Mohammad Rahman

STROKE PREDICTION USING MACHINE LEARNING TECHNIQUES

Thesis

CENTRIA UNIVERSITY OF APPLIED SCIENCES

INFORMATION TECHNOLOGY

DECEMBER 2023

ABSTRACT

Centria University of Applied Sciences	**Date** December 2023	**Author** Mohammad Rahman
Degree programme Information Technology		
Name of thesis STROKE PREDICTION USING MACHINE LEARNING TECHNIQUES		
Centria supervisor Aliasghar Khavasi		**Pages** 33 + 6

People today are affected by a wide range of diseases as an impact of the current state of the environment and human lifestyle choices. Early detection and prediction of such diseases are necessary to prevent them from progressing to their final stages. Stroke, a cerebrovascular illness, is one of the leading causes of death and a significant financial burden on patients. Health-related behavior, which is becoming an increasingly important focus of prevention, is one of the major risk factors for stroke. The risk of stroke has been predicted using a variety of machine learning algorithms, which also include predictors such as lifestyle variables to automatically diagnose stroke. Five supervised machine learning classifiers, including Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes, and K-Nearest Neighbor Algorithm are utilized in this study to predict strokes. The dataset, consisting of 5110 items with 10 attributes, is preprocessed to make it suitable for prediction, after which the aforementioned classifiers are trained on the data, and the confusion matrix is used to evaluate the performance of the classifiers. With an accuracy of 95.8%, the RF algorithm outperformed all others in the used dataset for predicting strokes based on several physiological parameters. The clinical estimation of stroke using machine learning algorithms can be more effective when compared to a person's medical background and physical activity. In addition to all of these diagnoses, stroke patients require ongoing intensive care, which can be offered by an interdisciplinary team.

Keywords

Confusion Matrix, Machine Learning, Random Forest Classifier, Stroke

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the ROC Curve
DT	Decision Tree
DNN	Deep Neural Network
EMG	Electromyography
HANN	Hybrid Artificial Neural Network
KNN	K-Nearest Neighbor
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NB	Naïve Bayes
RF	Random Forest
RLR	Regularized Logistic Regression
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence

Contents

FIGURES

TABLES

1 INTRODUCTION

A stroke, also known as a brain attack, happens when a blood vessel in the brain breaks or when something stops the flow of blood to a specific area of the brain. The brain is an organ that manages human bodily activities, retains their memories, and generates our ideas, feelings, and verbal expression. In addition, the brain regulates a variety of bodily processes, including respiration and digestion. We need oxygen for our brains to function correctly. All the areas of our brain receive oxygen-rich blood from the arteries. Brain cells begin to die within minutes of a blockage in blood flow because they are unable to receive oxygen causing a stroke. This can result in long-term impairment, permanent brain damage, or even death.

The second largest cause of death and the primary cause of disability worldwide is stroke. According to the Global Stroke Factsheet published in 2022, the risk of having a stroke over the course of a person's lifetime has increased by 50% in the past 17 years, with 1 in 4 people considered to be at risk. Stroke incidence, deaths from stroke, prevalence, and Disability Adjusted Life Years (DALY) all rose by 70%, 43%, 102%, and 143%, respectively, over the years 1990 and 2019. The most notable aspect is that lower- and lower-middle-income nations bear the heaviest burden of stroke globally (86% of stroke-related deaths and 89% of DALYs). Families with limited resources are facing a new struggle because of this disproportionate financial stress felt by lower- and lower-middle-income countries. (World Stroke Day 2022.)

In Europe, stroke is the leading cause of disablement among adults, and it can have an impact on several areas of daily life. It is estimated that 12 million people will have a stroke in Europe by 2040, a rise from the current nine million. This would place an even higher burden on social services, health care, families, and providers. The first-ever European Life After Stroke Forum took place in person on March 10, 2023, in Barcelona. Its goal is to draw attention to this underdeveloped region of the stroke care pathway as a means of increasing interest in and awareness of care and support for people who have had a stroke as well as giving it status on scale with acute care and rehabilitation. (World Stroke Organization 2023.)

The primary root cause of stroke is a clogged artery or a blood vessel that is leaky or bursts. There may only be a temporary reduction in blood flow to the brain in some patients, with no long-lasting effects. From infants to adults, anyone can have a stroke, even though some people are more at risk than others. About two-thirds of strokes occur in persons over the age of 65, making them more prevalent in later life. Stroke risk factors include illnesses like heart disease, diabetes, high blood pressure, smoking, excessive drinking, and several other health issues (Centers for Diseases Control and Prevention 2023). There are three different forms of stroke, with some being more harmful and likely to leave a person disabled than others. These include ischemic strokes, hemorrhagic strokes, and transient ischemic attacks (mini-strokes). (Holek, n.d.)

The most frequent type of stroke in the general population is an ischemic stroke, which is usually caused by a blood clot that stops or plugs a blood vessel in the brain. This results in a blood supply interruption, which prevents oxygenated blood from reaching certain areas of the brain. When the supply of blood and oxygen to the nerve cells becomes disrupted, they begin to malfunction and die. The various challenges that a stroke survivor could face because of the stroke are caused by these damaged areas of the brain. A hemorrhagic stroke occurs when a brain artery bursts or releases blood. Blood leakage exerts too much pressure on brain cells, significantly increases intracranial pressure, and causes the brain to swell or hemorrhage. This pressure results in some cells dying or being injured. Transient Ischemic Attacks are known as "mini-strokes" because they are usually less harmful and dangerous than other types of strokes. In this instance, the blockage of blood flow to the brain lasts no longer than five minutes. However, it is seen as an indicator of further strokes and should not be ignored.

Loss of brain function results from dying brain cells. It is possible that the patient will not be able to perform tasks that require that brain region. For instance, a stroke may impair a person's ability to move, speak, eat, think, and remember, as well as their capacity to control their emotions and other important bodily functions. The patient may have difficulty smiling, his face could drop to one side, he could be struggling to raise both arms and maintain them there, and his speech might sound slurred. These three are the most common symptoms of stroke. Other symptoms may include total paralysis of one side of the body, abrupt loss or blurring of vision, dizziness, confusion, difficulty understanding what others are saying, issues with balance and coordination, difficulty swallowing (dysphagia), a sudden and extremely severe headache accompanied by a blinding pain unlike anything previously felt, loss of consciousness. Some of the strokes are mentioned in FIGURE 1.

