Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/17/20240657

Exploratory analysis of the contributing factors of stroke

Xi Luo

Guangdong Country Garden School, Beijiao, Shunde, Foshan, Guangdong, 528312, China

luoxi2426078717@163.com

Abstract. Stroke is a medical condition that occurs when the blood supply to the brain is interrupted, which causes cell death. Previous studies have shown that the leading risk factors of stroke include elevated systolic blood pressure, poor diet, high body mass index, high fasting plasma glucose, ambient particulate matter pollution, smoking, high low-density lipoprotein, kidney dysfunction, alcohol use, and low physical activity. However, how these factors differentially contribute to the occurrence of stroke remains elusive. In this study, 5110 cases and 11 risk factors were investigated and machine learning models were built to predict the occurrence of stroke and the importance of each of the factors. It was found that the random forest model showed the best performance in predicting stroke and age, glucose level, and body mass index were the top three most important risk factors underlying stroke. These findings shed light on future research in the prevention and diagnosis of stroke.

Keywords: Stroke, Risk Factors, Machine Learning, Binary Classification, Predictive Modelling.

1. Introduction

Stroke is a medical condition that occurs when the blood supply to the brain is interrupted, which causes cell death. The signs and symptoms of a stroke include difficulties with speaking and understanding, problems with moving or feeling on one side of the body, dizziness, or loss of vision [1].

Stroke is a very widespread and influential disease. In 2010, the absolute number of people with first stroke was 16.9 million and had increased 68% since 1990, the absolute number of stroke survivors was 33 million and had increased 84% since 1990, the absolute number of stroke-related deaths was 5.9 million and had increased 26% since 1990, and the absolute number of disability-adjusted life-years (DALYs) lost was 102 million and had increased 12% since 1990 [2].

In 2016, there were 5·5 million deaths caused by stroke. In 2019, in both the age between 50 and 74 years and 75 years and older age groups, stroke was the second leading cause of disability-adjusted life-years worldwide [3].

From 1990 to 2019, the absolute number of cases had a 70% increase in incident strokes, 43% death due to stroke, 102.0% prevalent strokes, and 143.0% DALYs. Moreover, low- to middle-income countries had 89% of the global stroke deaths and disability combined. In 2019, the number of stroke survivors in Asia, The Americas, Africa, and Europe were 58.1 million, 15.5 million, 14.8 million, and 12.6 million respectively [2].

DOI: 10.54254/2753-8818/17/20240657

Strokes can be classified into ischemic stroke, which accounts for 87% of cases, and hemorrhagic stroke, which accounts for 13% of cases [1]. Ischemic stroke occurs when the brain lacks blood supply. Embolism (obstruction of blood vessels by embolus, which disrupts normal blood flow), thrombosis (formation of a blood clot within a cerebral vessel), systemic hypoperfusion (inadequate blood supply to body), and cerebral venous sinus thrombosis (formation of blood clots in the veins that drain the brain) are the mechanisms that can cause ischemic stroke. Hemorrhagic stroke occurs when there is a rupture of a blood vessel or abnormal vascular structure within the brain. Intracerebral hemorrhage and subarachnoid hemorrhage are two subtypes of hemorrhagic stroke. Intracerebral hemorrhage is bleeding within the brain when the rupture of a blood vessel occurs, which is caused by either intraparenchymal hemorrhage (bleeding into the brain tissue) or Intraventricular hemorrhage (bleeding within the ventricles of the brain). Subarachnoid hemorrhage is bleeding outside of the brain tissue, between the pia mater and the arachnoid mater.

The ten leading stroke risk factors in the world were elevated systolic blood pressure, which contributes to 56% of total stroke DALYs; poor diet, which contributes to 31% of total stroke DALYs; high body mass index (BMI), which contributes to 24% of total stroke DALYs; high fasting plasma glucose (FPG), which contribute to 20% of total stroke DALYs; ambient particulate matter pollution, which contribute to 20% of total stroke DALYs; smoking, which contributes to 18% of total stroke DALYs; high low-density lipoprotein (LDL) cholesterol, which contribute to 10% of total stroke DALYs; kidney dysfunction, which contribute to 8% of total stroke DALYs; alcohol use, which contribute to 6% of total stroke DALYs; and low physical activity, which contribute to 2% of total stroke DALY [2].

The common factors between these ten leading stroke risk factors and the features used for modeling were systolic blood pressure (related to hypertension), high body mass index, high fasting plasma glucose (related to average glucose level), and smoking status. The risk factors only included in the ten leading stroke risk factors were poor diet, ambient particulate matter pollution, high LDL cholesterol, kidney dysfunction, alcohol use, and low physical activity. The risk factors only included in the features used for the modelling in the present study were gender, work type, residence type, age, heart disease history, and marriage status.

In this study, 5110 cases and 10 risk factors were investigated and machine learning models were built to predict the occurrence of stroke and the importance of each of the factors. It was found that the random forest model showed the best performance in predicting stroke and age, glucose level, and BMI were the top three most important risk factors underlying stroke.

2. Materials and Methods

2.1. Data information

In the data, there were 5110 cases and 11 variables. These variables are gender, age, hypertension, “heart_disease”, “ever_married”, “work_type”, “Residence_type”, “avg_glucose level”, BMI, “smoking_status”, and stroke. The dataset was retrieved from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/).

2.2. Python packages

Many packages were used. *NumPy* is a powerful *Python* library that provides support for large, multi-dimensional arrays and matrices, along with a comprehensive collection of mathematical functions to operate on these arrays. It was used to serve as a fundamental package for scientific computing and data analysis. *Pandas* is an open-source *Python* library that provides powerful data manipulation and analysis tools. It was used for data analysis, data preprocessing, and data cleaning. *Matplotlib* is a *Python* plotting library that provides a wide range of tools for creating static, animated, and interactive visualizations. It was used to visualize data and produce plots with customizable appearance and layout. *Seaborn* is a *Python* data visualization library that provides a high-level interface for creating informative statistical graphics. It was used to explore statistical data and produce plots with a wide range of plot types and customization options. *Sklearn* (scikit-learn) is a popular *Python* open-source machine learning library.

DOI: 10.54254/2753-8818/17/20240657

It was used to provide a wide range of tools and algorithms for feature selection, data preprocessing, model training, evaluation, and prediction.

