Neuroscience Informatics 5 (2025) 100216

Contents lists available at ScienceDirect

Neuroscience Informatics

journal homepage: www.elsevier.com/locate/neuri

Original article

Predicting stroke with machine learning techniques in a sub-Saharan African population

Benjamin Segun Aribisala [a,b,c], Deirdre Edward [a], Godwin Ogbole [d], Onoja M. Akpa [e], Segun Ayilara [d], Fred Sarfo [f], Olusola Olabanjo [b], Adekunle Fakunle [g,h], Babafemi Oluropo Macaulay [b], Joseph Yaria [h], Joshua Akinyemi [e], Albert Akpalu [i], Kolawole Wahab [j], Reginald Obiako [k], Morenikeji Komolafe [l], Lukman Owolabi [m], Godwin Osaigbovo [n], Akinkunmi Paul Okekunle [o], Arti Singh [p], Philip Ibinaye [k], Osahon Osawata [e], Adeniyi Sunday [j], Ijezie Chukwuonye [q], Carolyn Jenkins [r], Hemant K. Tiwari [s], Okechukwu Ogah [e], Ruth Y. Laryea [f], Daniel T. Lackland [r], Oyedunni Arulogun [t], Omotolani Ajala [h], Rufus Akinyemi [u,v], Bruce Ovbiagele [w,1], Steffen Sammet [a,*,1], Mayowa Owolabi [h,v,**,1] on behalf of the SIREN and SIBS Genomics investigators

[a] Department of Radiology, University of Chicago, Chicago, IL, USA
[b] Department of Computer Science, Lagos State University, Lagos, Nigeria
[c] Department of Computer Science, Oduduwa University, Nigeria
[d] Department of Radiology, University of Ibadan, Ibadan, Nigeria
[e] Department of Epidemiology and Medical Statistics, University of Ibadan, Nigeria
[f] Department of Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[g] Department of Public Health, College of Health Sciences, Osun State University, Osogbo, Nigeria
[h] Department of Medicine, College of Medicine, University of Ibadan, Ibadan, Nigeria
[i] Department of Medicine, University of Ghana Medical School, Accra, Ghana
[j] Department of Medicine, University of Ilorin Teaching Hospital, Ilorin, Nigeria
[k] Department of Medicine, Ahmadu Bello University, Zaria, Nigeria
[l] Department of Medicine, Obafemi Awolowo University Teaching Hospital, Ile-Ife, Nigeria
[m] Department of Medicine, Aminu Kano Teaching Hospital, Kano, Nigeria
[n] Department of Medicine, University of Jos, Jos, Nigeria
[o] Department of Food and Nutrition, Seoul National University, Korea
[p] Department of Epidemiology and Biostatistics, Kwame Nkrumah University of Science and Technology, Ghana
[q] Department of Medicine, Federal Medical Centre, Umuahia, Nigeria
[r] Medical University of South Carolina, Charleston, USA
[s] University of Alabama at Birmingham, Birmingham, AL, USA
[t] Department of Health Promotion, University of Ibadan, Ibadan, Nigeria
[u] Department of Medicine, Federal Medical Centre, Abeokuta, Nigeria
[v] Centre for Genomic and Precision Medicine, College of Medicine, University of Ibadan, Nigeria
[w] Weill Institute for Neurosciences, School of Medicine, University of California San-Francisco, USA

ARTICLE INFO

Article history:
Received 28 March 2025
Received in revised form 9 June 2025
Accepted 12 June 2025

Keywords:
Stroke

ABSTRACT

Background: Stroke is the second leading cause of death and the third leading cause of disability globally, including Africa, which bears its largest burden. Accurate models are needed in Africa to predict and prevent stroke occurrence. The aim of this study was to identify the best machine learning (ML) algorithm for stroke prediction.

Methods: We assessed medical data of 4,236 subjects comprising 2,118 stroke patients and 2,118 controls from the SIREN database. Sixteen established vascular risk factors were evaluated in this study. These are

* Corresponding author at: Department of Radiology, University of Chicago Medicine, 5841 S. Maryland Avenue, MC2026 Chicago, IL 60637, USA.
** Corresponding author at: Center for Genomics and Precision Medicine, College of Medicine, University of Ibadan, Ibadan, University College Hospital, Ibadan, and Blossom Specialist Medical Center, Ibadan, Nigeria.
E-mail addresses: ssammet@uchicago.edu (S. Sammet), mayowaowolabi@yahoo.com (M. Owolabi).
[1] Joint senior authors.

https://doi.org/10.1016/j.neuri.2025.100216

SIREN
Sub-Sahara Africa
Risk factors
Machine learning
Artificial neural network

addition of salt to food at table during eating, cardiac disease, diabetes mellitus, dyslipidemia, education, family history of cardiovascular disease, hypertension, income, low green leafy vegetable consumption, obesity, physical inactivity, regular meat consumption, regular sugar consumption, smoking, stress and use of tobacco. From these, we also selected the 11 topmost risk factors using Population-Attributable Risk ranking. Eleven ML models were built and empirically investigated using the 16 and the 11 risk factors.

Results: Our results showed that the 16 features-based classification (maximum AUC of 82.32%) had a slightly better performance than the 11 feature-based (maximum AUC 81.17%) algorithm. The result also showed that Artificial Neural Network (ANN) had the best performance amongst eleven algorithms investigated with AUC of 82.32%, sensitivity of 71.23%, specificity of 80.00%.

Conclusion: Machine Learning algorithms predicted stroke occurrence employing major risk factors in Sub-Saharan Africa better than regression models. Machine Learning, especially Artificial Neural Network, is recommended to enhance Afrocentric stroke prediction models for stroke risk factor quantification and control in Africa.

1. Introduction

Stroke is the second leading cause of death and the third leading cause of disability worldwide [1], with over half of stroke survivors chronically disabled [2]. The management of stroke has improved greatly in the 21st century with the introduction of new evidence-based interventions for acute stroke [3]. However, the availability of these interventions for stroke treatment varies considerably across geographic regions, with suboptimal care in low and middle income settings, which bear most of the stroke burden [4].

