SciNexuses

Journal Homepage: sciencesforce.com/scin

SciNexuses Vol. 1 (2024) 16–27

Paper Type: Original Article

Revisiting Machine Learning for Predictive Modeling for Stroke from Electronic Health Records

Mohamed Refaat Abdellah [1,*], Rahma Elsayed Owaidah [2], Sarah Mohammed Selem [2] and Ayman H. Abdel-aziem [2]

[1] Computer Science, Misr University for Science and Technology, Egypt; mohamed.refaat@must.edu.eg.

[2] Faculty of Information Systems and Computer Science, October 6th University, Giza, 12585, Egypt;

Emails: r.owaidah19@gmail.com; sarahmselim9@gmail.com; ayman.hasanein.comp@o6u.edu.eg.

Received: 15 Nov 2023 **Revised:** 19 Jan 2024 **Accepted:** 18 Feb 2024 **Published:** 20 Feb 2024

Abstract

Stroke is a big worldwide threat with serious health and economic implications. To solve this, researchers are developing automated stroke prediction algorithms, allowing for early intervention and perhaps saving lives. Machine learning (ML) techniques have emerged as powerful tools for predicting stroke risk based on patient data. This study explores the application of various ML algorithms, including logistic regression, decision trees, support vector machines, and k-nearest neighbors, for stroke prediction. The models are trained on historical data from patients with known stroke outcomes and evaluated on their ability to classify individuals into stroke-risk categories accurately. ML techniques are being increasingly adapted for use in the medical field because of their high accuracy. The main contribution of this study is a stacking method that achieves a high performance that is validated by various metrics, such as AUC, precision, recall, F-measure, and accuracy. This study also shows how different ML models can predict strokes by analyzing factors such as age, gender, accompanying habits, and personal lifestyle. It is quite essential to understand the risk factors that make a patient more susceptible to strokes, thus some factors make stroke prediction much easier. This research offers an analysis of the factors that enhance the stroke prediction process based on electronic health records.

Keywords: Predictive Analytics, Machine Learning, Decision Tree, Electronic Health Records, Stroke.

1 | Introduction

Stroke continues to be one of the world's most common causes of death and permanent disability, placing a heavy strain on patients, families, and healthcare systems [1]. To reduce the incidence and severity of stroke-related consequences, it is imperative to identify individuals who are at high risk of stroke early on and to apply preventative measures and interventions. Predictive modeling methods that make use of cutting-edge ML algorithms have become increasingly attractive as tools for assessing stroke risk in recent years. Stroke continues to be one of the world's most common causes of death and permanent disability, placing a heavy strain on patients, families, and healthcare systems [2]. To reduce the incidence and severity of stroke-related consequences, it is imperative to identify individuals who are at high risk of stroke early on and to apply preventative measures and interventions. Predictive modeling methods that make use of cutting-edge machine

Corresponding Author: mohamed.refaat@must.edu.eg

https://doi.org/10.61356/j.scin.2024.1304

learning (ML) algorithms have become increasingly attractive tools for assessing stroke risk in recent years. These models predict a person's risk of having a stroke within a specified time by using a variety of clinical, demographic, lifestyle, and imaging variables. Stroke prediction models are becoming more and more popular, but there are still several obstacles to overcome [3]. These include the requirement for strong validation across a variety of populations, the addition of new biomarkers and imaging modalities, and integration with clinical practice. Furthermore, it's critical to guarantee these models' interpretability, generalizability, and ethical concerns to promote patient and healthcare provider trust. Our goal in this work is to present a thorough analysis of predictive modeling techniques for estimating the risk of stroke. By examining the most recent state-of-the-art methods, emphasizing their benefits, drawbacks, and uses in stroke prediction. These methods include logistic regression, decision trees, random forests (RF), support vector machines, neural networks, and ensemble approaches. 2- go over how crucial it is to do data preprocessing, feature selection, model evaluation, and validation techniques to improve the clinical usefulness and dependability of stroke prediction models. The structure of the paper is as follows. The remaining part of Section 1 provides an overview of the related work and describes the dataset used in our study. Section 2 covers the data analysis and understanding. The stroke prediction results are explained in Section 3. Finally, Section 4 concludes the paper and discusses future work.

1.1 | Related Work

Existing works in the literature investigated various aspects of stroke prediction. M. M. Islam [4], this paper uses ML techniques to predict strokes, try to reduce their occurrence, and focus on reducing stroke-related deaths by identifying them in some way using ML models. The aim is to develop a system that can alert individuals in advance, about the risk of a stroke using algorithms and healthcare data. With a dataset containing 5110 entries and 12 parameters related to health and demographics, the focus is on managing missing values through Exploratory Data Analysis (EDA) during data preprocessing. The interrelation between variables is analyzed for understanding. In dealing with imbalanced datasets, techniques like the Synthetic Minority Oversampling Technique (SMOTE) and feature engineering are applied. A comparison, among ML models including RF Ada Boost Algorithm and Naïve Bayes, shows that RF yields the outcomes. One of its advantages the paper provides a thorough analysis of stroke, highlighting its causes and the importance of early detection, which provides a solid foundation for the research. The study carefully evaluates several ML models, and results are shown using measures including F1 score, precision, and recall. Insufficient information is provided on the data-gathering method and user interface of mobile apps.

