Predicting Strokes using Logistic Regression and K-Nearest Neighbours ECS784

***Abstract*—Strokes are the second-leading cause of death in the world and if survived may lead to temporary or permanent impairments including paralysis. This makes early prediction crucial for timely medical intervention. This paper studies the application of machine learning algorithms in predicting stroke on the basis of ten varying features. Two supervised algorithms namely Logistic Regression and K-Nearest Neighbours are used.**

This report will also discuss the available literature related to the topic. The chosen dataset comprising five features of type object, three of type float64, and three of type int64 is preprocessed for model training and evaluation. The data is then normalized by applying the SMOTE technique to the target variable, followed by the feature selection process. The models are assessed using performance metrics such as accuracy, recall and F1 score. The report will also provide an analysis of the two machine learning algorithms employed and present the findings and conclusions.

Keywords—Logistic Regression, K-Nearest Neighbours, Brain Stroke, data analytics

I. INTRODUCTION

This study aims to use machine learning algorithms and data analysis techniques to predict the possibility of having a stroke when given data on various health metrics and lifestyle indicators.

Across the world, strokes are the second-leading cause of death and third-leading cause of death and disability [1]. Strokes happen when blood supply to a part of the brain is cut off, destroying brain cells in the process. They are of three main types: Ischaemic strokes, caused by a blockage to the brain obstructing blood supply; Haemorrhagic strokes caused due to bleeding in or around the brain, and Transient ischaemic attacks which are strokes lasting a shorter period of time. Strokes can happen to anyone of any age, making them difficult to predict. However, there are certain risk factors found to be associated with the chance of getting a stroke, such as age, smoking, high blood pressure or belonging to certain ethnic backgrounds.

The statistics related to stroke are alarming. Someone suffers from a stroke every five minutes, and around 100,000 people have a stroke each year. In the case where a stroke is survived, it can still have debilitating impacts on a patient's physical, emotional, cognitive and behavioural aspects [2].

Predicting the likelihood of a stroke attack has proven difficult as it is the result of many different pathophysiologies which are not accurately represented in the common risk factors associated with strokes [3]. Simultaneously, detecting a stroke early prevents death and severe brain damage in 85% of cases. Currently work is being done on the use of MRI and CT scan images to help classify diseases such as stroke, but this is an expensive approach which does not benefit underdeveloped nations where stroke attacks are most common. Therefore the need for an inexpensive, non-invasive method for predicting strokes is of utmost importance [4].

OBJECTIVES

- Analyse and improve the data to ensure it meets the requirements for an adequate machine learning dataset.
- Determine which features in the dataset are associated with predicting stroke attacks.
- Use common libraries in Python to visualise the data.
- Train two supervised machine learning models namely Logistic Regression and K-Nearest Neighbours classification to make the prediction.
- Compare the models, evaluate their accuracy and draw subsequent conclusions.

LITERATURE REVIEW

Research into the use of machine learning models for stroke prediction has been ongoing. The first paper, a 2023 report titled 'Machine Learning and the Conundrum of Stroke Risk Prediction' by Chahine et al. discusses what they refer to as the 'ongoing conundrum' of stroke risk prediction. After analyzing and comparing machine learning algorithms with conventional methods of stroke prediction, the study found machine learning algorithms show a higher rate of predictive accuracy compared to traditional statistical inferences due to their ability for multiscale, multiphysics computational modelling. The clinical and demographic variables employed by the old paradigm resulted in inadequate risk categories which were not precise and fine-grained enough. The study proposes solving this using machine learning techniques

capable of processing large datasets, complex features and various predictors [5].

