RESEARCH ARTICLE

DOI: 10.34256/ijcci2514

International Journal of Computer Communication and Informatics

A Comparative Analysis of Machine Learning Models for Stroke Prediction

Kripa Mary Jose [a,*], Nizar Banu [a], A. Melvin Infant [a]

[a] Department of Computer Science and Engineering, Christ Deemed to be University Bangalore, Karnataka, India

* Corresponding Author: josekripamary99@gmail.com

Received: 29-01-2025, Revised: 06-04-2025, Accepted: 17-04-2025, Published: 22-04-2025

Abstract: Stroke is a leading global health burden, and there is an urgent need for improvement in risk prediction and treatment. This paper examines the capability of several machine learning algorithms, including Decision Trees, Random Forests, Neural Networks, Support Vector Machines (SVMs), Elastic Nets, and Lasso, to predict stroke risk on four cardiovascular and stroke datasets. The results indicate that Decision Trees and Random Forests are always better than Neural Networks, although Neural Networks show promising accuracy. SVMs are consistent, while the Elastic Net and Lasso models give average results.

Keywords: Stroke, Machine Learning, Risk Prediction, Decision Trees, Random Forests, Neural Networks, Support Vector Machines, Elastic Net, Lasso, Neuroplasticity, Rehabilitation

1. Introduction

Severe medical condition-stroke-an instantaneous blockage within the system-the results of which, if nothing happens to reverse that blockage, can lead to serious devastations. According to definitions there are two distinct types: TIA- ischemic (and hemorrhagic). TIA-ischemic takes place when one's cerebral flow is diminished, primarily via a blood clot, lowering the oxygen reaching the cerebral cells. Hemorrhagic strokes, on the other hand, entail bleeding in the brain caused by the rupture of weaker blood vessels [1]. Both types of strokes require rapid medical intervention to prevent brain damage and enhance results. Mainly, symptoms include weak- ness or paralysis, numbness or tingling, trouble with speak- ing or speech understanding, impairment of vision, severe headaches, dizziness, and coordination lack [1]. Early detection is critical as it may substantially minimize the severity of the stroke, preventing further damage or even death. Many risk factors are involved in the development of stroke especially advancing ages such as high blood pressure, heart disease, diabetes, smoking, obesity, high cholesterol levels, physical inactivity, and excessive alcohol consumption. Prevention of stroke requires adequate management of these risk factors through lifestyle modification and pharmacological treatment.

ML refers to artificial intelligence technology in support of systems learning patterns or rules from given data. Modern progress made in machine learning enables us to make stroke prediction models far more accurate than previously seen. Pre- vious studies have reported that the SVM algorithm achieved a high score of 96. 74% precision. This paper uses machine-learning algorithms for their performance in stroke prediction through cross-validation of various datasets. Here, Decision Trees, Random Forests, Neural Networks, Support Vector Machines, Elastic Nets, and Lasso are carefully tested to understand their predictiveness along with the appropriateness for stroke risk analysis [1, 2].

2. Literature Review

In their study, Nojood Alageel, Rahaf Alharbi, and col- leagues investigate the advantages of machine learning in stroke prediction using multiple datasets. They used Kaggle and local hospital datasets for stroke prediction and machine learning models like Stacking, Decision Tree, and Random Forest. Notably, they discovered that the NB classifier had the lowest accuracy level (86%), but other algorithms achieved comparable accuracies, f1 scores, precision, and recall [3].

Elias Dritsas, Maria Trigk, et al. drew emphasis on the substantial impact of stroke, which affects millions of people each year and causes death and disability. Their research finds factors that increase stroke risk, such as age, hypertension, and smoking. Interestingly, their research shows that the stacking approach surpasses the RF, 3-NN, and DT models, with all models significantly outperforming recall, F measure, and precision [4].

Rishna Mridha et al. investigate how Explainable AI (XAI) techniques improve stroke prediction by offering human- readable explanations. The proposed machine learning tech- nique obtained up to 91% prediction accuracy. Their study, which employs SHAP and LIME explainable approaches, provides vital insights into model decision making. They conclude that complicated models beat simpler models in stroke prediction accuracy, with their top model scoring nearly 91% [5].

S. Sahriar , S. Akther et al. said that the disease stroke is currently on the increase in incidence globally. From statistics, some contributions are coming to morbidity and mortality all over the globe. ML provided promising tool for the early prediction or risk assessment of stroke, still facing some inconsistencies in available data [6].

T. Priyadarshini, A. Hameed et al. studied the application of machine learning algorithms for stroke risk forecasting. They showed that ensemble-based methods such as RF and Naive Bayes may be used to model stroke risks based on patient data and that ML methods outperform traditional statistical methods by significant margins in terms of both accuracy and prediction time[7]. A. Gupta , N. Mishra et al assessed multiple machine learning algorithms, including Stacking, Random Forest, and PCA-based methods. Their findings demonstrated that

these algorithms exhibited high pre- diction accuracy, with ensemble methods outperforming single classifiers by improving sensitivity and specificity metrics [8].