Y. M. Chen et al. [5] This paper presents an ML model for predicting the existence of a stroke in a patient, using an RF classifier that outperforms state-of-the-art models such as Logistic Regression, Decision Tree Classifier (DTC), and K-NN. The researchers conducted experiments on datasets with 5110 observations and 12 attributes, applying EDA for preprocessing and feature techniques for balancing the datasets. A cloud-based mobile app was developed to collect user data, analyze, and provide the possibility of stroke for alerting the person with an accuracy of 96% precision, recall, and F1-score. The user-friendly system can be a lifesaver, as the person gets an essential warning very easily by providing very little information from anywhere with a mobile device. Stroke is one of the fatal brain diseases that cause death in 3 to 10 hours. The study aims to improve stroke prediction and the quality of life for stroke patients by identifying the nature of the stroke and reacting promptly through smart health systems.

S. Dev, and H. Wang, [6] in Figure 1 provided a detailed analysis of various benchmarking algorithms in stroke prediction in this section. We benchmark three popular classification approaches — neural network (NN), decision tree (DT), and RF for stroke prediction from patient attributes Figure 1. The decision tree model is one of the popular binary classification algorithms. This method involves building a tree-like decision process with several condition tests and then applying the tree to the medical record dataset. Each node in this tree represents a test, and the branches correspond to the outcome of the test. The leaf nodes finally represent the class labels. The pruning ability of such an algorithm makes it flexible and accurate, which is required in medical diagnosis. We also benchmark the dataset on an RF approach. The flexibility and ease of use of the RF algorithm coupled with its consistency in producing good results, even with minimal tuning of the hyperparameters makes this algorithm valuable in this application. The possibilities of over-fitting are limited by the number of trees existing in the forest. Moreover, RF can also provide adequate indicators of the way it assigns significance to each of these input variables. the dataset contains a total of 29072 medical records. Out of this, only 548 records belong to patients with stroke conditions, and the remaining 28524 records have no stroke condition. This is a highly unbalanced dataset.

Y. Chahine, [7] ML is used to predict stroke risks and understand the mechanisms of these strokes. Upon observation, there is a huge difference in accuracy when using ML algorithms for diagnosis compared with conventional statistical models, which leads to more reliable and personalized predictions when using ML. The methodology is to review how we can benefit from ML algorithms for predicting stroke risk. The strengths of the paper are that it reviews and discusses the ability of ML to improve stroke risk prediction, while the weaknesses are that there is no development; it relies on the previously published paper. The evaluation of the methodology and data sources used generally involves the validity and reliability of the studies looked at, which include considerations such as the sample size, study design, statistical methods, and generalizability of the conclusion. Key findings of this paper include the more effective forecasting of ML compared to conventional statistical models in predicting stroke risk, especially in AF patients. Also, it strengthens the need for more research to verify these findings and analyze the feasibility of implementing ML models in clinical practice. The models used in this paper involve various ML algorithms, such as deep neural networks, RF s, and support vector machines. The authors of this research paper include multiple individuals from various institutions, as illustrated by the citations provided throughout the document. M. Daidone,[8] Multiple superior strategies have been applied to enhance the diagnosis, treatment, and results of stroke sufferers, with one of the maximum crucial of those strategies being ML. Different fashions of ML have proven sizable accuracy in imaging analysis, diagnosing stroke types, and assessing risks. These fashions consist of quite a few algorithms, consisting of RFs, Support Vector Machines, Recurrent Neural Networks, and Deep Neural Networks. For instance, an RF version educated on facts from sufferers with number one intracerebral hemorrhage carried out first-rate results. It should be expected practical results after one month with an accuracy of 83.1%, followed by an Area Under the ROC Curve (AUC) of 0.899.

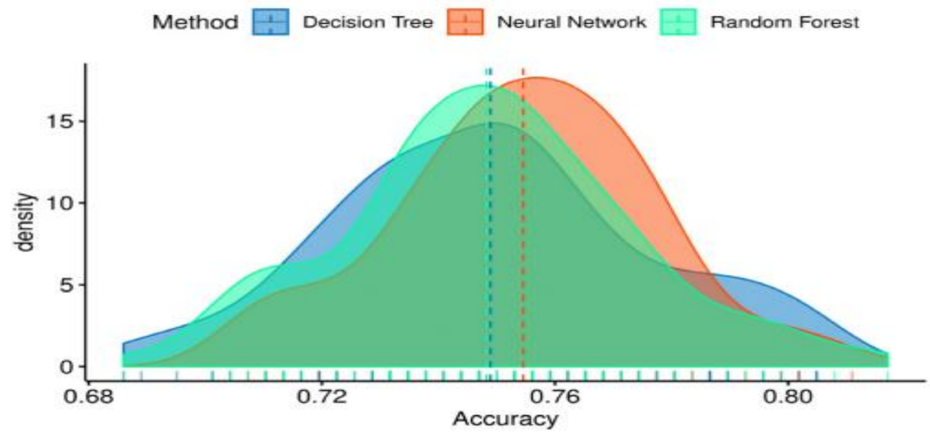

Figure 1. Histogram distribution of classification accuracies using top 4 features; age, heart disease, average glucose level, and hypertension for the benchmarking algorithms [6].

