsizing experiments were conducted to mitigate sampling bias. Subsequently, 70% of the dataset was allocated for training, with the remaining 30% reserved for testing purposes. Across the 1000 downsizing experiments, the neural network model demonstrated superior performance, achieving the highest accuracy at 75.02%. Following closely, the RF model attained an accuracy of 74.53%, and the DT model exhibited an accuracy of 74.31%.

In the study referenced as [11], the researchers opted for intricate algorithms like ADABOOST and XGB, achieving outcomes comparable to ours. However, our study achieved impressive results using simpler algorithms, a more preferable and efficient approach. In a study by Sailasya and Kumari [9], similar to the findings in reference [12], the Kaggle dataset was utilized along with various algorithms including DT, NB, SVM, RF, KNN, and LR. Their results demonstrated that DT outperformed the other algorithms, achieving the highest performance, followed by KNN with an accuracy of 96.3%.

The problem is that the rising incidence of strokes emphasizes the need for effective prediction models that accurately identify individuals at risk. While previous studies have tackled this issue using classical methods, these approaches have not produced satisfactory results.

Our contribution aims to investigate how ML and ensemble learning techniques can be used to predict strokes. By examining the contributions of each methodology and their synergies, the research seeks to provide a comprehensive understanding of how advanced computational techniques can be harnessed to enhance accuracy, interpretability, and clinical relevance in stroke prediction models.

The rest of this article is structured into several sections, including one that describes our method. Another section is dedicated to results and discussion, presenting research conclusions along with comparisons to other similar techniques. Finally, a section summarizes the findings and suggests directions for future research.

2. METHOD

In the sections that follow, a detailed description of the methods that was used for this work is provided. In section 2.1, the details of the dataset that was used are explained. Section 2.2 outlines the data preprocessing technique. In section 2.3, the proposed method is presented, while in section 2.4, the ML algorithms used for stroke prediction are explained in greater detail.

2.1. Description of dataset

The dataset used in our study is called 'stroke prediction dataset', it contains important information from medical records, like whether a patient has hypertension, heart disease, various physiological and environmental details. The dataset is organized into rows and columns, with each row representing a different patient.

The dataset has 5110 rows, and each row is info about one person. There are 12 columns. Ten tell us things about the people, like health conditions. One column has an ID, and another says if the person had a stroke (1) or not (0). The dataset is not balanced; 4861 people are normal, and 249 had a stroke. This imbalance might affect our models, so we're going to fix it during training to make things more even. You can get this dataset on Kaggle using this link: [stroke prediction dataset](#). We're studying this dataset to build good models for predicting strokes, keeping in mind the uneven number of normal and stroke cases.

2.2. Data preprocessing

Handle missing values: the way we fill in missing data depends on what kind of data it is and what the dataset is like. The BMI column has 201 and the smoking status has 1544 missing values. To handle these missing values, we have a more options. For us, because BMI and smoking status are important factors and we're missing quite a few values for them, it makes sense to fill in those missing values. For these we use the KNN algorithm to impute BMI missing values, for each missing value, find its KNN based on other features and for the smoking status we utilize the RF algorithm to impute missing values. Train a RF model on the subset of data with complete information, and predict the missing values based on other features.

Encode categorical variables: converting categorical variables to numerical format using a method called one-hot encoding is a common preprocessing step to create binary columns (0 or 1) for each category, effectively transforming it into a set of numerical features.

Imbalanced dataset handling: handling imbalanced datasets is crucial in ML, as models trained on such datasets might have a bias towards the majority class [13]. In the context of a stroke dataset, where strokes are likely to be a minority class, addressing the imbalance Figure 1 is important for creating a reliable and effective model. There are several methods to compensate for an imbalance of classes in a dataset. Depending on the amount of data available, we will then choose one or other of the following methods.

Data scaling: in stroke prediction models, the dataset may contain numerical features such as age, blood pressure, avg_glucose_level, and BMI Figure 2. These features can have vastly different scales and units, which can affect the performance of ML algorithms [14]. In this study, we use the min-max scaling (). This technique scales the features to a fixed range, typically between 0 and 1 Figure 3.