G. Barmparis , M. Marketou et al. used a Stacking ensemble method and achieved an AUC score of 98.9 in stroke risk prediction models. Their findings indicate that ensemble learning with multiple classifiers improves prediction performance by combining diverse feature representations [9]. N. Biswas et al. demonstrated that the Random Forest models were better than other machine learning classifiers, with up to 92.55% accuracy and an AUC of 98.15. The results indicate that adaptability of the RF classifier to medical datasets makes it one of the most effective tools for stroke prognosis [10]. K. Akash et al. have focused on the most critical issues when using ML for stroke prediction, such as robustness problems and data costs. In this regard, imaging data is very expensive and causes overfitting that negatively impacts generalization across various populations [11].

The researchers Y. Chahine et al. investigated behavioral risk factors for stroke by demonstrating how Naive Bayes and K-means clustering of ML could be used in risk assessments, and they highlighted the role of sensitivity and specificity as metrics to assess predictive accuracy [12]. M. Chun, R. Clarke et al discussed using the electronic health record (EHRs) as critical input in a ML predictive model of a stroke and concluded that evaluating history and demography from ML results into early detection and optimum distribution of health services towards risker patients [13]. J. Amann et al. showed that neural networks performed best for stroke prediction, with a high accuracy rate of 95.16%, as compared to other ML algorithms like Logistic Regression and KNN. Their study indicates the strength of neural networks in detecting nonlinear patterns in complex medical datasets [14]. A. Jamthikar, D. Gupta et al. reported that Random Forest models, when combined with advanced hyperparameter tuning techniques, achieved accuracy levels of 99.87%. Their work highlights the importance of optimizing hyperparameters for ML models' performance in real-world medical prediction tasks [15]. G. Sailasya et al., however, have given a rather wholesome idea of how the techniques applied by machine learning, namely, Random Forest, KNN, and Stacking Classifiers, have enhanced the accuracy of stroke prediction models. Here it was confirmed that the techniques such as SMOTE, dealing with class imbalances, has been the major part which improved the rates of predictions, and the models gained up to 96.7% accuracy[16].

3. Dataset and Preprocessing

3.1 About Datasets

The first is the Brain dataset, contains 11 columns and 4981 entries. The columns include information such as gender, age, hypertension, heart disease, and smoking status, as well as the goal variable "stroke." This dataset is likely to contain information regarding people's health and if they have had a stroke. The attributes include demographic and health- related data

such as gender, age, hypertension, heart disease, marital status, type of work,residence type, average glucose level, body mass index (BMI), smoking status, and stroke occurrence.

The second, Diabetics dataset consists of 18 columns and 28119 entries. Similar to the first dataset, it contains information about health and medical conditions, such as age, gender, cholesterol levels, and whether the individual has had a heart disease attack. The attributes include age, sex, high cholesterol (presence or absence), cholesterol check history, BMI (body mass index), smoking status, history of heart disease attack, physical activity level, fruit consumption, vegetable consumption, heavy alcohol consumption (presence or absence), general health status, mental health status, phys- ical health status, difficulty in walking, and stroke occurrence (presence or absence).

The third, Heart dataset contains 12 columns and 918 entries. It includes information such as age, gender, chest pain kind, resting blood pressure, and whether the person has cardiac disease. This dataset appears to focus on car- diovascular health markers and could be useful for predictive modeling or study of heart diseases. It contains attributes such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar level, resting electrocardiographic re- sults, maximum heart rate achieved during exercise, exercise- induced angina (presence or absence), oldpeak (ST depression induced by exercise relative to rest), slope of the peak exer- cise ST segment, and heart disease occurrence (presence or absence).

The fourth is the Heart Failure dataset, has 13 columns and 299 items. It covers variables such as age, anemia, diabetes, ejection fraction, and smoking status, as do the other databases. This dataset may be smaller in size, but it still contains valuable health-related information that can be utilized for a variety of analytical reasons, such as investigat- ing the association between these characteristics and health outcomes.The attributes consist of age, anemia (presence or absence), creatinine phosphokinase level, diabetes (presence or absence), ejection fraction, high blood pressure (presence or absence), platelets count, serum creatinine level, serum sodium level, sex, smoking status, time (follow-up period), and death event occurrence (presence or absence). The datasets are taken from Kaggle repository and reasearch papers.

3.2 Preprocessing

3.2.1 Brain Dataset

There were no null values in the data. But, upon closer inspection, it was found that the ”smok- ing status” column had ”unknown” entries. These were re- placed by mode since the variable was categorical. Label encoding was used on ”smoking status”, ”work type”, and ”ever married” columns and one-hot encoding was used on ”residence type” and ”gender”. Standard scaling was used for ”bmi”, ”avg gluce level”, and ”age” to normalize this data set as shown in Figure 1. An imbalance data was also observed, for which 4981 occurances of stroke =0 and only

13 instances of stroke =1" as shown in figure 2. The issues of imbalance were resolved by SMOTE and undersampling techniques, respectively.