The accuracy improved to 83.9% after six months, with a sensitivity of 72.5% and specificity of 90.6%, alongside an AUC of 0.917. Additionally, the Deep Neural Network version confirmed advanced overall performance as compared to standard fashions for predicting stroke results. Its ROC AUC (0.888) was that of the ASTRAL version (0.839), indicating advanced predictive capability.

Table 1. Summary of previous work.

Article	Method	Result	Phase	Domain	Strengths	Weakness
M. M. Islam, [3] Machine Learning and the Conundrum of Stroke Risk. (2021)	the methodology here is to use machine learning algorithms for predicting stock by using a healthcare database, Exploratory Data Analysis (EDA), mobile app interface, and Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalanced dataset, using a machine learning model such that Random Forest, Ada Boost	The effectiveness of the Random Forest model for stroke prediction is highlighted, and recommendations are made for improving accuracy through the use of learning approaches. The work offers perspectives on applying machine learning for stroke prediction, highlighting the significance of detection and preventive measures.	: Problem Identification, Literature Review, Data Collection and Preprocessing, Experiment Design, Model Training, Results Analysis, Discussion and Conclusion	The domain of the paper is healthcare, using machine learning to predict stroke.	The paper provides a thorough analysis of stroke, highlighting its causes and the importance of early detection, which provides a solid foundation for the research. The study carefully evaluates several machine learning models.	There is insufficient information provided on the data-gathering method and user interface of mobile apps
Y. M. Chen *et al.*,[4] Received: October 4, 2021 **Revised: December 8, 2021** **Accepted: December 15, 2021**	Applied EDA for preprocessing and feature techniques for balancing the datasets. Random forest classifier outperforms the state-of-the-art models, including Logistic Regression, Decision Tree Classifier (DTC), and K-NN. We conduct the experiments on datasets which has 5110 observations with 12 attributes.	The Random Forest models showed the best accuracy as precision at 96%, recall at 96%, and F1-score at 96%, outperforming the state-of-art models including logistic regression, decision tree classifier, and K-NN. The utilized dataset is imbalanced; therefore, SMOTE feature engineering is used to process the data.	applied EDA for preprocessing and feature techniques for the dataset and try to make data balance then build models and apply dataset on its Random Forest model outperforms the best accuracy.	Medical purpose by using technology.	many techniques can be employed, including Near Miss, SMOTE, Tomak Links, etc. This paper utilized the synthetic minority over-sampling technique (SMOTE) after preprocessing the datasets in the EDA step. and using some of the best models for classification (random forest classifier.	too small dataset and imbalanced data and not improving the efficiency of the models and fail to analyze the dataset using deep learning methods and try to enhance the accuracy.
S. Dev, H. Wang,[5]A predictive analytics approach for stroke prediction using machine learning and neural networks (2022)	Provided detailed analysis of three different algorithms: neural network (NN), decision tree (DT), and random forest (RF) They applied it to a dataset of 1096 after balancing the data	The neural network model was the highest in accuracy among the three with an accuracy of 0.77, and then the decision tree and random forest with an accuracy of 0.74	define problem, data preprocessing and analysis, modeling, results	Machine Learning Models in Healthcare Applications	The paper used great machine-learning models for stroke prediction and it did a detailed data analysis before the modelling process	The dataset is too small and it becomes even smaller after balancing the data, so the models didn't have enough data for the training process

Y. Chahine, M. J. Magoon,[2] This paper is an extensive review of existing studies on stroke risk prediction using machine learning algorithms which were published in 2022.	The methodology here is to review how we can benefit from ML algorithms for stroke risk prediction. This reviews the study of designs, sample sizes, statistical methods, and generalizability of the results.	The paper points out the more effective forecasting of machine learning algorithms compared to traditional models in predicting stroke risk.	The phase of the research can be considered as a review of existing literature in the field of stroke risk prediction using machine learning.	The domain of the research is healthcare, specifically stroke risk prediction.	extensive review of relevant literature, Discussion of the potential of machine learning in improving stroke risk prediction, ensure the need for further research to validate findings and figure out feasibility in clinical practice.	The authors of this research paper include multiple individuals from various institutions.
M. Daidone, S. Ferrantelli,[1] Predictive modeling for stroke chances the use of distinctive device mastering models. (2024)	Various machine learning getting to know strategies and fashions are noted inside the paper, including Random Forest (RF) Linear aid vector system (SVM) regression Deep neural network (DNN) Logistic regression Convolutional neural network (CNN) AdaBoost (ADA-B) Gradient boosting (GB)	Random Forest version educated on facts from sufferers with number one intracerebral hemorrhage. practical results after one month with an accuracy of 83.1%, followed by an Area Under the ROC Curve (AUC) of 0.899. The accuracy improved to 83.9% after six months, with a sensitivity of 72.5% and specificity of 90.6%, alongside an AUC of 0.917	diagnosis, assessment of prognosis, treatment planning, and outcome prediction	Machine learning techniques in various aspects of stroke care	These strategies can investigate big quantities of scientific statistics and pick out styles that might not be obvious via conventional methods. Machine mastering fashions can help clinicians make greater knowledgeable choices and enhance affected personal outcomes	Interpretability of device studying fashions may be challenging, making it hard for clinicians to recognize the underlying motives at the back of predictions. Generalization of device studying fashions to various affected person populations and healthcare settings can be limited, requiring validation and version in distinctive contexts.