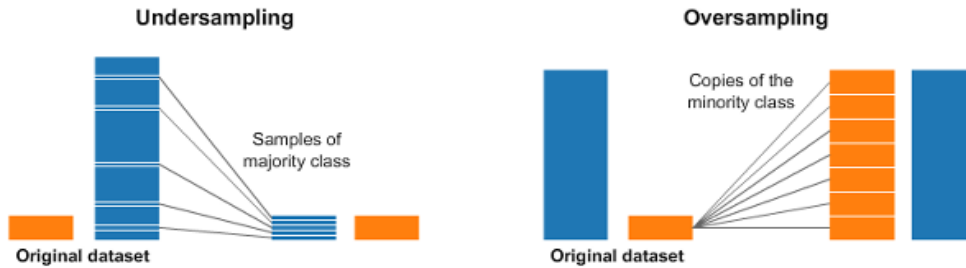


Figure 1. Undersampling and oversampling

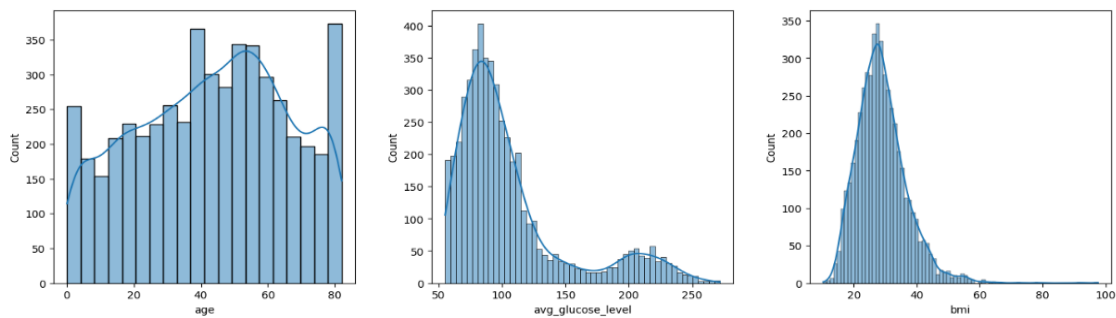


Figure 2. Original numerical features

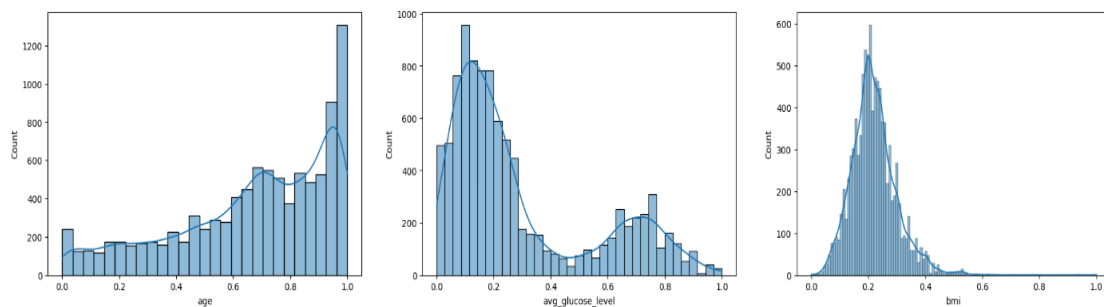


Figure 3. Dataset scaled after using MinMaxScaler

$$X_{scaled} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

It is calculated using the eq4 where X represents the original value of the feature, Xmin is the smallest value of the feature, and Xmax is the largest value of the feature. This preprocessing step enhances the performance and interpretability of the models, ultimately leading to better healthcare outcomes.

2.3. Proposed method

At the beginning of our experiment, presented in Figure 4 we carefully prepared the dataset to make sure it was good to use. After we impute and fixed any missing information, decode the categories variables to numbers using one hot encoding, and made sure all the data was in the same range.

Next step, we split our dataset into training dataset which had 80% of the data, and testing dataset which had 20%. Doing this helped us see how well the models worked. Then, we dealt with the problem of there being more of one type of data than the other in the training set. We used a technique called SMOTE to make more of the minority class (cases where people had strokes). This was to help the models learn better.

Once we had a balanced training set, we taught three different models RF, XGBoost, and SVM using that data. Each model learned the patterns in the data. To make the predictions even better, we used a method called model stacking. This means we put the three models together and let them learn from each other. This helped make the predictions more accurate by using the strengths of each model.

After stacking the models, we combined their predictions using a voting system. This let us use all the models together to make predictions. By doing this, we could make better predictions overall. Finally, we tested how well our combined model worked using the test set. We examined its accuracy, precision, recall, and F1-score to see how well it predicted the likelihood of someone having a stroke.

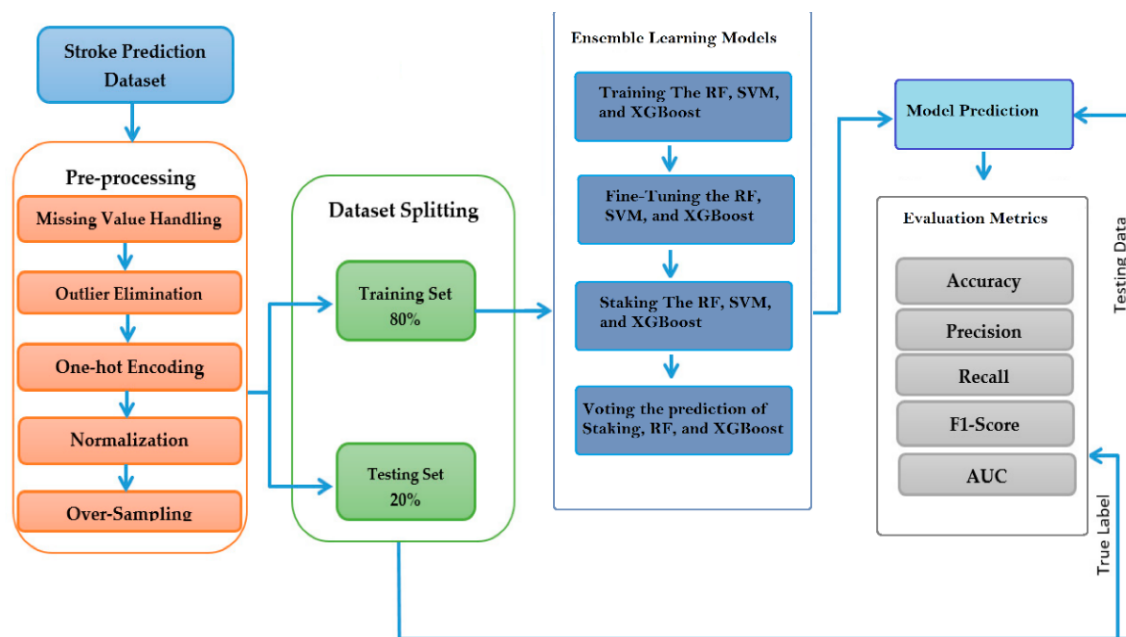


Figure 4. Illustrates the workflow for predicting strokes using the provided dataset