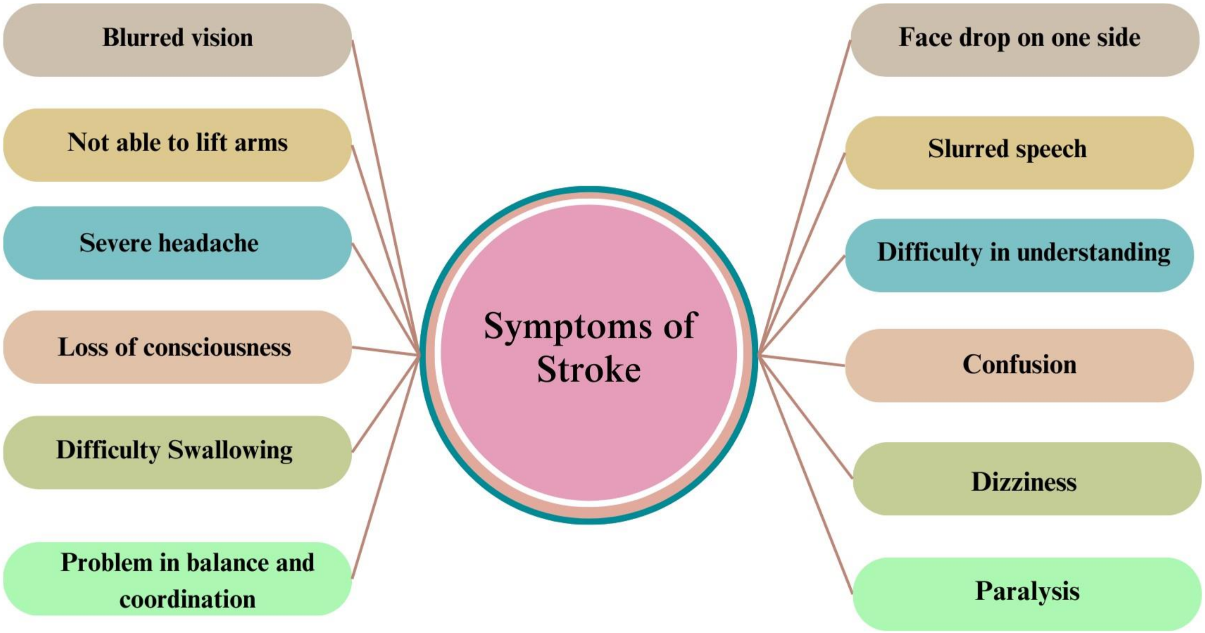

FIGURE 1. Symptoms of stroke

Anyone can experience a stroke at any time. A stroke is a serious life-threatening medical condition that happens when the blood supply to part of the brain is cut off (NHS 2022). However, certain people are more susceptible to suffering a stroke. While some stroke risk factors can be altered or managed, others cannot. High blood pressure, heart disease, diabetes, smoking, birth control pills (oral contraceptives), a history of TIAs (transient ischemic episodes), high blood cholesterol and lipids, and obesity, Inactivity, obesity, excessive alcohol use, irregular heartbeat, anatomical abnormalities of the heart, family history of stroke, and older age are risk factors for stroke that can be altered, treated, or medically controlled. The hospital's medical emergency team will attempt to identify the type of stroke the patient is suffering. The patient may undergo several blood tests, including those to determine how quickly blood clots, whether blood sugar levels are too high or low, and whether the patient has an infection. A computerized tomography (CT) scan produces a precise image of the brain using a series of X-rays. It may reveal brain hemorrhage, an ischemic stroke, a tumor, or other diseases. Strong radio waves and magnets are used in magnetic resonance imaging (MRI) to provide a detailed image of the brain. It can identify cerebral hemorrhage and ischemic stroke-related brain tissue damage. Carotid ultrasound is used to obtain precise pictures of the interior of carotid arteries. This test reveals blood flow in the carotid arteries as well as the accumulation of fatty deposits (plaques). An echocardiogram helps identify the origin of heart clots by producing precise images of the heart. Strokes may have been triggered by the clots' passage from the heart to the brain. The most common traditional diagnosis of stroke is mentioned in FIGURE 2.

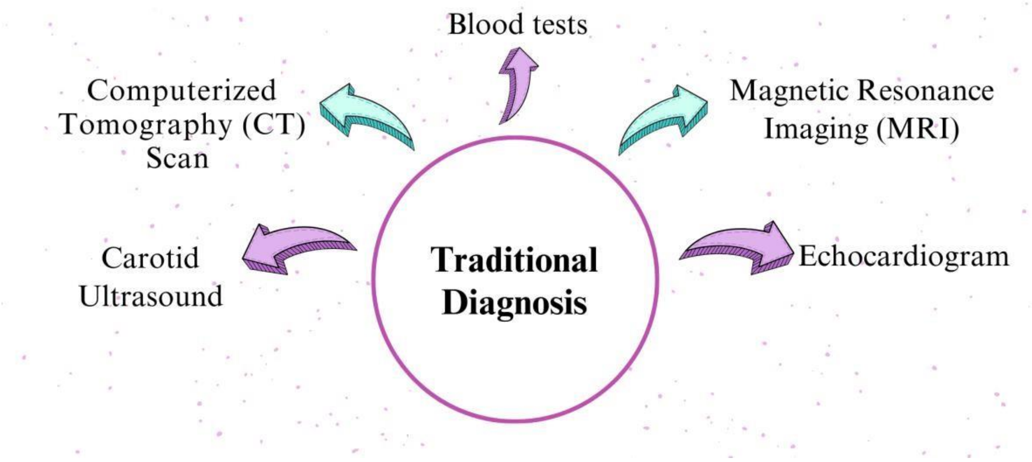
Blood tests
Computerized Tomography (CT) Scan
Magnetic Resonance Imaging (MRI)
Carotid Ultrasound
Traditional Diagnosis
Echocardiogram

FIGURE 2. Diagnosis of Stroke

2 LITERATURE REVIEW

A lot of researchers have already employed machine learning-based methods to forecast strokes. To estimate the stroke, the machine learning classification techniques Naive Bayes Classification, Support Vector Machine, Logistic Regression, Decision Tree Classification, Random Forest Classification, and K-Nearest Neighbors were used. Along with a web application, an HTML page and a Flask were also created to allow users to input values for predictions. 82% accuracy was obtained with this model. The researcher's use of textual data to train the model rather than actual brain images was a shortcoming in this study (Ojhai & Jha 2023). Deep neural networks, random forests, and logistic regression were used to forecast long-term outcomes for ischemic stroke patients. This study comprised 2604 patients in all, and 2043 (or 78%) of them had successful outcomes. The deep neural network model's area under the curve was noticeably larger than the ASTRAL score. (Yoon, Park & Kim 2019.)

For stroke prediction with unbalanced data, machine learning approaches with data balancing techniques are useful tools. Wu and Fang (2020) used regularized logistic regression (RLR), support vector machine (SVM), and random forest (RF) resulting in 78% accuracy. However, as this study's outcome variable was a self-reported stroke, there may be some subjective bias. A weighted voting classifier for predicting stroke was suggested utilizing factors like hypertension, body mass index, heart disease, average blood sugar, smoking status, prior stroke, and age. In comparison to the base classifiers, the weighted voting classifier performed better, with an accuracy of 97%. (Emon, Keya & Meghla 2020.)

Another study suggests a strategy for putting argumentation on top of ML to create Explainable Artificial Intelligence (XAI) models. XAI operates well and generates explanations that are understandable by humans. The accuracy was found to be 78% when the authors compared their findings to those of Random Forests and the SVM classifier that was considered best for the same dataset (Prentzas, Nicolaides & Kyriacou 2019). A nationwide disease registry was used to apply ML methods for 90-day stroke outcome predictions. The implementation and evaluation of SVM, RF, ANN, and a hybrid artificial neural network (HANN) used a 10-time repeated hold-out with 10-fold cross-validation. Using preadmission and inpatient data, ML methods provide over 0.94 AUC in both ischemic and hemorrhagic stroke. When follow-up data was included, the prediction accuracy increased to 0.97 AUC. (Lin, Hsu & Johnson 2020.)