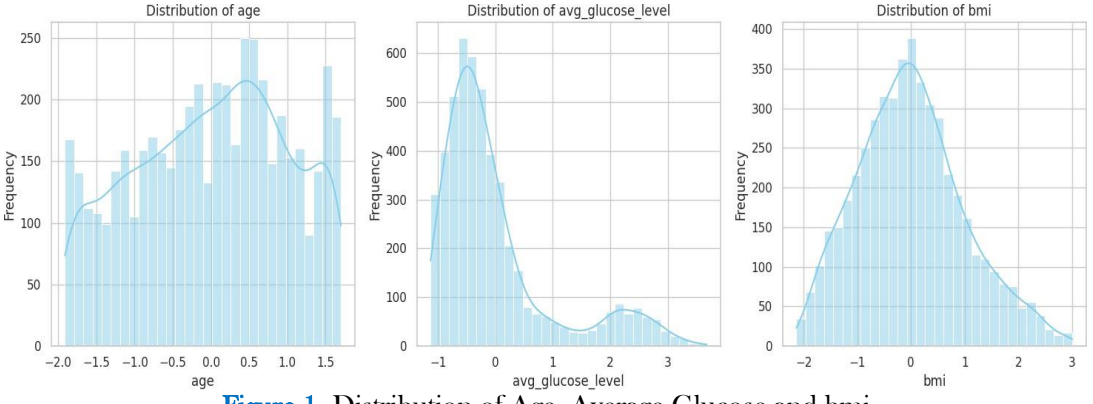

Figure 1. Distribution of Age, Average Glucose and bmi

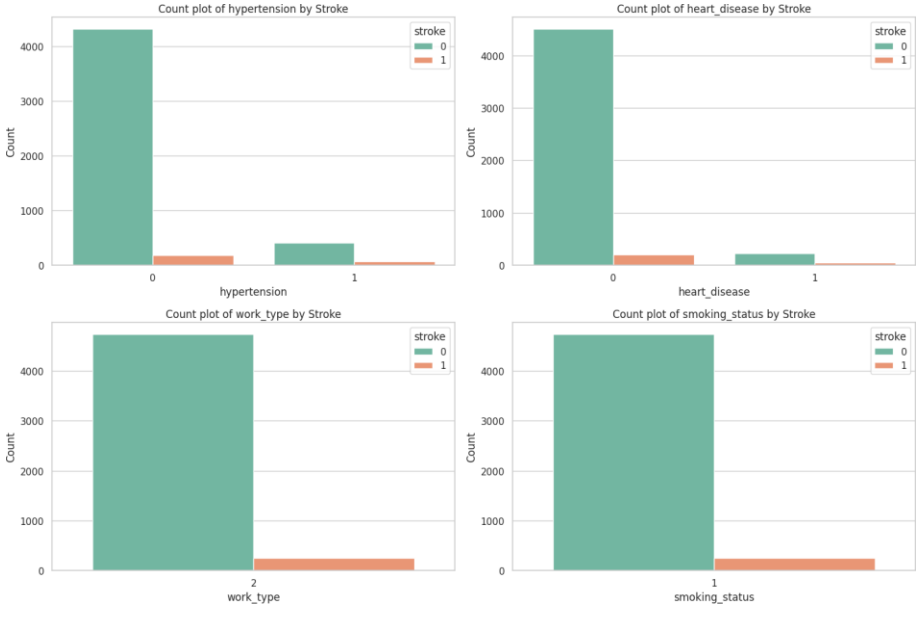

Figure 2. Count plot for hypertension, heart disease, work type and smoking status

3.2.2 Diabetics Dataset

No null values in the data. The corre- lation matrix analysis has revealed that "HighBP" and "Heart- DiseaseorAttack" are most negatively correlated with the target variable "stroke". However, the most positive correlation is for "Diabetes", "HighChol", and "HvyAlcoholConsump" with the target variable. "PhysHlth", "Veggies", "Fruits", "CholCheck", and "Sex" have weak correlations. Methods such as SMOTE and undersampling are applied

when dealing with the class imbalance between the 4395 cases of stroke = 0 and the 66297 cases of stroke = -1.

3.2.3 Heart Dataset:

There were no null values in the dataset. The categorical variables, which included "Sex", "ChestPainType", "RestingECG", "ExerciseAngina", and "ST_Slope" columns, are encoded using label encoding then standardized to the common scale. Perfect multicollinearity noted on the variable "ChestPainType," which was removed from the analysis to avoid bias in results.

3.2.4 Heart Failure Dataset:

There were no null values in the data. The age column was binned for further analysis, and then it was label encoded for the model's suitability. The dataset then passed through standardization for fea- tures. It was observed that there is no imbalance of data. Features that are taken by considering VIF and correlation matrix are "creatinine_phosphokinase", "anaemia", "diabetes", "high_blood_pressure", "serum_sodium", and "time" with "DEATH_EVENT" as the target variable.

4. Methodology

Machine learning transforms systems by allowing them to learn and improve autonomously from historical data, identifying patterns that guide future decision making without requiring human intervention.