2.4 Machine learning models

2.4.1. Random forest

RF is a strong computer method that combines many DT and combines their guesses to make better predictions overall. It's really good at dealing with lots of information and complicated connections between different parts of the data, which makes it great for predicting strokes. Our results corroborate previous findings demonstrating the efficacy of RF in healthcare applications [15]. By harnessing the ensemble nature of RF, we enhance the robustness and generalization capability of our predictive model, offering valuable insights into stroke risk assessment.

2.4.2. Support vector machines

SVM [16] are powerful supervised learning algorithms that find the optimal hyperplane to separate classes in the feature space. SVMs maximize the margin between the classes and can handle non-linear decision boundaries using kernel functions. SVM [17] are good at dividing data into two groups and are famous for being able to work well even with new data they haven't seen before.

2.4.3. Ensemble learning

Many studies in healthcare, as shown in references [18], [19], have used ensemble learning. These studies focus mainly on using the same type of machine-learning methods as their basic tools, which are often called weak learners. Three popular methods, bootstrap aggregating (bagging), stacking, and boosting, combine these weak learners.

Bagging works by training many copies of the same basic learning method on different parts of the training data Figure 5, chosen randomly but with replacement. When predicting strokes, bagging could be used with DT or RF. In this method, each tree in the group is trained on a different random sample of the dataset. Bagging helps make predictions more reliable and accurate by combining the predictions of multiple models [20].

$$\hat{f}_{\text{bagging}}(x) = \frac{1}{B} \sum_{b=1}^B f_b(x) \tag{2}$$

- $\hat{f}_{\text{bagging}}(x)$ is the ensemble prediction.
- B the number of basic learners.
- $f_b(x)$ refers to the prediction made by the b -th basic learner.

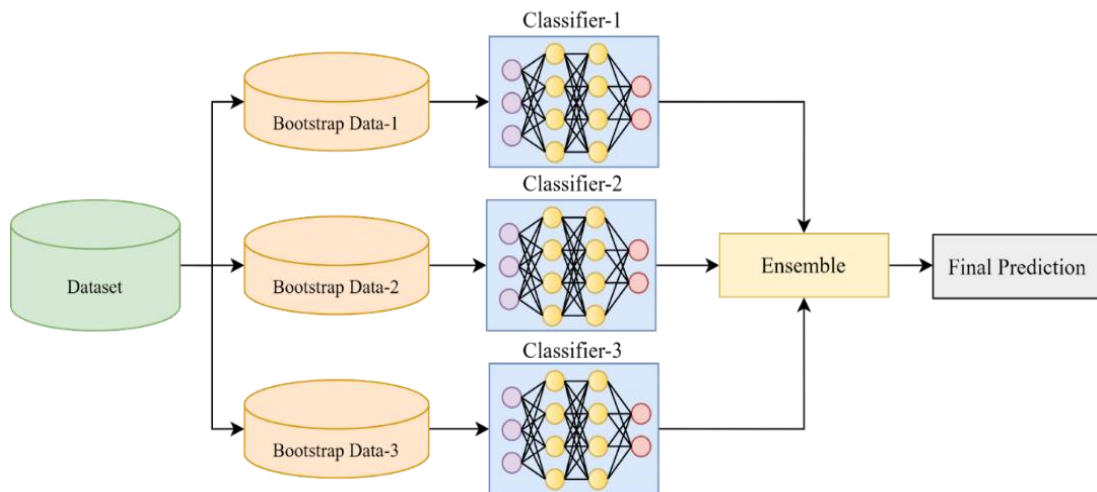


Figure 1. Bagging ensemble

The diagram in the Figure 4 illustrates how the Bagging ensemble method works. In Bagging, multiple processes happen simultaneously. The primary goal of Bagging is to decrease variability in the predictions made by the ensemble.

Boosting is a method where weak learners are trained one after another in a series of steps, where each new model focuses on the instances that were misclassified by the previous models Figure 6. Algorithms like adaptive boosting (AdaBoost) and gradient boosting machines (GBM) are commonly used boosting methods. In the context of stroke prediction, boosting algorithms could be applied to DT or other weak learners to iteratively improve the prediction accuracy by emphasizing difficult-to-classify instances related to stroke risk factors [21].

$$\hat{f}_{\text{boosting}}(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \dots + \alpha_B f_B(x) \tag{3}$$

- $\hat{f}_{\text{boosting}}(x)$ is the ensemble prediction.
- α_i is the weight assigned to the i -th base learner.

