University of Ibadan
Journal of Science and Logics in ICT Research

A Hybrid Approach to Developing a Stroke Prediction System

[1]✉Sotonwa, K. A., [2]Oluwatosin, E. A., [3]Raji-Lawal, H. Y., [4]Zubair, A. F. and [5]Alegbhele, I. B.

[1, 3, 4, 5]Lagos State University
[2]Bells University of Technology
kehinde.sotonwa@lasu.edu.ng, eaoluwatosin@bellsuniversity.edu.ng, halaw313@yahoo.com, adamfunsho@hotmail.com, ishaquabdulrahim18gmail.com

Abstract
The development of a stroke prediction system using machine learning algorithms offers a novel approach to identifying individuals at risk for stroke. By analyzing large datasets, it is possible to identify patterns and predictors of stroke that may not be apparent to human clinicians. This system has the potential to improve early detection and treatment of stroke, leading to better patient outcomes and helping to identify at-risk individuals who may benefit from preventive measures. Although single techniques have been employed to improve the accuracy and robustness of stroke prediction models, this study performs a hybrid technique using logistic regression (LR), random forest (RF), and support vector machines (SVMs) to enhance the accuracy and robustness of the proposed model. All three algorithms performed well in terms of accuracy, with random forest achieving the highest accuracy. However, LR and SVM were more efficient regarding training time and complexity. The overall conclusion was that RF is the best-performing algorithm for this particular task, but other algorithms may be more suitable for different applications. In conclusion, developing a stroke prediction system using machine learning algorithms is a promising approach for improving stroke prediction and patient outcomes. This study shows that machine learning algorithms can effectively identify individuals at risk for stroke and may have advantages over traditional risk factors. However, more research is needed to fully understand the potential of machine learning in this field and to determine the most effective algorithms and training methods.

Keywords: *Stroke Prediction System (SPS), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Machine learning.*

1. INTRODUCTION

A stroke, or CVA (cerebrovascular accident), is a serious medical condition in which the blood flow to the brain is disrupted. This could occur due to a blockage in one of the blood vessels that supply the brain or even a rupture of a blood vessel in the central nervous system. Stroke can damage the brain permanently, making one paralyzed and, sometimes, even leading to death. It is the second most common cause of death globally and is a prevalent cause of adult disability. There are two principal types of stroke: Ischemic stroke and haemorrhagic stroke [1]. Losing blood supply to the brain causes damage to the tissue inside the brain.

The damage in the brain from a stroke mirrors the corresponding parts of the body. The better a stroke condition is treated earlier, the better the recovery. Knowing the signs of stroke thus empowers a person to act fast. Stroke symptoms include weakness, paralysis, or numbness in the arms, face, or legs, particularly on one side of the body; difficulty talking or comprehending words; slurring speech; confusion, disorientation, or non-responsiveness; sudden behavioral changes, specifically increased agitation; visual disturbances, involving darkness or blurriness in one or both eyes, double vision; loss of balance or coordination; dizziness; severe and sudden headache with an unknown cause; and seizures. Prediction of stroke involves the identification of persons at risk of suffering stroke shortly [2].

This could mean the use of predictive models to predict who will develop the stroke, so the developing preventative measures could then be

Sotonwa K. A., Oluwatosin A. E., Raji-Lawal H. Y., and Olugbhele I. B. (2024). Development of Stroke Prediction System using Hybridized Techniques, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 5 No. 1, pp. 65 - 73

undertaken to reduce the said chances of suffering from a stroke. This may include behavioral change and medication, or surgery, if required. Ultimately, one would want to implement stroke prediction for the advancement of patients' outcomes either through a reduction in number of people with stroke affliction or by early detection and treatment of the affliction [1].

Stroke Prediction System (SPS) is a unique, developed computerized system conceived to predict stroke risk of individuals more reliably and accurately. SPS practices leverage clinical, imaging, and genetic data integration with a wide range of complementary machine learning and medical knowledge representation techniques to achieve the goal of reliably and accurately predicting stroke risk. Through machine learning, the computer goes through an ability to learn by example— an extraordinarily powerful tool, since with experience it can come close to understanding the SOURCE data complexity through vast amounts of data analysis with patterns that even human beings might not grasp [2].