In the second paper titled 'The Probability of Ischaemic Stroke Prediction with a Multi-Neural-Network Model' by Yan Liu et al., a convolutional neural network was employed based on a fusion of different models. The paper posits that the current issue with prediction models is the use of single-modal data, which was solved by extracting multi-modal data using multiple end-to-end neural networks. The model's effectiveness was tested when it was run on the data of 79 subjects, and the prediction accuracy was an impressive 98.53%. [6]

The third paper is the 'Ranking of stroke and cardiovascular risk factors for an optimal risk calculator design: Logistic regression approach' in which researchers use multivariate logistic regression to compute the odds ratio of ranking 26 cardiovascular risk factors. The paper successfully finds an optimal risk calculator for prediction using the AtheroEdge composite risk score (AECRS1.0). Using multivariate logistic regression helps the research group indicate the highest impact of carotid ultrasound image phenotypes in predicting cardiovascular risk [7].

The fourth paper is titled 'Binary Logistic Regression Model of Stroke Patients: A Case Study of Stroke Centre Hospital in Makassar' and aims to determine the risk factors significantly associated with stroke. A binary logistic regression algorithm is used with features such as age, sex, history of diseases etc. The parameters are estimated at times using Maximum Likelihood Estimation and covariates are combined using goodness-of-fit measures. According to the paper, a binary logistic model is the most accurate model for determining history of disease and blood sugar level as the most important determining factors [8].

The fifth paper, 'Comparative Analysis of KNN and Decision Tree Classification Algorithms for Early Stroke Prediction: A Machine Learning Approach' by Eldora et al. compares the use of K-Nearest Neighbour and Decision Tree classification algorithms to determine the most suitable one in predicting early stroke. The comparison is done after applying oversampling techniques during preprocessing to mitigate the issue of unbalanced classes. The paper concludes that K-Nearest Neighbours has a higher predictive accuracy at 97.1845% [9].

II. ANALYSIS

EXTERNAL LIBRARIES

Pandas – An open-source data manipulation and analysis tool which makes use of the Python language. It is fast, flexible and powerful.

Numpy – An open-source Python library used primarily to work with arrays, but also includes functions for working with matrices and linear algebra.

Scikit-learn – A popular Python library to work in the domain of data analysis and machine learning, featuring regression, classification and clustering algorithms. Allows for the easy implementation of the machine learning models used in this study.

Matplotlib – A library to create static, dynamic and interactive visualizations in Python such as graphs and charts.

Seaborn – A data visualization library for Python based on Matplotlib, which allows for the creation of visually robust statistical graphics.

DESCRIBING THE DATA

The dataset titled 'Brain stroke prediction dataset' is gathered from Kaggle [10]. The data consists of 11 features and 4981 rows of data instances. Six of the features are in binary format, including the target feature 'stroke' in the form of yes/no variables. Three features are in numerical format, and the remaining three comprise three different variables.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
2	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
3	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
4	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1

Figure 1 Sample of the original dataset

Furthermore, five of the features are of type object, three of type float64 and three of type int64. All object-type features are converted to numerical values using sci-kit's LabelEncoder() function.

```
RangeIndex: 4981 entries, 0 to 4980
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             4981 non-null   object
 1   age                4981 non-null   float64
 2   hypertension       4981 non-null   int64
 3   heart_disease      4981 non-null   int64
 4   ever_married       4981 non-null   object
 5   work_type          4981 non-null   object
 6   Residence_type     4981 non-null   object
 7   avg_glucose_level  4981 non-null   float64
 8   bmi                4981 non-null   float64
 9   smoking_status     4981 non-null   object
 10  stroke             4981 non-null   int64
dtypes: float64(3), int64(3), object(5)
```

Figure 2 Overview of the dataset

The dataset contains no duplicate features and no null values as shown using the isnull() function, therefore no further steps are necessary to deal with missing data.

```
gender              0
age                 0
hypertension        0
heart_disease       0
ever_married        0
work_type           0
Residence_type      0
avg_glucose_level   0
bmi                 0
smoking_status      0
stroke              0
```

Figure 3 Checking for any null values

Using the Matplotlib Python library, the data is represented visually to make the analysis more intuitive. In observing the target feature, it is found to contain a visible case of class oversampling with a vast imbalance between the number of patients who had a stroke (~5%) and those who didn't (~95%). As the study relies on binary classification, this will impact the training of the models and skew results. It is addressed using the SMOTE technique. Synthetic Minority Over-sampling Technique is a way to normalize data by generating synthetic samples for the minority class.