2.3. Python functions

Many functions and attributes were used. Function *read_csv* was used to read data from a Comma-Separated Values (CSV) file into a DataFrame. Function *train_test_split* was used to split a dataset into training and testing subsets. Function *fillna* was used to fill missing values in a dataframe with a specific value. Function *subplot* was used to create a grid of subplots within a single figure. Function *hisplot* was used to create a histogram that represents the distribution of numerical data. Function *describe* was used to obtain a summary of count, mean, standard deviation, minimum, medium, the first and third quartiles, and maximum. Function *boxplot* was used to create a boxplot that displays the dataset based on minimum, maximum, median, and the first and third quartiles. Function *violinplot* was used to create a violin plot which is a combination of box plot and density plot. Function *lineplot* was used to create a line plot that represents numerical data. Function *bar* was used to create a bar plot that represents categorical data. Function *get_dummies* was used to convert categorical variables into dummy variables which are binary columns representing the presence or absence of a particular category. Function *SMOTE*, which refers to Synthetic Minority Over-sampling Technique, was used to generate synthetic samples to balance the class distribution. Function *pipeline* was used to automate and organize the machine learning workflow. Functions *RandomForestClassifier*, *SVC*, *LogisticRegression*, and *GradientBoostingClassifier* were used to create a Random Forest Classifier, a Support Vector Classifier, a Logistic Regression Classifier, and a Gradient Boosting Classifier, respectively. Function *cross_val_score* was used to estimate the performance of the model by performing cross-validation. Function *fit* was used to train the model based on the given dataset. Function *predict* was used to predict new, unseen data using a trained machine learning model. Function *confusion_matrix* was used to evaluate the performance of the model by summarizing the counts of true positive, true negative, false positive, and false negative predictions. Function *classification_report* was used to evaluate the performance of the model by summarizing their precision, recall, f1-score, and support for each class, as well as the average values across all classes. Functions *f1_score*, *accuracy_score*, *recall_score*, *precision_score*, and *roc_auc_score* were used to calculate the f1, accuracy, recall, precision, and ROC-AUC scores of the models, respectively. The feature_importances_ attribute was used to retrieve the feature importance calculated by the model. The coef_ attribute was used to access the learned coefficients associated with each feature in the Logistic Regression and the SVM.

3. Results

3.1. Data inspection

To know the number of cases and variables in the data, the data was inspected. For gender, 2994 (58.60%) cases were female, 2115 (41.40%) cases were male, and 1 case was other (0.00%). For hypertension, 4612 (90.24%) cases did not have a hypertension history while 498 (9.76%) cases had a hypertension history. For “heart_disease”, 4834 (94.64%) cases did not have “heart_disease” history while 276 (5.36%) cases had “heart_disease” history. For “ever_married”, 3353 (65.62%) cases were married while 1757 (34.38%) cases were not married. For “work_type”, 2925 (57.24%) cases were “private”, 819 (16.03%) cases were “self-employed”, 687 (13.44%) cases were “children”, 657 (12.86%) cases were “government job”, and 22 (0.43%) cases were “never_worked”. For “Residence_type”, 2596 (50.80%) cases lived in urban areas, and 2514 (49.20%) cases lived in rural areas. For smoking status, 1892 (37.03%) cases were “never smoked”, 1544 (30.22%) cases were “unknown”, 885 (17.23%) cases were “formerly smoked”, and 789 (15.44%) cases were “smoked”. For stroke, 4861 (95.13%) cases had a stroke while 249 (4.87%) cases did not have a stroke.

DOI: 10.54254/2753-8818/17/20240657

3.2. Data split and processing

Then, each feature was examined to see whether there was missing data in the dataset, and 201 missing data was found in the column of BMI. Before processing the data, 80% data was assigned to the training set and 20% data was assigned to the testing set in the random state of 42. To deal with missing data, the mean BMI of training data was used to fill in the missing data of both training and test sets. After further inspection of data of each categorical variable, it was found that there was only 1 case belonging to the "Other" classification in the variable "gender". Therefore, this case was considered abnormal and was excluded from the dataset.

3.3. Data Visualization

At first, the numerical data (age, average glucose level, and BMI) was plotted into histograms to visualize the distribution of data of these features. The distribution of age was evenly distributed, and it had the highest percentage between the ages of 50 years and 60 years old (age: mean ± std = 43.2 years old ± 22.6 years old, min = 0 years old, max = 82 years old). The distribution of average glucose level was positively skewed, and it had the highest percentage between the average glucose level of 75mg/dl and 100mg/dl (average glucose level: 105.7 mg/dl ± 45.3 mg/dl, min = 55.0 mg/dl, max = 271.0 mg/dl). The distribution of BMI was positively skewed, and it had the highest percentage between the BMI of 20kg/m2 and 50kg/m2 (BMI: 28.4 kg/m2 ± 7.7 kg/m2, min = 10.0 kg/m2, max = 97.0 kg/m2) (Figure 1).

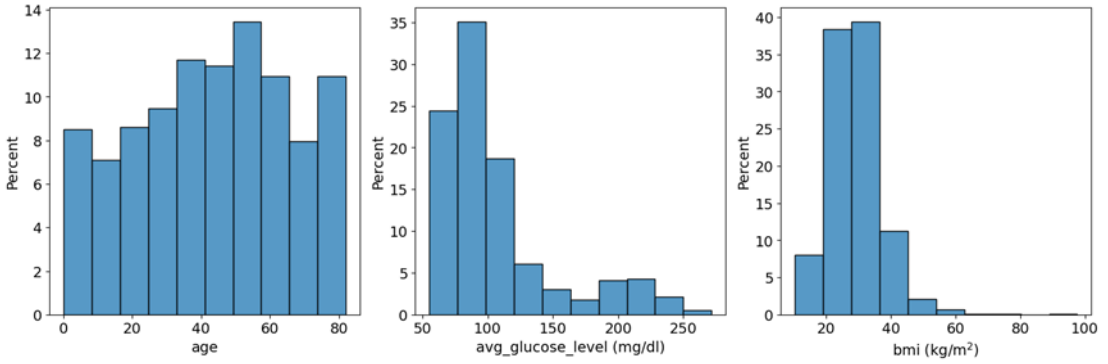

Figure 1. Percent distribution of age, average glucose level, and BMI.

Then these numerical data were separated into Stroke and No Stroke to analyze the effect of these variables on stroke.

Data of each feature was plotted into box plots and violin plots to visualize the difference in distributions of No Stroke data and Stroke against different features. In the box plot of age against No Stroke/ Stroke, the IQR and median of the No Stroke box are 37 years and 42 years respectively. The IQR and median of the Stroke box are 19 years and 70 years respectively. In the box plot of average glucose level against No Stroke/ Stroke, the IQR and median of the No Stroke box are 38 mg/dl and 91 mg/dl respectively. The IQR and median of the Stroke box are 110 mg/dl and 100 mg/dl years respectively. In the box plot of BMI against No Stroke/ Stroke, the IQR and median of the No Stroke box are 8.8 kg/m^2 and 28 kg/m^2 respectively. The IQR and median of the Stroke box are 5 kg/m^2 years and 27.5 kg/m^2 respectively (Figure 2).