2 | Material and Methods

2.1 | Dataset

The dataset is available from Kaggle,3 a public data repository for datasets [9-11]. The dataset contains the EHR records of 29072 patients. As in the Table 2. It has a total of 11 input attributes and 1 output feature. The output response is a binary state, indicating if the patient has suffered a stroke or not. The remaining 11 input features in EHR are patient identifier, gender (G), age (A), binary status if the patient is suffering from hypertension (HT) or not, binary status if the patient is suffering from heart disease (HD) or not, marital status (M), occupation type (W), residence (urban/rural) type (RT), average glucose level (AG), body mass index (BMI), and patient's smoking status (SS). The dataset is highly unbalanced concerning the occurrence of stroke events; most of the records in the EHR dataset belong to cases that have not suffered from stroke. The publisher of the dataset has ensured that the ethical requirements related to this data are met to the highest standards. In the subsequent discussion of this paper, we will exclude the patient identifier as one of the input features. We will consider the remaining 10 input features, and 1 response variable, in our study and analysis as in Table 2.

2.1.1 | Analyzing the Dataset

We conduct a correlation analysis of the features in this part. To conduct such analysis on the input attributes of the EHR records, we use the complete dataset. The following are some ways that correlation analysis is helpful for feature selection in Figure 4: When predicting the likelihood of a stroke, one variable with a high correlation can be disregarded since it adds no new information to the prediction model. Additionally, we assess each component's behavior both singly and collectively to determine the significance of each trait in predicting the likelihood of a stroke. To anticipate strokes, a thorough examination of the input feature space is essential. To lower the computational cost of modeling, it is crucial to identify the ideal and smallest set of predictive features. Figure 2 shows a target sample distribution from the original dataset.

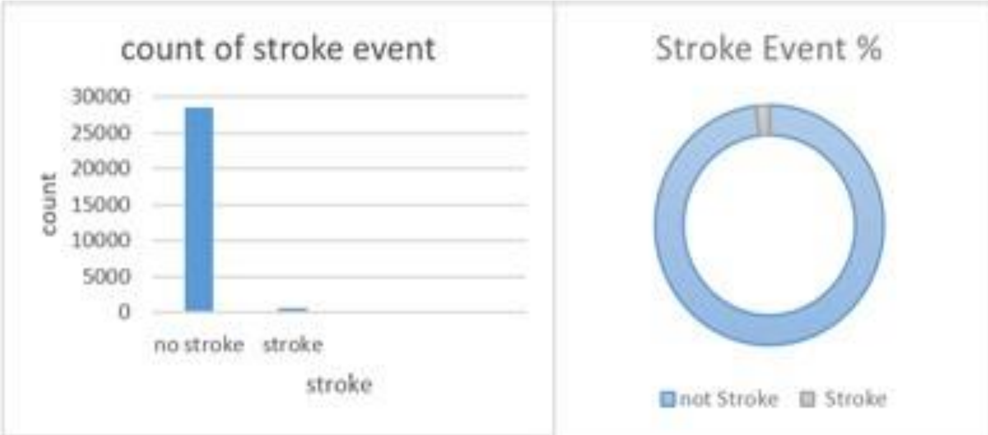

Figure 2. Target sample distribution from the original dataset.

Table 2. Dataset description.

Features	Description	Variable Type
Gender	Male, Female, Other	Categorial features
Ever Married	Yes or no	Categorial features
Work Type	Children, Govt_jov, never worked, Private, or Self-employed	Categorial features
Residence Type	Rural or Urban	Categorial features
Smoking Status	Formerly Smoked, Never Smoked, Smokes, Unknown	Categorial features
Hypertension	Yes= 1, No= 0	Binary Numerical features
Heart Disease	Yes= 1, No= 0	Binary Numerical features
Stroke	Stroke= 1, Healthy= 0	Binary Numerical features
Age	Age of the patient	Continuous numerical features
Average Glucose	The average glucose	Continuous numerical features
Level	Level in the blood	Continuous numerical features
BMI	Body mass index	Continuous numerical features

2.2 |The Correlation Between Features

The visualization helps to identify patterns and relationships between features in the dataset. Positive correlations (values closer to 1) indicate that the features tend to increase or decrease together, while negative correlations (values closer to -1) indicate an inverse relationship. Features with low correlation coefficients (values closer to 0) have weaker relationships as in Figure 3. Using the support vector machine in Figure 4 to show the feature importance. Figures 5 and 6 show the gender and smoking status features distribution.