This is done by first computing the difference between each feature value between the neighbour and the template sample. The difference is then multiplied by a random number between 0 and 1, and the result added to the corresponding feature value. As a result, a new synthetic feature value is generated. The results of performing SMOTE on the dataset can be expressed as follows, with 0 signifying non-stroke cases and 1 signifying stroke cases:

Ratio of stroke instances before SMOTE: 0: 3786, 1: 198

Ratio of stroke instances after SMOTE: 0: 3786, 1: 3786

Resampled dataset shape: (7572 rows, 11 columns)

Although a perfect 1:1 ratio is unrealistic in real-world scenarios, it is acceptable for the purposes of this study.

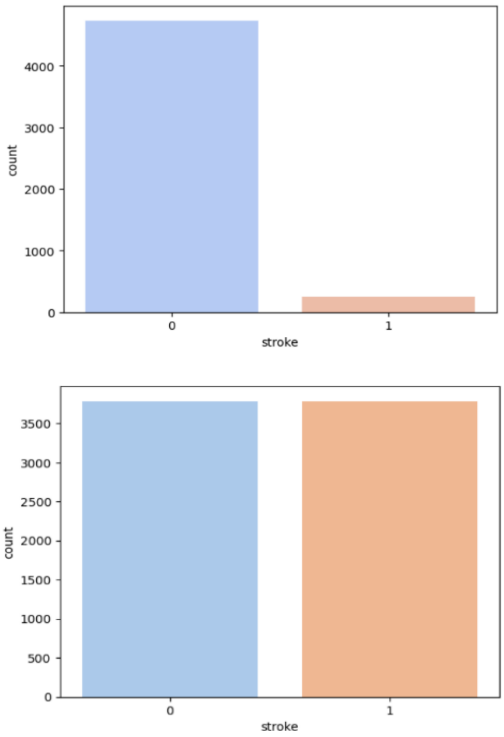


Figure 4 Distribution of classes before and after SMOTE

In studying the data for feature selection purposes, the greatest discrepancy in data is found in 'hypertension' and 'heart_disease' features where there is a class imbalance. However, these columns are still kept as they are important markers in stroke prediction. A Pearson Correlation Matrix is used to compute a heatmap showing the relationship between the different features. A correlation matrix is a table measuring the strength and direction of the linear relationship between a pair of variables. The Pearson Correlation Matrix assigns a value of -1 to +1 known as the Pearson Correlation coefficient indicating negative and positive correlation respectively.

Age is seen as having the strongest correlation with stroke, showing a Pearson Correlation coefficient of 0.58. As expected, age also has a correlation to being married or not. However the correlation is not strong enough to consider dropping one or the other feature.

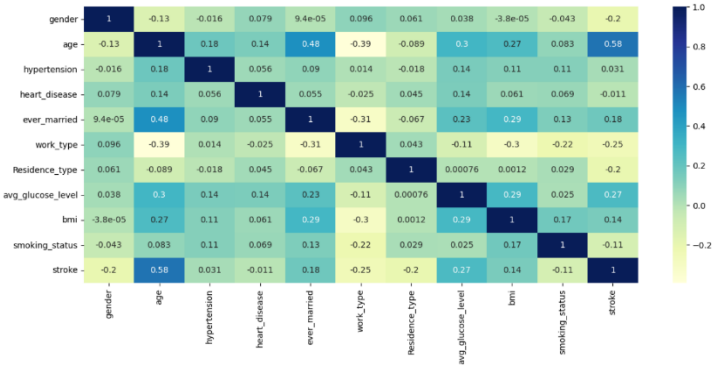


Figure 5 Correlation Heatmap

When the independent variables are correlated with the target variable, some of them show a negative correlation. In data analysis a fundamental belief is that 'correlation does not equal causation', therefore these features are kept to study their impact on stroke prediction. There are also not many features to begin with, so keeping the original number should not have any detrimental effects.