Techniques of medical knowledge representation encode the contexts and concepts of medical aspects in a planned and uniform manner. Currently, several technologies are in development for the formal encoding and representation of medical concepts and contexts, including ontologies and taxonomies, which will permit SPS to integrate and handle the enormous amounts of medical information for the prediction of stroke risk as high or low [3, 4].

The aggregation of many techniques, such as logistic regression, remarks how the independent variable is related to the dependent variable in [5-7]. Support vector machine is a technique that creates optimized lines or decision boundaries, which in turn isolate classes in n-dimensional space [8, 9]. Random forest trains many decision trees and aggregates their predictions to increase the accuracy of the model [10] and the k-nearest neighbour's technique is used in doing the work for both classification and regression problems [11]. Besides, it gave better insight into the accuracy and model performance than a single technique, which could improve stroke prediction [12, 13].

A deep learning approach for predicting stroke risks in patients that combined convolutional neural networks and long short-term memory networks. It achieved a good accuracy rate in the prediction of stroke risk and outperformed other state-of-the-art methods [14].

Also, integrated use of fuzzy logic in quantifying uncertainty in data, fuzzy rough sets in feature selection, and support vector machines in grouping patients in different risk groups with better prediction accuracy than a single machine learning technique [15]. At present, only a few of the state-of-the-art stroke prediction methods that have been based on the following techniques have taken into account a limited number of the important risk factors, such as age, gender, and blood pressure [16, 17], without considering other important ones like genetic markers and environmental factors. Hence, this research would develop SPS by considering other potential factors like genetic markers and environmental factors.

2. RELATED WORKS

Ali *et al*., [18] in their work combined fuzzy logic with support vector machines in a hybrid technique for the prediction of stroke. The present study observed the hybrid approach to be useful for the reason that not only did it improve model performance but also increased interpretability. However, the study was applied to the prediction of only ischemic stroke, which represents just one type of stroke.

A deep learning-based approach was used to classify a stroke patient into different risk groups using a Convolutional Neural Network and a Short-Term Memory network. The study pinpointed that this approach outperforms other machine learning algorithms and can classify patients into different risk groups with a relatively high degree of accuracy. Unfortunately, this was not a matter considered in this study, which focused on the prediction of the objurgate for stroke within 90 risks but not viewed for long-term risk or outcome [14].

Hybrid machine learning was suggested: a mixture of multiple support vector machines, random forests, and logistic regression algorithms for stroke prediction in this regard. The result demonstrated that the hybrid approach would predict a higher mix of stroke events than if one should use a single algorithm. However, this study included only a handful of risk factors so the features used would be just age, gender, and blood pressure and no other important

factors, like genetic markers and environmental factors [16, 17].

Abiodun and Wreford [19] employed two stacked ensemble machine learning (ML) algorithms, i.e., KNN-based and XGBoost-based, to develop the model and test its framework in predicting the long-term risk of stroke. Experimental results confirmed that the proposed stacked algorithm outperformed all the other used ensemble methods and obtained great accuracy of 97%, with a recall of 95% and 98%, precision of 98% and 95%, and f1 score of 97%; the scope of the study seeks to augment the ML framework with deep learning methods [19]. A model was developed using a combined CNN and BiGRU to collect patient data using the portable MUSE-2EEG device. This authorized individual's mobile phones through the sending of many messages, making the seeking of results very fast [20].