The World Health Organization (WHO) estimates that low- and middle-income countries (LMIC) account for 70% of stroke incidence [5]. Africa is experiencing a rise in the prevalence of risk factors and the most viable option to control the rising burden of stroke in Africa is prevention [6,7]. A risk calculator is the topmost recommendation for stroke prevention by the American Heart Association [6]. While several stroke risk calculation models are available for different regions around the world [8], none has been found to be valid in Africa or developed specifically for the African population. From the Stroke Investigative Research and Education Network (SIREN) study, we had previously developed a logistic regression model for stroke prediction, which we seek to improve through machine learning (ML) [9]. The SIREN study is a multicenter case-control project, encompassing fifteen sites across Nigeria and Ghana.

ML has a great potential to enable the development of mathematical models that may accurately predict and characterize stroke. ML is an application of Artificial Intelligence that enables systems to learn and improve from experience without being explicitly programmed. Previous research efforts directed at application of ML to stroke prediction have yielded promising results [10]. However, the limited availability of data regarding stroke in sub-Saharan Africa has restricted the productivity of computational stroke research in this region. Nevertheless, the SIREN data is now available to fill this gap [11]. The aim of this study was therefore to develop an ML technique for stroke prediction amongst Africans using SIREN data. Eleven commonly used ML algorithms in medical diagnosis were investigated and the algorithm that performed best was selected (Supplementary material S1).

2. Materials and methods

2.1. Study participants

We conducted this study using 4,236 participants with 2,118 stroke patients and 2,118 controls from the Stroke Investigative Research and Educational Network (SIREN) study. The enrollment protocol for SIREN has been published previously [11–13]. SIREN is the largest stroke study performed on subjects in Africa, with the most comprehensive clinical data from stroke patients and age- and sex-matched stroke-free controls in Nigeria and Ghana [14]. This study has identified contributions from sociodemographic, environmental and genetic factors to the risk and outcomes of stroke and its subtypes in Africa, as well as modifiable lifestyle risk factors, that predispose individuals to stroke [15].

The overall coordinating Institutional Review Board (IRB) at all study sites for the SIREN study was the University of Ibadan/University College Hospital Ibadan, Nigeria (IRB Approval No.: UI/EC/13/0105). This research conformed to the principles of the Helsinki Declaration and all respondents provided written informed consent before participating in the study. Stroke patients included in this study were consenting adults aged ($\geq$18 years) who presented within 8 days of stroke onset. In patients that were unconscious or aphasic, consent was obtained from their next of kin. Stroke status was ascertained using neuroimaging, with either Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), within ten days of symptom onset. Stroke risk factor evaluation were performed according to a standardized protocol at each site. Patients were recruited from hospitals to prevent inaccurate stroke phenotyping and minimize referral bias [16].

Controls were stroke-free adults who consented to participate in the SIREN study and were from the same communities, where stroke patients were recruited. Their stroke-free status was verified using the questionnaire for Verifying Stroke-Free Status (QVSFS) which had a 98% negative predictive value [17]. Controls were matched to case at a ratio of 1:1 using age (with a variation of $\pm$5 years), sex, and ethnicity to minimize potential confounding factors. Details of the risk factor definitions for stroke cases and controls are provided in the Supplementary file (S2).

2.2. Machine learning algorithms for stroke classification

In the context of stroke prediction, classification algorithms are functions that map measurable input variables to one of two categories: “stroke” or “control.” The resulting models draw conclusions about the presence of stroke in a subject, based on observations made about these input variables. Eleven binary classification algorithms which are mostly used in medical application of ML were investigated in this study: Logistic Regression, Nearest Neighbors, Linear Support Vector Machine (SVM), Radial Basis Function (RBF), Gaussian Process, Quadrant Discriminant Analysis (QDA), Naive Bayes, Decision Tree, Random Forest Classifier (RFC), Adaboost boosting (AdaBoost), Artificial Neural Network (ANN). A brief description of the 11 ML algorithms is provided below and more information is provided in the Supplementary Material S1.

Logistic regression uses the probability of odds to perform classification [18] and Nearest Neighbor uses a distance metric to

classify a data point based on the nearest class [19]. SVM does classification by computationally determining the hyperplane that maximally separates the datapoints into the target classes and it can be linear of kernel based [20]. The Gaussian process classifier [21] and QDA [19] use Gaussian probability [22] while the Naïve Bayes uses conditional probability [21]. Decision tree works by combining the series of related choices and mapping them to an outcome [23]. RFC is a family of ML classification algorithms that consists of large numbers of individual decision trees that operate in an ensemble [24]. Each tree in the RFC makes a prediction and vote, the class with the most vote is the model's prediction. Adaboost is also an ensemble ML but it works in phases. At the first phase a prediction is made by the multiple trees, then the incorrectly classified points are improved progressively in subsequent phases until an optimum result is achieved [25]. ANN was inspired by the neural networks of human brain and has been demonstrated to have good potentials at revealing hidden patterns in biological data [26].

2.3. Data description

The data used for this study have been published in one of our previous studies [18]. Briefly, the data contain basic demographic and lifestyle data including the ethnic region of the participants and their parents, socio-economic status, cardiovascular risk profile and dietary patterns. The original dataset consists of 30 potential risk factors for stroke in Nigeria and Ghana, which were carefully selected based on findings in the literature [27–30], our clinical understanding of stroke risk and empirical evidence from the statistical analysis of significant associations between risk factors and stroke status (i.e. case *vs* control). The statistical analysis helped to identify 16 explanatory variables and also to identify the topmost 11 of these 16. In the analysis, we assessed the predictive power of all the risk factors at both individual level and different combinations [18]. We used logistic regression to estimate odds ratios (ORs) and population-attributable risks (PARs) with 95% CIs. Our analysis identified 16 explanatory variables and using the top 11 were further selected based on their predictive powers as measured by the magnitude. The 16 explanatory variables are: Addition of salt to food at the table during eating, Cardiac Disease, Diabetes Mellitus, Dyslipidemia, Education, Family history of cardiovascular disease (CVD), Hypertension, Income, Low green leafy vegetable consumption, Obesity, Physical Inactivity, Regular meat consumption, Regular sugar consumption, Smoking, Stress, Use of tobacco and waist to hip ratio. Selection of the topmost 11 variables led to the removal of education, income, use of tobacco, regular sugar consumption and family history of CVD (Supplementary material S2). In the current study, we developed two ML models, the first used the entire 16 explanatory variables while the second used only the selected 11 explanatory variables. The selected variables are, however, also supported by existing literature [27–30]. The risk factors for the cases and control group were compared using chi-square and independent t-test, depending on whether they were continuous or categorical variables.