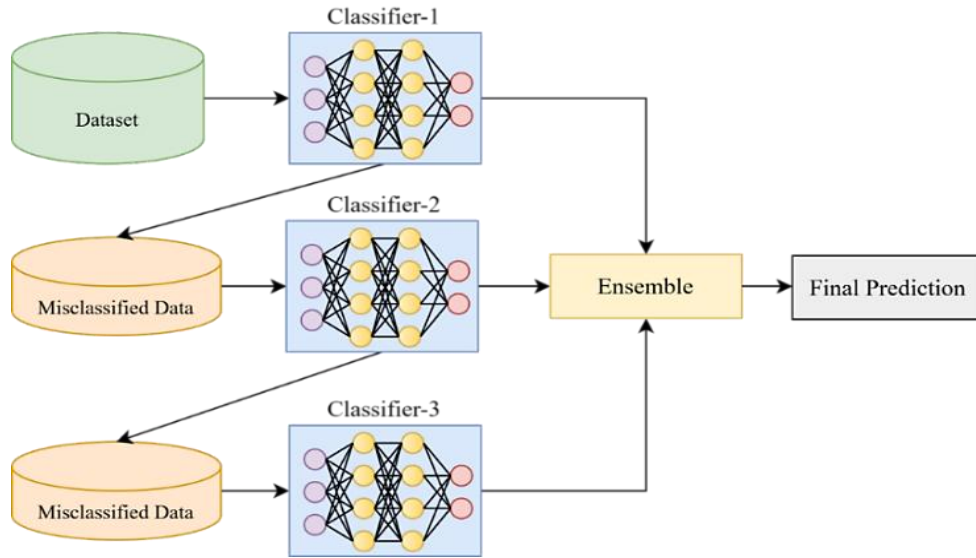


Figure 6. Illustrates the boosting ensemble method

The main goal of boosting is to decrease the mistakes made in the ensemble decision. So, the classifiers picked for the group usually should have less chance of being wrong but may be simpler, with fewer things to learn.

Stacking also known as stacked generalization, mixes the guesses from many different basic models using a special learner. These basic models can be various kinds of computer methods trained on the same data. For example, for predicting strokes, stacking might use different models like DT, LR, and SVM on the stroke data. Then, another special learner (like LR or another computer method) is trained on the guesses of these basic models to make the final guess [22].

$$\hat{f}_{\text{stacking}}(x) = g(\sum_{b=1}^B \alpha_b f_b(x)) \tag{4}$$

- $\hat{f}_{\text{stacking}}(x)$ is the ensemble prediction.
- g is the meta-learner.
- α_b is the weight assigned to the b -th base learner.
- $f_b(x)$ is the prediction from the b -th base learner.

In the diagram above Figure 7, we see one level of stacking. However, there are also more complex stacking methods with multiple layers of classifiers added in between.

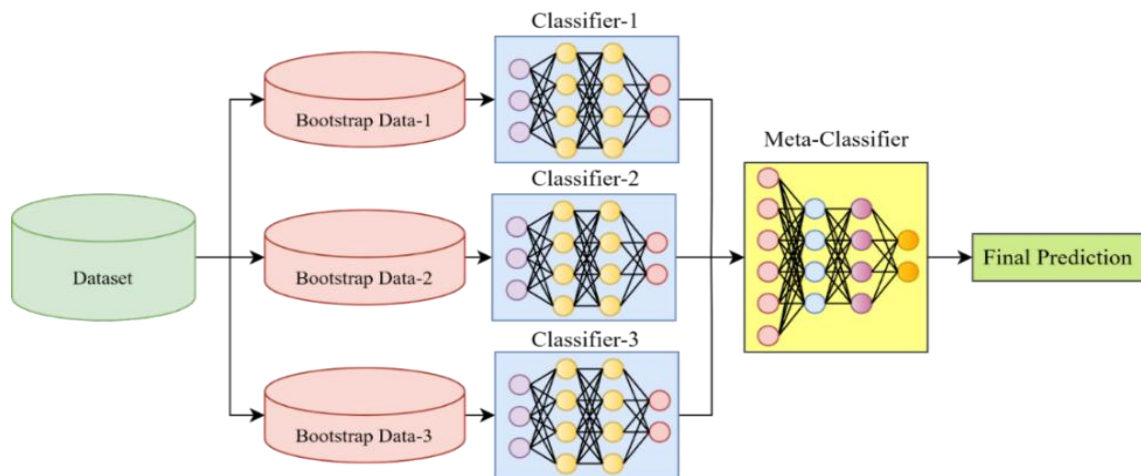


Figure 7. Illustrates how the stacking ensemble method works

2.5. Performance evaluation metrics

To see how well a classification algorithm is doing, we use different methods. One of them is called the confusion matrix shown in Figure 8. It's a table that shows how well a supervised learning algorithm is doing. Each row represents the actual instances of a class, and each column represent the predicted instances of a class. From this table, we can calculate all the metrics to evaluate the performance of the algorithm [23].

		Predicted 0	Predicted 1
Actual 0	TN	FP	
Actual 1	FN	TP	

Figure 8. Confusion matrix

Accuracy: the proportion of correct predictions made by the classifier.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{5}$$

Recall: known as the true negative rate, is calculated by using (6).

$$Recall = \frac{TP}{(TP+FN)} \tag{6}$$

Precision indicates the proportion of positive predictions that are actually correct. It measures how accurately the classifier identifies positive cases.

$$Precision = \frac{TP}{(TP+FP)} \tag{7}$$

The F1-score as shown in (8), it is calculated as the true positive divided by the sum of true positive and one-half of the sum of false positive and false negative.

$$F1 - Score = \frac{TP}{\left(TP + \frac{1}{2}(FP+FN)\right)} \tag{8}$$

Receiver operating characteristic (ROC) curves are a ROC curve, is a graphical plot which illustrates the performance of a binary classification algorithm as a function of true positive rate and false positive rate.

