National College of Ireland

Predicting Stroke at Adulthood Using Machine Learning Techniques

MSc Research Project
Data Analytics

Sudhir Clinton Manjunath
Student ID: x20247818

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sudhir Clinton Manjunath	
Student ID:	x20247818	
Programme:	M.Sc. Data Analytics	**Year:** 2022
Module:	Research Project	
Supervisor:	Vladimir Milosavljevic	
Submission Due Date:	1st February 2023	
Project Title:	Predicting Stroke at Adulthood Using Machine Learning Techniques	
Word Count:	5963	**Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sudhir Clinton Manjunath
Date:	31st January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Stroke at Adulthood Using Machine Learning Techniques

Sudhir Clinton Manjunath
x20247818

Abstract

Early identification and primary prevention of stroke is essential because it frequently causes death or severe disability. Hence, the targeted area for this research are the adults aged from 25-64 years. It is crucial to examine the relationships between the risk factors in patients' medical records and understand how each one contributes to the prediction of strokes, and this is performed by utilising feature selection technique. Since most of the electronic medical records are heavily imbalanced with majority of negative cases, it is hard to train the machine learning models (ML) because they will identify the negative cases most of the time. Hence different sampling techniques are employed to balance the positive and negative cases. The research aims to increase the certainty and dependability of the doctor's diagnosis. Consequently, a stroke prediction model that integrates ensemble learning in addition to current ML techniques together with the aforementioned approaches is built. Stacking classifier (SC) achieved the best prediction result with 76% recall and an area under the receiver operating curve (AUROC) score of 0.7078, this will aid medical professionals in detecting stroke in its initial stages using less computing time and effort.

1 Introduction

Stroke is a cerebrovascular condition in which the arteries that supply the brain with nutrients and oxygen are damaged, cutting off the blood supply to certain parts of the brain. As a result, the brain's blood cells are completely destroyed (Rajora et al. 2021). When the brain cells are damaged, two of the activities that the brain region controls are lost: muscle co-ordination and memory (Karthik et al. 2020). In 2019, stroke was the third most frequent cause of mortality and

disability (5.7% of all Disability-adjusted life years), while it remained the second leading cause of mortality globally (11.6% of all fatalities). To calculate the overall burden of disease, one uses disability-adjusted life years (DALYs). Total stroke incidence increased by 70% from 1990 to 2019, prevalent strokes increased by 85%, fatalities from stroke increased by 43%, and disability-adjusted life years (DALYs) increased by 32%. As people become older, their chance of having a stroke increases, and those over 50 are particularly at risk (Feigin 2021). According to the presented statistics, it is clear that thousands of lives are lost annually as a result of stroke, this places a heavy burden on the families of the victims as well as the nation as a whole owing to the population decline. One of the key determinants of a nation's economy, progress, and innovation is its people. In essence, a country's strength rests on its citizens, hence the goal of this research is to identify stroke in advance in order to save many lives each year. To accomplish this, a stroke prediction model is constructed utilizing two innovative approaches, one is age filtering to target a specific age group and the other is backward elimination to choose significant features leading to stroke.

Machine learning (ML) is a powerful technology in the healthcare industry that provides tailored therapeutic care for stroke patients. It makes predictions quickly and accurately (Sirsat et al. 2020). With its innovative applications, ML is helping to improve the healthcare industry. Due to required practices like electronic medical records, medical systems have already embraced big data tools for next generation data analytics. The value that ML tools will provide to this process is expected to increase. These enhance the level of automation and rational decision-making in treating patients and public healthcare systems (Javaid et al. 2022). To persuade clinicians that a specific ML-based algorithm is the most appropriate and effective tool for disease prediction and diagnosis, which can facilitate their standard practice without harming patients, explanations regarding patterns of features that the ML model has learned and the justification for the chosen model producing better results than other models are required. Patients will benefit from the outcomes that have been explained since they will better understand ML predictions and feel trusted and satisfied (Rasheed et al. 2022).

Stroke is a horrible disease that can be life threatening and typically strikes people aged 65 years and older, but nowadays it is occurring to those younger than that due to poor diet. If a stroke can be predicted in its early stages, it can be prevented (Singh et al. 2019). This leads to the research question: How well can machine learning predict stroke among the adults? Clinical experts will benefit

from this research's ability to predict strokes more accurately and quickly in their early phases. Furthermore, it will increase peoples' confidence in medical diagnoses.

2 Literature Review

This section compiles the most pertinent earlier works over the last six years that have been proposed and used for stroke prediction. The experimental works of the authors are subjected to critical investigation. After carefully weighing the advantages and disadvantages of these research works, a new approach is applied in a way that keeps the advantages of the prior research works while also addressing their disadvantages.

2.1 Stroke Predictions with Data Preprocessing and Balancing

The authors developed an integrated machine learning technique for predicting cerebral stroke for clinical diagnosis based on physiological data with class imbalance and missing values. On the unbalanced dataset, strokes were predicted using a deep neural network (DNN)-based automatic hyperparameter optimization (AutoHPO). Undersampling was employed in AutoHPO to reduce the dataset's imbalance ratio. The results demonstrate that the suggested technique significantly reduced the false negative rate to 19% while keeping a generally high accuracy, proving a successful reduction in the probability of wrong stroke diagnosis. The sensitivity is 67.4% and specificity is 33.1%, according to the results (Liu et al. 2019). The undersamplingclustering-oversampling algorithm (UCO algorithm), which was created by the authors to handle the unbalanced data, makes use of random undersampling (RUS), oversampling, and clustering approaches. The research made use of the MIT laboratory's Medical Information Mart for Intensive Care III (MIMIC-III) database. The findings showed that UCO(120) + Random Forest (RF) had the highest predictive performance in terms of precision, AUC, and accuracy of 70%, as well as 75% recall. The authors came to the conclusion that it is unknown how well UCO performs in relation to other stroke patient data sets because there is a shortage of data on stroke patients (Wang et al. 2021). The researchers established a framework to determine the most effective stroke prediction method in a Chinese hypertensive population using machine learning models such stepwise logistic regression (SLR), Logistic Regression (LR), RF, and Extreme Gradient Boosting (XGB). The dataset was gathered from the China

Stroke Primary Prevention Trial (CSPPT), and demographic characteristic data were employed both with and without laboratory variables as a technique of analysis. The whole nested case-control (NCC) dataset was used as an external test dataset to further assess the model's efficacy. Data balancing techniques including synthetic minority oversampling (SMOTE) and RUS were used to handle the uneven training set. According to the results, the RUS-applied RF model with laboratory variables, which employed AUC and sensitivity as its main metrics, had the best model performance with a score of 64% and 72%, respectively (Huang et al. 2022).