4.1 Proposed Workflow

The machine learning algorithms starting from the Decision tree, Random Forest , Neural Network, Support Vector Machine, Lasso and ElasticNet have been applied to the datasets and the workflow is shown in Figure3 80% of the datasets are divided into training and 20% to testing.

4.2 Implementation

4.2.1 Decision Tree

A decision tree is a widely used supervised learning algorithm for classification and regression tasks in machine learning. A decision tree constructs a model to predict a target variable by learning easy-to-understand decision rules from data attributes. It is a flowchart-like structure where each internal node is a feature or attribute, each branch is a decision based on that feature, and each leaf node is the predicted class or outcome.

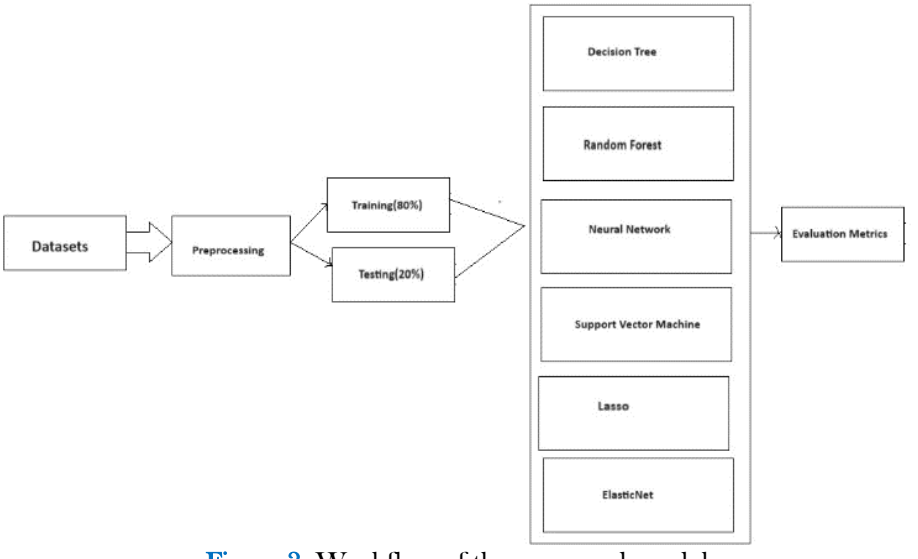

Figure 3. Workflow of the proposed model

The node at the top is termed as the root node and refers to the best predictor. Decision trees can support categorical as well as numerical data. Formula for classification trees:

$$Gini(D) = 1 - \sum_{i=1}^{c} (p_i)^2$$

Where:

- D is the dataset,
- c is the number of classes,
- p_i is the probability of class i. Entropy (for Classification Trees):

$$Entropy(D) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Where:

- D is the dataset,
- c is the number of classes,
- p_i is the probability of class i. Information Gain:

$$IG(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \times Entropy(D_v)$$

Where:

- D is the dataset,
- A is a feature,
- D_v is the subset of D for which feature A has value v,

- $|D|$ is the total number of examples in the dataset,
- $|D_i|$ is the number of examples in D_i.

The decision tree algorithm classifies data points step by step. The first is choosing the best feature to split, after which the algorithm checks for splits using metrics such as Gini impurity or entropy, then splits the data in recursive partitioning by taking a subset based on a feature selected until purity or a maximum depth is achieved. The assigned class labels of the majority class samples in the leaf nodes, then the algorithm makes predictions using the appropriate class label from the traverse tree by assessing feature values. It systematically approaches issues and solves them for good classification. This makes it very efficient with various data sets.

4.2.2 Random Forest

Random Forest is an ensemble learn-ing technique in which several decision trees are combined to increase the predictability accuracy and limit overfitting. Random Forest is an ensemble learning method which uses de- cision trees for increased prediction power as well as to prevent overfitting. There are several key steps to it: First, it makes use of bootstrapping, that is, sampling with replacement to create multiple subsets of the training data for each decision tree. Then, at every node of the decision trees, only a random subset of features is considered to split. Feature randomness decreases the correlation between trees and introduces diversity within the ensemble. Each decision tree is trained separately on the bootstrapped samples and feature subsets to the maximum depth or the minimum number of samples within a leaf node Once the decision trees are built, a Random Forest aggre- gates the results of these decision trees based on a voting mechanism if the task is classification and average if the task is regression. This ensures the final prediction will be sound and accurate, combining all the wisdom from any and all decision trees that make up the forest. Indeed, a combination of strengths from a number of trees and neutralization of respective individual weaknesses would make Random Forest seem a very effective and versatile tool to solve many machine learning applications.

4.2.3 Neural Networks

Neural networks are a machine-learning algorithm that is designed based on the structure and functioning of the human brain. A node or neuron, connected in several layers, namely, input layer, one or more hidden layers, and output layer, constitutes a neural network. Each link has a weight, which specifies the strength of that link.

It essentially consists of two major stages in the neural network, namely forward propagation and backpropagation. Forward propagation is the input data that feeds into the network and runs through the layers with the calculations to create a prediction. Error is computed as a comparison between the output and actual target values.