3. RESULTS AND DISCUSSION

This study introduced a new method to evaluate the effectiveness of four ML classification algorithms, along with one hybrid model, in predicting stroke. We assessed the performance of each model based on five key criteria: specificity, recall, precision, F1-score, and area under the curve (AUC). The performance evaluation criteria for the different ML algorithms are presented in Table 1.

Table 1. Model comparison for multiple metrics

Model	F1-score	Accuracy	Recall	Precision	ROC AUC
RF	98.9%	98.90%	100.00%	97.8%	98.8%
XGBoost	96.90%	96.80%	100.00%	94.00%	96.70%
SVM	90.40%	89.70%	94.8	86.40%	89.60%
LR	77.30%	76.20%	79.40%	75.30%	76.10%
Our proposed hybrid model using ML, Staking, and voting	99.70%	99.70%	100%	99.50%	99.70%

The RF model achieved an F1-score and accuracy of 98.90%, with a ROC AUC of 98.8%, indicating excellent performance. XGBoost followed closely with an F1-score of 96.90% and accuracy of 96.80%. The SVM model showed solid recall but lower precision, with a ROC AUC of 89.60%. LR had an F1-score of 77.30%, demonstrating reasonable performance but less effectiveness compared to the more complex models. The proposed hybrid model, utilizing ML, stacking, and voting techniques, outperformed all others, achieving an F1-score and accuracy of 99.70%, along with a ROC AUC of 99.70%, indicating superior performance.

To visualize these results, we create a bar plot for comparing the F1-score, accuracy, recall, precision, and ROC AUC for each model, this bar plot in Figure 9 represents the performance metrics of each model, enabling a clear comparison of their strengths and weaknesses, in Figure 10, we presented the ROC curve comparing the performance of different models. It's clear from the plot that the proposed hybrid model outperforms the other models across all metrics, indicating its superiority in classification tasks.

