A Hybrid Efficient Data Analytics Framework for Stroke Prediction

Hosam Alhakami, Shouq Alraddadi, Shurug Alseady, Abdullah Baz, Tahani Alsubait
hhhakam@uqu.edu.sa, s44180476@st.uqu.edu.sa, s44180584@st.uqu.edu.sa, aobaz01@uqu.edu.sa, tmsubait@uqu.edu.sa
College of Computer and Information Systems, Umm Al-Qura University, Saudi Arabia

Summary
Stroke is considered one of the most universal diseases. To understand this disease in medical sciences, large and complex datasets are collected and analyzed. The analysis of this data has been recognized as a big challenge in modern life. Therefore, there is a need to find effective techniques to deal with such huge datasets. To predict the stroke, researchers study the impact of various risk factors on the onset of stroke in an individual. Then, they use the analyzed data to predict the probability of stroke occurrence using machine learning algorithms and techniques like neural networks, decision tree, random forest, linear regression, etc. In this research, recent studies that proposed stroke prediction frameworks using data mining approaches have been reviewed, and a new hybrid framework is proposed to predict stroke disease using two main steps, clustering and classification. Enhanced Hierarchal Clustering is applied on the dataset, then five classifiers are evaluated and compared. The used algorithms are Logistic Regression, Random Forest, Support Vector Machine, Neural Network and XGBoost. All of them show good results according to accuracy and AUC. The best result which is (97%) has been achieved by Random Forest classifier.

Key words:
Stroke Prediction; Machine learning; Data mining; Big Data; Risk factors; Accuracy; Clustering.

1. Introduction

Stroke is an unexpected onset of central neurological deficits that lasts around 24 hours and is caused by an artery blocking. Stroke's signs appear suddenly but they often occur gradually. Strokes can cause disabilities and lead to a mental, physical and financial difficulties for patients, their families, and the community. However, the early prediction can prevent or help in the treatment intervention of a stroke [1]. The first step in any prediction process is to collect data about patients to identify the common risk factors between them [2]. Medical practitioners conduct many surveys to collect information of patients, which are describing patients with common risk factors. Then, data mining and machine learning approaches are used to predict the probability of stroke occurrence based on the risk factors, by studying the inter-dependency of different risk factors of stroke [3]. Data mining is the procedure of extracting key information from a large set of databases. The main goals of data mining are description and prediction of diseases. It is achieved through the processing of a set of attributes (variables) in the dataset and determining the upcoming states of the rest of variables [4]. Data mining use various machine learning techniques, which are playing vital roles in predicting stroke. For example, neural networks [5], clustering [6], support vector machine [7], random forest classifier [8] and logistic regression [9].

Prediction can prevent or help in the treatment intervention of a stroke. Data mining and machine learning are playing vital roles in predicting stroke. To predict the occurrence of stroke, a huge collected data like large amounts of medical records for patients is needed. This data is hard to be managed using traditional analysis methods, since this data can be in unstructured, structured or semi-structured formats. Most of the previous studies of stroke disease prediction have used only one of the machine learning algorithms to predict stroke, which is insufficient for huge data analysis [10]. Therefore, there is a need for a new efficient hybrid framework to process big data and predict the probability of stroke occurrence.

The main objective of this research is to propose a new framework to predict the occurrence of a stroke using the identified risk factors. The sub-objectives are as follows:

- To adopt and enhance a clustering algorithm to gather similar data into clusters.
- To apply classification on the clustered data, using machine learning algorithms
- To evaluate five machine learning algorithms based on Accuracy and Area Under the Curve (AUC).

2. Literature Review

To predict the stroke, researchers study the impact of various risk factors on the onset of stroke in an individual. Then, they use the analyzed data to predict the probability of stroke occurrence using machine learning algorithms like neural networks, decision tree, random forest, linear regression, etc.