It was shown from the box plots and violin plots (Figure 3) that people who got strokes were more likely to have higher age, average glucose, and relatively higher BMI than healthy people. Based on the plots, it seemed clear that age was a big factor in stroke patients.

DOI: 10.54254/2753-8818/17/20240657

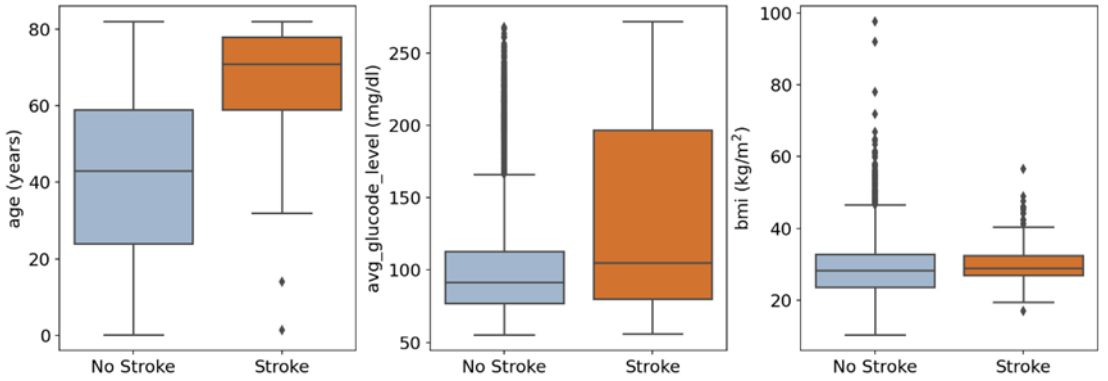

Figure 2. Box plots for No Stroke and Stroke cases based on age, average glucose level, and BMI.

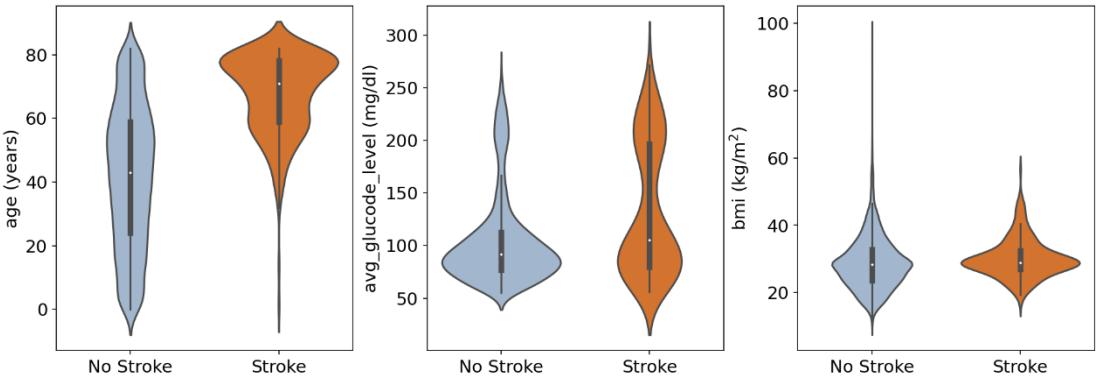

Figure 3. Violin plots for No Stroke and Stroke cases based on age, average glucose level, and BMI.

Line plots were drawn to analyze how risk rate changes by age, BMI, and average glucose level. The risk rate of stroke against three features was calculated by dividing the number of stroke cases less than a specific value of the feature by the number of all cases less than that value. These plots showed that the risk rate of having a stroke increased as age increased from 30 years to 80 years, risk increased when BMI increased from 10.3 kg/m2 to 30.3 kg/m2, and risk increased when average glucose level increased from 155.0 mg/dl to 255.0 mg/dl (Figure 4-6).

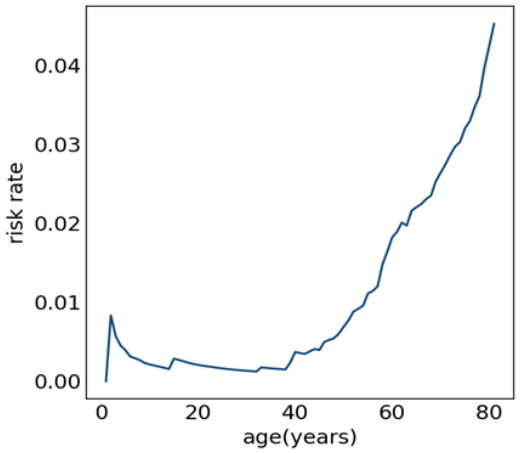

Figure 4. Line plot for rate change of age.

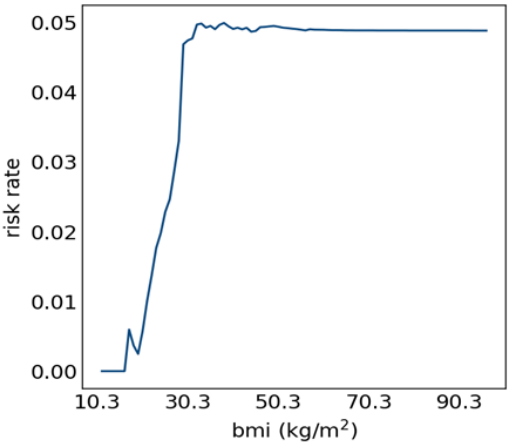

Figure 5. Line plot for rate change of BMI.

DOI: 10.54254/2753-8818/17/20240657

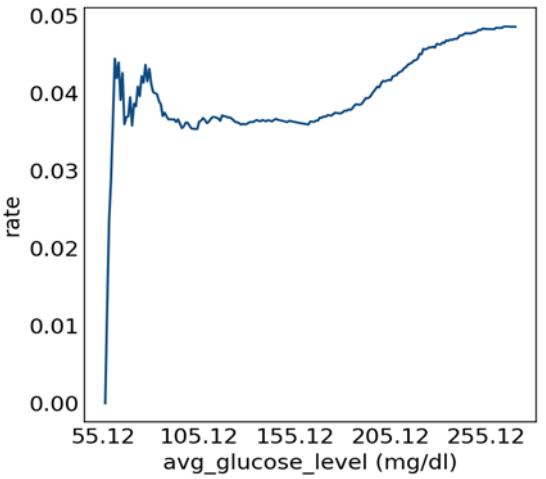

Figure 6. Line plot for rate change of average glucose rate.