3. METHODOLOGY

3.1 Data Collection

It exploits the Patients of 5110 who have suffered from stroke dataset in Kaggle. The dataset contains data related to the results of laboratory tests, the history of the patient's medical background, and demographic information of patients. In this paper, the developed model uses this dataset to predict the probability that a patient will suffer from a stroke, taking into consideration several input parameters like age, gender, whether they have certain diseases or not, and whether a patient smokes or not. Here is the inventory explaining each attribute column [21]:

- Id: This variable contains the numerical identifier for all patients. Data Type: int64
- gender: This variable contains the gender of every patient. Data Type: Object
- age: This variable carries the age of every patient. Data Type: float64
- hypertension: the variable shows if a patient has hypertension or not. Data Type: int64
- heart disease: It includes the status of heart sickness to a specific patient. Data Type: int64
- ever married: What is included in this variable is the state of marriage by every patient. Data Type: object
- work type: This is the variable that indicates current work categories by every patient. Data Type: object
- residence type: This variable reflects the patients who live either in an urban or rural area. Data Type: object
- average glucose: This feature represents the average glucose level of a patient. Data type: float64
- body mass index: this variable contains the body mass index of respective patients. Data type: float64
- smoking status: this variable represents whether the patient is a smoker or not. Data type: object
- stroke: it represents whether a patient has ever suffered a stroke or not. Data type: int64 [21-23].

3.2 Data Preprocessing

The data was pre-processed and cleaned to make it ready for analysis by removing the extra data like checking for duplicates in the dataset, reducing the missing values, encoding categories features, imbalanced data management, and modelling such as age, sex, blood pressure, should include genetic markers and environmental factors

3.3 Evaluation Parameters

The data was then hybridized with machine learning algorithms like logistic regression, random forest, and support vector machine to evaluate and compare their performance using these hybrid models.

4. RESULTS AND DISCUSSION

4.1 System Analysis and Design for SPS using LR, RF and SVM

Figure 1 is a Python script to display LR performance by importing necessary libraries. Again, the most commonly used library for performing LR is scikit-learn, including the class of LR. This package is imported by the following command: import sklearn.linear_model as lm, which makes the class of LR available as class lm. Logistic regression is to be used for specifying relevant parameters. Figures 1 and 2 show a very useful method that allows for the discovery of trends or patterns, difficult to determine through the simple raw observation of data. Categorical columns, which can represent factors such as age, gender, or different comorbidities, are shown here. In this way, there will be a possibility for the detection of inequalities or biases in data or to verify whether the machine learning model has been trained on a representative sample population.