Another study established a stroke prediction system that uses artificial intelligence (AI) and real-time bio signals to identify stroke. Real-time EMG (Electromyography) bio-signals from the thighs and calves were gathered, the key features were identified, and prediction models based on regular activities were created. For this suggested system, prediction accuracy values of 90.38% for Random Forest and 98.958% for LSTM were found (Yu, Park & Kwon 2020). The relevant data was extracted from the raw data using tagging and maximum entropy approaches. The processed dataset was retrieved using a brand-new stemming technique. By emphasizing the variation in the dataset, the kind of stroke was identified with respectable accuracy. ANN, SVM, DT, Logistic Regression (LR), Bagging, and Boosting are the methods that are utilized. With a classification accuracy of 95% and a standard deviation of 14.69, artificial neural networks trained with a stochastic gradient descent approach surpassed the other algorithms. (Govindarajan, Soundarapandian & Gandomi 2019.)

Multilayer perceptron (MLP) neural networks are used to analyze 584 patients with stroke to predict the mortality in these people. Quick Propagation (QP), Levenberg-Marquardt (LM), Back Propagation (BP), Quasi-Newton (QN), Delta Bar Delta (DBD), and Conjugate Gradient Descent (CGD) are six different MLP algorithms that were used. The highest accuracy (80.7%), sensitivity (78.4%), and specificity (81.3%) were all attained by QP (SÜT & ÇELİK, 2012). A stroke prediction model was created using data mining with a sample of 147 stroke patients and 294 non-stroke people using demographic data and medical screening data. The study included three classification algorithms: Artificial Neural Network (ANN), Decision Tree, and Naive Bayes. The ANN with integrated data was the algorithm that performed the best. This method had a 0.84 overall accuracy rate, a 0.90 AUC, a 0.12 FPR, and a 0.25 FNR. (Thammaboosadee & Kansadub, 2019.)

To assist in medical diagnosis, a novel method of segmenting the stroke using cranial CT scans is offered. The level set method is automatically initiated within the stroke region, and the stroke is segmented using a nonparametric estimate method based on the Parzen window. Using fuzzy C-means, the outcomes of the level set algorithms are compared to those of the suggested approach. The lowest standard deviation was 0.08% and the highest accuracy mean was 99.84%. (Rebouças, Marques & Braga, 2018.) In predicting motor outcome in the upper and lower limbs at 6 months following stroke, a deep neural network (DNN) model implemented 2 well-known ML methods, logistic regression, and random forest, were used. The Area Under the Curve (AUC) for the DNN model's prediction of upper limb function was 0.906 in this case. The AUC was 0.874 for the logistic regression model and 0.882 for the random forest model, respectively. The AUCs for the DNN, logistic regression, and random forest models for the prediction of lower limb function were 0.822, 0.768, and 0.802, respectively. (Kim, Choo & Chang, 2021.)

In another study, the use of machine learning algorithms to forecast Early Neurological Deterioration (END) in patients with acute mild stroke was studied. Boosted, deep neural networks, and logistic regression all produced results with an accuracy of 96.6%. (Sung, Kang & Cho, 2020.) For predicting functional outcomes in patients with acute ischemic stroke following endovascular therapy, machine-learning algorithms were examined for their predictive accuracy. The dataset was trained using a variety of machine learning and regression models, such as the Random Forest (RF), Classification and Regression Tree (CART), C5.0 Decision Tree (DT), Support Vector Machine (SVM), Adaptive Boost Machine (ABM), Least Absolute Shrinkage and Selection Operator (LASSO) and logistic regression models. When validated internally, AUC range = 0.65-0.72, MCC range = 0.29-0.42, and externally, AUC range = 0.66-0.71, MCC range = 0.34-0.42. Both logistic regression and machine learning models demonstrated similar predictive accuracy. (Alaka, Menon & Brobbey 2020.)

Another study also used Machine learning to develop a generic methodological pipeline for future analysis as well as to discover the best reliable stroke prediction approach in a Chinese hypertensive population. The Random Under Sampling (RUS)-applied RF model with laboratory variables showed the best model performance. Data balancing approaches enhanced overall performance with RUS compared to null models (sensitivity = 0, specificity = 100, and mean AUCs = 0.643), exhibiting a more satisfying outcome in the current study (RUS: sensitivity = 63.9, specificity = 53.7, and mean AUCs = 0.624). (Huang, Cao & Chen, 2022.)

TABLE 1. Summary of the related work in prediction of stroke using machine learning techniques.

Paper Reference	**Contribution**	**Methodology**	**Results**

(Ojhai & Jha 2023.)	A web application, an HTML page and a Flask were created to allow users to input values for predictions.	Naive Bayes Classification, Support Vector Machine, Logistic Regression, Decision Tree Classification, Random Forest Classification, and K-Nearest Neighbors	82% accuracy
(Yoon et al. 2019)	Long-term outcomes for ischemic stroke patients were predicted	Deep neural networks, random forests, and logistic regression.	AUC of Deep Neural Network (DNN) ASTRAL score
(Wu &Fang 2020)	Data balancing techniques are used to predict stroke with unbalanced data.	Regularized Logistic Regression (RLR), Support Vector Machine (SVM), and Random Forest (RF)	78% accuracy
(Emon et al. 2020)	A weighted voting classifier for predicting stroke was suggested	Voting classifier	97% accuracy
(Prentzas et al. 2019.)	A strategy for putting argumentation on top of ML was suggested to create Explainable Artificial Intelligence (XAI) models	Random Forests and the SVM classifier	78% accuracy

(Lin et al. 2020.)	To utilise ML techniques for 90-day stroke outcome predictions, a national illness registry was used.	SVM, RF, ANN, and a hybrid artificial neural network (HANN)	0.97 AUC
(Yu et al. 2020.)	A stroke prediction system was established that uses artificial intelligence (AI) and real-time biosignals to identify stroke	Random Forest, Long Short Term Memory (LSTM)	98.958% accuracy
(Govindarajan et al. 2019.)	Artificial neural networks trained with a stochastic gradient descent approach surpassed the other algorithms	ANN, SVM, DT, Logistic Regression (LR), Bagging, and Boosting	95% accuracy
(SÜT & ÇELİK, 2012)	Mortality in stroke patients are predicted using ML algorithms	Six types of MLP Neural Networks: Quick Propagation (QP), Levenberg-Marquardt (LM), Back Propagation (BP), Quasi Newton (QN), Delta Bar Delta (DBD), and Conjugate Gradient Descent (CGD	80.7% accuracy, 78.4% sensitivity and 81.3% specificity

(Thammaboosadee & Kansadub, 2019.)	Data mining technique was used to create a stroke prediction model.	Artificial Neural Network (ANN), Decision Tree, and Naive Bayes. Adaptive	84% accuracy, 0.90 AUC, 0.12 FPR, and 0.25 FNR
(Rebouças et al. 2018).	A novel method of segmenting the stroke using cranial CT scans is offered.	Parzen window, Fuzzy Cmeans	99.84% accuracy
(Kim et al. 2021.)	Motor outcome in the upper and lower limbs at 6 months following stroke was predicted.	Deep neural network (DNN) model, Logistic Regression and Random Forest	Upper Limb: Area Under the Curve (AUC) for DNN,LR and RF are 0.906, 0.874 and 0.882 respectively. Lower Limb: The (AUC) for DNN,LR and RF are 0.822, 0.768, and 0.802, respectively
(Sung et al. 2020.)	Early Neurological Deterioration (END) in patients with acute	Boosted trees, Deep neural network, and Logistic Regression	96% accuracy
	mild stroke was forecasted using ML		