To see whether there was a difference in percentage or distribution between the stroke group and no stroke in various features, side-by-side bar graphs or histograms of these features against Stroke/ No Stroke were plotted. The age graph plotted the distribution of age against Stroke/ No Stroke in the histogram. It showed that most people who had a stroke also have an age higher than 40 years, while the age of people who did not have a stroke is evenly distributed. The smoking status graph plotted the percentage of "smokes", "unknown", "formerly smoked", and "never smoked" against Stroke/ No Stroke in a side-by-side bar graph. It showed that stroke happened more in those who smoked or formerly smoked and happened less in those who never smoked. The gender graph plotted the percentage of males and females against Stroke/ No Stroke in a side-by-side bar graph. It showed that there was not much difference between the stroke percentage of males and females. The heart disease graph plotted the percentage of "No History" and "History" against Stroke/ No Stroke in a side-by-side bar graph. It showed that those who have a history of heart disease were more likely to have a stroke than those who did not. The average glucose level graph plotted the distribution of average glucose level against Stroke/ No Stroke in the histogram. It showed that people who had strokes tend to have higher average glucose levels than people who did not have strokes. BMI graph plotted the distribution of BMI against Stroke/ No Stroke in the histogram. It showed that people who had strokes tend to have higher BMI than people who did not have strokes. Work type graph plotted the percentage of "Govt_job", "Private", "Self-employed", "Never_worked", and "children" against Stroke/ No Stroke in a side-by-side bar graph. It showed that people who had strokes usually did private work and those who did not have strokes usually did self-employed work. The hypertension graph plotted the percentage of "No History" and "History" against Stroke/ No Stroke in a side-by-side bar graph. It showed that those who had a history of hypertension were more likely to have strokes than those who did not (Figure 7-13).

DOI: 10.54254/2753-8818/17/20240657

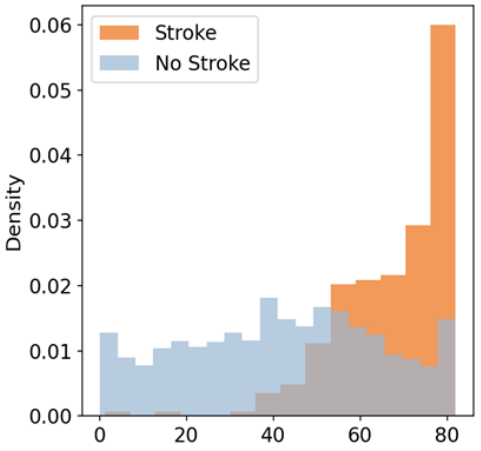

Figure 7. Histogram of distribution of age against Stroke/ No Stroke.

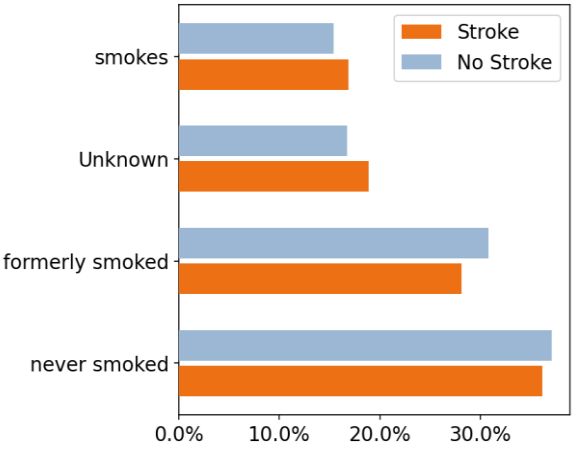

Figure 8. Bar graph of the percentage of Stroke against No-Stroke in cases of four smoking statuses.

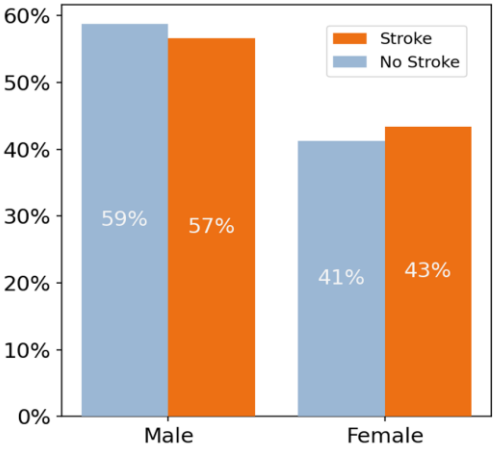

Figure 9. Bar graph of the percentage of Stroke against No-Stroke in cases of males and females.

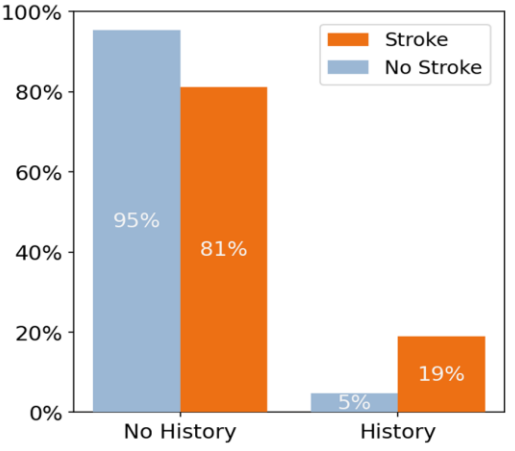

Figure 10. Bar graph of the percentage of Stroke against No-Stroke in cases without and with heart disease history.

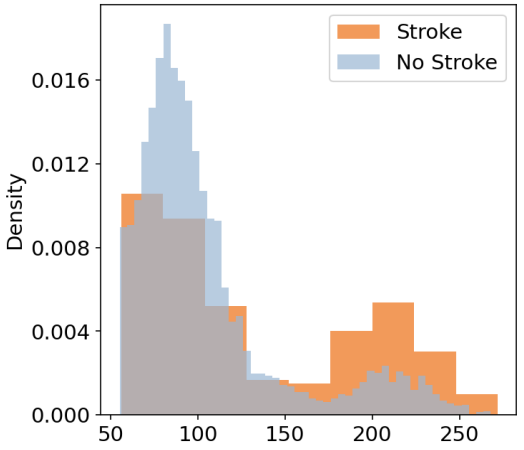

Figure 11. Histogram of distribution of average glucose level against Stroke/ No Stroke.

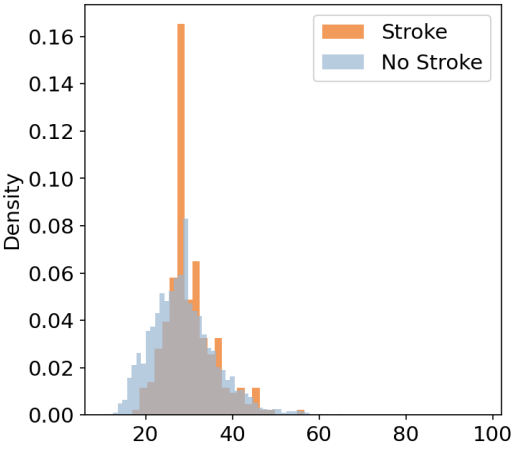

Figure 12. Histogram of distribution of age against Stroke/ No Stroke.