The study employed ML to pinpoint the major indicators of death for patients with embolic stroke in the intensive care unit (ICU). The authors used a multiple imputation methodology called multiple imputation by chained equation (MICE) to address missing data during the data preprocessing stage. For categorical predictors with several categories, dummy variables were generated. Data were scaled, centered, and underwent Box-Cox transformation since some of the models benefit from predictors that are on a same scale and have less skewness. Data were scaled, centered, and BoxCox transformed for some models because they benefit from having predictors with a consistent scale and less skewness. For tree and rule-based models, no data transformation was done. To address the class imbalance, the synthetic minority oversampling technique (SMOTE) was utilized. All models were then evaluated using three repetitions of 10-fold cross-validation. As per the results RF achieved the highest AUROC of 80% and 83% in the internal and external validation phase respectively (Liu et al. 2022). The authors created nine models to categorize the participants' individual stroke risk levels using data from the 2017 China National Stroke Screening and Intervention Program. The researchers created the ML models in a way that can enhance the current screening technique and mitigate the effects of unidentified values. The train set was built using 70% of the experimental dataset, and the test set was created using the remaining 30%. Then, to address the issue of data imbalance, SMOTE oversampling and RUS were implemented on the training set. Test sets with missing data were created to imitate the scenarios that happen in screening practice and were inspired by the concept of developing test sets with blocked regions for testing image recognition models. The findings indicate that, for both test sets, the boosting model with decision trees had the highest recall and random forest achieved the highest precision (Li et al. 2019).

2.2 Feature Selection for Stroke

The researchers presented a feature selection and dimensionality reduction approach to identify signs of heart disease to forecast whether an individual has heart disease. Cleveland, Hungarian, and a mixture of the two datasets from the UCI repository were used for the analysis. These datasets contain a total of 74 features. When system resources must be considered, using all features is not practical. As a result, the study used dimensionality reduction techniques to enhance the outcomes of the raw data. Principal component analysis (PCA) was applied as a dimensionality reduction technique, while chi-square (CHI) was implemented for feature selection. Three feature groups were chosen from the available 74 features, and those features delivered the best results. Cholesterol, maximal heart rate, chest pain, features associated with ST depression, and heart vessels were some of the significant anatomical and physiological parameters that were chosen from the analysis. The CHI-PCA with RF has the highest accuracy (99%) and recall (98%), according to the results (Gárate-Escamila et al. 2020). A range of statistical methods were used in the research, including PCA to select the most significant risk variables for stroke. This concept is similar to that of the previous paper and was used to establish the most critical parameters for stroke prediction. RUS was used to decrease the negative effects of this imbalance because the dataset is highly uneven with regard to the incidence of stroke. The following algorithms were used to predict stroke: NN, convolutional neural network (CNN), DT, and RF. According to the results, NN with a combination of 4 features—age, hypertension, average blood glucose level, and heart disease obtained through PCA outperformed the other algorithms with a recall of 71%, precision of 78%, and 19% miss rate. The researchers concluded that they could not more effectively train their neural networks because the dataset was not particularly complex (Dev et al. 2022).

For atrial fibrillation (AF)-related stroke and to assess the prediction accuracy and feature significance of ML models, a model that predicts early neurological deterioration (END) was constructed. In order to use 25% of the dataset as a test set for the final evaluation of the model's performance, it was randomly divided into groups according to the END stratification. The rest, 75% of the dataset was utilized as a train set for leave-one-out cross-validation training procedures and hyperparameter determination. The stratified random sampling technique was applied during the data splitting procedure. Selecting the top-k ranked features that improved the performance of entire model was done via recursive feature

elimination, which eliminates features below a predetermined threshold value. With an AUROC of 77% and a recall of 38%, the LightGBM model displayed the highest results (Kim et al. 2021). A strategy for classifying the severity of a stroke using symmetric gait features and cross-validated recursive feature elimination (RFECV) was presented in the research. Along with the general gait features, which did not entirely capture the patients' walking characteristics, symmetric gait features indicating the proportion between the right and left side values were employed as inputs. The best subset for separating the old people and stroke groups according to severity using RFECV was determined by the authors using four different ML techniques.RFE takes away traits until the appropriate features are left, then chooses the most pertinent subset from that subset. Feature selection algorithms for symmetric and general features were separately used in order to validate the efficacy of applying a feature selection method to categorize stroke patients. The RF-RFECV method produced the best classification result, with 95.32% sensitivity, 97% specificity, and 96% accuracy using an RF classifier built from symmetric features (Sung et al. 2022).

2.3 Ensemble Learning for Stroke Prediction

The authors proposed a model that can estimate the probability of 4 unique types of strokes while accounting for all aspects of risk variables. In the first stage, the input dataset is altered using FCM clustering, and the altered dataset is then utilized in second phase to estimate the risk of stroke by adopting an ensemble learning based classifier, which uses a number of base models to enhance performance. On the training subset, 10-fold cross validation was used to create the training and validation sets. The results based on accuracy, F1 and F2 scores show that the proposed model works well overall. According to the researchers, patient data now needs to be manually entered with a specific order of features, either as a Matlab or Excel file. This situation additionally makes the model less user-friendly by increasing the user's effort (Akyel 2022). In order to balance the data, the Random Over Sampling (ROS) technique was applied in the research, which provided a machine learning approach to accurately diagnose strokes with unbalanced data. Multiple classifiers were implemented and evaluated along with a VC that was built using RF and Support Vector Machine (SVM). Data preprocessing was done, including filling in missing values, label encoding, and dividing the data into train and test sets. To improve the outcomes, cross-validation and hyperparameter tuning were applied to each model. For a variety of hyperparameter settings, GridSearchCV was used to test the machine learning models. In order to

pick the optimal model, the best accuracy was obtained for each combination of hyperparameters. After that, machine learning classifiers underwent another evaluation, and with the use of cross-validation, the effectiveness of machine learning models was calculated. As per the results before optimization SVM and VC outperformed all other algorithms, whereas after optimization SVM achieved the highest score of 99% in terms of accuracy, recall, F-1 and precision (Biswas et al. 2022). To categorize the time since stroke (TSS) as less than or more than 4.5 h, the researchers presented an automated machine learning method. The stroke lesions were first precisely segmented from DWI and FLAIR images using a cross modal convolutional neural network. The features were then taken from DWI and FLAIR in accordance with the segmentation regions of interest (ROI). The attributes were then used to identify TSS by feeding them into machine learning models such as LR, ET, GBDT, RF and SVM. These five algorithms were used to create a voting classifier (VC), which provided the final prediction while also enhancing the classification robustness. According to the findings, VC outperformed human-derived DWI-FLAIR mismatch in terms of specificity (84%) and accuracy (80%) (Zhu et al. 2021).