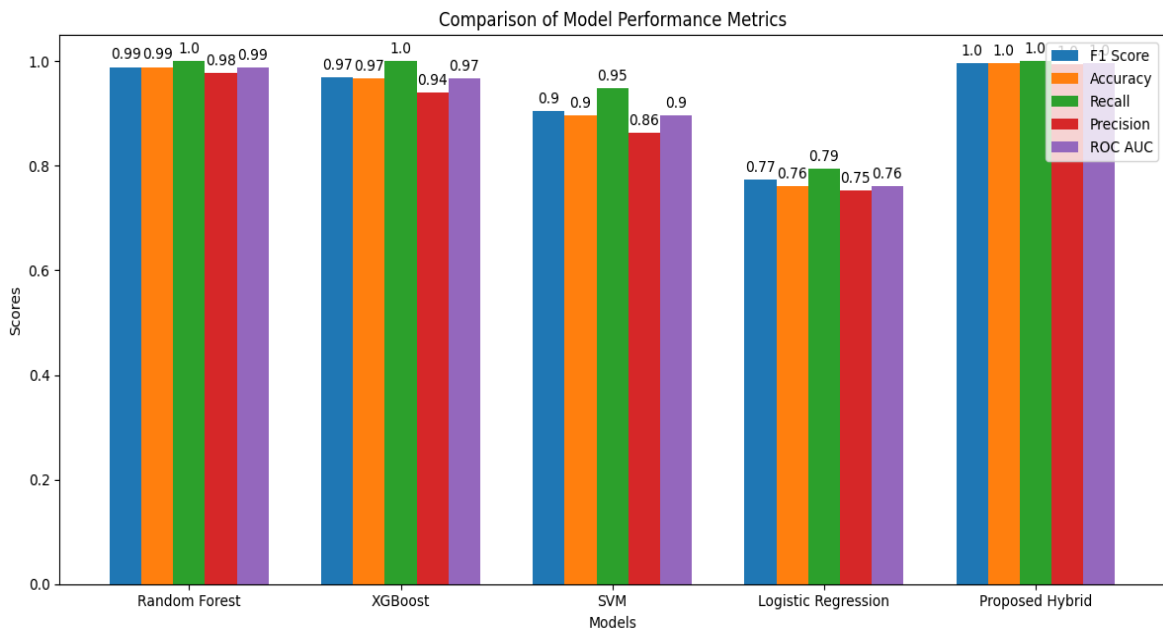


Figure 9. Comparison of model performance metrics

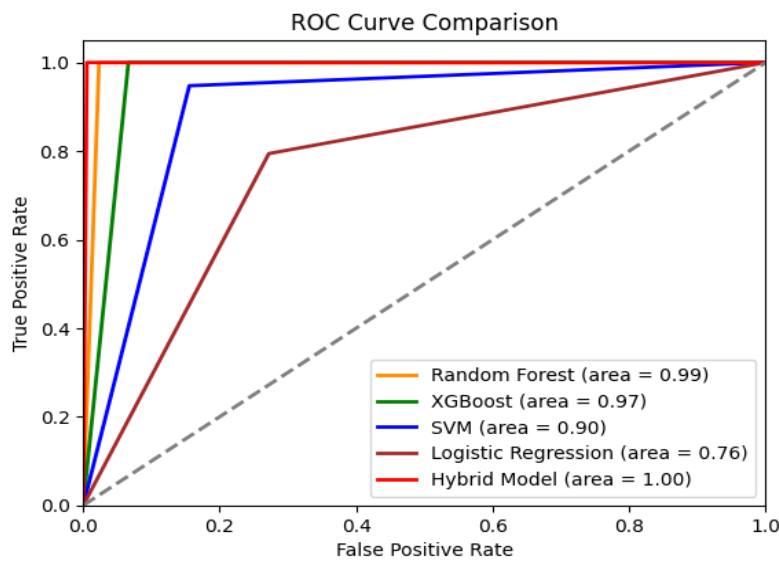


Figure 10. ROC curve

In our study, Figure 11 and Table 2 discusses and compares our proposed model with several related studies in the domain of stroke prediction. Firstly, Alamoudi and Abdallah [5] utilized LR, DT, RF, KNN, SVM, and NB classifiers, achieving an accuracy of 82% on a stroke prediction dataset. Similarly, Javale and Desai [19] employed RF, KNN, and LR, achieving a higher accuracy of 93.32%. Breiman *et al.* [20] utilized LR, DT, voting, and RF, attaining an impressive accuracy of 96%. Schapire *et al.* [21] explored RF, XGBoost, and LightGBM, achieving a high accuracy of 96.34%. Additionally, Drucker *et al.* [22] experimented with LR, SVM, artificial neural networks (ANN), XGBoost, and RF, obtaining a notable accuracy of 97%. In comparison to these studies, our first model, which used the base ML techniques like SVM, XGBoost, and RF, achieved an accuracy of 94.90%. Notably, this accuracy surpasses that of Alamoudi and Abdalla [5], Javale and Desai [19], although slightly lower than that achieved by Breiman [20], Schapire [21], and Drucker *et al.* [22], [23].

Table 2. comparison stroke prediction with related studies

Ref	Methodology	Accuracy	Dataset
Sailasya and Kumari [9]	LR, SVM, DT, KNN, RF, and NB was the best	82%	Stroke prediction dataset
Badriyah <i>et al.</i> [24]	RF, KNN, LR, RF which is the best	93.32%	Stroke prediction dataset
Tazin, <i>et al.</i> [25]	DT, LR, Voting, and RF was the best.	96%	Stroke prediction dataset
Alruily <i>et al.</i> [26]	RF, LightGBM, and XGBoost	96.34%	Stroke prediction dataset
Alhakami <i>et al.</i> [27]	LR, SVM, ANN, XGBoost, and RF, which was the best	97%	Stroke prediction dataset
proposed model using base ML	SVM, XGBoost, RF wich was the best	94.90%	Stroke prediction dataset
Proposed model using hybrid ML, ensemble learning	RF, SVM, XGBoost, staking, and voting	99.74%	Stroke prediction dataset

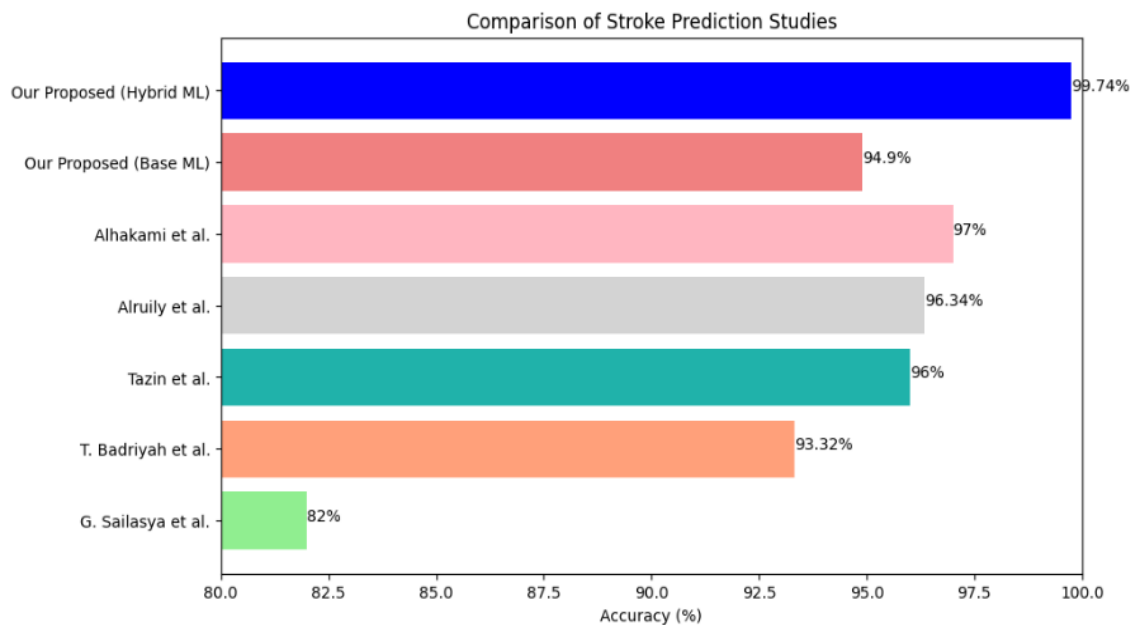


Figure 11. Comparison of stroke prediction studies

However, the important contribution of our study lies in the development of a novel hybrid ML approach, utilizing ensemble learning techniques such as stacking and voting. This hybrid model, combining RF, SVM, XGBoost, stacking, and voting, remarkably achieved an accuracy of 99.74%, outperforming all previous studies. The integration of ensemble learning techniques helped us make the most of the strengths of each model while compensating for their limitations, resulting in a highly accurate predictive model for stroke occurrence.