DOI: 10.54254/2753-8818/17/20240657

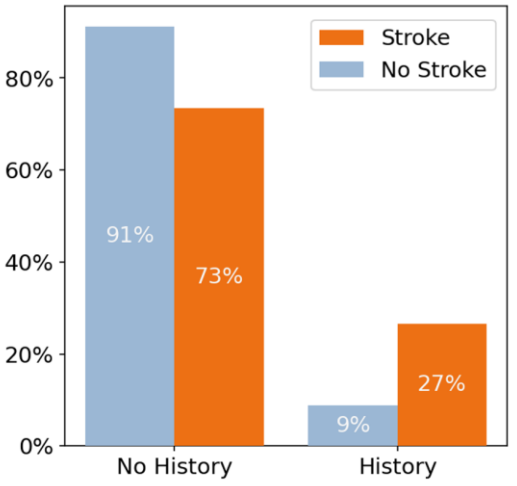

Figure 13. Bar graph of the percentage of Stroke against No-Stroke in cases without and with hypertension history.

3.4. Model development

Before model development, categorical data were converted to dummy variables by using the function *get_dummy* for further model training.

The baseline was calculated and the null accuracy of 0.5 and inverse of the null accuracy of 0.5 was obtained, which was used to examine the performance of the model.

Age, hypertension, "heart_disease", "ever_married", "avg_glucose_level", BMI, "gender_Male", "residence_Urban", "work_Never_worked", "work_Private", "work_Self_employed", "work_children", "smoking_formerly smoked", "smoking_never smoked", and "smoking_smokes" were used as features for model training. Stroke was used as a label. After that, *SMOTE* was applied to all the data to relieve the effect of biased data, because, during the process of data inspection, it was found that there was a great difference in the number of stroke cases (n = 249) and no-stroke cases (n = 4861).

After all the preparation, Random Forest, Support Vector Machine, Logistic Regression, and Gradient Boosting Classifier models were constructed.

3.5. Model Performance

To have a basic understanding of the performance of each model, their mean accuracy scores for training data were obtained.

Mean accuracy scores for Random Forest, SVM, Logistic Regression, and Gradient Boosting Classifier were 0.94, 0.89, 0.86, and 0.90 respectively. Random Forest had the highest mean accuracy score while Logistic Regression had the lowest mean accuracy score in four models.

A chart was organized to compare each model's performance based on their accuracy, precision, recall, f1 score, and ROC-AUC score. True positives (TP) are the correct predictions made that are labeled as positive. True negatives (TN) are the correct predictions made that are labeled as negative. False positives (FP) are the incorrect predictions made that are labeled as positive. False negatives (FN) are the incorrect predictions made that are labeled as negative. Accuracy is the proportion of the correct predictions out of all predictions made. It was calculated by using (TP + TN) / (TP + FN + FP + TN). Precision is the proportion of the correct predictions out of all positive predictions. It was calculated by using TP / (TP + FP). Recall is the proportion of correct predictions out of all positive classes. It was calculated by using TP / (TP + FN). The f1 score is the harmonic mean of precision and recall. It was calculated by using (2 × precision × recall) / (precision + recall). ROC curve is the plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, which illustrates the performance of a binary classification model. The ROC-AUC score is the area under the ROC

DOI: 10.54254/2753-8818/17/20240657

(Receiver Operating Characteristics) curve. For Random Forest, SVM, Logistic Regression, and Gradient Boosting Classifier, their accuracies were 0.95, 0.9, 0.87, and 0.89 respectively, their precisions were 0.94, 0.89, 0.86, and 0.87 respectively, their recalls were 0.96, 0.91, 0.87, and 0.91 respectively, their f1 scores were 0.95, 0.9, 0.87, and 0.89 respectively, and their ROC-AUC Scores were 0.95, 0.9, 0.87, and 0.89 respectively. Random Forest had the best performance on these five, while Logistic Regression had the worst performance on these five indexes (Figure 14).

Model Comparison

	Accuracy	Precision	Recall	F1	ROC AUC Score
Random Forest Score	0.95	0.94	0.96	0.95	0.95
Support Vector Machine Score	0.9	0.89	0.91	0.9	0.9
Logistic Regression Score	0.87	0.86	0.87	0.87	0.87
Gradient Boosting Classifier Score	0.89	0.87	0.91	0.89	0.89

Figure 14. Comparison of the performance of four models.

3.6. Importance of the risk factors

Feature importance was further analyzed to see how each feature affected the prediction made by models. Plots below were drawn to evaluate the absolute value of feature importance for each model. For the Random Forest, age, “avg_glucose_level”, and BMI were the three most important features. For the SVM, “smoking_never smoked”, “work_Self-employed”, and “smoking_formerly smoked” were the three most important features. For the Logistic Regression, age, “smoking_never smoked”, and “work_Self-employed” were the three most important features. For the Gradient Boosting Classifier, age, “smoking_never smoked”, and “avg_glucose_level” were the three most important features (Figure 15-18).

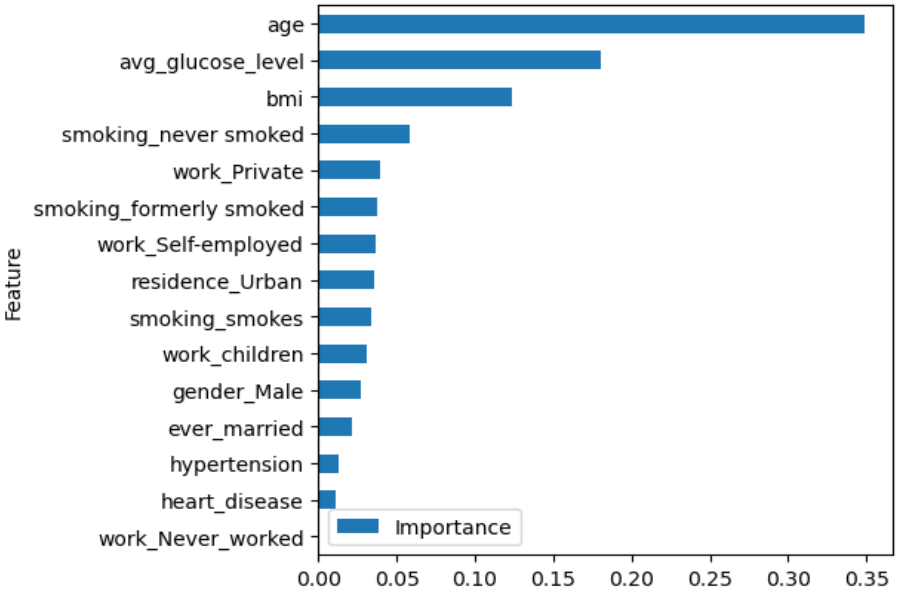

Figure 15. Feature importance—Random Forest.