Figure 1: Import libraries

Figure 2: Visualizing categorical columns

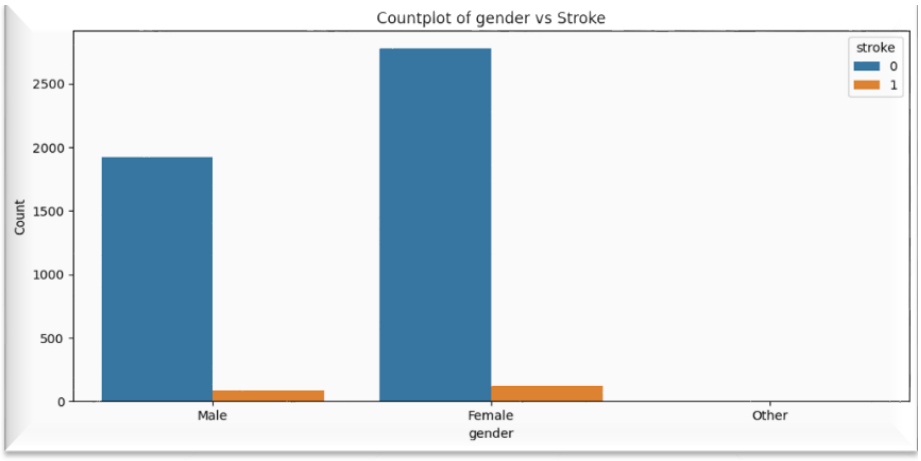

Figure 3a: Countplot of gender vs stroke

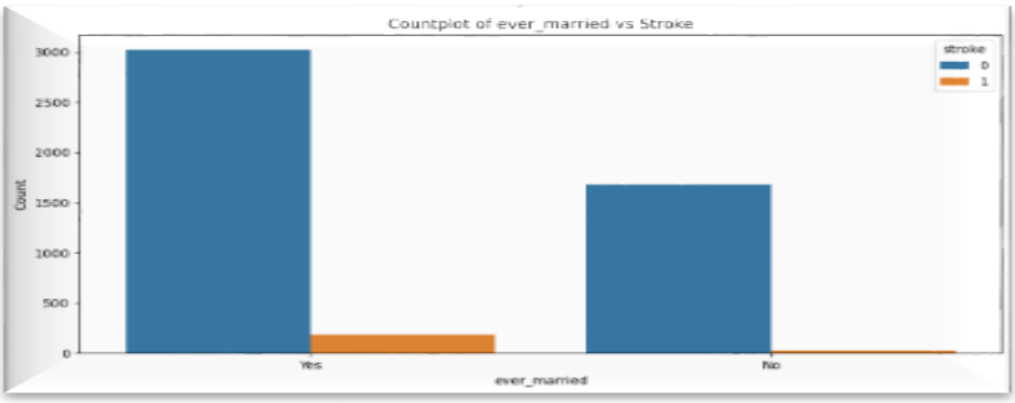

Figure 3b: Countplot of ever_married vs stroke

Figure 3a is the frequency of each gender about whether the patient had a stroke, which would be useful in identifying any gender-based disparities in stroke incidence while Figure 3b is the frequency of ever_married about whether they had a stroke, which could be used in investigating marital status as a risk factor for stroke.

Figure 3c, the frequency of different types of jobs considering the presence or absence of stroke that may relate certain occupations to a higher risk for stroke. Figure 3d shows the frequency of different types of residence, such as urban, suburban, and rural, concerning the presence or absence of stroke, to see whether having residence in a particular type of area predisposes one to stroke.

Figure 3e is the graph that plots the smoking status formerly smoked, never smoked, smokes, unknown against whether the subject had a stroke or not. This could help in identifying whether smoking indeed is one of the risk factors for stroke.