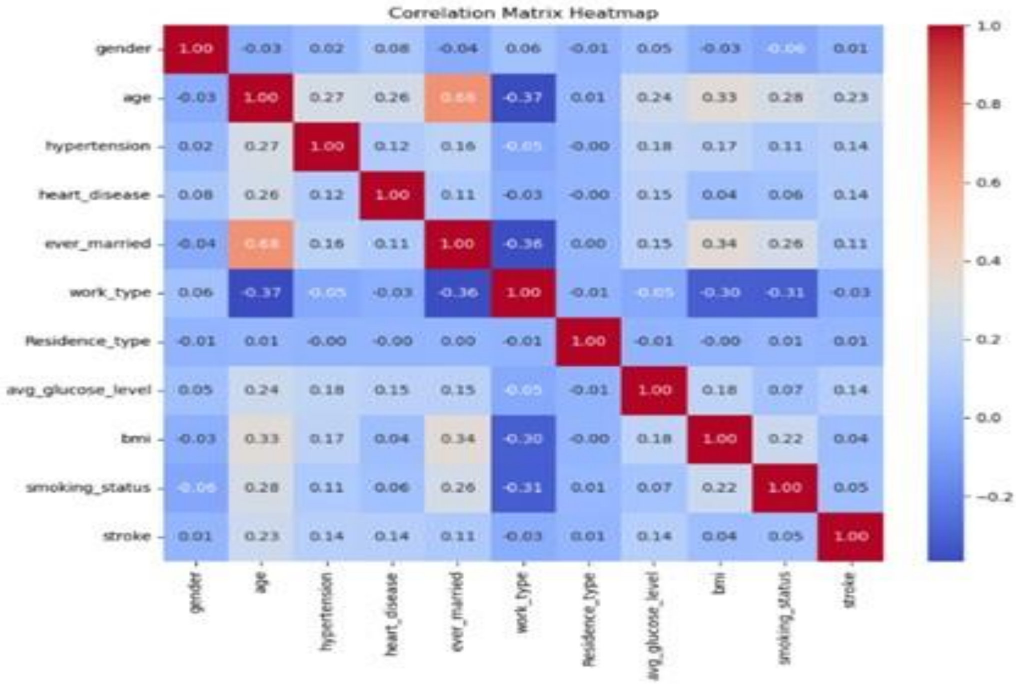

Figure 3. Heatmap that shows the correlation between features.

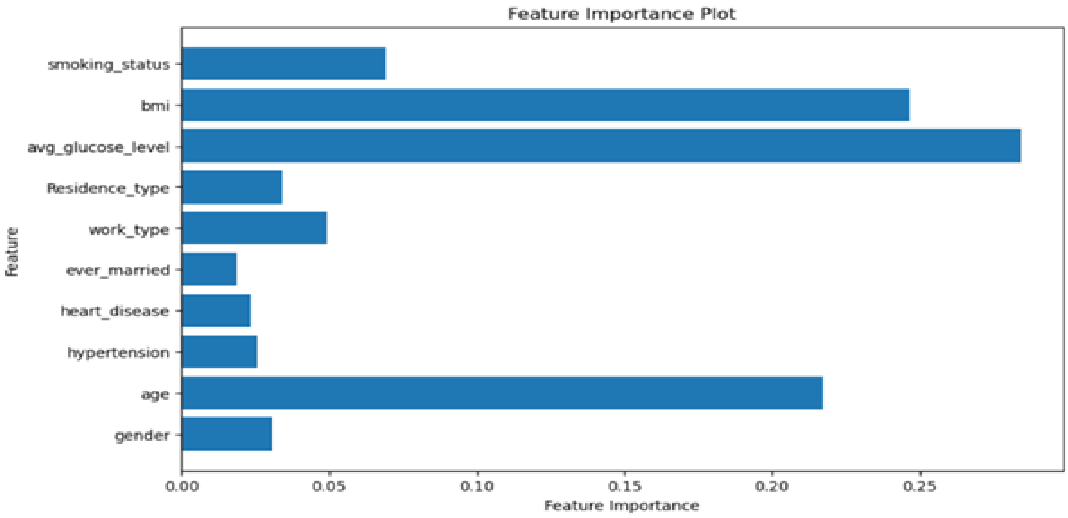

Figure 4. Importance of each feature using a support vector machine model.

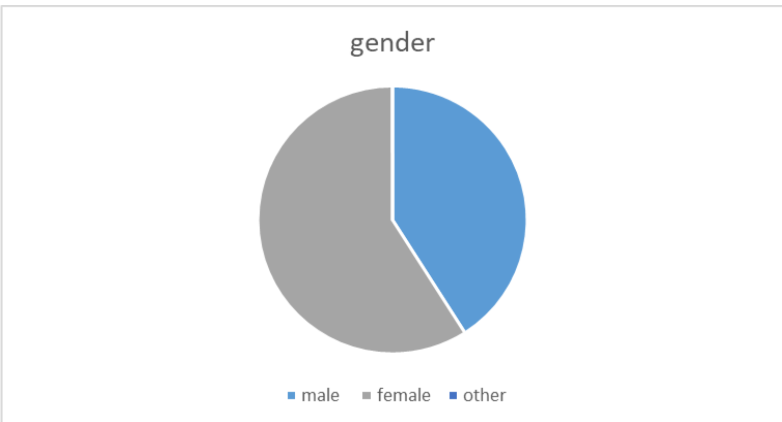

Figure 5. Stroke patient's gender

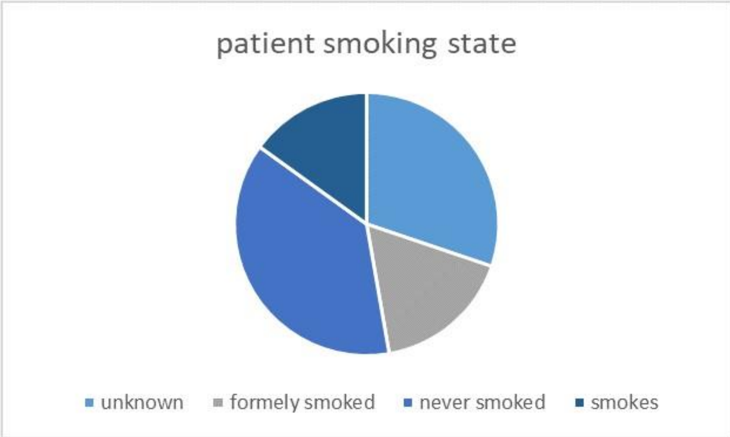

Figure 6. Stroke patient's smoking status.