During the process of backpropagation, an error is propa- gated backward from the network, and weight adjustments on the connections minimize the error. This goes on in an iterative forward and backward propagation until the network becomes satisfactory in its performance.

The formula for the output of a neuron in a neural network is typically represented as:

$$z = \sum_{i=1}^{n} w_i \cdot x_i + b$$

where:

- z is the weighted sum of inputs and biases,
- w_i represents the weights associated with each input x_i,
- x_i denotes the input values,
- b is the bias term,
- n is the number of inputs.

The output of the neuron is then calculated using an activation function $f(z)$:

$$y = f(z)$$

The loss function used here is Binary Crossentropy, which is perhaps one of the most commonly used loss functions for binary classification problems. It is the difference between true labels and the predicted probabilities in binary classification tasks, and Adam is being used as the optimizer.

4.2.4 Support Vector Machine

Support Vector Machine is perhaps a supervised machine-learning algorithm that will be used for classification and regression tasks. While SVM seeks the best of the hyperplanes, which maximally separates the data, points into different classes, it maximizes the margin between classes in classification. 1. Linear SVM for Binary Classifi- cation: For linearly separable data, the decision boundary or simply called hyperplane is represented as: wT x + b = 0 where:

- w is the weight vector. - x is the input feature vector. – b is the bias term.

The distance between a data point and the decision boundary is given by:

$$\frac{|\mathbf{w}^T\mathbf{x} + b|}{||\mathbf{w}||}$$

4.2.5 Lasso

Lasso is an acronym for Least Absolute Shrinkage and Selection Operator. It is a technique of linear re- gression for feature selection and regularization. It introduces a penalty term to the

regular linear regression objective function. This penalty prevents overfitting and favors simpler models by shrinking the coefficients of the less important features towards zero. The Lasso regression objective function is stated as:

$$\text{minimize } \left(\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \alpha \|\beta\|_1 \right)$$

- *y* represents the target variable vector,
- **X** represents the feature matrix,
- $\boldsymbol{\beta}$ represents the coefficient vector to be estimated,
- $\|\cdot\|_2$ denotes the L2 norm (Euclidean norm) of a vector,
- $\|\cdot\|_1$ denotes the L1 norm (absolute value sum) of a vector,
- α is the regularization parameter that controls the strength of the penalty term.

4.2.6 ElasticNet

Elastic Net is the method of regularized regression, where both penalties of Lasso, known as L1 regularization and Ridge, also called L2 regularization, are merged to try to overcome the weaknesses in both of them. These will have a balance between variable selection from Lasso and coefficient shrinkage in Ridge. The Elastic Net objective function is represented by:

- *y* represents the target variable vector,
- **X** represents the feature matrix,
- $\boldsymbol{\beta}$ represents the coefficient vector to be estimated,
- $\|\cdot\|_1$ denotes the L1 norm (absolute value sum) of a vector,
- $\|\cdot\|_2$ denotes the L2 norm (Euclidean norm) of a vector,
- α is the regularization parameter that controls the overall strength of the penalty term, and ρ is the mixing parameter that controls the balance between Lasso (when ρ = 1) and Ridge (when ρ = 0) penalties.

4.3 Evaluation Metrics

In classification problems, the over- all performance of a predictive model is often evaluated using accuracy as a common assessment metric. It reflects the number of correct classifications made out of all instances in the dataset. Accuracy is computed using:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Mathematically, if TP denotes the count of actual positive instances correctly classified, TN denotes the count of actual negative instances correctly classified, FP denotes the count of actual negative instances misclassified as positive and FN denotes the count of actual positive instances misclassified as a negative one, then accuracy can be defined as:The formula for Accuracy is given by:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

where

- TP (True Positives): The number of instances that are correctly predicted as positive.
- TN (True Negatives): The number of instances that are correctly predicted as negative.
- FP (False Positives): The number of instances that are incorrectly predicted as positive but actually negative.
- FN (False Negatives): The number of instances that are incorrectly predicted as negative but actually positive.

5. Result and Analysis

The study used various figures 4-7 datasets to determine how effective different machine learning algorithms were at predicting strokes. Overall, Decision Trees and Random Forests performed the best, with excellent accuracy on most of the datasets. These are ideal for solving complicated classification issues in medical data. Their ability to spot subtle patterns within datasets makes them useful for predicting strokes in many populations and data distributions. The study also considered the influence of different sampling methods-smote and undersampling-on the performance of the models. SMOTE has a tendency to enhance prediction for rarely occurring outcomes-minority class-most of the time. However, it was not steady in its effectiveness on accuracy. For example, performance of a Support Vector Machine did not improve significantly when using dataset 1 with SMOTE. In parallel, Decision Trees and Random Forests demonstrated extraordi- nary robustness across numerous datasets, confirming their versatility and dependability in identifying nuanced patterns and trends in stroke data. Their ability to accommodate non-linear correlations and complicated feature interactions makes them indispensable in real-world applications where data distributions may be diverse. In contrast, the evaluation of regression models using Lasso and ElasticNet regularization approaches, as measured by mean squared error (MSE) for continuous variables related to stroke occurrence, revealed higher MSE values, indicating suboptimal model fit. This implies that linear regression is insufficient for capturing the subtle interactions between characteristics and stroke risk, emphasizing the importance of considering nonlinear relationships in predictive models for stroke prediction.