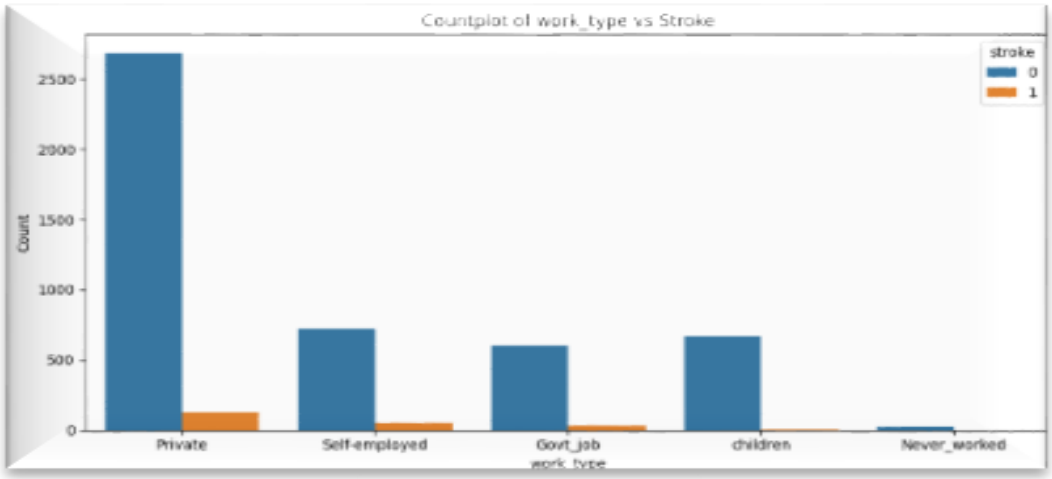

Figure 3c: Countplot of work_type vs stroke

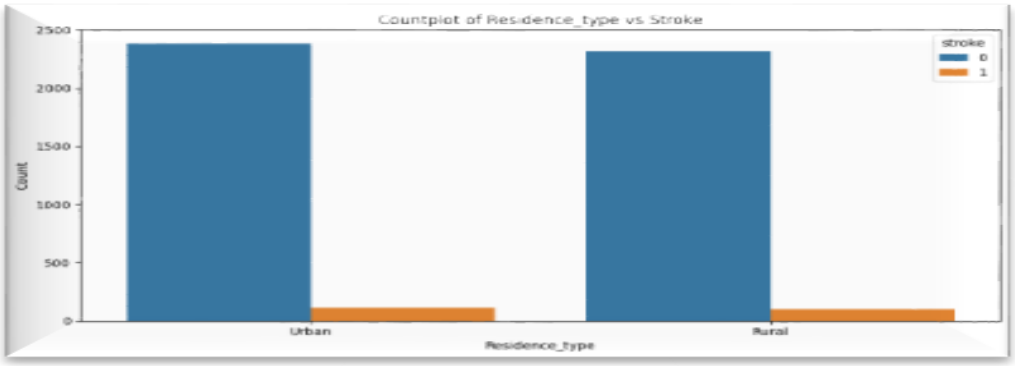

Figure 3d: Countplot of residence_type vs stroke

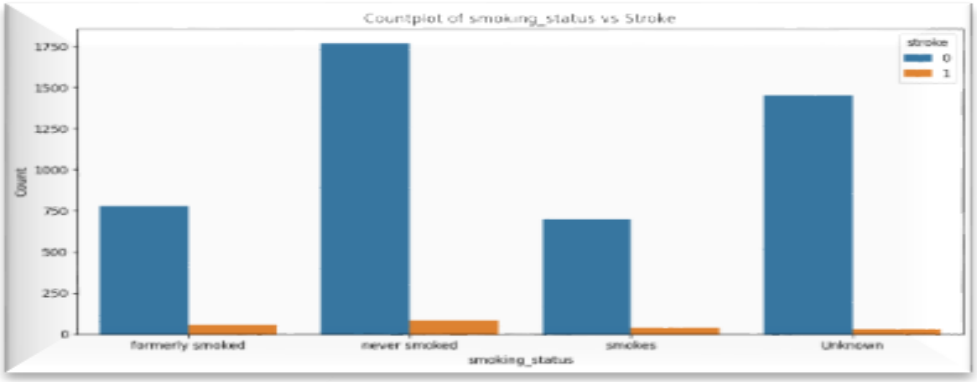

Figure 3e: Countplot of smoking_status vs stroke

Figure 4a is the distribution of age, and a stroke or no stroke used to understand the distribution of ages for strokes and the age-related patterns while Figure 4b is: the distribution of average glucose levels about whether the person suffered a stroke, by which to highlight whether there is a risk that high blood glucose levels contribute towards having a stroke. Finally, Figure 4c - body mass index (BMI) distribution to log10 scale when considering stroke or no stroke. This would allow investigation of whether obese persons are at higher risk of getting a stroke.

4.2 Training the Hybridized techniques (LR, FR and SVM)

To develop a stroke prediction application model in both LR and FR and supported by SVM, data is divided into training and test sets, as depicted in Figure 5 the sets for training are computed to train the model from a set of data, whereas the test set will be applied to validate the model and access its performance. The LR used the Logistic Regression class in Python scikit-learn, RF the Random Forest Classifier class in Python scikit-learn, and SVM used the SVC class in scikit-learn in Python.