(Alaka et al. 2020.)	Machine-learning algorithms were evaluated for their predictive efficacy in predicting functional outcomes in patients with acute ischemic stroke after endovascular therapy.	Random Forest (RF), Classification and Regression Tree (CART), C5.0 Decision Tree (DT), Support Vector Machine (SVM), Adaptive Boost Machine (ABM), Least Absolute Shrinkage and Selection Operator (LASSO) and logistic regression models	internally (AUC range = [0.65-0.72]; MCC range = [0.29-0.42]) and externally (AUC range = [0.66-0.71]; MCC range = [0.34-0.42])
(Huang et al. 2022.)	The best reliable stroke prediction approach in a Chinese hypertensive population and a generic methodological pipeline for future analysis are developed using machine learning.	Random Under Sampling (RUS) applied RF	RUS: sensitivity = 63.9, specificity = 53.7, and mean AUCs = 0.624

3 METHODOLOGIES

This chapter is divided into five parts, these are data description, data preprocessing, machine learning algorithms, Evaluation matrices and working flowchart. The implantation procedure is explained in detail along with necessary figures in this section.

3.1 Data Description

This dataset has been collected from the website Kaggle to estimate whether a patient is likely to suffer from a stroke. It is the document of 5110 people's information including 10 attributes (FEDESORIANO 2021). One of the features is Gendre. This attribute refers to a person's gender. It is categorical data involving Male, Female, or Other. Another one is Age. This attribute refers to a person's age. It is numerical data. Another one is hypertension. This attribute means whether a person is hypertensive or not. 0 is for the patient who does not have hypertension, and 1 is for the patient who has hypertension. It is numerical data. Another one is heart disease. This attribute means whether a person has heart disease or not. 0 is for the patient who does not have any heart disease, and 1 is for the patient who has a heart disease. It is numerical data. Another one is ever married; this attribute represents a person's marital status. It is categorical data involving No or Yes. Another one is work type. This attribute represents the person's work scenario. It is categorical data involving children, Gov-jobs, never worked, Private or Self-employed.

Another one is the Residence type. This attribute represents the person's living scenario. It is categorical data involving Rural or Urban. Another one is the glucose level. This attribute represents the average glucose level in blood of a patient level in blood. It is numerical data. Another one is smoking status. This attribute represents a person's smoking status. It is categorical data involving formerly smoked, never smoked, smokes, or Unknown. Another one is stroke. This attribute means a person previously had a stroke or not. It is numerical data. 1 means that the patient had a stroke and 0 is the opposite. Among all these attributes, stroke is the decision class, and the rest of the attribute is the response class.

3.2 Data Preprocessing

In data preprocessing there are several steps which must go through, the 1st one is Null checking. In programming, there can be events in which a variable may not be assigned a value. Such events are frequently represented by the unique value null. The dataset's null or missing values can be checked using the Pandas function is null. Only the rows with null values are presented since the null values are assigned to True values. In this study, some null values are found in the dataset. These null values are removed using the dopna function. Another step is Converting categorical data into numerical data. The data set's categorical variables should be transformed into numerical values. Several machine learning algorithms can support categorical values naturally, but many more algorithms cannot. For this reason, Label Encoding and the One Hot Encoding approach are used for these transformation procedures. (Analytics Vidhya, 2020.)

Another step is the Train-Test Split. The Train-Test Split technique was adopted to evaluate the performance of a model. It divides some percentage of the data for training the model and the rest for testing the model so that it can compare my machine-learning model to the actual machine-learning. The model selection module is provided in the sci-kit-learn library in which the function train_test_split resideds (Scikit-learn.org, 2018). In our model, we have split the training set into 80% and the test set into 20% of the dataset. The last step is Feature scaling. A technique for normalizing the number of independent variables or features in data is called feature scaling. It is typically carried out during the data preprocessing step and is sometimes referred to as data normalization in the context of data processing. Each feature's value in the data is standardized so that it has a zero mean and unit variance. The fundamental approach to computation is to identify the distribution mean and standard deviation for each feature, then use the formula below to generate the new data point. (Vashisht.R, 2021.)

3.3 Machine Learning Algorithms

Predictive analytics and machine learning are both methods for making predictions that use historical data to predict future events. The performance of the trained prediction models is then evaluated using new data by comparing the actual values with the estimated outcomes. The algorithms used in this study are, Decision Tree, a decision support method that resembles a tree is called a decision tree. It has three parts decision nodes, leaf nodes, and a root node. A training dataset is divided into branches by a decision tree algorithm, which then further splits the branches into other branches. This pattern keeps on until a leaf node is reached. Further separation of the leaf node is not possible. The attributes utilized to forecast the outcome are represented by the nodes in the decision tree. Links to the leaves are provided by the decision nodes. (Engineering Education (EngEd) Program Section, n.d.)

The most important thing to keep in mind while developing a machine learning model is to select the optimal method for the dataset and task at hand. The two main benefits of adopting a decision tree are that it frequently simulates human decision-making processes, making it simple to understand and that the decision tree's justification is clear because it has a tree-like structure. (javaTpoint, 2021.) By importing the DecisionTreeClassifier class from the sklearn.tree library, the model is fit to the training set. The code is below.

```
from sklearn.tree import DecisionTreeClassifier
classifier= DecisionTreeClassifier(criterion='entropy',
random_state=0)   classifier.fit(x_train, y_train)
```

Code 1.

The classifier object was formed in the code above, and two key parameters were passed into it. One is criterion='entropy'. Criterion measures the quality of split, which is determined by information gained by entropy. Another one is random_state='0': for producing the random states (javaTpoint, 2021). Another algorithm is the Random Forest Classifier. In a random forest algorithm, there are many different decision trees. The random forest algorithm creates a forest that is trained via bagging or bootstrap aggregation.

Based on the predictions made by the decision trees, this algorithm determines the outcome. It makes predictions by averaging out the results from different trees. The accuracy of the result grows as the number of trees increases. The decision tree algorithm's shortcomings are eliminated with a random forest. It improves precision and lowers dataset overfitting. (Engineering Education (EngEd) Program | Section, n.d.)

Many different decision trees are used in a random forest system. Each decision tree has a root node, leaf node, and decision node. The result generated by a particular decision tree is represented by the leaf node of each tree. The majority voting method is used to choose the result. The final output of the system in this scenario is the output that most of the decision trees have selected (Engineering Education (EngEd) Program Section, n.d.). Random Forest systems are used by medical practitioners to diagnose patients. Patients are diagnosed based on an evaluation of their past medical experience. To determine the appropriate dosage for the patients, prior medical records are examined. By importing the Random Forest Classifier class from sklearn. ensemble library, the Random Forest algorithm is fit to the training data. The code is below.

```
from sklearn.ensemble import RandomForestClassifier
classifier           RandomForestClassifier(n_estimators=           10,
criterion="entropy")   classifier.fit(x_train, y_train)
```

Code 2.

The classifier object was formed in the code above, and two key parameters were passed into it.