DOI: 10.54254/2753-8818/17/20240657

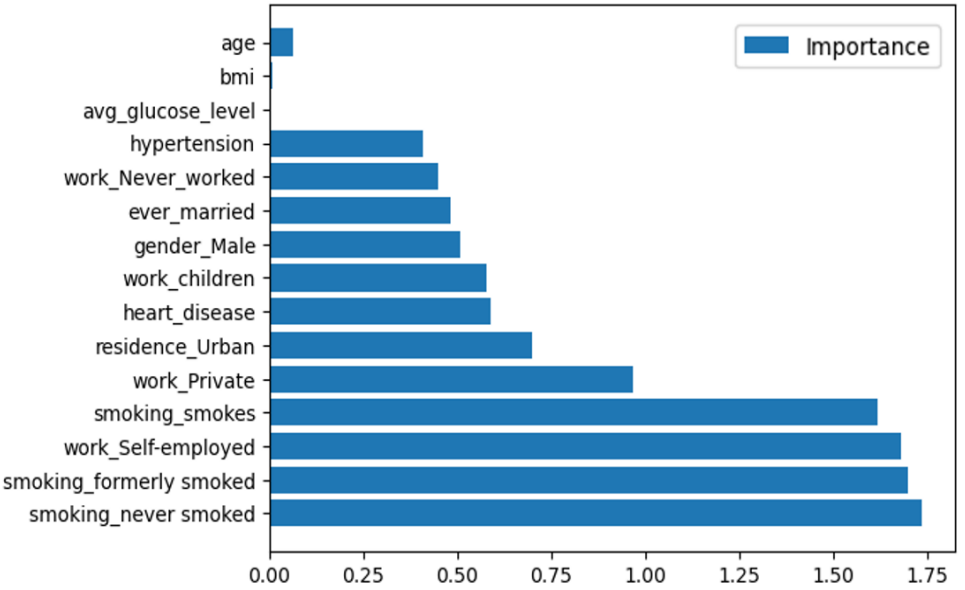

Figure 16. Feature importance—Support Vector Machine.

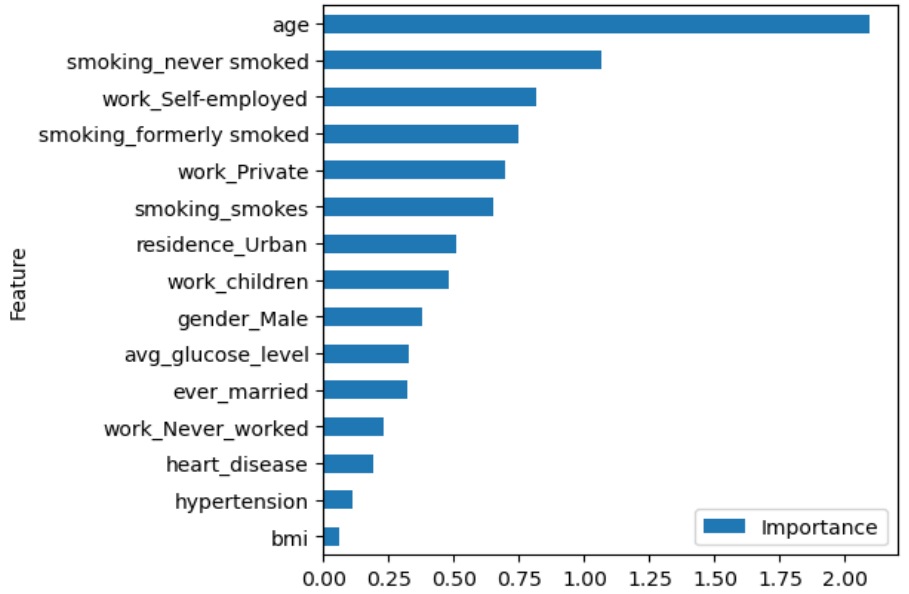

Figure 17. Feature importance—Logistic Regression.

DOI: 10.54254/2753-8818/17/20240657

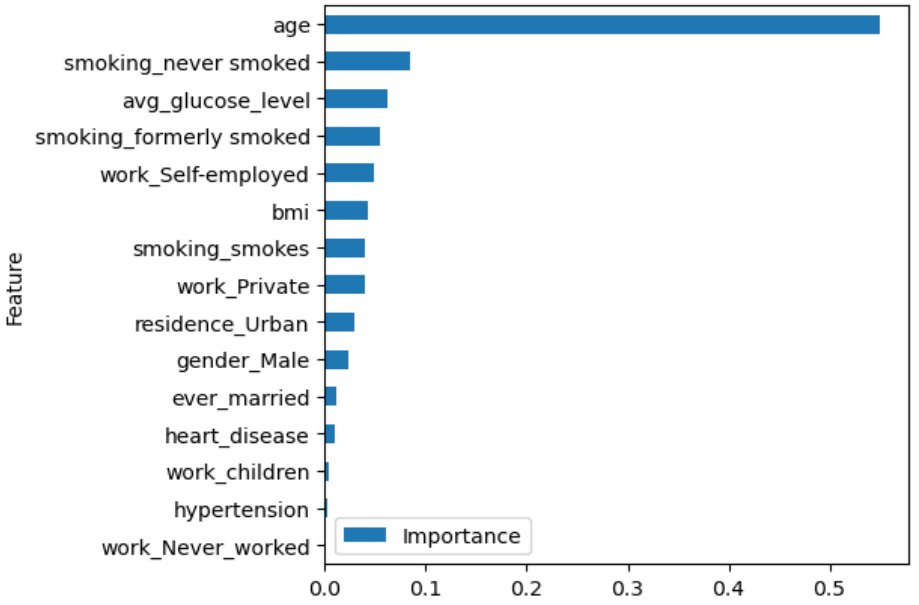

Figure 18. Feature importance—Gradient Boosting Classifier.