3 | Results

3.1 | The Stroke Prediction

In this part, we offer a thorough examination of several stroke prediction methods. We compare four widely used categorization techniques: k-nearest neighbors (KNN), decision trees (DT), support vector machines (SVM), and logistic regression (LR). When there are only two possible outcomes for the outcome variable in a binary classification job, the statistical technique known as logistic regression is employed. Regression analysis of this kind analyses the likelihood that a class or event will occur based on one or more predictor variables. The logistic function, sometimes referred to as the sigmoid function, is used in logistic regression to represent the relationship between the predictor variables and the probability of the result. Using this function, a probability score between 0 and 1 is generated from the linear combination of the predictor variables. The logistic function has the following mathematical definition:

$$(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

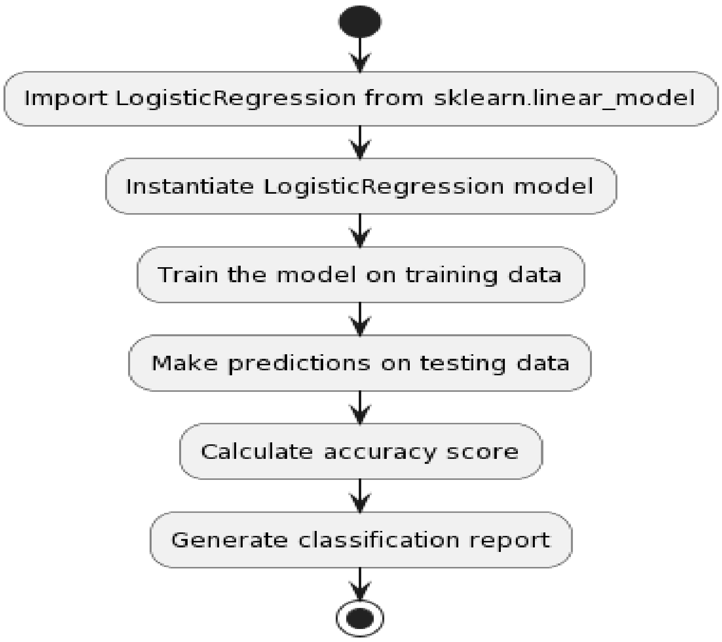

Figure 7. Logistic regression flowchart.

3.1.1 | Logistic Regression

Logistic regression in Figure 7 is used to estimate the coefficients that, given the predictor variables, maximize the chances of detecting the actual outcome during training. Usually, optimization procedures like gradient descent are used to do this. Based on the predictor variables, the logistic regression model can be trained to estimate the likelihood of the outcome for fresh data. The code performs logistic regression classification, trains the model, makes predictions on the test data, and evaluates the model's performance using an accuracy score and classification report.

3.1.2 | Decision Tree

In supervised learning, the decision tree model in Figure 8 is a potent technique that is utilized for both classification and regression tasks. With each internal node representing a feature, each branch representing a choice based on that feature, and each leaf node representing the result or forecast, it builds a structure resembling a tree. Utilizing a decision tree model involves the following essential steps:

- Selecting the optimal feature: divide the data at each node is known as feature selection. This process is usually carried out using criteria such as variance reduction for regression jobs or Gini impurity or information gain for classification tasks.
- Data splitting: involves dividing the data into subsets according to the chosen feature and then repeating this procedure recursively until a set of predetermined conditions are satisfied.
- Building the Tree: Building the decision tree by dividing the data into child nodes, picking features repeatedly, and stopping when the stopping criteria are met.
- Making Predictions: By following the judgments made by the root node down to the leaf nodes, one may use the trained decision tree to generate predictions on fresh data.

For hyperparameter tuning. It utilizes Grid Search CV for hyperparameter optimization using cross-validation with 2 folds.