One is n_estimators= This refers to the minimum quantity of trees needed for the Random Forest. 10 is the default value. I am free to choose any number, but I must address the overfitting problem. Another one is criterion = Its purpose is to examine the split's precision. Here, "entropy" has been used to measure knowledge gain (JavaTpoint, n.d.). Another algorithm is Naïve Bayes**,** a widely used supervised machine learning approach for classification applications like text classification is the Naive Bayes classifier. This method is predicated on the idea that given the class, the properties of the input data are conditionally independent, enabling the algorithm to predict outcomes rapidly and precisely. Naive Bayes classifiers are regarded as straightforward Bayes theorem-applied probabilistic classifiers in statistics. With the aid of this method, the classifier can perform better in challenging situations where the data distribution is ill-defined by estimating the probability density function of the input data using a kernel function. The naive Bayes classifier is thus an effective machine learning tool, particularly in text categorization, spam filtering, and sentiment analysis, among other applications. It is simple to create and is very beneficial for really large sets of data. Naive Bayes is well renowned for its efficiency and superior performance in many real-world applications, in addition to its simplicity. (Chauhan, 2022.). The Naïve Bayes model is fit to the training set. The code is below.

```
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(x_train, y_train)
```

Code 3.

The GaussianNB classifier is used to fit it to the training dataset. (JavaTpoint, n.d.)

Another algorithm that has been used in this study is Support Vector Machine**.** Support vector machines are thought of as a classification method. To distinguish between various classes, SVM creates a hyperplane in multidimensional space. It iteratively develops the best hyperplane, which is then utilized to reduce error. The main goal of SVM is to identify the maximum marginal hyperplane (MMH) that best separates the dataset into classes. Linear hyperplane cannot always be used to address an issue. The kernel trick, a method used by SVM, turns nonseparable issues into separable problems by giving them more dimensions. It is most helpful in problems with non-linear separation. A classifier can be created that is more accurate by using a kernel trick (Datacamp.com, n.d.). The SVM classifier is fitted to the training set. Importing SVC class from the Sklearn.svm library allows us to build the SVM classifier. The code is below.

```
from sklearn.svm import SVC # "Support vector classifier"
classifier = SVC(kernel='linear', random_state=0)
classifier.fit(x_train, y_train
```

Code 4.

I used kernel='linear' in the code above since I am generating an SVM for data that can be separated linearly. For non-linear data, I can alter it. The classifier is then fitted to the training dataset (x_train, y_train). (JavaTPoint, n.d.)

The classification and regression algorithms in the K-Nearest Neighbors (KNN) family are sometimes referred to as memory-based learning techniques. As a result of its incredibly accurate predictions, the KNN algorithm can compete with the most precise models. In applications that need great accuracy but refrain from contacting a model that can be read by humans, the KNN algorithm can be employed. 'Feature similarity' is a technique the KNN algorithm uses to forecast the values of any new data points. This implies that a value is assigned to the new point based on how much it resembles the points in the training set. The KNN parameter 'k' denotes the number of nearest neighbors. First, the distance between each training point and the new point is determined using the Euclidean, Manhattan, and Hamming distances, depending on whether the distance is continuous or categorical. Based on proximity, the nearest k data points are chosen. The last prediction for the new point is based on the average of these data points. (Analytics Vidhya, 2019.). By importing the KNeighborsClassifier class from the Sklearn Neighbours library, the K-NN classifier is fitted to the training set of data.

```
from sklearn.neighbors import KNeighborsClassifier   classifier=
KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2 )
classifier.fit(x_train, y_train)
```

Code 5.

After importing the class, the Classifier object of the class is created. The Parameters of this class are n_neighbors: It specifies the algorithm's necessary neighbors. It typically takes 5, metric='minkowski': The distance between the points is determined by this default setting and p=2: It is the same as the traditional Euclidean metric. (JavaTpoint, 2021)

3.4 Evaluation Metrics

Metrics for model evaluation play a role in evaluating the performance and accuracy of a model. These metrics help in evaluating the model's performance and in comparing it to other models or algorithms. In this work, the model is evaluated using the Confusion Matrix and its helpful measurements, such as Accuracy, Precision, Recall and F1-score. In Confusion Matrix, the evaluation of a machine learning model's performance on a set of test data is summarized by a confusion matrix. It is frequently used for measuring how well categorization models work. These models try to predict a categorical label for each input event. The matrix shows how many true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) the model generated using the test data (GeeksforGeeks, 2018). True denotes a precise prediction of the values, while False denotes an incorrect prediction. (Simplilearn.com, n.d.)

In True Positive, the number of times our real positive values match our positive predictions. In False Positive, the number of times a model mispredicts positive values as negatives.
Although it was expected a negative value, the outcome is positive. In True Negative, the number of times our real negative values match our negative predictions. In False Negative, the number of times a model predicts positive values for negative values in error. It is positive, contrary to what I had predicted. In accuracy, the percentage of values that were correctly categorized is determined using accuracy. It is the result of dividing the sum of all true values by all values. (Simplilearn.com, n.d.)

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN}$$

where, TP, TN, FP, and FN stand for True Positive, True Negative, False Positive and False Negative respectively. In precision, the model's accuracy in classifying positive values is determined by precision. It is the ratio of the true positives to all the predicted positive values. (Simplilearn.com, n.d.)

$$\text{Precision}= \frac{TP}{TP+FP}$$

where, TP and FP stand for True Positive and False Positive respectively.

In Recall, it is a measurement of the percentage of true positive cases (or actual positive cases) that were correctly predicted as positive. Another name for the recall is sensitivity. This suggests that there will be a further percentage of actual positive cases that are mistakenly forecasted as negative (and are hence sometimes referred to as the false negative). A false negative rate can also be used to demonstrate this. (Simplilearn.com, n.d.)

$$\text{Recall}= \frac{TP}{TP+FN}$$

where, TP and FN stand for True Positive and False Negative respectively.

A higher recall number would indicate a higher true positive and a lower false negative. Lower recall would translate to lower true positive and larger false negative values. Models with high sensitivity are preferred for the healthcare and finance industries. In F1-Score, Recall and precision are effectively combined to form this. When both precision and sensitivity must be considered, it is helpful (Simplilearn.com, n.d.).

$$\text{F1-Score}= \frac{2*Precision*Recall}{Precision+Recall}$$

The concept of correlation explains the relationships between one or more variables. These factors could be characteristics of the raw data that were utilized for predicting our target variable. An accurate correlation analysis improves data understanding. It plays a significant role in locating the important variables. There are several different formulas for calculating correlation coefficient, but Pearson's correlation also known as Pearson's R, which is frequently used for linear regression, is one of the most well-known. The letter "R" stands for the Pearson's correlation coefficient. The correlation coefficient formula gives a result that ranges from -1 to 1. (GeeksforGeeks, 2022.)

A score of -1 indicates a very poor relationship. A score of 1 indicates extremely positive connections. A score of 0 means there is no association at all. It measures how well classification models perform when they make predictions based on test data and determines the level of accuracy of a classification model. It not only identifies the classification error but also the specific form of error, such as type-I or type-II error. The confusion matrix can be used to calculate a variety of model characteristics, including accuracy and precision. (Javatpoint, n.d.)