Our proposed hybrid model, which combines ML techniques such as stacking and voting, represents a significant advancement in stroke prediction compared to previous studies. By utilizing the strengths of multiple base models of ML and our approach, we achieved a remarkable accuracy of 99.74%, surpassing all previous studies in stroke prediction accuracy.

4. CONCLUSION

In this study, we've explored how ML and ensemble learning algorithms can be combined for predicting strokes, resulting in the development of a new hybrid model. Using a large stroke dataset, we've shown that our approach is effective in identifying individuals at risk of stroke accurately by combining different ML models and leveraging their unique strengths. However, our hybrid model achieves exceptional predictive accuracy, surpassing previous benchmarks in stroke prediction. Additionally, ensemble learning is essential in our hybrid model. Methods like stacking and voting help us merge insights from different models, which helps overcome the limitations of individual algorithms and enhances overall predictive accuracy. By using ensemble learning, we maximize the potential of our predictive model, offering strong and dependable stroke risk assessments for clinical decision-making. Finally, our study presents an innovative way to predict strokes by combining ML and ensemble learning techniques. We found that our hybrid model performs better than previous methods, offering precise risk assessments. Ensemble learning greatly improves the model's performance, showing how important it is to merge different algorithms for better accuracy.

In future work, we plan to implement association rules to further justify our results and explore deep learning models for image analysis to predict strokes with enhanced performance. This could significantly advance the use of predictive analytics in clinical settings, enabling healthcare professionals to make informed decisions based on robust data.

ACKNOWLEDGEMENTS

There are no significant findings to report. We are sharing this article solely for the benefit of the scientific community.




REFERENCES

- [1] V. L. Feigin *et al.*, "Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the global burden of disease study 2019," *The Lancet Neurology*, vol. 20, no. 10, pp. 795–820, Oct. 2021, doi: 10.1016/S1474-4422(21)00252-0.
- [2] V. Rajinikanth and S. C. Satapathy, "Segmentation of ischemic stroke lesion in brain MRI based on social group optimization and fuzzy-tsallis entropy," *Arabian Journal for Science and Engineering*, vol. 43, no. 8, pp. 4365–4378, Aug. 2018, doi: 10.1007/s13369-017-3053-6.
- [3] S. V. Shetty, H. Sarojadevi, S. Ankalaki, C. Dedepeya, S. Shreeraksha, and N. Ganavi, "Early detection of stroke using MRI images - A machine learning approach," in *AIP Conference Proceedings*, 2024, vol. 3122, no. 1, p. 080020, doi: 10.1063/5.0216076.
- [4] M. Rudiansyah, T. A. Sardjono, and R. Mardiyanto, "Segmentation of the intracerebral hemorrhagic strokes (Bleeds) from brain CT image based on GVF snake," in *Proceeding - 2018 International Seminar on Intelligent Technology and Its Application, ISITIA 2018*, Aug. 2018, pp. 465–470. doi: 10.1109/ISITIA.2018.8711155.
- [5] A. Alamoudi and Y. Abdallah, "Characterization of brain stroke using image and signal processing techniques," in *Biomedical Signal and Image Processing*, IntechOpen, 2021. doi: 10.5772/intechopen.96288.
- [6] W. S. Alazawee, Z. H. Naji, and W. T. Ali, "Analyzing and detecting hemorrhagic and ischemic stroke-based on bit plane slicing and edge detection algorithms," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 25, no. 2, pp. 1003–1010, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp1003-1010.
- [7] N. S. I. M. Rafei, R. Hassan, R. D. R. Saedudin, A. F. M. Raffei, Z. Zakaria, and S. Kasim, "Comparison of feature selection techniques in classifying stroke documents," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 14, no. 3, pp. 1244–1250, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1244-1250.
- [8] H. Maheshwari, D. Yadav, and U. Chandra, "Brain stroke prediction using the artificial intelligence," in *Communications in Computer and Information Science*, vol. 1742 CCIS, 2022, pp. 1–11. doi: 10.1007/978-3-031-23647-1_1.
- [9] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.
- [10] C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John, "Predicting stroke from electronic health records," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Jul. 2019, pp. 5704–5707. doi: 10.1109/EMBC.2019.8857234.
- [11] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, "Performance analysis of machine learning approaches in stroke prediction," in *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, Nov. 2020, pp. 1464–1469. doi: 10.1109/ICECA49313.2020.9297525.
- [12] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, Jun. 2022, doi: 10.3390/s22134670.
- [13] A. K. Hamoud *et al.*, "A prediction model based machine learning algorithms with feature selection approaches over imbalanced dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 28, no. 2, pp. 1105–1116, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp1105-1116.
- [14] T. A. Assegie, V. Elanangai, J. S. Paulraj, M. Velmurugan, and D. F. Devesan, "Evaluation of feature scaling for improving the performance of supervised learning methods," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 12, no. 3, pp. 1833–1838, Jun. 2023, doi: 10.11591/eei.v12i3.5170.
- [15] A. R. Chowdhury, T. Chatterjee, and S. Banerjee, "A random forest classifier-based approach in the detection of abnormalities in the retina," *Medical and Biological Engineering and Computing*, vol. 57, no. 1, pp. 193–203, Jan. 2019, doi: 10.1007/s11517-018-1878-0.