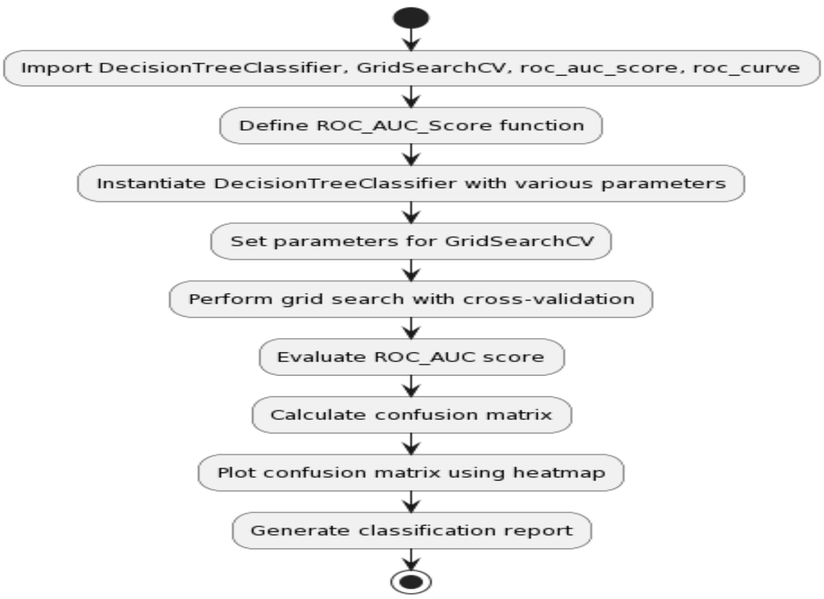

Figure 8. A decision Tree classifier with predefined hyperparameters and a parameter grid.

3.1.3 | Support Vector Machine

Support Vector Machines (SVMs) in Figure 9 are powerful algorithms used for classification and regression tasks. They aim to find a hyperplane that maximizes the margin between classes, allowing for effective separation even in high-dimensional spaces. SVMs can handle non-linear separation using kernel functions and are robust to overfitting due to regularization. While computationally expensive, SVMs are versatile and suitable for various tasks, especially when there is a clear separation between classes or non-linear decision boundaries are needed. This code trains a linear Support Vector Machine classifier on the training data, predicts labels for the test data, and evaluates its accuracy and classification performance using the accuracy score and classification report.

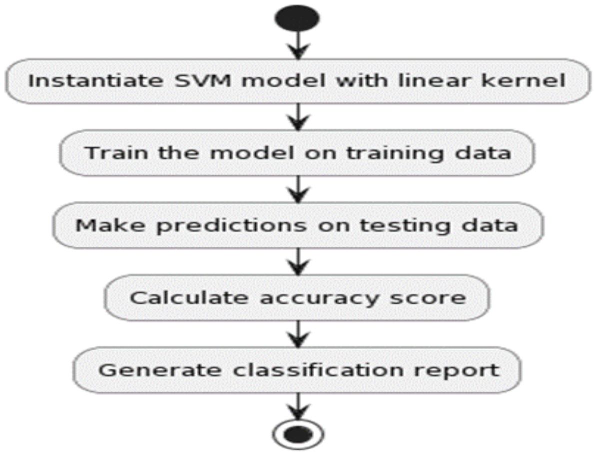

Figure 9. A linear Support Vector Machine classifier evaluation method.

3.1.4 | K-Nearest Neighbors

K-Nearest Neighbours (KNN) in Figure 10 is a simple yet effective supervised learning algorithm used for classification and regression tasks. It makes predictions based on the majority vote (for classification) or the average (for regression) of the k nearest data points in the feature space. KNN does not explicitly learn a model; instead, it stores the entire training dataset and computes distances to determine the closest neighbors during prediction. The choice of k (the number of neighbors) and the distance metric (e.g., Euclidean distance) are key parameters influencing KNN's performance and computational efficiency. KNN is intuitive, but it can be sensitive to noise and irrelevant features in high-dimensional spaces. Code snippet to train a K-Nearest Neighbors classifier with Euclidean distance metric predicts labels for test data and evaluates its accuracy and classification performance using an accuracy score and classification report.

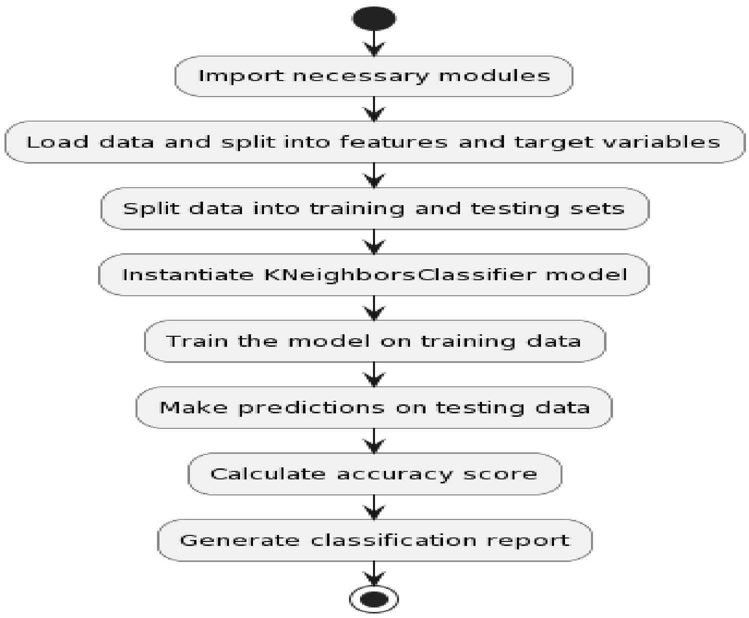

Figure 10. K-Nearest Neighbors classifier Evaluation method.

3.2 | Comparison of Four Classifiers

In Figure 11 the accuracies of the four models are DT 96%, LR 95%, KNN 95%, and SVM 94%. the DT model has the highest accuracy. Table 3 shows the performance evaluation for four metrics for four classifiers.