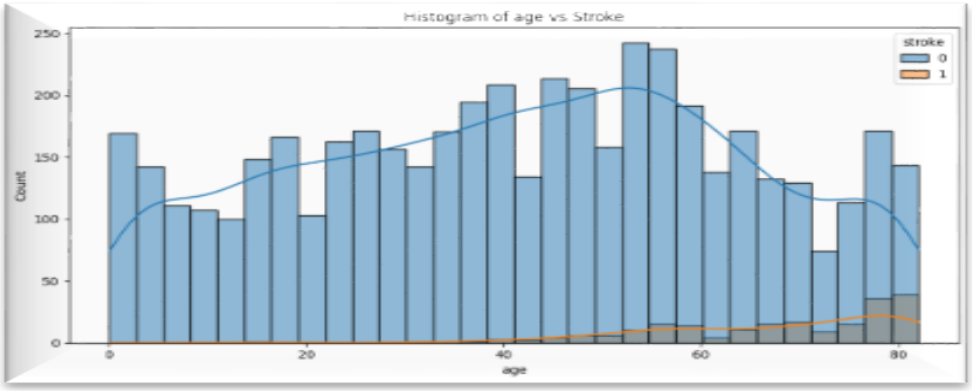

Figure 4a: Histogram of age vs stroke

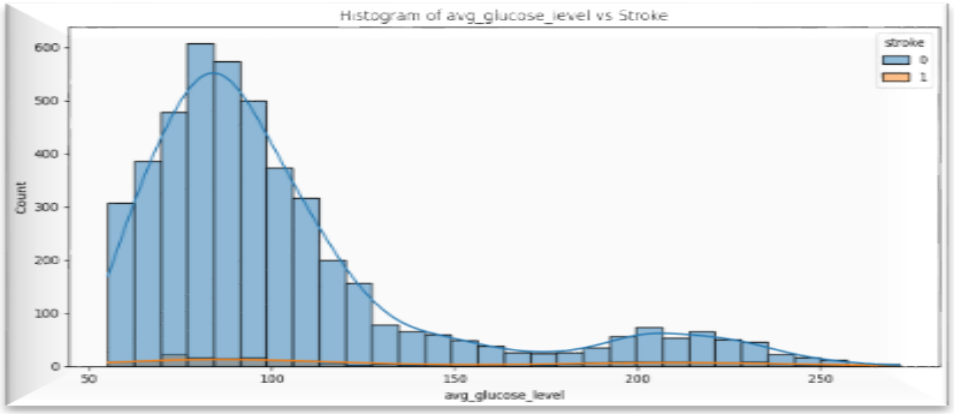

Figure 4b: Histogram of avg_gluclose_level vs stroke

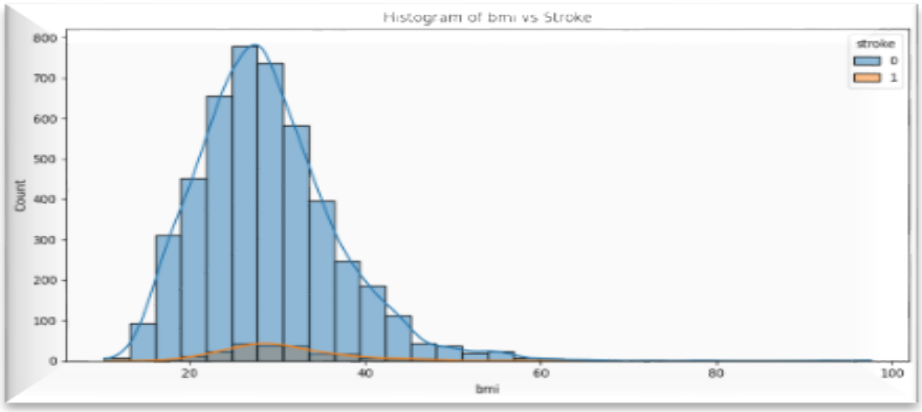

Figure 4c: Histogram of bmi vs stroke

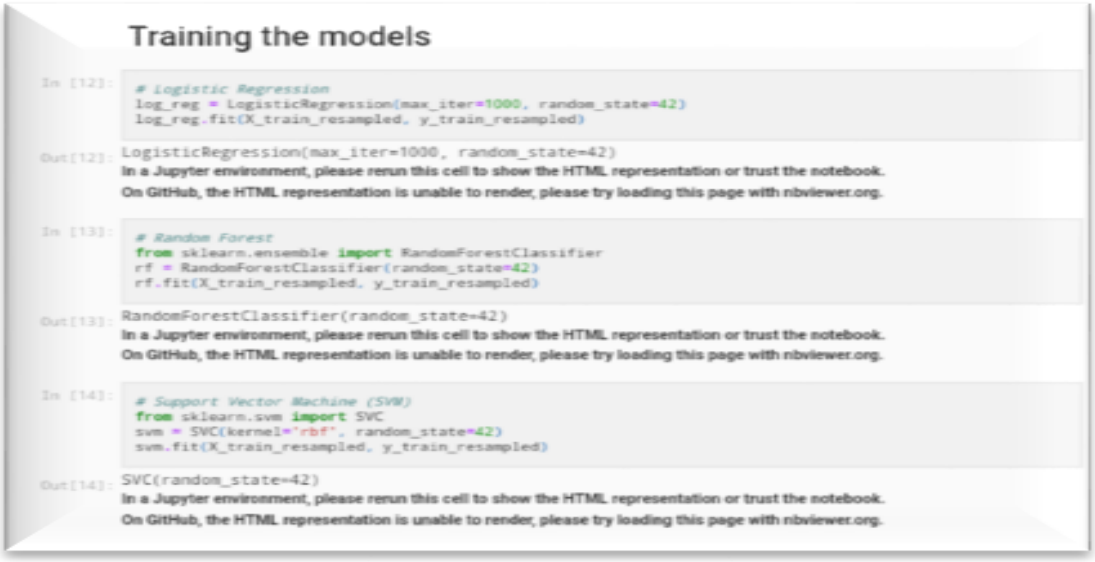

Figure 5: Training the models

4.3 Evaluating the Hybridized techniques (LR, FR and SVM)

Performance evaluations like accuracy (Acc), precision, recall, accuracy average (avg), and weighted average (avg) were used to evaluate the three algorithms. Generally, as summarized in Table 1, of the three kinds of algorithms, RF showed the best accuracy in stroke prediction. This is because RF is less likely to over fit and deals with high-dimensional data well. The best-performing model will vary for a particular dataset, depending on the hyper-parameters used for training the model. Feature importance scores are provided by RF, while the LR model gives the baseline accuracy, and the SVM is for dealing with high-dimensional data, as shown in Figure 6.

Figure 6: Hybrid prediction

Table1: Classification results for hybridized model

Logistic Regression				
	Precision	Recall	FI	Support
0	0.98	0.76	0.85	960
1	0.17	0.76	0.28	62
Acc			0.76	1022
Acc avg	0.57	0.76	0.57	1022
Weighted avg	0.93	0.76	0.82	1022
Random Forest				
0	0.94	0.97	0.96	960
1	0.20	0.11	0.14	62

Acc			0.92	1022
Acc avg	0.57	0.54	0.55	1022
Weighted avg	0.90	0.92	0.91	1022
Support Vector Machine				
0	0.96	0.79	0.87	960
1	0.14	0.53	0.22	62
Acc			0.78	1022