- [16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.
- [17] B. X. W. Liew, F. M. Kovacs, D. Rügamer, and A. Royuela, "Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain," *European Spine Journal*, vol. 31, no. 8, pp. 2082–2091, Aug. 2022, doi: 10.1007/s00586-022-07188-w.
- [18] V. H. A. Ribeiro, G. Reynoso-Meza, and H. V. Siqueira, "Multi-objective ensembles of echo state networks and extreme learning machines for streamflow series forecasting," *Engineering Applications of Artificial Intelligence*, vol. 95, p. 103910, Oct. 2020, doi: 10.1016/j.engappai.2020.103910.
- [19] D. P. Javale and S. S. Desai, "Machine learning ensemble approach for healthcare data analytics," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 28, no. 2, pp. 926–933, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp926-933.
- [20] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/bf00058655.
- [21] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, Jun. 1990, doi: 10.1007/bf00116037.
- [22] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and other ensemble methods," *Neural Computation*, vol. 6, no. 6, pp. 1289–1301, Nov. 1994, doi: 10.1162/neco.1994.6.6.1289.
- [23] A. Arabiat and M. Altayeb, "An automated system for classifying types of cerebral hemorrhage based on image processing techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 2, pp. 1594–1603, Apr. 2024, doi: 10.11591/ijece.v14i2.pp1594-1603.
- [24] T. Badriyah, N. Sakinah, I. Syarif, and D. R. Syarif, "Machine learning algorithm for stroke disease classification," in *2nd International Conference on Electrical, Communication and Computer Engineering, ICECCE 2020*, Jun. 2020, pp. 1–5, doi: 10.1109/ICECCE49384.2020.9179307.
- [25] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. M. Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–12, Nov. 2021, doi: 10.1155/2021/7633381.
- [26] M. Alruily, S. A. El-Ghany, A. M. Mostafa, M. Ezz, and A. A. A. El-Aziz, "A-tuning ensemble machine learning technique for cerebral stroke prediction," *Applied Sciences (Switzerland)*, vol. 13, no. 8, p. 5047, Apr. 2023, doi: 10.3390/app13085047.
- [27] H. Alhakami, S. Alraddadi, S. Alseady, A. Baz, and T. Alsubait, "A hybrid efficient data analytics framework for stroke prediction," *International Journal of Computer Science and Network Security*, vol. 20, no. 4, pp. 240–250, 2020.

BIOGRAPHIES OF AUTHORS






Outmane Labaybi    born on October 2nd, 1986, he received his master's degree in Big Data Analytics and Smart System (BDSaS) from the Sidi Mohamed Ben Abdellah University in Fez, Morocco in 2022, he is currently pursuing a Ph.D. in Data Science at Sidi Mohamed Ben Abdellah University in Fez, Morocco. Passionate about data-driven research, Outmane's academic journey has focused on leveraging advanced data analytics and machine learning techniques to solve complex problems. With a strong commitment to academic excellence and a keen interest in the intersection of data science and healthcare, Outmane aims to contribute significantly to the field of predictive analytics and improve healthcare outcomes through innovative data-driven approaches. He can be contacted at email: outmane.labaybi@usmba.ac.ma






Mohamed Bennani Taj    received his Master's degree in Computer Science and Networking from Science Faculty of Tangier, Tangier in 2011. He received his Ph.D. in 2019 (Computing Science and Networking) from Science Faculty of Tangier. At present, he is working as Prof. in Faculty of Science of Dhar el Mahraz Fez since 2019. He is qualified university professor since June 2023. He can be contacted at email: taj.bennani@usmba.ac.ma.






Khalid El Fahssi    received the Ph.D. degrees in mathematics and computer science from the Faculty of Sciences Ain Chock, Hassan II University, Casablanca, Morocco, in 2018. Since 2019, he has been with the Faculty of Sciences Dhar el Mehraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco, as an associate professor. He is currently with the Department of Computer Science, Faculty of Sciences Dhar el Mehraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco. His research interests are related to based image retrieval (CBIR) and artificial intelligence for medical applications. He can be contacted at email khalid.elfahssi@usmba.ac.ma.






Said El Garouani    He is a qualified professor of Faculty of science Dhar el Mahraz Fez. He works on SIG data analysis multidimensional deep learning, image analysis and Arabic Chatbots Challenges and Solutions. He can be contacted at email said.elgarouani@usmba.ac.ma.



Mohamed Lamrini    received the Ph.D. degree from the Claude Bernard –Lyon University in 1993. He is currently a professor of computer science at USMBA-Fez University. He is also a member of the LPAIS Laboratory. His research interests include software quality (CMMI, Six sigma, ISO 9001...), Industrial Engineering (Methods and statistical tools). He can be contacted at email: mohamed.lamrini@usmba.ac.ma.



Mohamed El Far    is a professor of the Sidi Mohamed Ben Abdellah University in Fez, Morocco. He teaches computer science. His research focus on data science, big data and artificial intelligence. He had published many researchers. He can be contacted at email: mohamed.elfar@usmba.ac.ma.