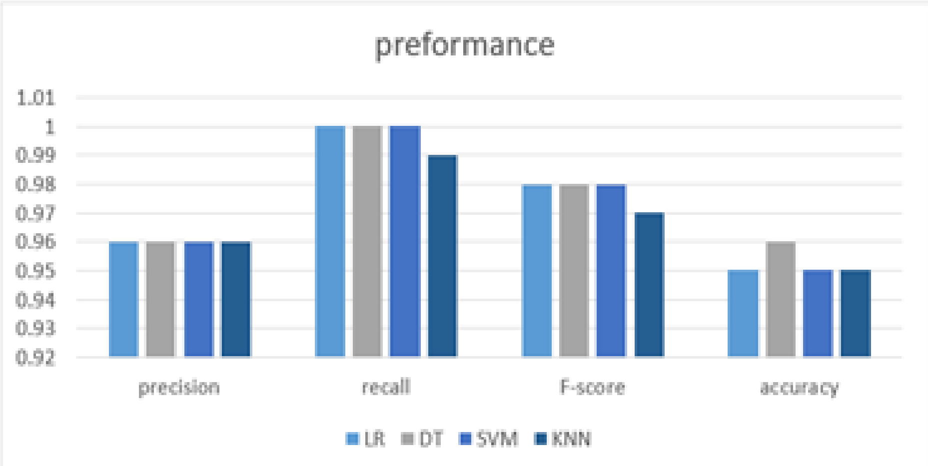

Figure 11. The performance of each model shows that the DT model has the highest accuracy among the four tested models.

Table 3. Performance evaluation of the four used models.

Classifier	Precision	Recall	F-score	Accuracy
LR	0.96	1	0.98	0.95
DT	0.96	1	0.98	0.96
SVM	0.96	1	0.98	0.95
RF	0.96	0.99	0.98	0.94

4 | Conclusion and Future Work

This study presents the implementation of four Machine learning algorithms on various feature and primary component setups. We discovered that a feature combination of A, HD, HT, and AG yields the best results for decision trees. 96% of the data is accurate. We have observed encouraging outcomes with just four elements. There are two main reasons why the accuracy of the perceptron model cannot be further enhanced: the absence of a discriminatory feature set and a dataset. We found that the majority of the dataset's features are highly associated with one and another do not, therefore, contribute any new information to the original feature space. More data will also allow us to train our model more effectively. We intend to gather institutional data for our upcoming research because, after balancing the data, we were only able to extract 1096 cases out of 12072.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributions

"Conceptualization, M.R. and A.H.; Methodology, M.R.; Software, A.H.; Validation, M.R., A.H., and S.S.; formal analysis, R.O.; investigation, S.S.; resources, R.O.; data maintenance, A.H.; writing-creating the initial design, A.H.; writing-reviewing and editing, M.R.; visualization, S.S.; monitoring, R.O.; project management, A.H.; funding procurement, M.R.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] M. Daidone, S. Ferrantelli, A. Tuttolomondo, M. Daidone, and M. Daidone, "Machine learning applications in stroke medicine: Advancements, challenges, and future prospective," Neural Regeneration Research, vol. 19, no. 4. Wolters Kluwer Medknow Publications, pp. 769–773, Apr. 01, 2024. doi: 10.4103/1673-5374.382228.

[2] Y. Chahine, M. J. Magoon, B. Maidu, J. C. del Álamo, P. M. Boyle, and N. Akoum, "Machine Learning and the Conundrum of Stroke Risk Prediction," Arrhythmia Electrophysiol. Rev., vol. 12, 2023, doi: 10.15420/aer.2022.34.

[3] Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H., Patan, R., Jayaraman, P. and Manikandan, R., 2020.} Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3), pp.817-828.

[4] M. M. Islam, S. Akter, M. Rokunojjaman, J. H. Rony, A. Amin, and S. Kar, "Stroke Prediction Analysis using Machine Learning Classifiers and Feature Technique," Int. J. Electron. Commun. Syst., vol. 1, no. 2, pp. 57–62, 2021, doi: 10.24042/ijecs.v1i2.10393.

[5] Y. M. Chen et al., "Predicting new-onset post-stroke depression from real-world data using machine learning algorithm," Front. Psychiatry, vol. 14, no. June, pp. 1–7, 2023, doi: 10.3389/fpsyt.2023.1195586.

[6] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," Healthc. Anal., vol. 2, no. February, p. 100032, 2022, doi: 10.1016/j.health.2022.100032.

[7] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin, and M. F. Mridha, "Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention," IEEE Access, vol. 11, no. June, pp. 52288–52308, 2023, doi: 10.1109/ACCESS.2023.3278273.

[8] Emon, M.U., Keya, M.S., Meghla, T.I., Rahman, M.M., Al Mamun, M.S. and Kaiser, M.S.}, 2020, November. 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.

[9] Dritsas, E. and Trigka, M., 2022.} Stroke risk prediction with machine learning techniques. Sensors,22(13),p.4670.

[10] Sailasya, G. and Kumari, G.L.A., 2021.} Analysing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications,12(6).

[11] Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I.J., Rudd, A.G., Wang, Y., Douiri, A., Wolfe, C.D. and Bray, B., 2020.} A systematic review of machine learning models for predicting outcomes of stroke with structured data. PloS one, 15(6),p.e0234722.