3.5 Working Flowchart

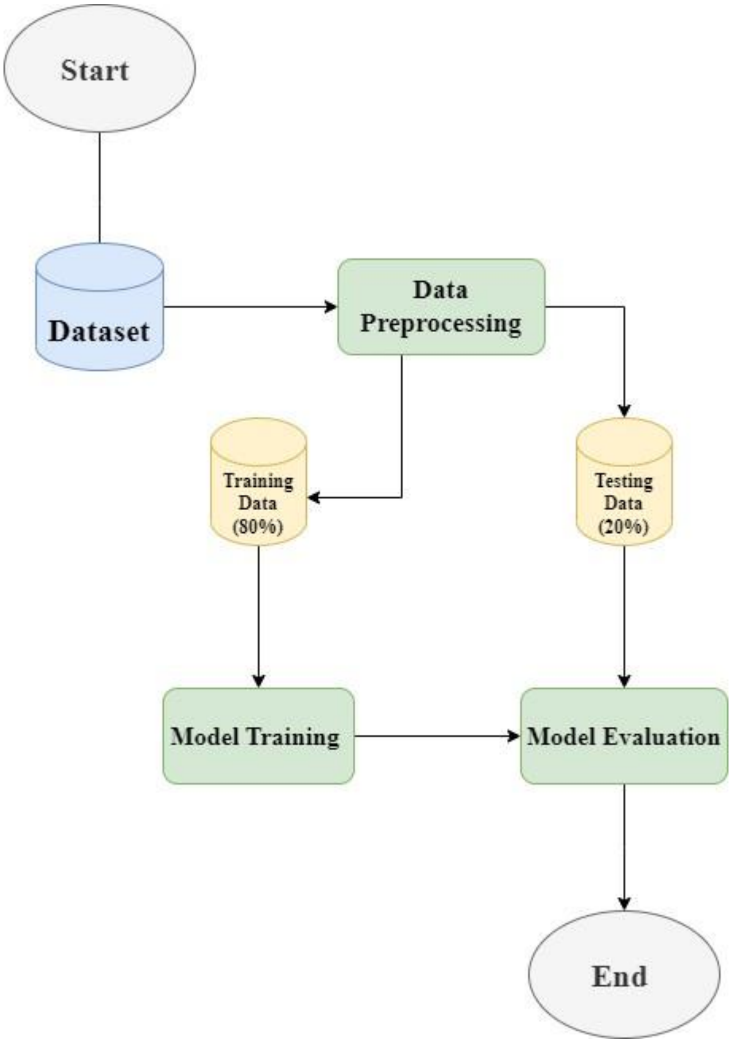

FIGURE 3. Working Flowchart of this study

This coding part of this work was done in Python. Thus, all the necessary python and Scikit-learn libraries are utilized. After collecting the data, the preprocess is to make it suitable for further use. After that, the dataset is split into two parts, i.e. training data and testing. The training dataset is trained using the five classifiers. The testing data is evaluated using the same classifiers to measure its performance. After analyzing the result, the best algorithm is identified.

4 RESULTS AND ANALYSIS

This chapter shows the results of the algorithms and draws comparisons among them. The outcomes are analyzed and explained in detail along with necessary figures and tables in this section.

4.1 Correlation Results

The effects of feature attributes on the target attribute are demonstrated by the Pearson connection's results. The relationship between the stroke characteristic and other attributes is depicted in Figure 4. Age, hypertension, heart disease, average glucose level, and smoking status are among the parameters that have a significant impact on stroke risk. Ever married, work type, and Residence type have the least impact.

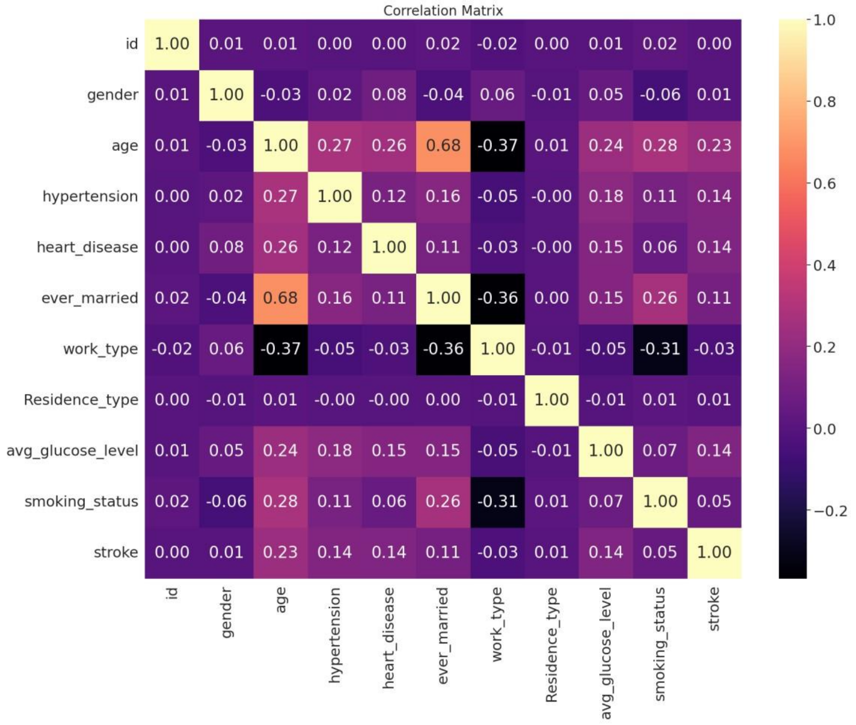

FIGURE 4. Correlation matrix of 10 attributes of the dataset.

4.2 Graphical Representation

The algorithms used in this thesis work are compared with the help of bar chart as the following. Five classifiers are used in this study which are trained and tested with 4909 and 982 data entries respectively. The algorithms are Decision Tree, Random Forest Classifier, Naïve Bayes, Support Vector Machine and K-Nearest Neighbor.

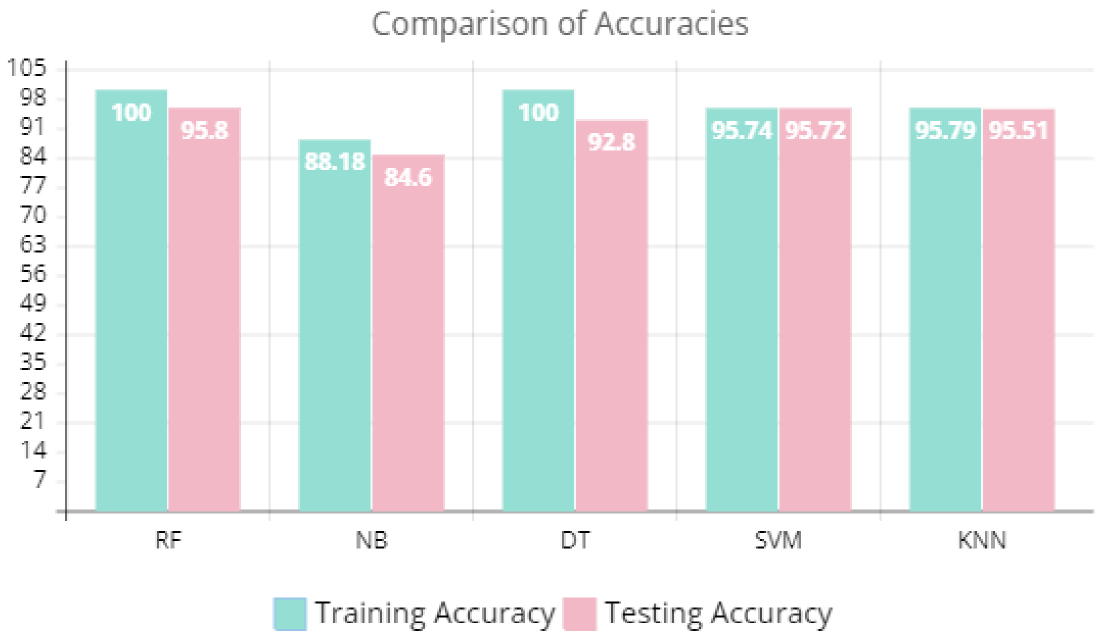

FIGURE 5. Graphical representations of accuracies of different algorithms

The comparison of training and testing accuracies of the aforementioned algorithms are shown in Figure 5. The RF and DT algorithms achieve the highest training accuracy. The highest testing accuracy, however, was achieved by RF and is 95.8%. SVM follows this percentage with a result of 95.72%. The accuracy that is obtained by NB is the lowest at 84.6%. In this research, it has been considered the most commonly considered parameters such as accuracy score, precision score, Recall and F1-score. The outputs of the aforementioned metrics are summarized in TABLE 2.