2.4. Classification of stroke data using eleven machine learning techniques

Data preprocessing included data cleaning, conversion and feature selection. Age was recorded in years so it was a continuous variable, but we converted to categorical data of 0 and 1 using age 50 years as a threshold [16,31]. Similarly, income was recorded in dollars and as a continuous variable, but we converted to categorical data of 0 and 1 using a threshold of $100. This conversion is to ensure internal consistency and to ensure fair comparison of the ML methods and logistic regression. All other features were categorical data with 0 representing no and 1 representing yes to a given risk factor. Features were selected based on literature [27–30], and our previous study [15]. Two sets of features were used, the first was all 16 risk factors of stroke mentioned in section 2.3 and the second was the 11 features selected from the 16 mentioned in Supplementary Material S2. The selected features were consequently used for developing the classification models.

After feature selection, the dataset was split into a ratio of 85 to 15 for the training and testing set, respectively. The training set was used to build the model for each of the eleven ML algorithms, while the testing case was used for model evaluation. In the modelling, we experimentally tuned the relevant parameters on the training split in a cross-validation fashion to get the best performance of each of the algorithms. For the final model, the number of neighbors for Nearest Neighbors was 3, the leaf size was 30, the metric was Minkowski and other parameters were set as default. For logistic regression, the penalty was L2 while other parameters were set to default. For Linear and RBF SVM, the maximum iteration was set to 1000 and the penalty was set to L2 while the random state was set to zero.

For Gaussian Process, we used Limited-Memory-Broyden-Fletcher-Goldfarb-Shannon algorithm as the optimizer. QDA and Naive Bayes used the default setting while Decision Tree and RFC used 5 as the maximum depth of the tree and 10 estimators. The number of estimators for Adaboost was set to 50.

ANN used a Multi-Layer Neural Perceptron with 100 hidden layers and the maximum number of iterations was set to 1000. Activation used RELU and Adam was used as the solver while all other parameters used the default settings.

The performance of each model was assessed using sensitivity analysis and the values of sensitivity, specificity and AUC were recorded. The feature importance of each feature for each of the ML was computed to determine the contributions of each feature. Comparative analysis was based on these performance metrics. We also compared the machine models with logistic regression using net reclassification index (NRI) [32]. All eleven algorithms were implemented in Python (version 3.9.7) programming language (Python Software Foundation, Wilmington, DE, USA) running on a Windows operating system (Microsoft Cooperation, Redmond, WA, USA). Some of the libraries used are numpy arrays, pandas and scikit-learn.

3. Results

3.1. Subjects

Computational experiments were performed on 4,236 study participants, comprising 2,118 controls and 2,118 stroke patients. Of the controls, 1,193 (56%) were male while of the stroke group, 1193 (56%) were male (Table 1). The mean age and standard deviation (STD) for the control group were 57.80 (13.70) years and 59.00 (13.80) years for the stroke cases. Most of the control group did not have diabetes mellitus (87%), cardiac disease (95%), family history of cardiovascular disease (71%), or stress (85.55%), but a high proportion of the controls had hypertension (57%), and dyslipidemia (61%). Most of the control group were educated (80%), ate meat (80%), and were older than 50 years (73%) and consumed vegetables at least once a month (82%) but did not add extra salt to their food (94%). About 94% of the stroke patients had hypertension and about 78% had dyslipidemia. The result of the chi-square and t-test comparison show that the risk factors for cases were significantly different from those of the control group (Table 1, all $p < 0.05$), except gender as expected because the data was gender-matched.

Table 1
Descriptive Statistics and Risk Factors for the study population.

Stroke Risk Factors	Control N = 2118	Case N = 2118
Sex		
Female	925(43.7)	925(43.7)
Male	1193(56.3)	1193(56.3)
Age in years (Mean ± SD)	57.8 ± 13.7	59.0 ± 13.8**
<50	579(27.3)	515(24.3)***
≥50	1539(72.7)	1603(75.7)***
Education		
No education	416(19.6)	343(16.2)***
Some education	1698(80.2)	1761(83.1)
Monthly Income		
≤$100	1165(55.0)	894(42.2)
>$100	915(43.2)	1168(55.1)
Physical Inactivity	49(2.3)	98(4.6)***
Current tobacco use	28(1.3)	71(3.4)***
Stressed in the last 2 weeks	297(14.0)	457(21.6)***
Hypertension	1216(57.4)	1999(94.4)***
Diabetes Mellitus (Type 2)	283(13.4)	800(37.8)***
Dyslipidemia	1297(61.2)	1659(78.3)***
Cardiac Disease	110(5.2)	248(11.7)***
Raised waist to hip ratio	1407(66.4)	1584(74.8)***
Family History of CVD	611(28.8)	850(40.1)
Add salt at table	120(5.7)	171(8.1)
Low green leafy vegetable consumption	372(17.6)	572(27.0)
Regular sugar consumption	724(34.2)	611(28.8)
Regular meat consumption	1635(77.2)	1647(77.8)

** = P < 0.001.
*** = P < 0.0001.
Note. Age is expressed as mean and standard deviation and other variables are expressed as N and %.

3.2. Results of classification using eleven machine learning algorithms

Table 2 and Fig. 1 show the results of the eleven classification algorithms using all sixteen and eleven risk factors (features), respectively. The results show that the sixteen features-based classification (maximum AUC of 82.32%) had a better performance than the eleven features-based classification (maximum AUC of 81.17%). The result also shows that ANN had the best performance amongst all the eleven algorithms investigated with AUC of 82.32% while RBF SVM gave the least performance with AUC of 73.6%. ANN had a sensitivity of 70.29% and a specificity of 79.02% when all features were used in the classification. Logistic regression gave AUC of 81.48%.

Table 3 shows the contributions of each of the features in each of the ML model. For brevity, the feature importance of the ML models with the 5 highest AUC were recorded. The result shows that hypertension, vegetable consumption, diabetes, dyslipidemia and use of salt at table time gave the highest contribution to all the models.