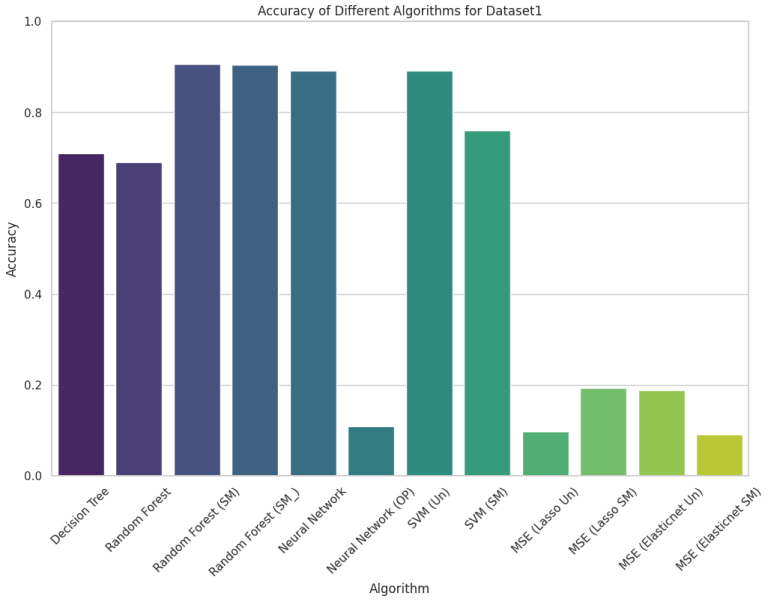

Figure 4. Brain Dataset

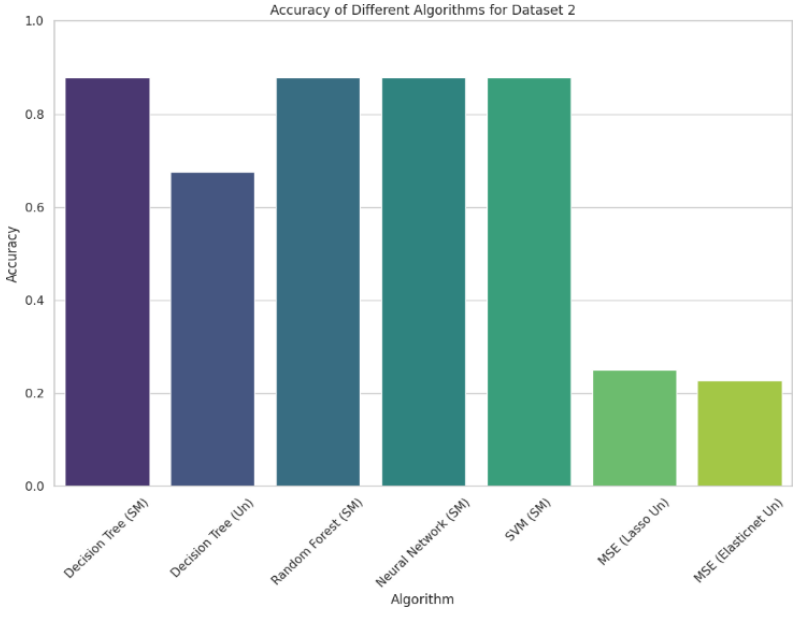

Figure 5. Diabetics Dataset

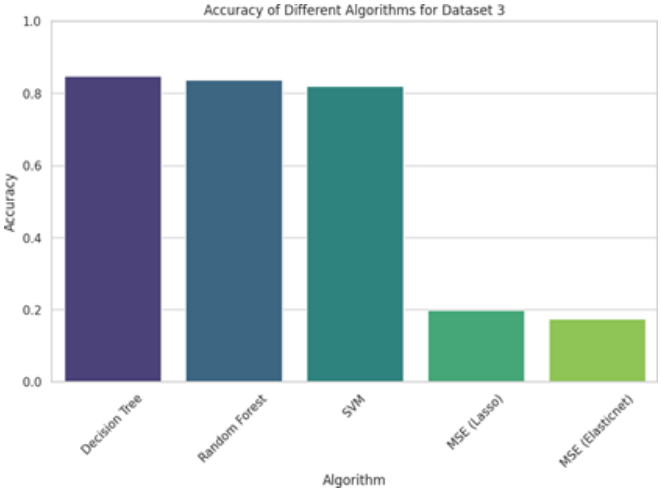

Figure 6. Heart Dataset

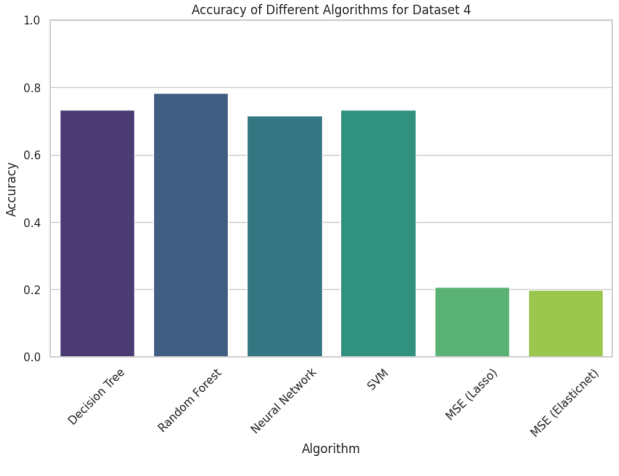

Figure 7. Heart Failure Dataset

Table 1. Accuracy and Mse For Brain Dataset

Model	Accuracy	MSE
Decision Tree (Undersampling)	0.71	-
Decision Tree (SMOTE)	0.89	-
Random Forest (Undersampling)	0.69	-
Random Forest (SMOTE)	0.91	-
Random Forest (SMOTE 1)	0.90	-
Neural Network	0.89	-
SVM (Undersampling)	0.89	-

SVM (SMOTE)	0.76	-
MSE (Lasso, SMOTE)		0.19
MSE (Lasso, Undersampling)		0.10
MSE (ElasticNet, Undersampling)		0.19
MSE (ElasticNet, SMOTE)		0.09

Table 2. Accuracy And Mse For Diabetics Dataset

Model	Accuracy	MSE
Decision Tree (SMOTE)	0.88	-
Decision Tree (Undersampling)	0.68	-
Random Forest (SMOTE)	0.88	-
Neural Network (SMOTE)	0.88	-
SVM (SMOTE)	0.88	-
MSE (Lasso, Undersampling)	-	0.25
MSE (ElasticNet, Undersampling)	-	0.23

Table 3. Accuracy And Mse For Heart Dataset

Model	Accuracy	MSE
Decision Tree	0.85	-
Random Forest	0.84	-
SVM	0.66	-
MSE (Lasso)	-	0.82
MSE (ElasticNet)	-	0.20

Table 4. Accuracy and Mse For Heart Failure Dataset

Model	Accuracy	MSE
Decision Tree	0.73	-
Random Forest	0.78	-
Neural Network	0.72	-
SVM	0.21	-
MSE (Lasso)	-	0.21
MSE (ElasticNet)	-	0.20

The accuracy and the mean squared details for all the datasets are shown in the tables1, 2, 3, 4.

6. Conclusion

Based on careful inspection of many datasets, there are sev- eral key general results about the performance and sensitivity of machine learning algorithms at the task of stroke prediction. First, Decision Trees and Random Forests always attain the best possible accuracies with essentially all datasets, making clear that they are remarkably capable of capturing very complex forms in the data. This kind of consistency talks about their potential to be implemented in real-world applications regarding assessing stroke risk. Moreover, as Neural Net- works are performing competitively for identifying complex correlations and patterns of the data, though performance may be inconsistent, SVM leads to consistent performance in different datasets, which reveals their reliability in dealing with changing types of data distributions.