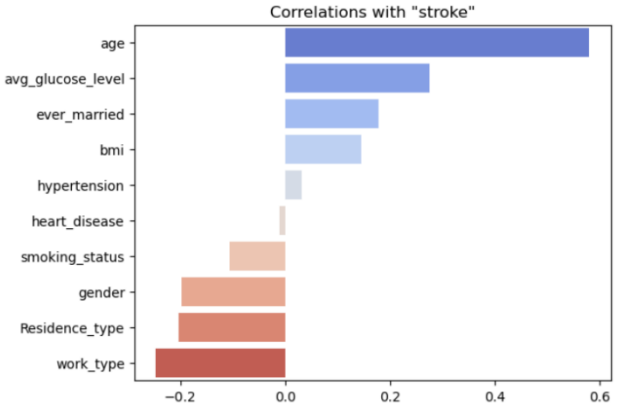
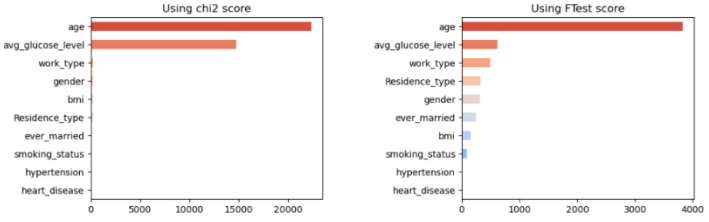


Figure 6 Correlation between features with 'stroke'

Four further methodologies are employed to evaluated the relationship between the independent variables and the target feature, namely the Chi2 score, FTest score, MI test score and ExtraTreesClassifier. The computed graphs illustrate that 'heart_disease' and 'hypertension' have the lowest impact on 'stroke', however this may also be because they have the lowest number of instances. For this study, none of the columns have been dropped as both 'heart_disease' and 'hypertension' are important contributing factors to a brain stroke.



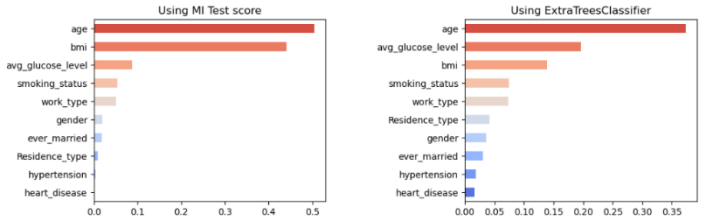


Figure 7 Correlation with target feature 'stroke'

METHODOLOGIES

As the dataset contains labelled features, this study takes a supervised learning approach. Labelled data is that which contains example data points alongside correct outputs. Supervised learning is a technique in machine learning using labelled datasets to train artificial intelligence algorithms for identifying relationships and underlying patterns between the input features and their outputs. Supervised learning techniques solve two problem categories: regression and classification. The process undertaken in this study is to create models from each category that can predict outputs by taking in real-world data as inputs.

The steps to a supervised machine learning process are as follows:

1. Identifying the training data type to use for training the model. The training data must be similar to the data intended to be input into the model for processing when it is complete.
2. Assembling the data and ensuring the dataset is correctly labelled. Data must also be checked for imbalance which could affect model performance.
3. The creation of three sets of data: training, validation, and test data. Validation data is used to assess the training in case it requires further fine-tuning and testing evaluates the final model. For the purposes of this study, the existing training dataset will be split to accommodate the training and testing data.
4. Choosing a suitable machine learning algorithm for creating the model.
5. Validating and testing the model.
6. Observing the model performance to maintain accuracy and finetune it as required [11].

LOGISTIC REGRESSION

Logistic regression is a statistical model often used in predictive and classification analytics. This model estimates the probability of an event occurring based on a dataset of independent variables. As the outcome is a probability, the dependent variable is

computed between 0 and 1 – in other words, it is designed to predict binary outcomes. In logistic regression, the probability of success is divided by the probability of failure, which is the application of a logit transformation on the odds. This logistic function is expressed by the formulas:

$$Logit(pi) = 1/(1+ exp(-pi))$$

$$ln(pi/(1-pi)) = Beta\_0 + Beta\_1*X\_1 + ... + B\_k*K\_k$$

Here, logit(pi) is the dependent variable and x is the independent variable.