We compared all ML algorithms with logistic regression using AUC and NRI. We found that three ML algorithms, ANN (AUC = 82.32), Gaussian process (AUC = 82.30) and Adaboost (AUC = 81.62) outperformed logistic regression (Table 2, Fig. 2), but not significantly. Supplementary Table S3 shows the results of the NRI for the best four ML models using logistic regression as the reference model. The results show that ANN correctly classified 1.71% more control than logistic regression but 1.75% fewer cases than logistic regression, totaling an overall NRI of −0.045% for ANN. Similarly, the Gaussian process correctly classified 3.15% fewer cases than logistic regression and 1.14% fewer control than logistic regression, totaling an overall NRI of −4.30% for Gaussian process. AdaBoost correctly classified 2.12% more cases than logistic regression but 37.89% fewer control than logistic regression, totaling an overall NRI of −35.79% for Adaboost.

Although the AUC of some of the ML techniques performed better than logistic regression, none of the ML models investigated individually performed better than logistic regression when performance is split into the percentage of positive or negative cases. The combination of ANN and AdaBoost (see Table S1) correctly classified 2.11% more cases than logistic regression and 1.71% more control than logistic regression, totaling an overall NRI of 3.82%.

4. Discussion

In this study, we used data from Africa's largest stroke study to develop ML models for stroke prediction. We used sixteen stroke risk factors obtained from 4,236 participants for the full model, and we also analyzed the eleven topmost stroke risk factors and compared their performances. Nearest Neighbors, Linear Support Vector Machine (SVM), Radial Basis Function (RBF) based SVM,

Table 2
Performance of the Classification Algorithms investigated in the study.

Algorithm	Accuracy (%)	Positive Predictive Value (%)	Negative Predictive Value (%)	Sensitivity (%)	Specificity (%)	AUC (%)
All the 16 Features						
Artificial Neural Network	75.16	81.43	69.30	71.23	80.00	82.32
Gaussian Process	74.84	81.31	68.88	70.66	80.00	82.30
AdaBoost	75.16	82.49	68.73	69.8	81.75	81.62
Logistic Regression	75.47	82.61	69.14	70.37	81.75	81.48
Random Forest	71.54	81.95	64.05	62.11	83.16	80.97
Linear SVM	74.53	84.36	67.04	66.1	84.91	80.72
Quadratic Discriminant Analysis	70.44	72.70	67.51	74.36	65.61	78.89
Decision Tree	71.23	81.82	63.71	61.54	83.16	78.77
Naive Bayes	70.28	72.38	67.52	74.64	64.91	77.56
Nearest Neighbors	71.54	75.6	67.11	71.51	71.58	77.16
RBF SVM	65.41	79.11	57.91	50.71	83.51	73.6
Selected 11 Features						
Linear SVM	74.21	83.27	67.04	66.67	83.51	81.17
Gaussian Process	73.74	84.07	66.12	64.67	84.91	81.13
AdaBoost	73.58	83.27	66.21	65.24	83.86	81.01
Neural Net	73.43	82.04	66.48	66.38	82.11	80.93
Logistic Regression	73.74	83.58	66.3	65.24	84.21	80.86
Decision Tree	71.54	83.46	63.61	60.4	85.26	78.5
Random Forest	70.60	76.80	64.85	66.95	75.09	78.35
Quadratic Discriminant Analysis	69.03	70.48	66.92	75.50	61.05	77.84
Naive Bayes	68.71	69.90	66.93	76.07	59.65	77.51
RBF SVM	71.70	80.43	64.79	64.39	80.70	75.07
Nearest Neighbors	67.92	73.79	62.39	64.96	71.58	70.22

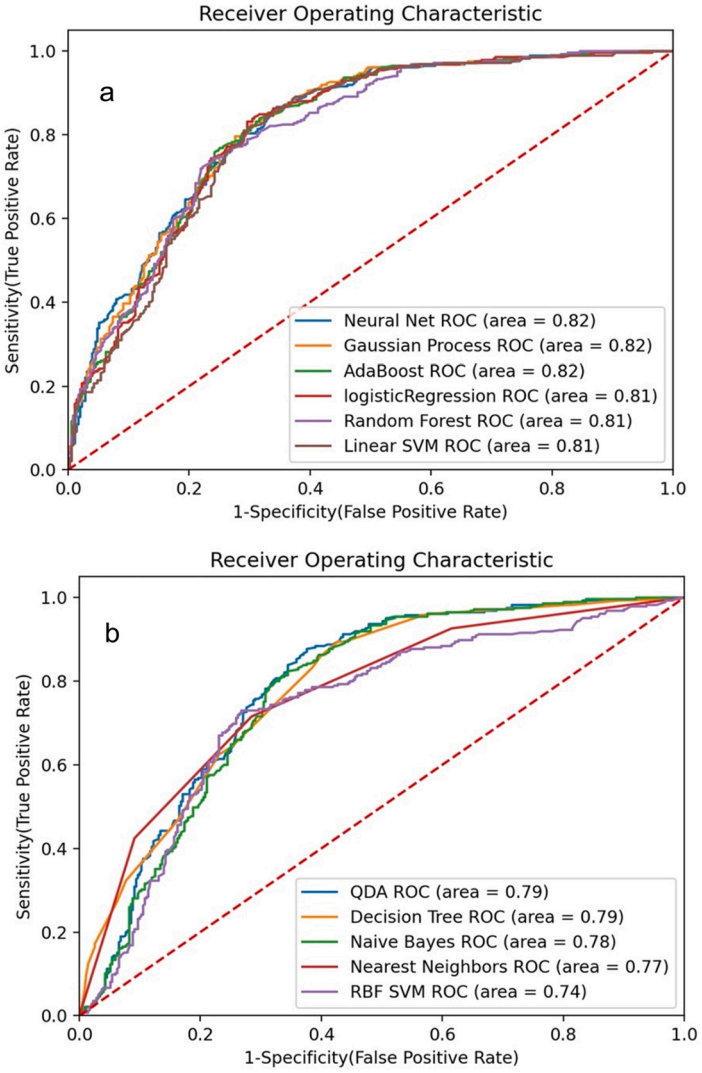

Fig. 1. The Receiver Operating Characteristics (ROC) curves of the eleven algorithms investigated showing their performance at predicting stroke. (a) contains the best 6 algorithms based on AUC and (b) contains the other 5 algorithms.

Gaussian Process, Decision Tree, Random Forest, Multi-Layers Perceptron Artificial Neural Network (ANN), Adaboost, Nave Bayes, Quadratic Discriminant Analysis, and Logistic regression were implemented and compared.