Additionally, the importance of SMOTE and under- sampling techniques with the performance of the model by sampling is noteworthy. For instance, even though in general terms SMOTE increases the precision levels of minor class prediction by a sizeable amount, that does not happen at the model performance all the time; it tends to differ while choosing between such methods appropriate to a dataset type and nature. In addition, Lasso and ElasticNet regression models may be applied in analyzing the predictive capability of continuous variables with respect to stroke occurrence by making use of MSE. However, their somewhat lower accuracy ratings may pose some challenges in capturing the complex interactions between characteristics and stroke risk.

In conclusion, each of these models-Decision Trees, Ran- dom Forests, Neural Networks, SVMs-performs exceptionally in stroke prediction tests. Appropriate evaluation of sample methodologies and feature selection strategies is critical to improvise the accuracy and generalization of the model. In conclusion, further research efforts intended at refining existing models along with novel methodologies are more than necessary to enhance accuracy in stroke prediction and patients' outcomes in the hospital setting.

7. Research Gap

The gap in research lies in the less explored potential of marrying machine learning with neuroplasticity datasets to establish individualized stroke recovery plans at critical time windows. Moreover, not enough studies have been conducted about the long-term impacts of rehabilitation interventions and the application of novel technologies like virtual reality. Inequitable and culturally insensitive practice developments are further limited by under researched socioeconomic and cultural factors influencing stroke rehabilitation outcomes. Addressing these gaps is therefore fundamental to the advance- ment of stroke rehabilitation and improved patient outcomes.