The top 10 important features of each model were then summarized. Based on this chart, "smoking_never smoked", "work_Private", "work_Self-employed", "smoking_formerly smoked", "residence_Urban", "smoking_smokes", and "gender_Male" were the top 10 important features in all four models. Age and "avg_glucose_level" were the top 10 important features in three of four models. BMI and "work_children" were the top 10 important features in two of the four models. "Heart_disease" and "ever_married" were the top 10 important features in only the SVM. (Figure 19)

Feature Importance				
	Random Forest	Support Vector Machine	Logistic Regression	Gradient Boosting Classifier
1	age	smoking_never smoked	age	age
2	avg_glucose_level	work_Self-employed	smoking_never smoked	smoking_never smoked
3	bmi	smoking_formerly smoked	work_Self-employed	avg_glucose_level
4	smoking_never smoked	smoking_smokes	smoking_formerly smoked	smoking_formerly smoked
5	work_Private	work_Private	work_Private	work_Self-employed
6	work_Self-employed	residence_Urban	smoking_smokes	work_Private
7	smoking_formerly smoked	heart_disease	residence_Urban	bmi
8	residence_Urban	gender_Male	work_children	smoking_smokes
9	smoking_smokes	work_children	gender_Male	residence_Urban
10	gender_Male	ever_married	avg_glucose_level	gender_Male

Figure 19. Feature importance in four models.

The ranking of each feature was calculated by dividing the sum of the rank of a feature in each model by the number of these models. For example, "work_Self-employed" was calculated by (6+2+3+5)/4).

DOI: 10.54254/2753-8818/17/20240657

The top three features were age (overall rank = 1), “smoking_never smoked” (overall rank = 2.25), and “work_Self-employed” (overall rank = 4).

4. Discussion

In this study, 5110 cases and 11 risk factors were investigated and machine learning models were built to predict the occurrence of stroke and the importance of each of the factors. It was found that the random forest model showed the best performance in predicting stroke and age, glucose level, and BMI were the top three most important risk factors underlying stroke.

There is more that can be done to improve this project. Now, the feature importance of each model was evaluated by calculating their weight directly. However, this approach had some flaws. First, it is not model-agnostic, which means it could just obtain explicit weights provided by some particular models, such as the Linear Regression and the Decision Tree. It cannot be applied to models like the Random Forest and the Gradient Boosting which have complex internal structures and do not provide explicit feature weights. Moreover, calculating explicit weight cannot provide local interpretability, which means it is not able to explain how models make predictions for each case. However, these flaws could be overcome by using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to analyze feature importance. SHAP applies game theory to assign importance values to each feature in a prediction and finally explains the contribution of each feature towards the predicted outcome. On the contrary, LIME trains a simpler, interpretable model to approximate the complex model in the vicinity of the instance of interest based on the perturbed samples it generated. SHAP’s advantage over Lime is that it provides Shapley value which could be used for various comparison explanations, and it properly attributes the contribution made by each feature. However, SHAP’s calculation is time-consuming, and it does not work well when correlations exist between features. By applying both SHAP and LIME, a relatively accurate feature importance for each model could be obtained.

The four models used have different advantages and disadvantages. The advantages of the Random Forest are: it is a robust algorithm that can handle noisy data and outliers. It is less likely to overfit the data, which means it can generalize well to new data; it is one of the most accurate machine learning algorithms. It can handle both classification and regression problems and can work well with both categorical and continuous variables; it is fast and can handle large datasets. It can also be easily parallelized to speed up training; it provides a measure of feature importance, which can help in feature selection and data understanding. Disadvantages of the Random Forest are: it will overfit the data when the number of trees in the forest is too high or if the trees are too deep; the Random Forest can be less interpretable than a single decision tree because it involves multiple trees. it can be difficult to understand how the algorithm arrived at a particular prediction; the training time of the Random Forest can be longer than other algorithms. Especially if the number of trees and the depth of the trees are high; it requires more memory than other algorithms because it stores multiple trees. This can be a problem if the dataset is large.

The advantages of the Support Vector Machine are: it performs well in high-dimensional space and has excellent accuracy; it requires less memory because it only uses a portion of the training data; it performs reasonably well when there is a large gap between classes; Disadvantages of the Support Vector Machine are: it requires a long training period, so it is not practical for large datasets; the inability of the SVM classifiers to handle overlapping classes is another drawback; large data sets are not a good fit for the SVM algorithm.

The advantages of the Logistic Regression are: the Logistic Regression is easier to implement, interpret, and very efficient to train; it can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions; it is very fast at classifying unknown records. Disadvantages of the Logistic Regression are: if the number of observations is lesser than the number of features, it should not be used, otherwise, it may lead to overfitting; it constructs linear boundaries; it can only be used to predict discrete functions. Hence, the dependent variable of the Logistic Regression is bound to the discrete number set.

DOI: 10.54254/2753-8818/17/20240657

The advantages of the Gradient Boosting Classifier are: it has high accuracy and strong predictive performance. It can handle complex patterns and interactions in the data, making it suitable for a wide range of tasks, including regression and classification; it can work well with a mixture of data types, including numerical and categorical features. It can handle missing values and automatically handle categorical variables without requiring explicit encoding; the algorithm provides a measure of feature importance, which helps in understanding the relative contribution of each feature in the prediction. This information can be valuable for feature selection, identifying important variables, and gaining insights into the problem domain. Disadvantages of the Gradient Boosting Classifier are: the training process of the Gradient Boosting Classifier can be computationally expensive, especially with large datasets or complex models. It involves iteratively building an ensemble of weak learners, which can be time-consuming compared to simpler algorithms like logistic regression or decision trees; the training process of the Gradient Boosting Classifier can be computationally expensive, especially with large datasets or complex models. It involves iteratively building an ensemble of weak learners, which can be time-consuming compared to simpler algorithms like logistic regression or decision trees; finding the optimal combination of hyperparameters can be challenging for the Gradient Boosting Classifier. Tuning multiple hyperparameters requires careful experimentation and can be time-consuming. Additionally, the optimal hyperparameters may vary depending on the dataset and problem domain.

From best to worst, the rank of the four models' accuracy was Random Forest, Support Vector Machine, Gradient Boosting Classifier, and Logistic Regression; the rank of the four models' precision was Random Forest, Support Vector Machine, Gradient Boosting Classifier, and Logistic Regression; the rank of the four models' recall was Random Forest, Support Vector Machine at the same level with Gradient Boosting Classifier, and Logistic Regression; the rank of the four models' f1 score was Random Forest, Support Vector Machine, Gradient Boosting Classifier, and Logistic Regression; the rank of the four models' ROC-AUC score was Random Forest, Support Vector Machine, Gradient Boosting Classifier, and Logistic Regression.

Consistency of different features was also displayed by the four models. Feature "smoking_never smoked" was included in all four models. Feature age was included in three models, which were Random Forest, Logistic Regression, and Gradient Boosting Classifier. Feature BMI was included in two models, which were Random Forest, and Gradient Boosting Classifier. Feature "heart_disease" was only included in the Support Vector Machine. The 5 most important features were age (overall rank = 1), "smoking_never smoked" (overall rank = 2.25), "work_Self-employed" (overall rank = 4), "smoking_formerly smoked" (overall rank = 4.5), and "avg_glucose_level" (overall rank = 5).