With the aid of machine learning, the researchers developed and assessed a number of models to produce a strong framework for predicting the risk of stroke incidence. Utilizing MLP, LR, RepTree, NB, J48, RF, and KNN, ensemble approaches like stacking and majority voting were implemented. The dataset was gathered from Kaggle, and during data preprocessing, SMOTE was used as a resampling approach to solve the class imbalance between cases with and without strokes. The dataset's participant count dropped from 5110 to 3254 as a result of the authors focusing on participants who were above 18 years old and the dataset had no missing values. Therefore, no further data imputation or cleaning was carried out. It is evident from the results that the stacking model which was applied to the selected base classifiers was the most reliable, achieving an accuracy of 98% with an AUC score of 98.9%, f1 score of 97.4%, and 97.4% of recall and precision (Dritsas & Trigka 2022). Using only a limited collection of clinical data from the first International Stroke Trial's hyperacute-phase, stacking ensemble ML was used in the study to construct a reliable model for 6-month death prediction in patients with acute ischemic stroke (AIS) who did not receive reperfusion therapy. The stacking classifier's base learners included KNN, SVM, XGB, RF, NB, and LR, while ANN served as the meta learner. Grid and randomized search were employed to fine-tune the parameters of these algorithms and stratified fivefold cross-validation was performed to assess them. Results revealed that the SC out-

performed all the individual algorithms in terms of AUROC, accuracy, sensitivity, and specificity with a score of 78.3%, 71.6%, 72.3%, and 70.9% respectively (Hwangbo et al. 2022).

The Literature Review section examines a number of research methodologies and tactics that have been proposed for employing deep learning and machine learning models to predict stroke and assesses their merits. Since the data will be consistent and simple to process, preprocessing has been found to be necessary. It is established that balancing the datasets provide positive findings since it removes the biasing impact. This was notably true for the research publications that used the SMOTE technique and had favourable prediction outcomes. Furthermore, it is evident that ensemble learning, cross validation, and feature selection all help the model perform better at predicting strokes.

3 Methodology for Creating a Stroke Prediction Model

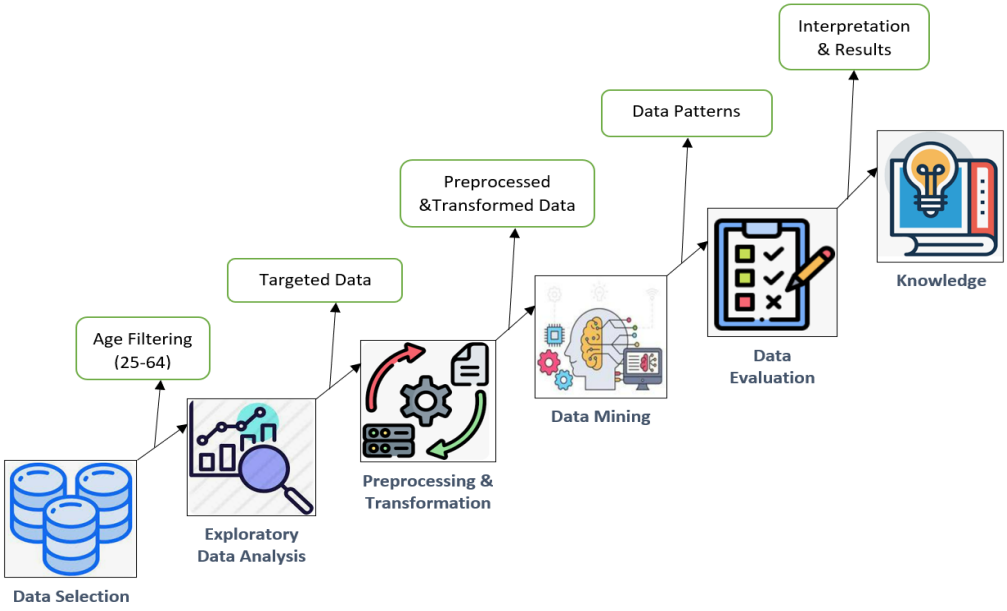

Figure 1: Modified KDD methodology for Stroke Prediction

This research project uses the knowledge discovery in databases (KDD) methodology, a way of evaluating data from databases that incorporates programming

and analytical approaches in order to extract relevant and useful information. After deciding on the KDD technique, it is then slightly changed to fit it in the requirements of this project. Figure 1, portrays the modified KDD methodology implemented in this project.

4 Project Design and Implementation

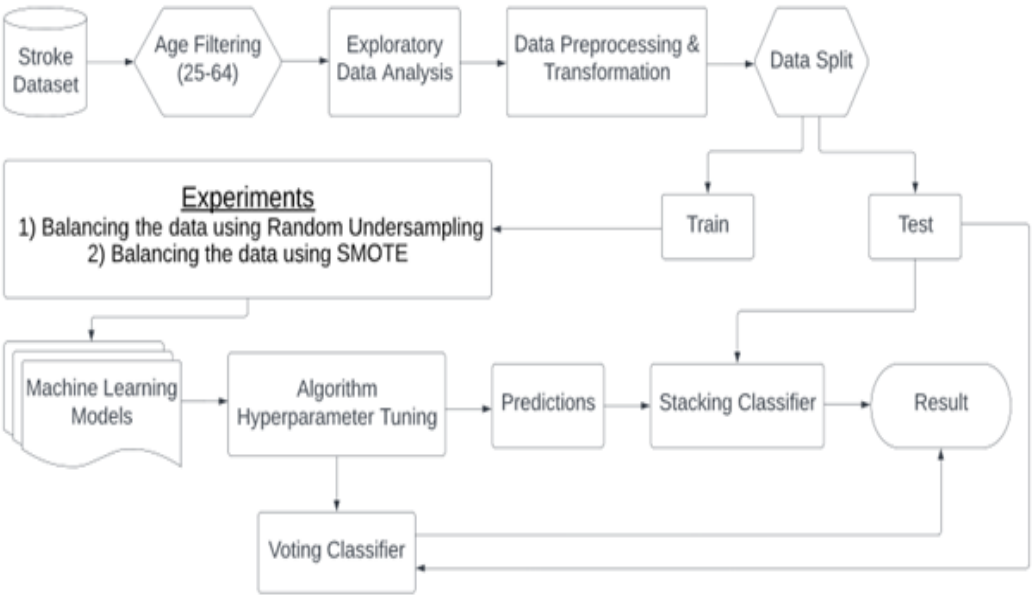

Figure 2: Workflow of the Process

Figure 2 shows the overall process flow of the research project using the modified KDD methodology, in which the stroke dataset is age filtered (25-64). The filtered dataset is then preprocessed and transformed so that the machine learning algorithms can comprehend it. The dataset is then partitioned using stratified sampling into two portions: train and test. RUS and SMOTE are two experiments that are run independently on the training set. Each algorithm's parameters are fine-tuned to maximize performance. The VC is built after the models' parameters are fine-tuned, the test set that is kept aside is used on the VC for evaluation and the VC produces the final prediction result. Once the parameters of the models are hyper tuned, the predictions from those models are fed into the SC as a training set. After the SC is trained using the outputs of the algorithms, it is tested using the

separately allocated test set, and the result of this test determines how effectively the model can identify stroke events.