Our findings revealed that the Artificial Neural Network algorithm built with all features performed best, with 72.2% sensitivity, 80.0% specificity, and 82.3% AUC in predicting stroke from risk factors. The eleven feature-based classifiers performed less than all feature-based classifiers combined.

The model performance result shows that the Artificial Neural Network model for stroke prediction is very promising. Our result compares very well with previous studies conducted in high income countries, for example, the study on the Cardiovascular Health Cohort dataset [33], which had an AUC of 77.7% using the SVM classifier, and the study on acute stroke [34], which had an AUC of 85.0% using RFC, and a study on a Chinese cohort which had an AUC of 78.0% [35] using RFC.

Our observation that ANN performed better than the other ML algorithms of this study (Nearest Neighbors, Linear SVM, RBF SVM, Gaussian Process, Decision tree, Random Forest, Adaboost, Naïve Bayes, Quadratic Discriminant Analysis and Logistic regression) is not surprising, since other studies have made similar observations [33–36]. ANN can learn and model non-linear and complex relationships, which is very common in real life. This applies to stroke, in which many relationships between predictor variables and outcome variables are non-linear or complex. Additionally, ANN can model the hidden relationship because it is built the upon human brain which can understand and detect hidden relationships. Another strength of ANN is that it does not impose any constraint on the input variables (e.g. distribution) and it also has a great ability to handle heteroscedasticity and collinearity in input variables [37,38].

Our results show that the predictors of stroke in this sub-Saharan African study, such as hypertension, stress, obesity, age, smoking and heart disease agree with stroke predictors of other studies, although the effect sizes differ [39–42].

This study has many strengths. First, this study can predict the likelihood of immediate occurrence of stroke using risk factors with an AUC of 82% and a positive predictive value of 81.4%. This is the first-ever of such prediction methods, as other prediction techniques foresee the future risk of stroke (usually 5-year, 10-year risk), which may not provide a sense of urgency for primary stroke prevention at the individual level since the risk is futuristic. This study is important for people in Africa where prevention is critical due to limited access to acute and chronic care facilities.

The second major strength of this work is that we have empirically compared eleven ML algorithms (logistic regression, decision tree, SVM and ANN) before determining which one to use. Our

Table 3
Feature Importance Showing the Contribution of Each Feature.

Features	Artificial Neural Network	Gaussian Process	AdaBoost	Logistic Regression	Random Forest
All the 16 Features					
Hypertension	0.1097	0.1252	0.1000	1.0607	0.4258
Low green leafy vegetable consumption	0.0233	0.0220	0.0800	0.3445	0.0528
Diabetes Mellitus	0.0204	0.0248	0.0600	0.4975	0.1115
Dyslipidemia	0.0173	0.0283	0.0600	0.3089	0.0767
Add salt at table	0.0157	0.0082	0.1000	0.2378	0.0382
Age	0.0154	0.0104	0.0800	0.1881	0.0123
Family History of CVD	0.0104	0.0072	0.0800	0.0797	0.0390
Waist to hip ratio	0.0097	0.0107	0.0600	0.1927	0.0747
Stressed in the last 2 weeks	0.0088	0.0094	0.0400	0.1946	0.0189
Cardiac Disease	0.0075	0.0035	0.0400	0.1802	0.0284
Regular sugar consumption	0.0069	0.0031	0.0400	0.0186	0.0302
Regular meat consumption	0.0066	0.0164	0.0600	0.1871	0.0156
Physical inactivity	0.0028	0.0057	0.0400	0.1157	0.0082
Income	0.0006	0.0019	0.0600	0.2026	0.0514
Use of tobacco	0.0013	0.0013	0.0400	0.1539	0.0085
Education	0.0013	0.0072	0.0600	0.1120	0.0079
features	Artificial Neural Network	Gaussian Process	AdaBoost	Logistic Regression	Random Forest
Selected 11 Features					
Hypertension	0.1292	0.1289	0.1000	1.0275	0.222
Low green leafy vegetable consumption	0.0236	0.0302	0.0800	0.3371	0.0621
Dyslipidemia	0.0223	0.0274	0.0800	0.3116	0.1711
Diabetes Mellitus	0.0176	0.022	0.0800	0.4878	0.1622
Waist to hip ratio	0.0119	0.0189	0.1000	0.1633	0.1246
Stressed in the last 2 weeks	0.0088	0.011	0.0600	0.2305	0.0394
Regular meat consumption	0.0057	0.0044	0.0400	0.2055	0.0272
Add salt at table	0.0050	0.0082	0.2800	0.2327	0.0580
Cardiac Disease	0.0041	0.0035	0.0600	0.1732	0.0803
Use of tobacco	0.0025	0.0013	0.0600	0.1658	0.0198
Physical inactivity	0.0016	0.0022	0.0600	0.0958	0.0333

results show that ANN outperformed the other ten algorithms. Non-linear relationships, complex relationships, hidden relationships, collinearity and even heteroskedasticity are all handled by ANN. The AUC for Logistic regression was 81%, for decision tree it was 79% and for Linear SVM it was 80%. ANN, on the other hand, provided an AUC of 82.32%.

The third major strength of our study is the fact that the performance analysis of ML algorithms was not limited to AUC comparison. Other similar studies compared ML algorithms like SVM, ANN Adaboost, etc. [43–47] using AUC only, but we used both AUC and NRI. NRI is known to outperform AUC in performance evaluation [32,48–50]. Our results that ML algorithms perform better than logistic regression agrees with those of other studies outside of Africa [43–47]. Although, the difference in performance is marginal, the result is promising and could improve if the sample size is increased.

The fourth major strength of this study is the uniqueness of the participants. We studied Africans, from 15 different sites in Nigeria and Ghana [24]. To the best of our knowledge, this is the first study to develop an ANN model for accurate prediction of stroke using risk factors in indigenous Africans.

Our study also has a limitation. The recent technological advancement and the promising results from deep learning suggest that deep learning could perform better than our model. Our future plan is to include polygenic risk scores and novel proteomic biomarkers and then use that to build a deep learning model. We hope such an approach will improve the AUC and other predictive characteristics. We also planned to validate our model on data from other African countries to make it generalizable. Also, as part of our future plan, we will convert the model to a mobile phone app and make it available to end users. We believe that such app has the potential to improve stroke risk assessment and diagnosis.