Manuscript received April 5, 2020
Manuscript revised April 20, 2020

2.1 Identifying the Risk Factors

Meschina et al. [3] studied the main factors that can cause stroke which are blood pressure, lipids, revascularization, anticoagulants, cigarette smoking cessation, and diet, and show that they can be broadly applicable to the general public. They showed that optimization of stroke prevention for an individual needs systems of care that consider all these risk factors. Harmsen et al. [10] showed that diabetes and high blood pressure are still the key risk factors for stroke prediction for a long term. A family history of cardiovascular disease does not show a significant relationship with stroke occurrence. Stress, transient ischemic attacks, smoking, atrial fibrillation, and a history of chest pain were related to stroke only for the first 1 or 2 periods. Antihypertensive medication and high body mass index are emerged as risk factors in the second and third decade. Schneider et al. [11] presented two types of risk factors, demographic like age, race, gender, and health records as hypertension, smoking, stress, cholesterol, high blood pressure, obesity, diabetes, lack of exercise, poor eating, alcohol test, and family history of stroke as the risk factors which can be used to predict the probability of stroke occurrence [12].

2.2 Analysis of Risk Factors

Nwosu et al. [2] performed a systematic analysis of patients' electronic health records to investigate the impact of different factors on stroke. The risk factors are presented in health records as patients' attributes. Principal component analysis (PCA) is used to study the sub-space of 10 attributes classified into two principal components. The (horizontal axis) of Fig. 1 represents the first component which is the patient's smoking status versus the remaining attributes. The vertical axis, which is the second component represents the patient's gender, glucose level, hypertension, heart disease status and body mass index with their marital status, age and work type. As shown in Fig 1, the marital status and patient's age have the highest correlation to the two-principle component, this shows that the older married patients with hypertension or heart disease and high glucose level do not smoke. While less aged patients with high glucose and hypertension are not married. The patient's residence has the lowest contribution. The result of analysis shows that the two principal components represent only 31.4% of the total data, which proves that the attributes of patients are not highly correlated, and the factors for a predicting framework cannot be reduced without loss of information.

Therefore, all patients' attributes should be used as input variables for prediction.

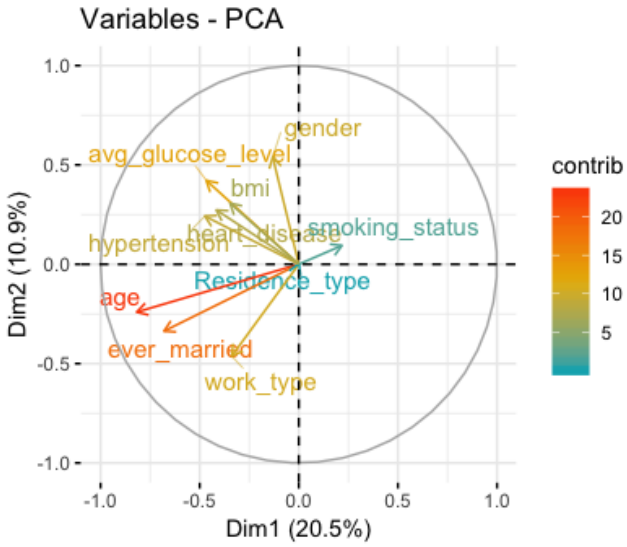

Fig. 1 Representation of the Patient Attributes Projected on Two Principal Components [2]

2.3 Stroke Prediction Using Machine Learning

Machine learning algorithms have been proposed for predicting stroke occurrence based on the risk factors [13] [14] [15] [16] [10] [2]. There are several works in literature that aim to predict the probability of stroke occurrence by employing machine learning techniques on health records. Shanthi et al. [13] applied Artificial Neural Networks (ANN) to predict the Thromboembolic stroke disease. The researcher used backpropagation to train the ANN architecture and the same was tested for different sorts of stroke disease. After optimizing the input parameters, the ANN was trained and tested, the overall predictive accuracy was 89%.

Hanifa and Raja [14] use polynomial and radial basis functions applied in a non-linear support vector classification to improve the accuracy of predicting stroke risk. The results are calculated using confusion matrix. The achieved classification accuracies using the kernel functions of Radial Basis Function (RBF) is 98% and Polynomial (Poly) is 92%. Polynomial kernels are suitable for problems of normalized training data. Cheng et al. [17] have used two ANN models for predicting ischemic stroke on the dataset from Sugam Multispecialty Hospital, Kumbakonam, Tamil Nadu, India. And the researchers achieved the accuracy rates between 79.2% and 95.1%.