4.1 Data Selection of Stroke Patients

The dataset [1] used for ML model building to predict stroke is obtained from Kaggle website and it is an open source public dataset in the form of comma separated values (CSV) file. It consists of 319795 instances with 18 featured attributes. This dataset is highly imbalanced which makes it tough for the ML model to learn the patterns and predict appropriately. Here, the targeted audiences for stroke prediction are mainly adults, that is, the age groups from 25-64. Hence the age filtering is done accordingly and as a result the number of records in the dataset decreased to 176711 which is a big sample to train the models.

4.2 Exploratory Data Analysis

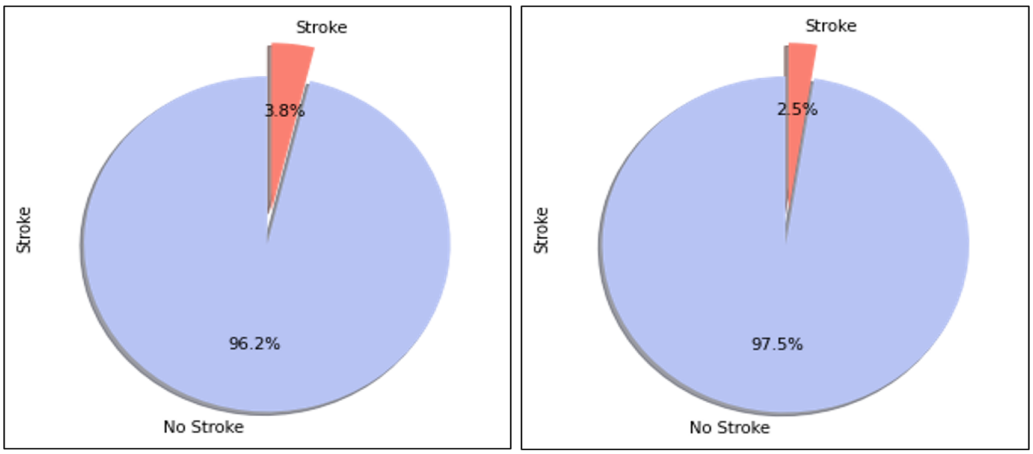

Figure 3: Proportion of stroke after age filtering

Figure 3 consists of pie charts that represents the proportion of patients with and without stroke. It can be observed from the first pie chart (left) that there 96.2% people without stroke and 3.8% people who have suffered from stroke. The second pie chart (right) is a result of age filtering, which shows that among the targeted audiences aged from 25 to 64, only 2.5% people have stroke which has

[1] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

further reduced from the first chart, which shows that out of 176711 records the minority class has only 4403 records. This indicates that the dataset is highly imbalanced and can lead the ML models in not identifying the minority classes accurately. Hence to resolve this issue, balancing techniques such as RUS and SMOTE are applied.

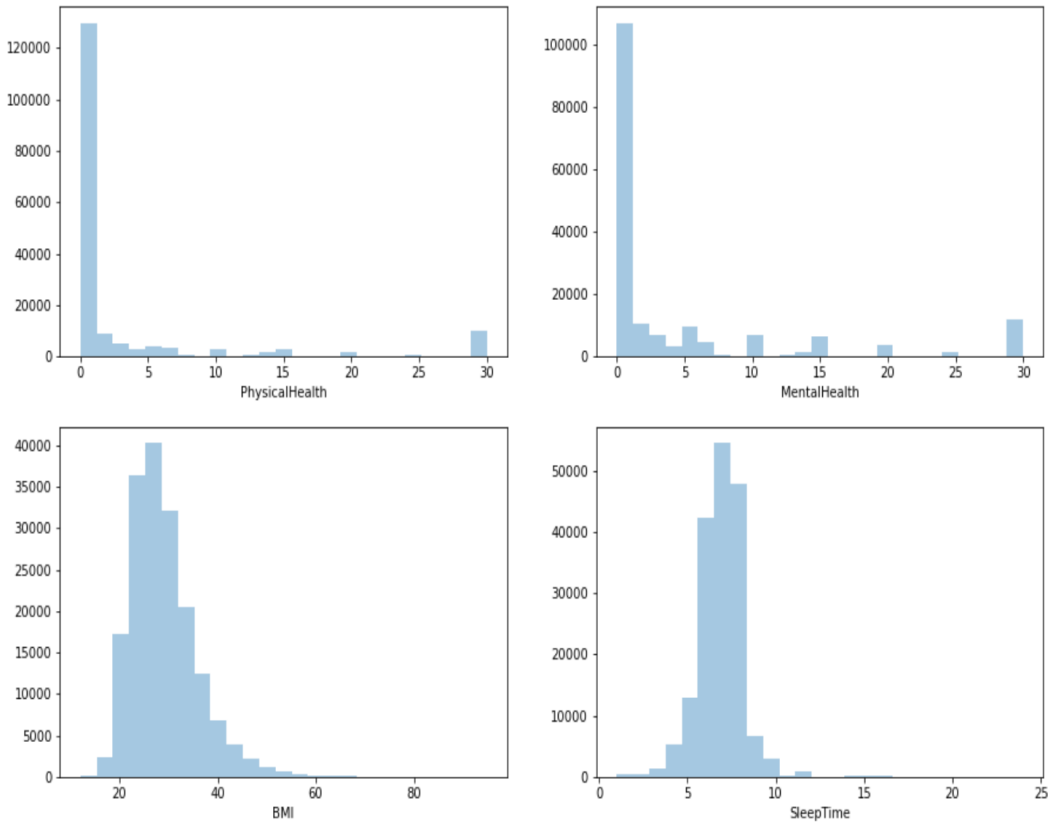

Figure 4: Distribution of numerical features

Figure 4 displays the histogram of numerical features in the dataset. It is observed that the variables PhysicalHealth and MentalHealth have most of the values in 0 and 30, few values are unevenly distributed between them, which shows that they are not normally distributed. The feature BMI is slightly skewed on the right and the attribute SleepTime is almost close to a normal distribution with some outliers above 10 hours. To overcome this situation, outliers treatment and feature scaling

is implemented to make the distribution of the data normal.