Acc avg	0.55	0.66	0.55	1022
Weighted avg	0.91	0.78	0.83	1022

5. CONCLUSION

The objective of this work is to present a machine-learning model that will predict the risk of stroke based on the demographic and health data obtained from the questionnaire. The data were sourced from varying sources, after which the model was trained and tested with machine learning techniques. It showed good accuracy in delineating the risk of stroke and, therefore, may potentially help in improving patient care and hence reducing the burden of stroke. This could be further developed in the future by augmenting the dataset from a larger number of patients, utilizing more advanced machine learning techniques such as deep learning, and developing a user interface that would bring the model closer to being accessible to clinicians.

REFERENCES

[1] Boehme, A.K., C. Esenwa, and M.S. Elkind (2017), Stroke risk factors, genetics, and prevention. *Circulation Research*, **120**(3): p. 472-495.

[2] Feigin, V.L., *et al.* (2016), Prevention of stroke: a strategic global imperative. *Nature Reviews Neurology*, **12**(9): p. 501-512.

[3] Li, X., *et al.* (2019), Using machine learning models to improve stroke risk level classification methods of China national stroke screening. *BMC medical informatics and decision making*, **19**: p. 1-7.

[4] Zu, W., *et al.* (2023), Machine learning in predicting outcomes for stroke patients following rehabilitation treatment: A systematic review. *Plos one*,. **18**(6): p. e0287308.