Sung et al. [18] analyzed patients' data with acute ischemic stroke (AIS) from hospital-based stroke records associated with a nationwide claims database. Researchers estimated the stroke severity index (SSI) based on patients' data. Real

stroke severity was measured with the National Institutes of Health Stroke Scale (NIHSS) and the modified Rankin Scale (mRS) tested the functional outcomes, which were retrieved from stroke records. The validity of the predictive model was calculated by correlating mRS with SSI. They use Logistic regression models to predict mortality. The limitations of this study, that the study patients were from only 2 hospitals, one is a regional hospital and the other is a medical center, which might not be sufficient to represent the overall population. In Taiwan, approximately 70 % of stroke patients are registered in regional hospitals and medical centers, with the rest recorded in district hospitals [19]. Jeena et al. [16] proved that there is a correlation between the probability of stroke occurrence and the total count of risk factors. They proposed a regression technique to examine the relationship between a risk factor and its impact. They collected data from International Stroke Trial database and tested it using Support Vector Machine (SVM). They have implemented SVM with different kernel functions and obtained accuracy of 90% using the linear kernel. Table 1 summarizes the studies predicting stroke using various machine learning techniques, their contributions and the achieved accuracy rates.

Table 1: Studies that applied machine learning techniques to predict stroke

Authors	# of Dataset	Contribution	Accuracy
[13]	50	Artificial Neural Networks (ANN)	89%
[14]	100	Poly and RBF of SVM	92-98%
[20]	5	Cox proportional hazards model, SVM	85% from SVM
[17]	82	Two Models of ANN	79.2% - 95.1%.
[18]	3,577	KNN, multiple LR	75%, 73% respectively
[15]	68	naive Bayes, ANN and DT	72, 74,75% respectively
[16]	19,435	SVM with different kernel functions	90% from linear kernel
[21]	192	PLR, SGB, SVM	98% from SVM
[22]	400	KNN, DT	95, 98% respectively
[2]	1096	DT, RF, ANN	74-75%
[10]	43	SVM, RL, DT, RF	76, 76, 79, 90% respectively

2.4 Comparison between Machine Learning Techniques

Khosla et al. [20] have compared between the Cox proportional hazards model and the machine learning method (SVM) for stroke prediction on the Cardiovascular Health dataset. The result concluded that support vector machine (SVM) achieved a higher performance according to Receiver operating characteristics (ROC) curve than the Cox proportional hazards model. Kansadub et al. [15] have used naive Bayes, decision trees (DTs) and ANN to predict stroke on the healthcare dataset stroke data. They worked on huge dataset collected from the Faculty of Physical Therapy, Mahidol University, Thailand from 2012-2015. The collected data includes more than 68,000 patients. The researchers reported that DT was the best classifier among the other applied methods. Adam et al. [22] have also compared two algorithms decision tree and k-nearest neighbor (KNN) for the stroke classification on the dataset of 400 patients from Sugam Multispecialty Hospital, Kumbakonam, Tamil Nadu, India, and the researchers showed that the classification of decision tree has better performance than KNN algorithm. Arslan et al. [21] applied three machine learning techniques on a collected data set (80 patients and 112 healthy individuals) from TurgutOzal Medical Centre, Inonu University, Malaya, Turkey. They compared between penalized logistic regressions (PLR), Stochastic Gradient Boosting (SGB) and SVM in predicting stroke. The findings of the research proved that SVM achieved the highest accuracy of 98%. Author in [10] proposed an architecture to predict the stroke using Apache Spark platform which is a big data platform. This platform includes an MLlib library. MLlib is an API combined with Spark to run machine learning techniques. Four machine learning classification algorithms were used to build the stroke prediction model; Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT) and Random Forest Classifier (RF). The cross-validation and hyperparameter tuning were integrated with machine learning algorithms to improve results. Precision, Accuracy, F1-measure and Recall were used as evaluation metrics of the proposed system. The proposed framework was applied in [34] with more than 43,000 patients. Fig 2 shows the framework of the proposed stroke prediction system [10]. This system includes five stages as follows: 1) loading stroke dataset 2) data pre-processing, 3) Cross-validation and Hyperparameter Tuning, 4) Classifiers, and 5) Evaluating Classifiers.