TABLE 2. Comparison of Accuracy, Precision, Recall and F1-score of five classifiers.

Algorithm	**Accuracy**	**Precision**		**Recall**		**F1-score**	
RF	95.8%	0	0.958	0	1.000	0	0.978
		1	1.000	1	0.023	1	0.046
DT	92.8%	0	0.964	0		0	0.962
		1	0.195	1	0.214	1	0.204

NB	84.6%	0	0.966	0	0.869	0	0.915
		1	0.102	1	0.333	1	0.156
SVM	95.72%	0	0.957	0	1.000	0	0.978
		1	0.000	1	0.000	1	0.000
KNN	95.51%	0	0.957	0	0.997	0	0.977
		1	0.000	1	0.000	1	0.000

In this study, precision refers to the proportion of correctly predicted positive outcomes, out of all positive outcomes, indicating a patient would have a stroke. Increasing precision will reduce the number of false positives. The classifier DT obtained the highest precision value for class 0 in the table, while class 1's precision value is 1which is the highest. Recall indicates how many patients are properly identified as having a stroke out of all those who had one. An acceptable or perfect F1 score is produced when a strong recall balances out a weak precision. The highest recall value 1 in the aforementioned table is attained by RF and SVM. For the objective of this study, we can say that getting a high recall is more essential than getting a high precision, in other words, we want to find as many stroke patients as we can. High precision is preferred for some other models, such as those that determine whether or not a bank customer is a loan defaulter, as the bank wouldn't want to lose clients who were turned down for a loan because the model predicted that they would default.

Furthermore, there are multiple scenarios in which precision and recall are both equally important. For our approach, if the doctor tells us, for instance, that the patients who were incorrectly identified as having strokes are equally essential since they could be suggestive of some other disease, then we would aim for not only a high recall but also a high precision. F1score is crucial in this situation. A low F-score is 0.0, and a perfect F-score is 1.0, like precision and recall. The result of the RF and SVM classifier is 0.978, which is both the value that is closest to 1 and the highest at the same time. As a result of our analysis of the data, we can conclude that RF and SVM both outperform the other classifiers at predicting the existence of stroke.

4.3 Confusion Matrix Results

TABLE 3. Confusion Matrices of Five Classifiers.

Algorithm	**Predicted ❼**	No stroke	Stroke
	Actual ↓		
RF	No stroke	940	0
	Stroke	41	1
DT	No stroke	903	37
	Stroke	33	9
NB	No stroke	817	123
	Stroke	28	14
SVM	No stroke	940	0
	Stroke	42	0
KNN	No stroke	938	2
	Stroke	42	0

Confusion matrices of the stroke estimation using the five different classifiers for measuring the performance of prediction are represented in TABLE 3. True Negative means Patient with No Stroke Correctly Detected It means that the predicted value is equal to the actual value of No stroke. False Positive means Incorrectly Detected. It means the outcome is positive, but it is predicted wrong. False Negative means Stroke Patient Missed, it means the outcome is negative, but it is predicted as positive. True Positive means Stroke Detected; it is the outcome where the actual value is equal to the predicted value of the Stroke detected. The classifier RF and SVM give the best True Negative outcome. On the other hand, the false positive value or incorrect detection given by NB is the highest in number which is very ineffective. Thus, we can say that RF and SVM classifiers outperform the other algorithms.

5 CONCLUSION

Stroke causes a significant number of fatalities and is increasing every day. Several stroke risk factors are responsible for different types of strokes. The design of a machine learning model can aid in the early detection of stroke and minimize its severe effects. Based on the comprehensive study of the complete research, a great deal of data may be gathered to help understand how machine learning techniques can help predict this disease with a high degree of accuracy. In this thesis work, five classifiers are used namely, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest and Neighbors (KNN) Algorithm to train the dataset with 5110 data having the attributes gender, age, hypertension, heart disease, glucose level, smoking status, marital status, job type and residence type. The model is then evaluated with the help of a confusion matrix along with metrics such as Accuracy, Precision, Recall, and F1-score. This study has demonstrated that RF and SVM classifiers are more accurate at producing predictions than the others whereas NB shows the least ineffective result.

The conclusions drawn from this study show that stroke can be predicted by machine learning techniques. The maximum accuracy is given by the RF classifier (95.8%) followed by SVM (95.72%) and the most suitable approach for the prediction of stroke is the RF and SVM classifier. It is strongly assumed that the suggested method can lower the risk of stroke by diagnosing them sooner and can also minimize the cost of diagnosis, treatment, and doctor consultation. In the future, I would like to combine an existing model that will improve performance indicators with deep learning-based imaging, such as brain CT scan and MRI.

REFERENCES

Alaka, S.A., Menon, B.K., Brobbey, A., Williamson, T., Goyal, M., Demchuk, A.M., Hill, M.D. and Sajobi, T.T. 2020. Functional Outcome Prediction in Ischemic Stroke: A Comparison of Machine Learning Algorithms and Regression Models. *Frontiers in Neurology*, 11. doi:https://doi.org/10.3389/fneur.2020.00889.

Analytics Vidhya. 2020. *Categorical Encoding One Hot Encoding vs Label Encoding*. Available at: https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vshttps://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/label-encoding-using-scikit-learn/.

Scikit-learn.org. 2018. *sklearn.model_selection.train_test_split — scikit-learn 0.20.3 documentation*. Available at: https://scikithttps://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.htmllearn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

Analytics Vidhya. 2019. *A Practical Introduction to K-Nearest Neighbor for Regression*. Available at: https://www.analyticsvidhya.com/blog/2018/08/khttps://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/nearest-neighbor-introduction-regression-python/.

Chauhan, N. 2022. *Naïve Bayes Algorithm: Everything You Need to Know*. Available at: https://www.kdnuggets.com/2020/06/naive-bayes-algorithmhttps://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.htmleverything.html.

Centers for Diseases Control and Prevention. 2023. *Know Your Risk for Stroke.* Available at: https://www.cdc.gov/stroke/risk_factors.htm#:~:text=Not%20getting%20enough%20physical%20activity,lower%20your%20chances%20for%20stroke

Datacamp.com. (n.d.). *Scikit-learn SVM Tutorial with Python (Support Vector Machines)*. Available at: https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python

Emon, M.U., Keya, M.S., Meghla, T.I., Rahman, Md.M., Mamun, M.S.A. and Kaiser, M.S. 2020. *Performance Analysis of Machine Learning Approaches in Stroke Prediction*. doi:https://doi.org/10.1109/ICECA49313.2020.9297525

Engineering Education EngEd Program Section. (n.d.). *Introduction to Random Forest in Machine Learning*. Available at: https://www.section.io/engineeringeducation/introduction-to-random-forest-in-machinelearning/#:~:text=The%20(random%20forest)%20algorithm%20establishes Accessed 26 Jun. 2023.