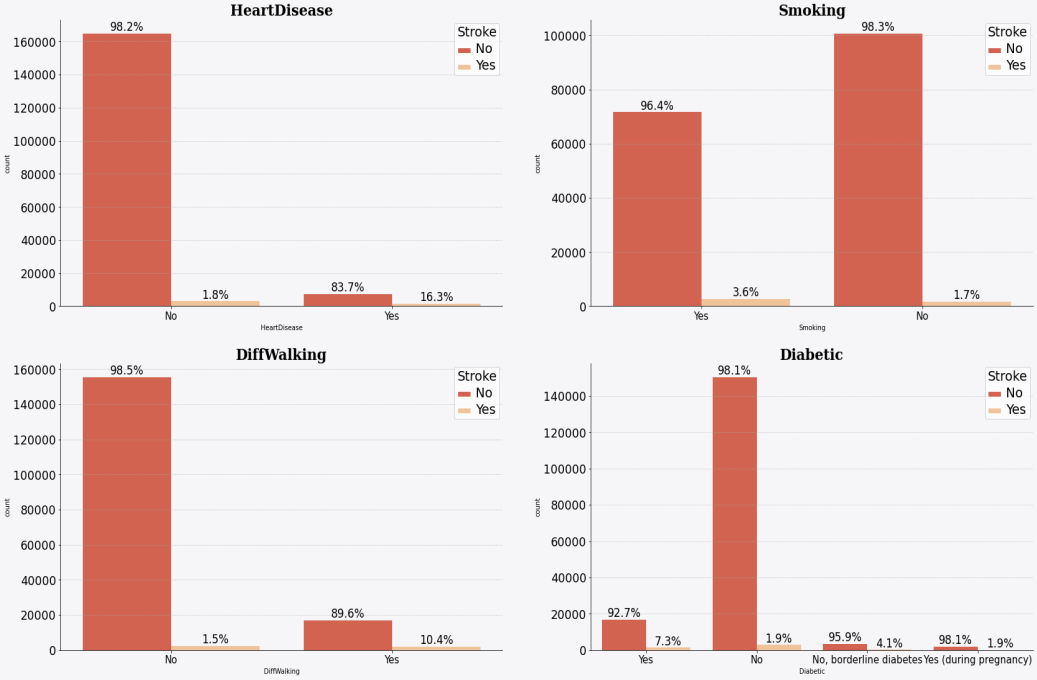

Figure 5: Bivariate analysis of categorical features with Stroke

Figure 5 displays the bivariate analysis of categorical features with respect to stroke. The clear distinction between those with and without heart disease demonstrates the significance of this feature in determining whether or not a person has suffered a stroke. People who smoke frequently have a noticeable increase in the risk of stroke. The difficulty to walk is a significant element that contributes to a stroke situation, just as it was with the feature heart disease. It can be seen that people who have difficulty in walking have almost 10% higher risk of getting a stroke. It can also be noticed that people with diabetes are more inclined towards getting a stroke.

Figure 6 shows the boxplots that highlights the outliers of several attributes in the dataset. It is clearly evident that variables BMI, PhysicalHealth, MentalHealth and SleepTime have the most number of outliers which may cause the ML models to not perform well in identifying and predicting the stroke instances. To solve this problem, the outliers are removed in the data preprocessing stage.

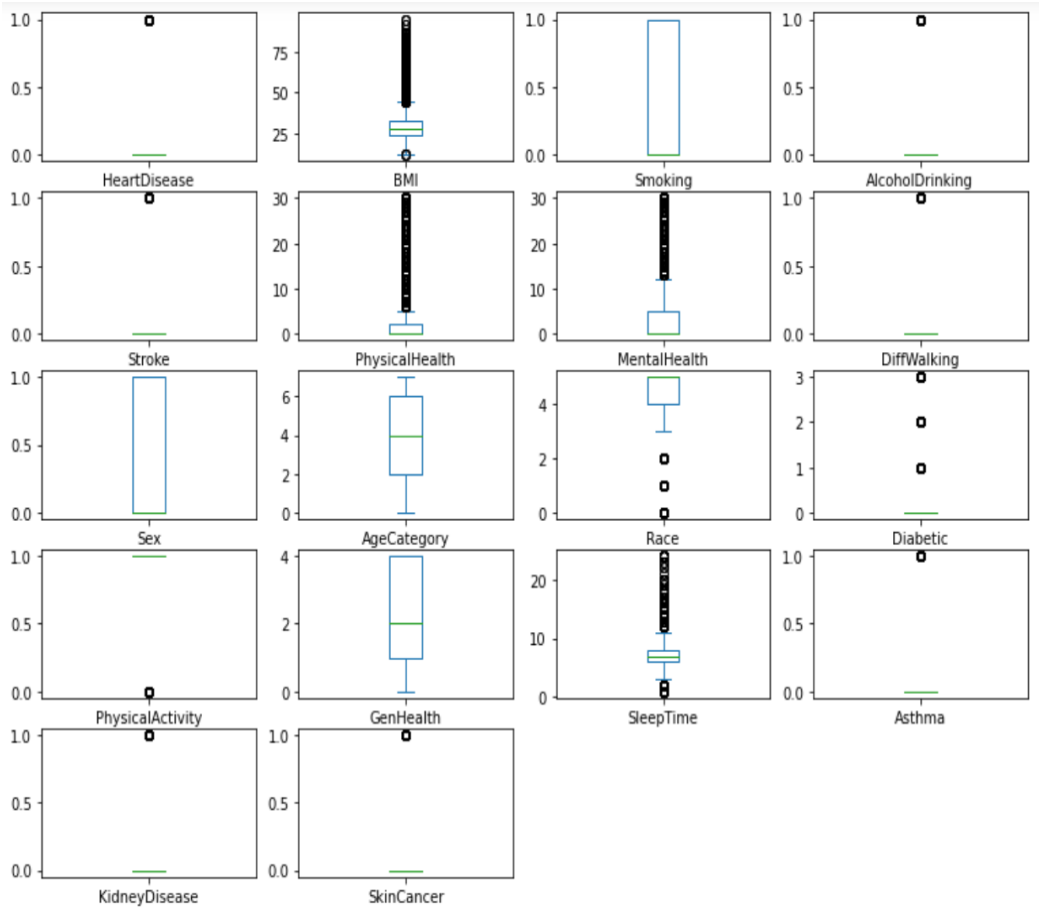

Figure 6: Outliers in the features

4.3 Data Preprocessing and Transformation

Checked for duplicated records and removed them using the drop_duplicates() function. Then checked for missing/null values and found that the data is complete. The categorical data is converted into a numeric form using the function LabelEncoder() to make them machine-readable. Outlier detection and removal is an important step in the pre-processing stage, hence through boxplots the outliers of different features are detected and visualized. In the dataset, 4 attributes have many outliers and those were BMI, PhysicalHealth, MentalHealth, and SleepTime. These are the anomalous observations that distort the data distribution and, if ignored, would result in the creation of less accurate models. In order

to remove the outliers, a statistical method known as interquartile range (IQR) is used. Statistics based outlier detection techniques imply that the regular data points would show up in high the probability areas of a stochastic model, whereas outliers would emerge in low probability areas. For the purpose of identifying outliers, the IQR technique computes the lower bound (first quartile/25th percentile) and the upper bound (third quartile/75th percentile) and these values are removed from the dataset.