[5] Monteiro, M., *et al.*, (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM transactions on computational biology and bioinformatics*, **15**(6): p. 1953-1959.

[6] Mohanty, R., *et al.* (*2017)* Machine learning-based prediction of changes in behavioral outcomes using functional connectivity and clinical measures in brain-computer interface stroke rehabilitation. in Augmented Cognition. Neurocognition and Machine Learning: *11th International Conference, AC, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 11*. 2017. Springer.

[7] Nathaniel, T.I., *et al.* (2016), Co-morbid conditions in use of recombinant tissue plasminogen activator (rt-PA) for the treatment of acute ischaemic stroke. *Brain injury*, **30**(10): p. 1261-1265.

[8] Torres-Riera, S., *et al.* (2024), Predictive Clinical Factors of In-Hospital Mortality in Women Aged 85 Years or More with Acute Ischemic Stroke. *Cerebrovascular Diseases*, p. 1-1.

[9] Taylor-Rowan, M., *et al.* (2019), Pre-stroke frailty is independently associated with post-stroke cognition: a cross-sectional study. *Journal of the International Neuro-psychological Society*, **25**(5): p. 501-506.

[10] Lee, W.H., *et al.* (2020), Development of a novel prognostic model to predict 6-month swallowing recovery after ischemic stroke. *Stroke*, **51**(2): p. 440-448.

[11] Sari, W.J., *et al.* (2024), Performance Comparison of Random Forest, Support Vector Machine and Neural Network in Health Classification of Stroke Patients. *Public Research Journal of Engineering, Data Technology and Computer Science*, **2**(1): p. 34-43.

[12] Bandi, V., D. Bhattacharyya, and D. Midhunchakkravarthy (2020), Prediction of Brain Stroke Severity Using Machine Learning. *Rev. d'Intelligence Artif.*,. **34**(6): p. 753-761.

[13] Sailasya, G. and 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).

[14] Chiu, I.-M., *et al.* (2021), Using a multiclass machine learning model to predict the outcome of acute ischemic stroke requiring reperfusion therapy. *Diagnostics*,. **11**(1): p. 80.

[15] Yuan, K., *et al.* (2020), *A nomogram for predicting stroke recurrence among young adults.* Stroke, **51**(6): p. 1865-1867.

[16] Elbagoury, B.M., et al. (2023), A hybrid stacked CNN and residual feedback GMDH-LSTM deep learning model for stroke prediction applied on mobile AI smart hospital platform. *Sensors*, 23(7): p. 3500.
[17] Fang, G., Z. Huang, and Z. Wang (2022), Predicting ischemic stroke outcome using deep learning approaches. *Frontiers in genetics*, 12: p. 827522.
[18] Kariasa, I.M., E. Nurachmah, and R.A. Koestoer (2019), Analysis of participants' characteristics and risk factors for stroke recurrence. *Enfermeria clinica*, 29: p. 286-290.
[19] Watila, M., *et al.* (2012), Risk factor profile among black stroke patients in Northeastern Nigeria. *J Neurosci Behav Health*, 4(5): p. 50-8.
[20] Ali, R., *et al*. (2019), Adaptive neuro-fuzzy inference system for prediction of surgery time for ischemic stroke patients. *International Journal of Integrated Engineering*, 11(3).
[21] Foroozanfar, Z., *et al.* (2020), Sex differences in 28-day mortality of ischemic stroke in Iran and its associated factors: a prospective cohort study. *Journal of Stroke and Cerebrovascular Diseases*, 29(8): p. 104896.
[22] Abiodun, O.J. and A.I. Wreford (2023), *Stroke* Prediction Using Smote for Data Balancing, XGBoost and KNN Ensemble Algorithms. *Journal of Applied Physical Science International*, 15(1): p. 42-53.
[23] Sawan, A., *et al*. (2024), Hybrid deep learning and metaheuristic model based stroke diagnosis system using electroencephalogram (EEG). *Biomedical Signal Processing and Control*, 87: p. 105454.