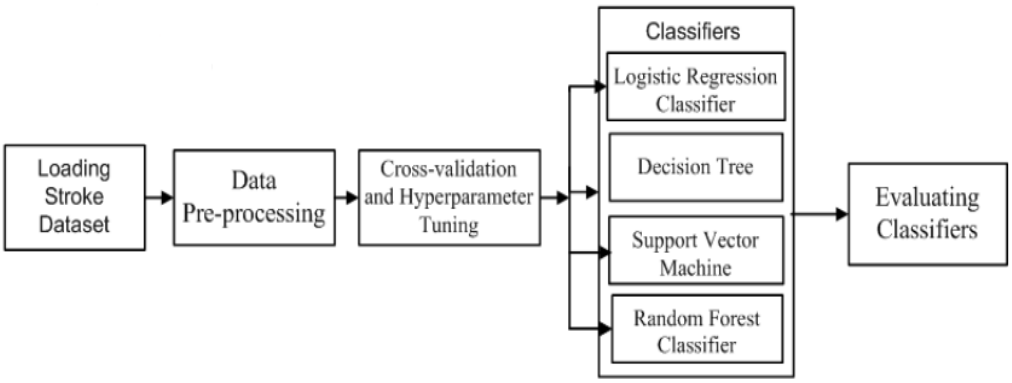

Fig. 2 The architecture of the stroke prediction system [10]

Fig 3 shows the accuracy of applying SVM, DT, LR, and RF. The highest accuracy of 90% recorded using random forest. The second-highest accuracy at 79% was recorded using decision tree. The logistic regression techniques and support vector machine recorded the same accuracy at 77%. Nwosu et al. [2] compared the performance of three state-of-art machine learning algorithms, random forests, neural networks and decision tree for stroke prediction. They formed a balanced dataset of 1096 patients and used 70% of the dataset for training the algorithms and 30% of the dataset for testing. The classification accuracy is used as the metric to evaluate the performance of the machine learning approaches. To remove sampling bias, they perform 1000 random down sampling experiments. Table II illustrates the average classification accuracy for the three models.

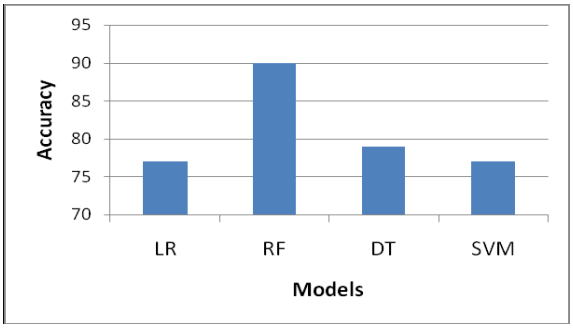

Fig. 3 Accuracy of Applying Machine Learning Algorithms [10]

Table 2: Prediction Accuracy of Decision Tree, Random Forest and Neural Network

Approach	Accuracy
Decision Tree	74.31%
Random Forest	74.53%
Neural Network	75.02%

As shown in Table 2, the performance of random forest and decision tree are similar. The best accuracy result of 75.02% is obtained from the feed-forward multi-layer perceptron model. Researchers also compare the density distribution of classification accuracy for each method over the 1000 experiments. Figure 4 shows this distribution. It is observed that the three methods overlap with each other around their mean values, and consequently they have similar classification accuracy [2]. Table 3 shows a comparison between different machine learning techniques in predicting stroke.

Table 3: Comparison Between Machine Learning Techniques in Predicting Stroke

Author	ML Techniques	Best Result
[20]	Cox proportional hazards model and SVM	SVM
[15]	naive Bayes, DTs and ANN	DT
[22]	DT and KNN	DT
[21]	PLR, SGB and SVM	SVM
[10]	LR, RF, DT, SVM RF	RF
[2]	RF, DT, NN	NN

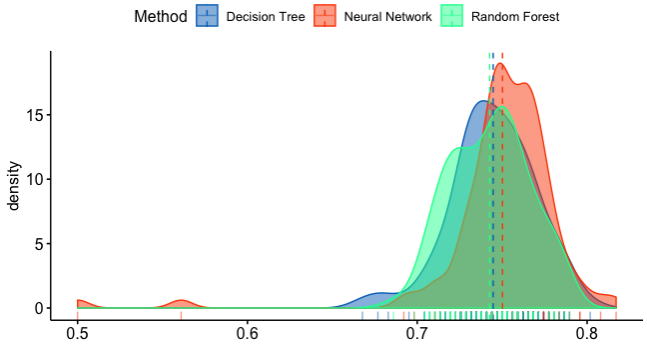

Fig. 4 Distribution of Classification Accuracy of the Three Algorithms Over 1000 Experiments [2]

2.5 Discussion

From Table 1, studies of [14] [17] [21] [22] achieved the highest accuracy rate (95-98%) among other researchers. However, those studies have a limitation that the dataset they used is small (80-100 patients). The highest accuracy rate for researchers used big dataset (more than 43,000 patients) is achieved in [10] which is 90% using the

Random Forest Classifier. This shows that when the dataset increased, the accuracy rate of the results decreased.
From Table 2, Decision Tree, Support Vector Machine, Random Forest and Neural Network achieved higher accuracy rates than other machine learning techniques in predicting stroke. This proved that these machine learning techniques (DT, SVM, RF and NN) can work better with different input variables, and noisy data than other techniques.