FEDESORIANO. 2021. *Stroke Prediction Dataset*. Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H., Patan, R., Jayaraman, P. and Manikandan, R. 2019. Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications*, 32(3), pp.817–828. doi:https://doi.org/10.1007/s00521-019-04041-y Accessed 26 Jun 2023.

GeeksforGeeks. 2018. *Confusion Matrix in Machine Learning – GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

GeeksforGeeks. 2022. *Pearson Correlation Coefficient*. Available at: https://www.geeksforgeeks.org/pearson- correlation-coefficient/

Huang, X., Cao, T., Chen, L., Li, J., Tan, Z., Xu, B.Y., Richard Huan Xu, Song, Y.S., Zhou, Z., Wang, Z., Wei, Y., Zhang, Y., Li, J., Huo, Y., Qin, X., Wu, Y., Gold, R., Wang, H., Cheng, X. and Xu, X. 2022. Novel Insights on Establishing Machine Learning-Based Stroke Prediction Models Among Hypertensive Adults. 9. doi:https://doi.org/10.3389/fcvm.2022.901240

Holek, I.M. (n.d.). *What Are the 3 Types of Stroke? - CBC Health cbchealth.de*. Available at: https://cbchealth.de/en/three-types-of-stroke/

Heo, J., Yoon, J.G., Park, H., Kim, Y.D., Nam, H.S. and Heo, J.H. 2019. *Machine Learning– Based Model for Prediction of Outcomes in Acute Stroke*. *Stroke*, 50(5), pp.1263–1265. doi:https://doi.org/10.1161/strokeaha.118.024293

JavaTPoint (n.d.). *Support Vector Machine (SVM) Algorithm – Javatpoint*. Available at: https://www.javatpoint.com/machine-learning-supporthttps://www.javatpoint.com/machine-learning-support-vector-machine-algorithmvector-machine-algorithm

JavaTpoint. 2021. *K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint*. Available at: https://www.javatpoint.com/k-nearest-neighboralgorithm-for-machine-learning

JavaTpoint. 2021. *Machine Learning Decision Tree Classification Algorithm - Javatpoint*. Available at: https://www.javatpoint.com/machine-learninghttps://www.javatpoint.com/machine-learning-decision-tree-classification-algorithmdecision-tree-classification-algorithm

JavaTpoint (n.d.). *Machine Learning Random Forest Algorithm - Javatpoint*. Available at: https://www.javatpoint.com/machine-learning-randomhttps://www.javatpoint.com/machine-learning-random-forest-algorithmforest-algorithm

JavaTpoint (n.d.). *Naive Bayes Classifier in Machine Learning - Javatpoint*. Available at: https://www.javatpoint.com/machine-learning-naive-bayeshttps://www.javatpoint.com/machine-learning-naive-bayes-classifier

JavaTpoint (n.d.). *Confusion Matrix in Machine Learning - Javatpoint*. Available at: https://www.javatpoint.com/confusion-matrix-in-machine-learning

Kim, J.K., Choo, Y.J. and Chang, M.C. 2021. Prediction of Motor Function in Stroke Patients Using Machine Learning Algorithm: Development of Practical Models. *Journal of Stroke and Cerebrovascular Diseases*, 30(8), p.105856. Available at: doi:https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105856

Lin, C.-H., Hsu, K.-C., Johnson, K.R., Fann, Y.C., Tsai, C.-H., Sun, Y., Lien, L.-M., Chang, W.L., Chen, P.-L., Lin, C.-L. and Hsu, C.Y. 2020. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Computer Methods and Programs in Biomedicine*, 190, p.105381. doi:https://doi.org/10.1016/j.cmpb.2020.105381

NHS. 2022. *Overview-Stroke.* Available at: https://www.nhs.uk/conditions/stroke/ Accessed 14 December 2023

Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A. and Pattichis, C. 2019. *Integrating Machine Learning with Symbolic Reasoning to Build an Explainable AI Model for Stroke Prediction*. doi:https://doi.org/10.1109/BIBE.2019.00152

Raghav Vashisht 2021. *Machine Learning: When to perform a Feature Scaling? - Atoti Community*. Available at: https://atoti.io/articles/when-to-perform-afeature-scaling/#:~:text=What%20is%20Feature%20Scaling%3F Accessed 26 Jun 2023

Rebouças, E. de S., Marques, R.C.P., Braga, A.M., Oliveira, S.A.F., de Albuquerque, V.H.C. and Rebouças Filho, P.P. 2018. New level set approach based on Parzen estimation for stroke segmentation in skull CT images. *Soft Computing*, 23(19), pp.9265–9286. doi:https://doi.org/10.1007/s00500-018-3491-4

Simplilearn.com. (n.d.). *What is a Confusion Matrix in Machine Learning?* Available at:

https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machinehttps://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning

SÜT, N. and ÇELİK, Y. 2012. Prediction of mortality in stroke patients using multilayer perceptron neural networks. *Turkish Journal of Medical Sciences*. doi:https://doi.org/10.3906/sag-1105-20

Sung, S.M., Kang, Y.J., Cho, H.J., Kim, N.R., Lee, S.M., Choi, B.K. and Cho, G. 2020. Prediction of early neurological deterioration in acute minor ischemic stroke by machine learning algorithms. *Clinical Neurology and Neurosurgery*, 195, p.105892.
Available at: doi:https://doi.org/10.1016/j.clineuro.2020.105892

Trailokya Raj Ojha and Ashish Kumar Jha 2023. Analyzing the Performance of the Machine

Learning Algorithms for Stroke Detection. 13 (2), pp.27–35. Available at: doi:https://doi.org/10.5815/ijeme.2023.02.04

Thammaboosadee, S. and Kansadub, T. 2019. Data mining model and application for stroke prediction: A combination of demographic and medical screening data Approach. *Interdisciplinary Research Review*, 14(4), pp. 61–69. Available at: https://ph02.tcithaijo.org/index.php/jtir/article/view/221565. Accessed: 26 June 2023

WHO.int.(n.d.). *World Stroke Day 2022*. Available at: https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-dayhttps://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-20222022#:~:text=Stroke%20is%20the%20leading%20cause Accessed 25 June 2023.

World Stroke Organization. (n.d.). *European Life After Stroke Forum 2023*. Available at: https://www.world-stroke.org/news-and-blog/blogs/european-life-after-stroke-forum-2023 Accessed 26 Jun 2023.

Wu, Y. and Fang, Y. 2020. Stroke Prediction with Machine Learning Methods among Older Chinese. *International Journal of Environmental Research and Public Health*, 17(6), p.1828. doi:https://doi.org/10.3390/ijerph17061828

Yu, J., Park, S., Kwon, S.-H., Ho, C.M.B., Pyo, C.-S. and Lee, H. (2020). AI-Based Stroke Disease Prediction System Using Real-Time Electromyography Signals. *Applied Sciences*, 10(19), p.6791. doi:https://doi.org/10.3390/app10196791