Logistic regression is chosen as the first machine learning algorithm as the study relies on binary outcomes, namely 'does not have stroke' (0) and 'has stroke' (1). It is also fast and effective on larger samples of data. The disadvantage of logistic regression is that the model can be prone to overfitting, but this has been addressed in a previous section of the study [12].

K-Nearest Neighbours

K-Nearest Neighbours is a supervised learning algorithm which makes classifications or predictions about how an individual data point will be grouped based on proximity. It can be used for both regression and classification problems, but is typically used for the latter. In this algorithm, the k value defines how many neighbours are checked in determining the classification of the target. For instance, a value of k=1 denotes that the instance will be assigned to the same class as its single nearest neighbour.

Figure 8 K-Nearest Neigbours Classification

K-Nearest Neighbours has been chosen as the second machine learning algorithm as it is easy to implement and requires relatively few hyperparameters. Since it is a 'lazy algorithm' it is prone to taking up more memory and data storage. That will not be an issue in this study as there are >7000 data samples which can be easily handled by the model [13].

Testing

After analyzing the data and correcting the problem of oversampling using the SMOTE technique, the training dataset is prepared for testing with the two models. For both models, the dataset is first split in a 80:20 ratio for training and testing purposes respectively.

```
# Splitting the data into training and testing sets (80% train, 20% test):
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

Figure 9 Splitting of dataset

The features are standardized using Sci-kit's inbuilt StandardScaler() function, and both models accuracy score is determined. A classification report is produced, highlighting the precision, recall, F1 score and support scores, which are metrics which will help evaluate the predictive output of each model. Finally, a confusion matrix is computed to provide a visual representation of the model's predictive accuracy.

Using Sci-kit Learn to perform Logistic Regression, the following outputs are generated:

```
Accuracy: 0.8805

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.82      0.87       758
           1       0.84      0.94      0.89       757

    accuracy                           0.88      1515
   macro avg       0.89      0.88      0.88      1515
weighted avg       0.89      0.88      0.88      1515
```

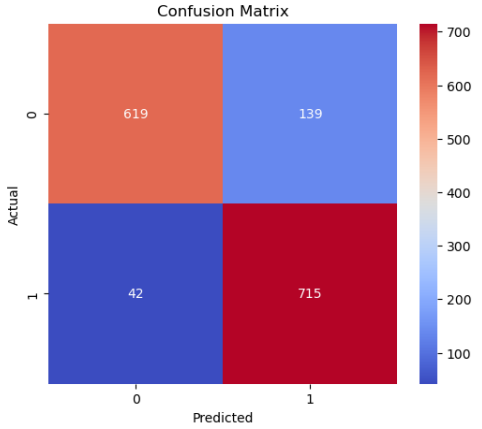


Figure 10 Classification report and confusion matrix for Logistic Regression

The findings from this output are as follows:

- The model has an accuracy score of 0.8033, meaning it correctly predicts the class ~80% of the time.
- Precision 0.80: This signifies when the model predicts a class, it is correct 80% of the time.

- Recall 0.80: Signifies the model correctly identifies 80% of actual positives.

- F1-score 0.80: This is the harmonic mean. Since both precision and recall are 0.80, it indicates a balanced model.

- The confusion matrix indicates significantly lower instances of false positives and false negatives, which means the model is correctly predicting true positive and true negative in more instances.

- Overall the model performance is decent, but there is still further room for improvement.

Next, Sci-Kit Learn is used again to perform K-Nearest Neighbours algorithm, and the following outputs are generated:

```
Accuracy: 0.8033

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.80      0.80       758
           1       0.80      0.80      0.80       757

    accuracy                           0.80      1515
   macro avg       0.80      0.80      0.80      1515
weighted avg       0.80      0.80      0.80      1515
```

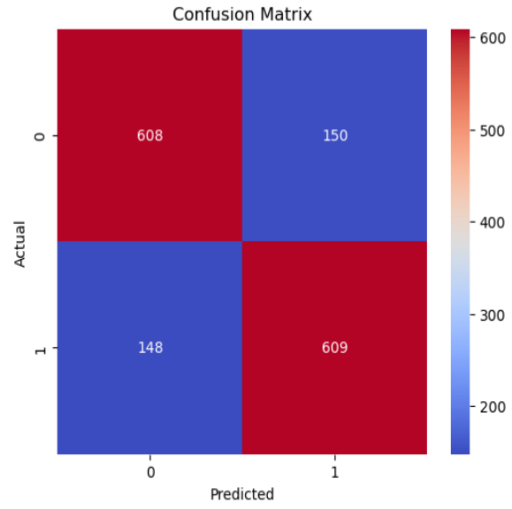


Figure 11 Classification report and confusion matrix for K-Nearest Neighbours

The findings from this output are as follows:

- The model has an accuracy score of 0.8805, meaning it correctly predicts the class ~88% of the time.

- Precision 0, 0.94 and 1, 0.82: This signifies when the model predicts a negative class, it is correct 94% of the time. When it predicts a positive class, it is correct 82% of the time. The model is better at predicting negative classes.

- Recall 0, 0.82 and 1, 0.94: Alternatively, the model correctly identifies actual '0' cases 82% of the time and actual '1' cases 94% of the time, showing it is better at correctly identifying positive classes.

- F1-score 0.88 and 0.89: The F1 score is strong for both classes and indicative of a balanced model.

- The confusion matrix indicates significantly lower instances of false positives and especially false negatives, and conversely achieves better prediction scores for true positive and true negative cases.

- The higher accuracy score and improved overall metrics of this model suggests it may be capturing better decision boundaries and performs better than Logistic Regression. There is, however, still room for further improvement.

CROSS VALIDATION

As confusion matrix and the classification report are more suited to classification use cases, a cross-validation technique is also implemented which is optimized for regression use cases. Using k-fold cross validation method with a value of cv=5, the following results are achieved:

```
Linear Regression R² Score: 0.4326 ± 0.0184
KNN Regression R² Score: 0.6717 ± 0.0239
```

Figure 12 Output for cross validation

R^2 is the coefficient of determination which measures how well a regression model is explaining variability in a target model. A higher value means the model is a better fit. The findings from this output are as follows:

- K-Nearest Neighbours has a significantly better R^2 score than Logistic Regression, indicating that the relationship between the features for predicting stroke are non-linear.

- The K-Nearest Neighbours model is better suited to capturing complex and non-linear patterns in data.

- For this particular dataset, classification is a better fit than regression.

III. Conclusion

This study set out to train and test two machine learning models to predict brain stroke given a set of independent variables. When treated as a classification problem, both models perform fairly well on the given data, achieving an accuracy score above 80% with K-Nearest Neighbours giving better results.

The statistics on stroke attacks is alarming, and its unpredictability is a major concern in the domain of health and medicine. This study has attempted to provide a relatively cheap, easy and quick means of predicting stroke attacks across a wide range of age groups with varying health. This is significant, as the vast majority of stroke attacks take place in underdeveloped nations which may not have the means to perform MRIs and CT scans. It has also given some interesting insights, such as the correlation of getting a stroke attack to whether an individual is married or not.

Limitations

The key limitation was the dataset itself which had extremely unbalanced class distribution in the target feature. When the dataset is tested without applying any normalization techniques such as SMOTE, both models fail despite having accuracy scores of over 90%. This is because they fail to predict the positive class entirely and are biased towards the majority negative class.

Another limitation is the absence of features in the dataset such as a history of stroke attacks, family history of stroke, ethnicity etc.

Improvements

The models can be improved with more stringent feature engineering such as removing features that do not contribute to the prediction or generating new ones. More attention should have been given into identifying potential outliers using methods such as boxplots. Normalization techniques such as SMOTE could also be applied to other features such as 'hypertension' and 'heart_disease' which show imbalanced classes. Gathering more training samples or using data augmentation techniques to increase training data will also lead to improvements in model performance.

As the features do not have a linear relationship in predicting the target, more sophisticated models such as Gradient Boosting (XGBoost or CatBoost) and Neural Networks (MLP) can be applied to capture and analyze the dataset in more complicated ways.

References

[1] Feigin VL, Brainin M, Norrving B, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. International Journal of Stroke. 2022;17(1):18-29. doi:10.1177/17474930211065917

[2] Stroke Association. (n.d.). Available at: https://www.stroke.org.uk/ [Accessed 2 Mar. 2025].

[3] Ryding, S. (2020). Biomarkers for Prediction of Stroke. [online] News-Medical. Available at: https://www.news-medical.net/health/Biomarkers-for-Prediction-of-Stroke.aspx#:~:text=Stroke%20can%20be%20very%20hard,%2C%20thrombosis%2C%20and%20oxidative%20stress. [Accessed 2 Mar. 2025].

[4] Lee, M., Ryu, J. and Kim, D.-H. (2019). Automated epileptic seizure waveform detection method based on the feature of the mean slope of wavelet coefficient counts using a hidden Markov model and EEG signals. Etri Journal, [online] 42(02), pp.217–229. Available at: https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2018-0118#citedby-section [Accessed 2 Mar. 2025].

[5] Yaacoub Chahine, Matthew J Magoon, Bahetihazi Maidu, Juan C del Álamo, Patrick M Boyle, Nazem Akoum, Machine Learning and the Conundrum of Stroke Risk Prediction, Arrhythmia & Electrophysiology Review 2023;12:e07. https://doi.org/10.15420/aer.2022.34

[6] Liu Y, Yin B, Cong Y. The Probability of Ischaemic Stroke Prediction with a Multi-Neural-Network Model. Sensors (Basel). 2020;20(17):4995. Published 2020 Sep 3. doi:10.3390/s20174995

[7] Cuadrado-Godia, E., Jamthikar, A.D., Gupta, D., Khanna, N.N., Araki, T., Maniruzzaman, M., Saba, L., Nicolaide, A., Sharma, A., Omerzu, T., Suri, H.S., Gupta, A., Mavrogen, S., Turk, M., Laird, J.R., Protogerou, A., Sfikakis, P., Kitas, G.D., Viswanathan, V. and Suri, J.S. (2019). Ranking of stroke and cardiovascular risk factors for an optimal risk calculator design: Logistic regression approach. Computers in Biology and Medicine, 108, pp.182–195.

[8] Annas, S., Aswi, A., Abdy, M. and Poerwanto, B. (2022). Binary Logistic Regression Model of Stroke Patients: A Case Study of Stroke Centre Hospital in Makassar. Indonesian Journal of Statistics and Its Applications, 6(1), pp.161–169.

[9] Eldora, K., Fernando, E., & Winanti, W. (2024). Comparative Analysis of KNN and Decision Tree Classification Algorithms for Early Stroke Prediction: A Machine Learning Approach. Journal of Information Systems and Informatics, 6(1), 313-338. https://doi.org/10.51519/journalisi.v6i1.664

[10] Turkalp Akbasli, I. (2022). https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset/data. [online] Kaggle. Available at: https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset/data [Accessed 1 Mar. 2025].

[11] Belcic, I. and Stryker, C. (2024). What is supervised learning? [online] IBM.com. Available at: https://www.ibm.com/think/topics/supervised-learning [Accessed 6 Mar. 2025].

[12] IBM (2024). https://www.ibm.com/think/topics/logistic-regression. [online] IBM.com. Available at: https://www.ibm.com/think/topics/logistic-regression [Accessed 7 Mar. 2025].

[13] IBM (2024b). What is the KNN algorithm? [online] IBM.com. Available at: https://www.ibm.com/think/topics/knn [Accessed 7 Mar. 2025].