To convert numerical features with diverse ranges into the same scale, feature scaling is implemented. It enhances numerical input stability, prevents learning process failure, shortens the learning period for the predictive model, and can significantly enhance the model performance (Nkikabahizi et al. 2022). The scaling method utilized is standardization, where the attribute's mean is set to zero and the resulting distribution has a standard deviation of one.

Feature selection using a statistical method is performed, in order to find the most pertinent and contributing factors for stroke. A Generalized linear model (GLM) is built for selecting the important features and the family type is set as binomial which is an indication to the GLM that the target variable is binary. Backward elimination technique is used to remove the insignificant variables, this is done by checking the significance level (p-value) of all the variables. The variable whose p-value is more than 0.05 is removed and the GLM is built using the remaining variables. This process is repeated until all the insignificant variables are eliminated. This method is similar to RFE, but here the variables with less significance are manually removed. Multicollinearity lowers the independent variables' statistical significance and hence for the purpose of determining the level of multicollinearity, the variance inflation factor (VIF) is examined. VIF above 5 means that the independent variables are highly correlated (multicollinearity exists), therefore the variable with a VIF greater than 5 is removed and the final GLM consists of 10 significant features for stroke prediction and those are: AgeCategory, PhysicalActivity, Smoking, PhysicalHealth, Diabetic, Asthma, AlcoholDrinking, DiffWalking, HeatDisease, and KidneyDisease.

The data is split into 80% for training and 20% for testing. To remove sampling bias in the test set, stratified sampling is used which involves selecting data at random from the total population so that each potential sample has an equal likelihood of occurring. It enables the creation of a test set that most accurately represents the population as a whole and an unbiased data is acquired as a result,

which is used for further processing and to develop ML models. Then the train set is experimented on the 2 balancing methods (RUS and SMOTE).

4.4 Data Mining Pattern for Stroke

The ML algorithms implemented for the research work are RF, Bernoulli Naive Bayes (BNB), Stochastic Gradient Descent (SGD), Adaptive Boosting (AB), LR and SVM. These algorithms were chosen because they were employed in the majority of related research papers and delivered successful results. In the field of healthcare, these algorithms are often employed (Verma & Verma 2022). The best parameters for tuning the algorithms are selected based on the implementation on randomized search using 10-fold cross validation.

RF: using a collection of decision trees, it can examine complex relationships between clinical traits and offer very accurate classification. It also calculates the significance of the classification's variables (Ooka et al. 2021). On tuning the algorithm, the optimum results are obtained when n_estimators is '70', min_samples_split is '4', max_features is 'sqrt', max_depth is '5', criterion is 'gini', bootstrap is 'true'. If False, each tree is constructed using the entire dataset.

BNB: when the data is in a binary pattern with the output label being present or absent, it is particularly advantageous to utilize this algorithm because it focuses mainly on looking for binary vector characteristics. The alpha value is set to 15 which is a parameter for additive smoothing, and the algorithm increases the recall value by 1% after tuning the parameter.

SGD: It is employed to lower the error rates in huge datasets and to improve the sensitivity, convergence rate, and execution speed of the system (Deepa et al. 2021). The best results are obtained after fine-tuning the algorithm when penalty is 'l2', loss is set as 'log', learning_rate is 'optimal', eta0 is '100', alpha is '0.01'.

AB: this classifier's primary objective is to increase the weight of uncategorized points and reduce the weight of categorized ones. To put it another way, incorrectly categorized instances have their weights increased for the upcoming rounds, whereas successfully classified examples have their weights reduced (Ogunseye et al. 2022). Even after tuning the parameters, where n_estimators is '50' and learning_rate is '1', the algorithm produces the same result.

LR: it is employed when an event's occurrence is the primary focus of the research approach. It is frequently employed in research in the health sciences because it is particularly suited for models that incorporate disease state (healthy or diseased) and decision-making (Boateng & Abaye 2019). The recall value increases by 1% by tuning the parameters where solver is 'newton-cg', max_iter is '150', and C value is '0.1'.

SVM: it has the ability to analyse huge datasets, notably those having binary classes utilized in classification and regression procedures. Clinical data set classification is regarded as a standard SVM approach (Khalaf et al. 2019). When the gamma value is set as '0.01' and C value as '100', the recall value decreases by 1% and the specificity increases by 1%, hence the base version of the algorithm is used.

VC: since it incorporates the predictions of various models, it delivers good overall outcomes than other base models. It compiles the results of any classified vote and predicts the output class based on a big majority of votes (Kumari et al. 2021). Based on the individual results of the above algorithms in terms of recall, AUROC, and specificity, this classifier is built using AB, SGD, and LR.

SC: it combines the abilities of multiple highly effective models to produce predictions that outperform any individual model in the ensemble on a classification problem. It overcomes the limitations of a single classifier by utilizing diverse classifiers (Lu & Uddin 2022). The algorithms used to build the VC are utilised for building this classifier, where AB and SGD are the base learners, and LR is the meta-learner.

4.5 Evaluation of the Stroke Prediction Model

The following metrics are considered while assessing the ML model because they were previously utilized by other research studies to gauge how accurately the model predicted the occurrence of strokes. In the medical field, detecting and diagnosing the condition depends greatly on recall, specificity, and AUROC.

Recall: it is also known as sensitivity. When comparing all the actual positive cases, it determines the percentage of precisely predicted positive instances. The

recall score is calculated using Eq (1).

$$Recall = TruePositive/(TruePositive + FalseNegative) \quad (1)$$

TruePositive is when an individual has stroke and the ML model correctly identified as stroke, whereas FalseNegative is when an individual has stroke, but the ML model incorrectly identified as no stroke.

Specificity: it determines the percentage of wrongly predicted negative cases. It works in a manner similar to recall, but only for negative classes. The specificity score is calculated using Equation (2).

$$Specificity = TrueNegative/(TrueNegative + FalsePositive) \quad (2)$$

TrueNegative is when a person does not have stroke and the ML model correctly identified as no stroke whereas FalsePositive is when a person does not suffer from stroke, but the ML model incorrectly identified as stroke.

AUROC: it is the models' capacity to accurately differentiate between stroke patients and non-stroke instances. It has a scale of 0 to 1, with 1 being a perfect model with 100% differentiating ability.

Based on these evaluation metrics, when picking the best model, recall is given the top priority. This is because it is at the utmost importance to identify the patient who is likely to suffer from a stroke rather than a patient who will not, because if informed then the patient will take immediate actions to prevent the occurrence of stroke and can be saved. However, AUROC is also important, because it is able to correctly identify the stroke and non-stroke patients.