References

[1] W. J. Powers, A.A. Rabinstein, T. Ackerson, O.M. Adeoye, N.C. Bambakidis, K. Becker, José Biller, Michael Brown, Bart M. Demaerschalk, Brian Hoh, Edward C. Jauch, Chelsea S. Kidwell, Thabele M. Leslie-Mazwi, Bruce Ovbiagele, Phillip A. Scott, Kevin N. Sheth, Andrew M. Southerland, Deborah V. Summers, David L. Tirschwell, American Heart Association Stroke Council. (2019). Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, 50(12), e344-e418. https://doi.org/10.1161/STR.0000000000000211

[2] M. J. O'Donnell, S. L. Chin, S. Rangarajan, D. Xavier, L. Liu, H. Zhang,., P. Rao-Melacini, X. Zhang, P. Pais, S. Agapay, P. Lopez-Jaramillo, A. Damasceno, P. Langhorne, M.J. McQueen, A. Rosengren, M. Dehghan, Graeme J Hankey, A.L.Dans, A. Elsayed, A. Avezum, C. Mondo, H.C. Diener, D. Ryglewicz, A. Czlonkowska, N. Pogosova, C. Weimar, R. Iqbal, Rafael Diaz, K. Yusoff, A. Yusufali, A. Oguz, X. Wang, E. Penaherrera, F. Lanas, O.S. Ogah, A. Ogunniyi, H.K. Iversen, G. Malaga, Z. Rumboldt, S. Oveisgharan, F. Al Hussain, D. Magazi, Y. Nilanont, J. Ferguson, G. Pare, S. Yusuf, (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. The lancet, 388(10046), 761-775. https://doi.org/10.1016/S0140-6736(16)30506-2

[3] N. Alageel, R. Alharbi, R. Alharbi, M. Alsayil, L.A. Alharbi, (2023). Using machine learning algorithm as a method for improving stroke prediction. *International Journal of Advanced Computer Science and Applications*, 14(4). 738- 744.

[4] E. Dritsas, M. Trigka, (2022). Stroke risk prediction with machine learning techniques. *Sensors*, 22(13), 4670. https://doi.org/10.3390/s22134670

[5] 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, in IEEE Access, 11 (2023), 52288-52308. https://doi.org/10.1109/ACCESS.2023.3278273

[6] S. Sahriar, S. Akther, J. Mauya, R. Amin, M.S. Mia, S. Ruhi, M.S. Reza, (2024). Unlocking stroke prediction: Harnessing projection-based statistical feature extraction with ML algorithms. *Heliyon*, 10(5), e27411. https://doi.org/10.1016/j.heliyon.2024.e27411

[7] T.S. Priyadarshini, M.A. Hameed, (2025). Collaboration of clustering and classification techniques for better prediction of severity of heart stroke using deep learning. *Measurement: Sensors*, 37, 101405. https://doi.org/10.1016/j.measen.2024.101405

[8] Gupta, A., Mishra, N., Jatana, N., Malik, S., Gepreel, K. A., Asmat, F., & Mohanty, S. N. (2025). Predicting stroke risk: an effective stroke prediction model based on neural networks. *Journal of Neurorestoratology,* 13(1), 100156. https://doi.org/10.1016/j.jnrt.2024.100156

[9] E. Dritsas, M. Trigka, Stroke Risk Prediction with Machine Learning Techniques. Sensors, 22(13), (2022) 4670. https://doi.org/10.3390/s22134670

[10] N. Biswas, K.M.M. Uddin, S.T. Rikta, S.K Dey, (2021) A Comparative Analysis of Machine Learning Classifiers for Stroke Prediction: A Predictive Analytics Approach. *Healthcare Analytics*, 2, 100116.

[11] K. Akash Mahesh, H. N. Shashank, S. Srikanth S, A.M. Thejas, Prediction of Stroke Using Machine Learning. *published via ResearchGate Institute of Technology (CMRIT),* Bangalore

[12] Y. Chahine, et al., "Arrhythmia Risk and Stratification," IEEE Reviews in Biomedical Engineering, vol. 15, pp. 211-225, 2022.

[13] M. Chun, R. Clarke, B.J. Cairns, D. Clifton, D. Bennett, Y. Chen, Y. Guo, P. Pei, J. Lv, C. Yu, L. Yang, (2021) Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *Journal of the American Medical Informatics Association,* 28(8), 1719-1727. https://doi.org/10.1093/jamia/ocab068

[14] Amann, J. (2022). Machine learning in stroke medicine: Opportunities and challenges for risk prediction and prevention. *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues,* 57-71.

[15] A. Jamthikar, D. Gupta, L. Saba, N.N. Khanna, T. Araki, K. Viskovic, S. Mavrogeni, J.R. Laird, G. Pareek, M. Miner, P.P. Sfikakis, A. Protogerou, V. Viswanathan, A. Sharma, A. Nicolaides, G.D. Kitas, J.S. Suri, (2020). Cardiovascular/stroke risk predictive calculators: a comparison between statistical and machine learning models. *Cardiovascular Diagnosis and Therapy,* 10, 919–938. https://doi.org/10.21037/cdt.2020.01.07

[16] S. Gangavarapu, G.L.A. Kumari, (2021) Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications*, 12(6), 539–545. https://doi.org/10.14569/IJACSA.2021.0120662

Funding

No funding was received for conducting this study.

Conflict of interest

The Author's have no conflicts of interest to declare that they are relevant to the content of this article.

About The License