5. Conclusion

In this unique study, we evaluated stroke patients and stroke-free controls from sub-Saharan Africa in the SIREN study. We successfully developed eleven machine learning models: Nearest Neighbors, Linear SVM, RBF SVM, Gaussian Process, Decision tree, Random Forest, Artificial Neural Network, Adaboost, Naïve Bayes, Quadratic Discriminant Analysis and Logistic regression. Our results show that Artificial Neural Network outperformed the other ten algorithms investigated. Most importantly, our results show that risk factors alone have a high predictive value in predicting stroke, even before the manifestation of symptoms. We established the fact that machine learning approaches predict stroke occurrence by employing the major stroke risk factors in Sub-Saharan Africa. Machine learning algorithms, especially artificial neural network, is recommended to enhance Afrocentric stroke prediction models for stroke risk factor quantification and control in sub-Saharan Africa.

Author contributions

MOO, SS and BO are senior authors. BSA, DE, SS, GO and MO conceptualized and designed the study. BSA, SS, and OMA curated, analyzed and interpreted the data. BSA, SS, DE and MO wrote the first draft of the manuscript. All authors edited and approved the final draft. MO, BO and SS contributed to the funding acquisition.

Ethics approval and consent to participate

The Stroke Investigative Research and Educational Network (SIREN) study is a multi-centre study, and Institutional Review Board (IRB) at all study sites provided ethical approval for the study. The overall coordinating IRB for the SIREN study was the University of Ibadan/University College Hospital Ibadan, Nigeria (IRB Approval No.: UI/EC/13/0105). This research conformed to the

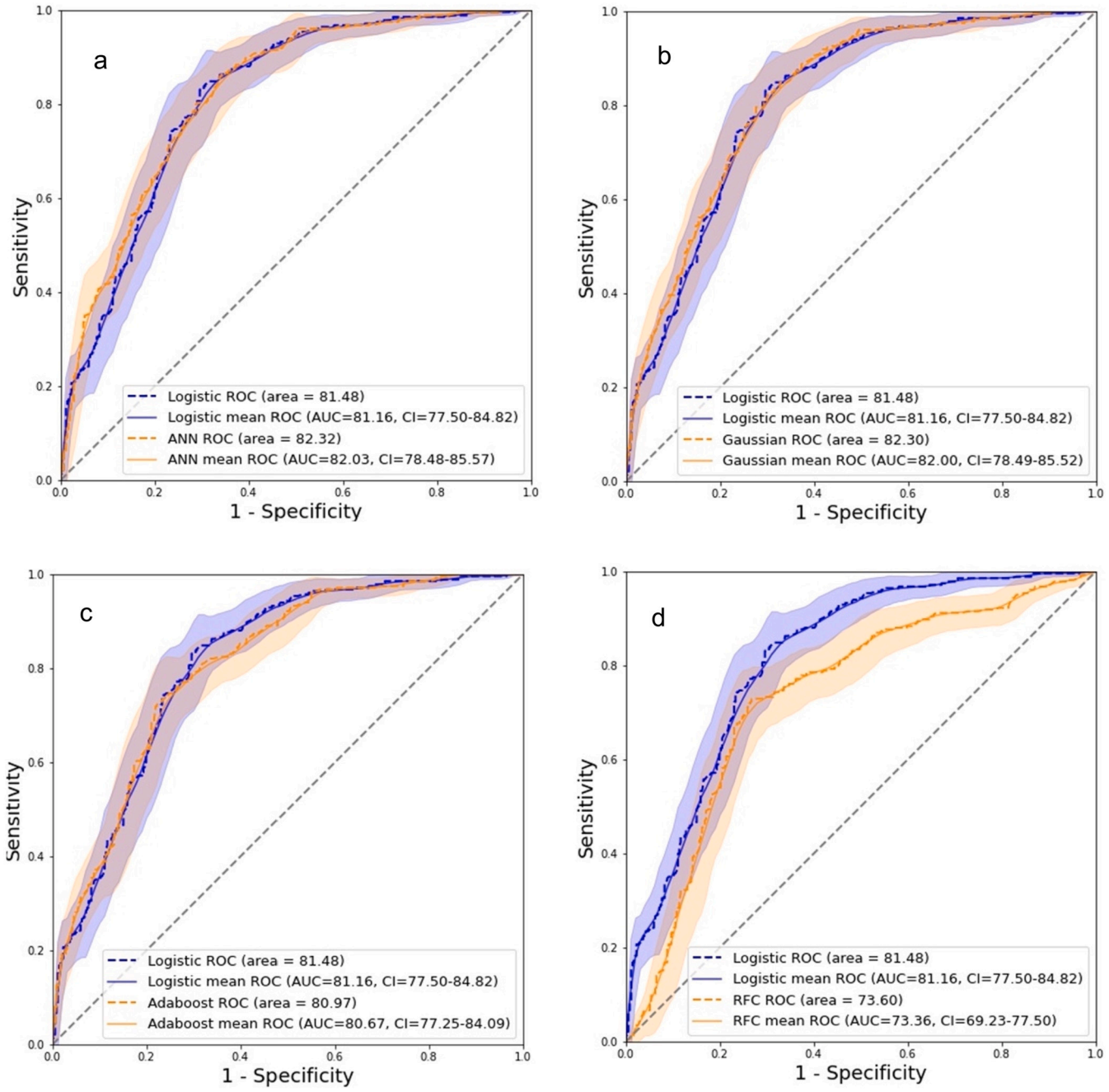

Fig. 2. ROC curves of the best four performing ML algorithms versus Logistic Regression showing the AUC and Confidence interval. (a) Displays the ROC of ANN vs. Logistic regression with confidence interval. (b) Shows the ROC curves of Gaussian process vs. Logistic regression with confidence intervals. (c) Displays the ROC curves of AdaBoost vs. Logistic regression with confidence interval. (d) Shows the ROC of Random Forest vs. Logistic regression with confidence interval. The blue color shows the logistic regression, while the orange color shows ML models. The shaded portion is the confidence plot of the confidence interval after bootstrapping for 1000 times.

principles of the Helsinki Declaration and all respondents provided written informed consent before participating in the study.