These features were also supported. Most stroke cases happen in people with ages greater than 65 years old, and compared with younger patients, aged patients have a higher death rate and poorer quality of life after stroke [4].

Cigarette smoking contributes to about one-quarter of all strokes. The risk of stroke rapidly and considerably decreases after smoking cessation [5].

Moreover, approximately one-third of stroke cases have diabetes. Acute diabetes and hyperglycemia were associated with higher mortality, poorer neurological and functional outcomes, stroke recurrence, longer hospital stays, and higher readmission rates after ischemic or hemorrhagic strokes [6].

Except for the features used for model training, there are other risk factors for stroke. Air pollution is an emerging risk factor for stroke and is estimated to contribute to 14% of all stroke-associated deaths. Pollutants can trigger autonomic respiratory reflex arcs, which leads to increased vascular resistance, arrhythmias, and hypertension. This can result in cardioembolic ischaemic stroke. Short-term exposure to air pollution can increase the risk of intracerebral hemorrhage. Long-term exposure to pollutants can lead to systemic inflammation and reactive oxygen species formation, which is associated with accelerated progression of atherosclerosis. These mechanisms can cause obesity or diabetes after prolonged exposure, which in turn increases the risk of ischaemic stroke [7].

Alcohol use is also associated with stroke. From 1990 to 2019, alcohol use attributed 6% to the total number of stroke-related DALYs [8]. Low alcohol consumption is linked to a decreased likelihood of stroke morbidity and mortality, whereas excessive alcohol consumption is associated with an elevated

DOI: 10.54254/2753-8818/17/20240657

risk of overall stroke. The correlation between alcohol intake and stroke morbidity and mortality follows a J-shaped pattern [9].

Moreover, Chronic kidney disease (CKD) can lead to sodium dysregulation, increased sympathetic nervous system, and alterations in renin angiotensin aldosterone system activity. These mechanisms can contribute to hypertension which is an important risk factor for stroke [10].

5. Conclusions

In conclusion, in the present study, 5110 patients and 10 risk factors (gender, age, hypertension, "heart_disease", "ever_married", "work_type", "Residence_type", "avg_glucose level", BMI, and "smoking_status") were investigated. Four machine learning models (Random Forest, Support Vector Machine, Logistic Regression, and Gradient Boosting Classifier) were built to predict the occurrence of stroke and determine the importance of each factor. It was found that the Random Forest model showed the best performance in predicting stroke. Age, glucose level, and BMI were the top three most important risk factors underlying stroke occurrence.

Based on the findings of the present study, there can be multiple future directions as follows. First, researchers are exploring advanced feature extraction techniques, such as deep learning and natural language processing, to extract meaningful information from complex and unstructured data sources like medical images, clinical notes, and genomic data from stroke cases. These techniques have shown promise in uncovering hidden patterns and relationships relevant to stroke prediction. Second, longitudinal analysis of stroke patient data over time can offer insights into the progression of risk factors and the dynamic nature of stroke occurrence. Machine learning models that can capture temporal patterns and changes in risk factors may enable personalized risk assessments and intervention strategies. Third, given the critical nature of stroke prediction, there is a growing emphasis on developing interpretable and explainable machine learning models, which enables healthcare professionals to trust and integrate these models into their decision-making processes. Fourth, machine learning models can facilitate personalized risk stratification by considering individual characteristics, genetics, lifestyle factors, and comorbidities. This strategy can lead to tailored preventive interventions that address the specific needs of individuals, potentially improving stroke outcomes. Finally, unsupervised learning techniques, such as clustering algorithms, can be employed to identify subtypes or phenotypes within stroke populations. By grouping individuals based on common patterns, these techniques can aid in understanding the heterogeneity of strokes and enable more personalized treatments.

References

[1] Donnan, Geoffrey A., et al. "Stroke." The Lancet, vol. 371, no. 9624, Elsevier BV, May 2008, pp. 1612-23. Crossref, https://doi.org/10.1016/s0140-6736(08)60694-7.

[2] Feigin, Valery L., et al. "Global and Regional Burden of Stroke During 1990-2010: Findings From the Global Burden of Disease Study 2010." The Lancet, vol. 383, no. 9913, Elsevier BV, Jan. 2014, pp. 245-55. Crossref, https://doi.org/10.1016/s0140-6736(13)61953-4

[3] Lanas, Fernando, and Pamela Seron. "Facing the Stroke Burden Worldwide." The Lancet Global Health, vol. 9, no. 3, Elsevier BV, Mar. 2021, pp. e235-36. Crossref, https://doi.org/10.1016/s2214-109x(20)30520-9.

[4] Roy-O'Reilly, Meaghan, and Louise D. McCullough. "Age And Sex Are Critical Factors in Ischemic Stroke Pathology." Endocrinology, vol. 159, no. 8, The Endocrine Society, July 2018, pp. 3120-31. Crossref, https://doi.org/10.1210/en.2018-00465.

[5] Hankey, Graeme J. "Smoking and Risk of Stroke." European Journal of Cardiovascular Risk, vol. 6, no. 4, Oxford UP (OUP), Aug. 1999, pp. 207-11. Crossref, https://doi.org/10.1177/204748739900600403.

[6] Lau, Lik-Hui, et al. "Prevalence of Diabetes and Its Effects on Stroke Outcomes: A Meta-analysis and Literature Review." Journal of Diabetes Investigation, vol. 10, no. 3, Wiley, Oct. 2018, pp. 780-92. Crossref, https://doi.org/10.1111/jdi.12932.

DOI: 10.54254/2753-8818/17/20240657

[7] Verhoeven, Jamie I., et al. “Ambient Air Pollution and the Risk of Ischaemic and Haemorrhagic Stroke.” The Lancet Planetary Health, vol. 5, no. 8, Elsevier BV, Aug. 2021, pp. e542-52. Crossref, https://doi.org/10.1016/s2542-5196(21)00145-5.

[8] Feigin, Valery L., et al. “World Stroke Organization (WSO): Global Stroke Fact Sheet 2022.” International Journal of Stroke, vol. 17, no. 1, SAGE Publications, Jan. 2022, pp. 18-29. Crossref, https://doi.org/10.1177/17474930211065917.

[9] Zhang, Chi, et al. “Alcohol Intake and Risk of Stroke: A Dose-response Meta-analysis of Prospective Studies.” International Journal of Cardiology, vol. 174, no. 3, Elsevier BV, July 2014, pp. 669-77. Crossref, https://doi.org/10.1016/j.ijcard.2014.04.225.

[10] Hamrahian, Seyed Mehrdad, and Bonita Falkner. “Hypertension in Chronic Kidney Disease.” Advances in Experimental Medicine and Biology, Springer International Publishing, 2016, pp. 307-25. Crossref, https://doi.org/10.1007/5584_2016_84.