5 Experiments

The following two sampling techniques are used for experimenting on the aforementioned eight ML models. A total of 16 models are built as a result.

5.1 Random Undersampling

The majority dataset is randomly reduced in size until it is the same size as the minority dataset. It randomly chooses and eliminates samples from the majority class, thereby lowering the proportion of instances in the majority class in the training set and discarding some essential information (Huang et al. 2022).

5.2 SMOTE

Increased data in the minority class is achieved by its use. The SMOTE algorithm's fundamental concept is to examine samples from the minority class and create new samples using those samples as a basis. It can avoid over-fitting to some extent because it is not merely replicating data from minority classes (Li et al. 2019).

6 Results

Table 1: Results of all the ML models under RUS and SMOTE

Sr No.	Sampling Technique	ML Model	Recall (%)	AUROC	Specificity (%)
1	Random Undersampling	Random Forest	69	0.7044	72
2	Random Undersampling	BernoulliNB	55	0.6805	82
3	Random Undersampling	Stochastic Gradient Descent	72	0.6955	67
4	Random Undersampling	AdaBoost	66	0.7028	75
5	Random Undersampling	Logistic Regression	67	0.7037	74
6	Random Undersampling	Support Vector Machine	70	0.6938	69
7	Random Undersampling	Voting Classifier	67	0.7018	73
8	Random Undersampling	Stacking Classifier	73	0.6975	66
9	SMOTE	Random Forest	71	0.6997	69
10	SMOTE	BernoulliNB	60	0.6816	76
11	SMOTE	Stochastic Gradient Descent	70	0.6929	69
12	SMOTE	AdaBoost	68	0.7063	73
13	SMOTE	Logistic Regression	68	0.7076	73
14	SMOTE	Support Vector Machine	68	0.6925	70
15	SMOTE	Voting Classifier	67	0.7046	74
16	SMOTE	Stacking Classifier	**76**	**0.7078**	66

Table 1 displays the performance of sixteen classifiers under RUS and SMOTE. It is observed that the same algorithm produces different results when implemented with 2 different sampling techniques, which implies that the right balancing technique for the train data must be selected based on the size and other attributes of the dataset. Out of the eight algorithms, five of them (BNB, AB, LR, VC,

and SC) have better stroke prediction results when implemented with the SMOTE technique, whereas three of them (RF, SGD, and SVM) have better predictions of stroke instances when RUS is used. This shows that majority of the algorithms implemented here, provide the desired results when equipped with SMOTE, and this implies that SMOTE is the most preferred sampling technique for this dataset. Section 2.3 emphasizes the importance of ensemble learning models for stroke predictions, especially SC being the most preferred and reliable model. From table 1, it is evident that SC is better than VC, and outperforms every other classifier used in this research with the highest recall of 76% and an AUROC score of 0.7078, which makes it suitable for predicting stroke cases among the targeted audience.

7 Discussion

This research mainly focuses on predicting stroke among adults aged from 25-64 years because this is the phase where most of them are working, married and have children. This means that they are under constant stress and family burdens which can lead them to depression or any other health issues. This is the perfect age range to predict stroke, so that appropriate health measures can be taken in advance and lead a happy, risk-free, and a healthy life when they grow older. One of the main actions taken is selecting the important features that contribute to stroke by implementing the backward elimination technique in which the insignificant features are manually removed. This is the first research to use such an approach in stroke prediction as per section 2.

It is clear from table 1, that SC is the best model which is able to accurately predict patients with stroke at rate of 76%, which is crucial in the health industry. It is also observed that SC outperforms all other ML models regardless of the balancing methods employed. RUS discards some pertinent information. This technique is not particularly effective because minority class' records are very low when compared to those of the majority class. However, SMOTE adds information by generating a synthetic instance at a point in feature space between two examples that is randomly selected and utilizing a randomly selected neighbor as its neighbor, which is not just copying the samples. This relates to a more reliable technique. This is justified with higher recall and AUROC scores. Additionally, it is noted that the specificity of the SC produces good results when compared to the previous research works in table 2.

Table 2: Comparison between this research and previous works

Work	ML Model	Sampling Technique	Recall (%)	AUROC	Specificity (%)
(Liu et al. 2019)	DNN-based AutoHPO	Random Undersampling	67.4	-	33.1
(Wang et al. 2021)	UCO(120) + RF	Random Undersampling & Oversampling	75.59	0.6977	63.95
(Huang et al. 2022)	Random Forest	Random Undersampling	72	0.64	52.8
This research	Stacking Classifier	SMOTE	**76**	**0.7078**	**66**

Based on the importance of SC in section 2.3, the comparison between previous related works that have not implemented any ensemble models (such as stacking or voting) and this research project is presented in table 2. When all their models are compared to the SC which is implemented in this research, it is clear that SC has the highest performance in terms of predicting the stroke positives and stroke negatives with 76% and 66% respectively, along with the model's ability to distinguish between people with and without stroke having a score of 0.7078 and outperforms the previous work. This proves that when the data is balanced using SMOTE and an ensemble technique is used, the predictions are much more effective.

8 Conclusion and FutureWorks

Since the damage caused by strokes cannot be undone, the only treatment for them is prevention. After carefully examining previous research articles on this topic, it is clarified that the SMOTE balancing technique yields the desired outcomes and that using ensemble learning techniques, particularly the stacking approach, improves the model's performance in predicting stroke (see section 2). Accordingly, the same methodology is used. Therefore, this research presents a solution by developing a stroke prediction model utilizing ensemble learning together with other ML techniques, so that people can take preventative measures to lower their chance of having a stroke, and this provides an answer to the research question (refer section 1). The research will benefit adults especially from the ages of 25 to 64, who will develop confidence in the healthcare system. Additionally, it will help doctors identify stroke at an early stage and diagnose patients within the targeted age group more quickly and efficiently through less computational work.

Due to time constraints, other sampling methods could not be used in this study, otherwise, the efficiency of various sampling strategies could be determined. Additionally, since Americans make up the dataset used, the study is more appropriate for them. The traits of an American will differ from those of other nationalities, hence a dataset with a mixed population will be employed in the future.

Acknowledgement

I thank my professor Vladimir Milosavljevic for his efforts in constantly helping and clarifying my doubts which helped me to complete this research project. I thank my parents and friends for their support.

References

Akyel, A. (2022), 'Accurate estimation of stroke risk with fuzzy clustering and ensemble learning methods', *Biomedical Signal Processing and Control* **77**, 103764–103774.

Biswas, N., Uddin, K. M. M., Rikta, S. T. & Dey, S. K. (2022), 'A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach', *Healthcare Analytics* **2**, 100116–100129.