Funding

This study was supported by the University of Chicago Pritzker School of Medicine, the University of Chicago Center for Global Health, the National Institutes of Health NIH/NINDS R25NS080949, SIREN (U54HG007479), SIBS Genomics (R01NS107900), and SIBS Gen (R01NS107900-02S1). The SIREN investigators are further supported by NIH grants ARISES (R01NS115944-01) H3Africa CVD Supplement (3U24HG009780-03S5), CaNVAS (1R01NS114045-01), Sub-Saharan Africa Conference on Stroke (SSACS) 1R13NS115395-01A1 Training Africans to Lead and Execute Neurological Trials & Studies (TALENTS) D43TW012030 and ELSI grant 1U01HG010273. Benjamin Aribisala was supported by the Institute of International Education and THE FULBRIGHT PROGRAM (PS00322782) as a visiting scholar at the University of Chicago. Godwin Ogbole is supported by grant number 2021-240505 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

Declaration of competing interest

We declare no competing interests.

Acknowledgement

We would also like to thank Oluremi Adeyemo and Mayowa Adeyemi from Lagos State University for their assistance with the protocol and data analysis, and the entire SIREN team for their support.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neuri.2025.100216.

References

[1] T.A. Gaziano, Economic burden and the cost-effectiveness of treatment of cardiovascular diseases in Africa, Heart 94 (2) (2008) 140–144.
[2] E.S. Donkor, Stroke in the century: a snapshot of the burden, epidemiology, and quality of life, Stroke Res. Treat. (2018) 2018.
[3] M. Kelly-Hayes, Influence of age and health behaviors on stroke risk: lessons from longitudinal studies, J. Am. Geriatr. Soc. 58 (2010) S325–S328.
[4] R. Khatib, Y.A. Arevalo, M.A. Berendsen, S. Prabhakaran, M.D. Huffman, Presentation, evaluation, management, and outcomes of acute stroke in low- and middle-income countries: a systematic review and meta-analysis, Neuroepidemiology 51 (1–2) (2018) 104–112.
[5] W. Johnson, O. Onuma, M. Owolabi, S. Sachdev, Stroke: a global response is needed, Bull. World Health Organ. 94 (9) (2016) 634.
[6] L.A. Coke, C.D. Himmelfarb, Guideline on the primary prevention of cardiovascular disease: let's get it into practice!, J. Cardiovasc. Nurs. 34 (4) (2019) 285–288.
[7] M.O. Owolabi, D. Adu, M. Ramsay, B. Ovbiagele, Understanding the rise in cardiovascular diseases in Africa: harmonising H3Africa genomic epidemiological teams and tools: cardiovascular topic, Cardiovasc. J. Afr. 25 (3) (2014) 134–136.
[8] I. Graham, D. Atar, K. Borch-Johnsen, G. Boysen, G. Burell, R. Cifkova, et al., European guidelines on cardiovascular disease prevention in clinical practice: executive summary: fourth joint task force of the European society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts), Eur. Heart J. 28 (19) (2007) 2375–2414.
[9] O. Akpa, F.S. Sarfo, M. Owolabi, A. Akpalu, K. Wahab, R. Obiako, et al., A novel afrocentric stroke risk assessment score: models from the siren study, J. Stroke Cerebrovasc. Dis. 30 (10) (2021) 106003.
[10] K.L. Chien, T.C. Su, H.C. Hsu, W.T. Chang, P.C. Chen, F.C. Sung, et al., Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan, Stroke 41 (9) (2010) 1858–1864.
[11] A. Akpalu, F.S. Sarfo, B. Ovbiagele, R. Akinyemi, M. Gebregziabher, R. Obiako, Phenotyping stroke in sub-Saharan Africa: stroke investigative research and education network (SIREN) phenomics protocol, Neuroepidemiology (2015).
[12] O. Popoola, B. Ovbiagele, O. Arulogun, J. Akinyemi, R. Akinyemi, E. Uvere, et al., African rigorous innovative stroke epidemiological surveillance: protocol for a community-based mobile-health study, Neuroepidemiology 56 (1) (2022) 17–24.
[13] M. Nichols, O. Arulogun, A. Singh, O. Olorunsogbon, E. Uvere, S. Melkiam, et al., Community insight on the ethical, legal, and social implications of stroke genomic and biobanking research in sub-Saharan Africa, J. Stroke Cerebrovasc. Dis. 31 (4) (2022) 106356.
[14] A.P. Kengne, B.M. Mayosi, Modifiable stroke risk factors in Africa: lessons from SIREN, Lancet Glob. Health 6 (4) (2018) e363–e364.
[15] M.O. Owolabi, F. Sarfo, R. Akinyemi, M. Gebregziabher, O. Akpa, A. Akpalu, et al., Dominant modifiable risk factors for stroke in Ghana and Nigeria (SIREN): a case-control study, Lancet Glob. Health (2018) e436–e446.
[16] F.S. Sarfo, B. Ovbiagele, M. Gebregziabher, K. Wahab, R. Akinyemi, A. Akpalu, et al., Stroke among young West Africans: evidence from the SIREN (stroke investigative research and educational network) large multisite case–control study, Stroke 49 (5) (2018) 1116–1122.
[17] F. Sarfo, M. Gebregziabher, B. Ovbiagele, R. Akinyemi, L. Owolabi, R. Obiako, et al., Multilingual validation of the questionnaire for verifying stroke-free status in West Africa, Stroke 47 (1) (2016) 167–172.
[18] S. Suzuki, T. Yamashita, T. Sakama, T. Arita, N.N. Yagi, Comparison of risk models for mortality and cardiovascular events between machine learning and conventional logistic regression analysis, PLoS ONE 14 (2019).
[19] M.Z.N. Al-Dabagh, M.M. Alhabib, F. Al-Mukhtar, Face recognition system based on kernel discriminant analysis, k-nearest neighbor and support vector machine, Int. J. Res. Eng. 5 (3) (2018) 335–338.
[20] V.K. Chauhan, K. Dahiya, A. Sharma, Problem formulations and solvers in linear SVM: a review, Artif. Intell. Rev. 52 (2) (2019) 803–855.