Boateng, E. Y. & Abaye, D. A. (2019), 'A review of the logistic regression model with emphasis on medical research', *Journal of Data Analysis and Information Processing* **07**, 190–207.

Deepa, N., Prabadevi, B., Maddikunta, P. K., Gadekallu, T. R., Baker, T., Khan, M. A. & Tariq, U. (2021), 'An ai-based intelligent system for healthcare analysis using ridge-adaline stochastic gradient descent classifier', *The Journal of Supercomputing* **77**, 1998–2017.

Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B. & John, D. (2022), 'A predictive analytics approach for stroke prediction using machine learning and neural networks', *Healthcare Analytics* **2**, 100032–100040.

Dritsas, E. & Trigka, M. (2022), 'Stroke risk prediction with machine learning techniques', *Sensors* **22**, 4670–4683.

Feigin, V. L., S. B. A. J. C. O. . M. G. A. R. C. J. L. (2021), 'Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019', *The Lancet Neurology* **20**, 795–820.

Gárate-Escamila, A. K., Hassani, A. H. E. & Andrès, E. (2020), 'Classification models for heart disease prediction using feature selection and pca', *Informatics in Medicine Unlocked* **19**, 100330–100340.

Huang, X., Cao, T., Chen, L., Li, J., Tan, Z., Xu, B., Xu, R., Song, Y., Zhou, Z., Wang, Z., Wei, Y., Zhang, Y., Li, J., Huo, Y., Qin, X., Wu, Y., Wang, X., Wang, H., Cheng, X., Xu, X. & Liu, L. (2022), 'Novel insights on establishing machine learning-based stroke prediction models among hypertensive adults', *Frontiers in Cardiovascular Medicine* **9**, 1–11.

Hwangbo, L., Kang, Y. J., Kwon, H., Lee, J. I., Cho, H.-J., Ko, J.-K., Sung, S. M. & Lee, T. H. (2022), 'Stacking ensemble learning model to predict 6-month mortality in ischemic stroke patients', *Scientific Reports* **12**, 17389–17397.

Javaid, M., Haleem, A., Singh, R. P., Suman, R. & Rab, S. (2022), 'Significance of machine learning in healthcare: Features, pillars and applications', *International Journal of Intelligent Networks* **3**, 58–73.

Karthik, R., Menaka, R., Johnson, A. & Anand, S. (2020), 'Neuroimaging and deep learning for brain stroke detection - a review of recent advancements and future prospects', *Computer Methods and Programs in Biomedicine* **197**, 105728–105745.

Khalaf, M., Hussain, A. J., Alafandi, O., Al-Jumeily, D., Alloghani, M., Alsaadi, M., Dawood, O. A. & Abd, D. H. (2019), An application of using support vector machine based on classification technique for predicting medical data sets, *in* 'International Conference on Intelligent Computing', Springer, pp. 580–591.

Kim, S.-H., Jeon, E.-T., Yu, S., Oh, K., Kim, C. K., Song, T.-J., Kim, Y.-J., Heo, S. H., Park, K.-Y., Kim, J.-M., Park, J.-H., Choi, J. C., Park, M.-S., Kim, J.-T., Choi, K.-H., Hwang, Y. H., Kim, B. J., Chung, J.-W., Bang, O. Y., Kim, G., Seo, W.-K. & Jung, J.-M. (2021), 'Interpretable machine learning for early neurological deterioration prediction in atrial fibrillation-related stroke', *Scientific Reports* **11**, 20610–20618.

Kumari, S., Kumar, D. & Mittal, M. (2021), 'An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier', *International Journal of Cognitive Computing in Engineering* **2**, 40–46.

Li, X., Bian, D., Yu, J., Li, M. & Zhao, D. (2019), 'Using machine learning models to improve stroke risk level classification methods of china national stroke screening', *BMC Medical Informatics and Decision Making* **19**, 261–267.

Liu, T., Fan, W. & Wu, C. (2019), 'A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset', *Artificial Intelligence in Medicine* **101**, 101723–101732.

Liu, W., Ma, W., Bai, N., Li, C., Liu, K., Yang, J., Zhang, S., Zhu, K., Zhou, Q., Liu, H., Guo, J. & Li, L. (2022), 'Identification of key predictors of hospital mortality in critically ill patients with embolic stroke using machine learning', *Bioscience Reports* **42**, 1–18.

Lu, H. & Uddin, S. (2022), 'Explainable stacking-based model for predicting hospital readmission for diabetic patients', *Information* **13**, 436–449.

Nkikabahizi, C., Cheruiyot, W. & Kibe, A. (2022), 'Chaining zscore and feature scaling methods to improve neural networks for classification', *Applied Soft Computing* **123**, 108908–108916.

Ogunseye, E. O., Adenusi, C. A., Nwanakwaugwu, A. C., Ajagbe, S. A. & Akinola, S. O. (2022), 'Predictive analysis of mental health conditions using adaboost algorithm', *ParadigmPlus* **3**, 11–26.

Ooka, T., Johno, H., Nakamoto, K., Yoda, Y., Yokomichi, H. & Yamagata, Z. (2021), 'Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in japan', *BMJ Nutrition, Prevention Health* **4**, 140–148.

Rajora, M., Rathod, M. & Naik, N. S. (2021), Stroke prediction using machine learning in a distributed environment, *in* 'International Conference on Distributed Computing and Internet Technology', Springer, pp. 238–252.

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A. & Qadir, J. (2022), 'Explainable, trustworthy, and ethical machine learning for healthcare: A survey', *Computers in Biology and Medicine* **149**, 106043–106065.

Singh, M. S., Choudhary, P. & Thongam, K. (2019), A comparative analysis for various stroke prediction techniques, *in* 'International Conference on Computer Vision and Image Processing', Springer, pp. 98–106.

Sirsat, M. S., Fermé, E. & Câmara, J. (2020), 'Machine learning for brain stroke: A review', *Journal of Stroke and Cerebrovascular Diseases* **29**, 105162.

Sung, J., Han, S., Park, H., Hwang, S., Lee, S. J., Park, J. W. & Youn, I. (2022), 'Classification of stroke severity using clinically relevant symmetric gait features based on recursive feature elimination with cross-validation', *IEEE Access* **10**, 119437–119447.

Verma, V. K. & Verma, S. (2022), 'Machine learning applications in healthcare sector: An overview', *Materials Today: Proceedings* **57**, 2144–2147.

Wang, M., Yao, X. & Chen, Y. (2021), 'An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients', *IEEE Access* **9**, 25394–25404.

Zhu, H., Jiang, L., Zhang, H., Luo, L., Chen, Y. & Chen, Y. (2021), 'An automatic machine learning approach for ischemic stroke onset time identification based on dwi and flair imaging', *NeuroImage: Clinical* **31**, 102744–102752.