[21] X. Feng, S. Li, C. Yuan, P. Zeng, Y. Sun, Prediction of slope stability using naive Bayes classifier, KSCE J. Civ. Eng. 22 (3) (2018) 941–950.
[22] R.B. Gramacy, Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences, Chapman and Hall/CRC, 2020.
[23] N. Kappelhof, L. Ramos, M. Kappelhof, H.J. van Os, V. Chalos, K. van Kranendonk, et al., Evolutionary algorithms and decision trees for predicting poor outcome after endovascular treatment for acute ischemic stroke, Comput. Biol. Med. 133 (2021) 104414.
[24] C. Fernandez-Lozano, P. Hervella, V. Mato-Abad, M. Rodríguez-Yáñez, S. Suárez-Garaboa, I. López-Dequidt, et al., Random forest-based prediction of stroke outcome, Sci. Rep. 11 (1) (2021) 1–12.
[25] Y. Zhang, M. Ni, C. Zhang, S. Liang, S. Fang, R. Li, et al., Research and application of AdaBoost algorithm based on SVM, in: IEEE, 2019, pp. 662–666.
[26] C.C. Peng, S.H. Wang, S.J. Liu, Y.K. Yang, B.H. Liao, Artificial neural network application to the stroke prediction, in: IEEE, 2020, pp. 130–133.
[27] D. Woo, M. Haverbusch, P. Sekar, B. Kissela, J. Khoury, A. Schneider, et al., Effect of untreated hypertension on hemorrhagic stroke, Stroke 35 (7) (2004) 1703–1708.
[28] H. Jerrgensen, H. Nakayama, J. Reith, H. Raaschou, T.S. Olsen, Stroke recurrence: predictors, severity, and prognosis. The Copenhagen stroke study, Neurology 48 (4) (1997) 891–895.
[29] B. Boden-Albala, R.L. Sacco, Lifestyle factors and stroke risk: exercise, alcohol, diet, obesity, smoking, drug use, and stress, Curr. Atheroscl. Rep. 2 (2) (2000) 160–166.
[30] R. Marinig, G. Lip, N. Fiotti, C. Giansante, D. Lane, Age as a Risk Factor for Stroke in Atrial Fibrillation Patients, 2010.
[31] J. Putaala, Analysis of 1008 consecutive patients aged 15 to 49 with first-ever ischemic stroke: the Helsinki young stroke registry, Stroke 40 (2009) 1195–1203.
[32] M.J. Pencina, E.W. Steyerberg, R.B. D'Agostino Sr, Net reclassification index at event rate: properties and relationships, Stat. Med. 36 (28) (2017) 4455–4467.
[33] A. Khosla, Y. Cao, C.C.Y. Lin, H.K. Chiu, J. Hu, H. Lee, An Integrated Machine Learning Approach to Stroke Prediction, 2010, pp. 183–192.
[34] H. JoonNyung, G.Y. Jihoon, P. Hyungjong, D.K. Young, S.N. Hyo, H.H. Ji, Machine learning–based model for prediction of outcomes in acute stroke, Stroke (2019).
[35] Y. Wu, Y. Fang, Stroke prediction with machine learning methods among older Chinese, Environ. Res. Public Health 17 (6) (2020).
[36] M. Rajora, M. Rathod, N.S. Naik, Stroke prediction using machine learning in a distributed environment, in: Distributed Computing and Internet Technology: ICDCIT 2021, in: Lecture Notes in Computer Science, 2021, p. 12582.
[37] Y. Gu, Z. Bao, Z. Rui, Complex lithofacies identification using improved probabilistic neural networks, Petrophys., SPWLA J. Form. Eval. Reserv. Descr. 59 (02) (2018) 245–267.
[38] C. Paul, G.K. Vishwakarma, Back propagation neural networks and multiple regressions in the case of heteroskedasticity, Commun. Stat., Simul. Comput. 46 (9) (2017) 6772–6789.
[39] M.S. Elkind, R.L. Sacco, Stroke Risk Factors and Stroke Prevention, © 1998 by Thieme Medical Publishers, Inc., 1998, pp. 429–440.
[40] T. Omae, Stroke risk factors and stroke prevention, J. Stroke Cerebrovasc. Dis. 2 (1) (1992) 45–46.
[41] K. Wang, Z. Lv, P. Xu, Y. Cui, X. Zang, D. Zhang, et al., Factors related to the risk of stroke in the population with type 2 diabetes: a protocol for systematic review and meta-analysis, Medicine 101 (3) (2022).
[42] K.M. Rexrode, T.E. Madsen, A.Y. Yu, C. Carcel, J.H. Lichtman, E.C. Miller, The impact of sex and gender on stroke, Circ. Res. 130 (4) (2022) 512–528.
[43] K.J. Ottenbacher, P.M. Smith, S.B. Illig, R.T. Linn, R.C. Fiedler, C.V. Granger, Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke, J. Clin. Epidemiol. 54 (11) (2001) 1159–1165.
[44] S.K. Jang, J.Y. Chang, J.S. Lee, E.J. Lee, Y.H. Kim, J.H. Han, et al., Reliability and clinical utility of machine learning to predict stroke prognosis: comparison with logistic regression, J. Stroke 22 (3) (2020) 403.
[45] S. Qu, M. Zhou, S. Jiao, Z. Zhang, K. Xue, J. Long, et al., Optimizing acute stroke outcome prediction models: comparison of generalized regression neural networks and logistic regressions, PLoS ONE 17 (5) (2022) e0267747.
[46] Y. Liang, Q. Li, P. Chen, L. Xu, J. Li, Comparative study of back propagation artificial neural networks and logistic regression model in predicting poor prognosis after acute ischemic stroke, Open Med. 14 (1) (2019) 324–330.
[47] J. Heo, J.G. Yoon, H. Park, Y.D. Kim, H.S. Nam, J.H. Heo, Machine learning–based model for prediction of outcomes in acute stroke, Stroke 50 (5) (2019) 1263–1265.
[48] K.F. Kerr, Z. Wang, H. Janes, R.L. McClelland, B.M. Psaty, M.S. Pepe, Net reclassification indices for evaluating risk-prediction instruments: a critical review, Epidemiology 25 (1) (2014) 114.
[49] M.S. Pepe, J. Fan, Z. Feng, T. Gerds, J. Hilden, The net reclassification index (NRI): a misleading measure of prediction improvement even with independent test data sets, Stat. Biosci. 7 (2) (2015) 282–295.
[50] M.J. Leening, M.M. Vedder, J.C. Witteman, M.J. Pencina, E.W. Steyerberg, Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide, Ann. Intern. Med. 160 (2) (2014) 122–131.