3. Research Methodology

The proposed research methodology includes four main steps, as shown in Fig 5. The first step is collecting the dataset. The second step is pre-processing includes preparing the data and normalization to enhance the clustering algorithm accuracy. The third step is an agglomerative hierarchal clustering. The last step is classification using machine learning algorithms.

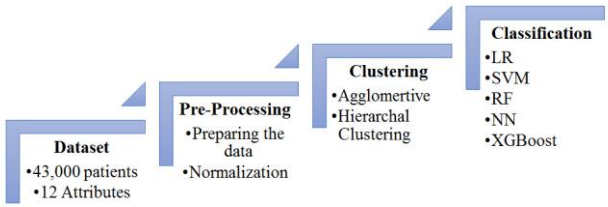

Fig. 5 The proposed Research Methodology

3.1 Stroke Dataset

The data used for this research is collected from an open source Healthcare Dataset Stroke Data. The test dataset consists of 43,400 data items and each item contains 12 attributes.
The data attributes are id, age, gender, hypertension, heart disease, residence, marital status, work type, glucose level, BMI (Body mass index), smoking status and stroke history. The data attributes are listed in Table 4.

Table 4: Dataset Attributes And Description.

No.	Variable	Definition
1	id	Patient ID
2	age	Age of patient
3	gender	Gender of Patient
4	hypertension	0-No hypertension, 1- suffering from hypertension
5	Heart_disease	0-No heart disease, 1- suffering from heart disease
6	ever married	Yes/ No
7	Work_type	Type of Occupation
8	residence type	Area type of residence (Urban/Rural)
9	Avg_glucose_level	Average Glucose level (measured after meal)
10	bmi	Body mass index
11	Smoking_status	Patient's smoking status
12	stroke	0-no stroke, 1- suffered stroke

- Hypertension (high blood pressure): presents as 0/1 values. 1 stands for hypertension, while the 0 stands for no hypertension.
- Heart Disease (Cardiovascular disease): Data consists of 0/1 values. The 1 value presents the suffering from heart disease, the 0 value stands for the person who do not suffer from heart disease.
- Work Type: categorical data including Self-employed, Private, Children, Government job and Never worked. These five categories are represented by 0, 1, 2, 3 and 4 respectively.
- Average Glucose Level: A numerical data measured in unit of mg/dL. The normal average glucose level is below 125 mg/dL [23]. When average glucose level is lower than 70 mg/dL, the person has Low blood sugar [24], when average glucose level is higher than 200 mg/dL, the person has High blood sugar [25].
- BMI (body mass index): A numerical data calculated from the height and mass of the person. The normal range of BMI is from 18.5 to 25. The World Health Organization (WHO) regards a BMI greater than 25 as overweight and above 30 as obese, and less than 18.5 as underweight,
- Smoking Status: A categorical data of never smoked, formerly smoked, smokes and never smoked. These categories are encoded by as 2, 1,0 respectively.
- Stroke: A data of 0/1 (False/True) values. The 1 value represents the person who has suffered from stroke. 0 represents the person who has never suffered from stroke.

3.2 Normalization

To enhance the adopted clustering algorithm, a normalization technique is used to enhance the Euclidean distance between clusters. Normalization will enhance the efficiency and accuracy of clusters by determining more accurate center points of clusters [26].

3.3 Agglomerative Hierarchal Clustering

The Clustering technique is used widely in the data mining. The objective of clustering is to summarize a very large dataset X with a smaller representative set of points C= ci— i=1, 2, 3 ..... k called as centroids. Several clustering algorithms like k-means, EM algorithm, hierarchical, Self-organizing Maps, etc. are used to make a of representatives.

In this research, Hierarchal clustering is used. Hierarchical Clustering [35] is used as a data mining method by clustering data into a hierarchy of disjointed groups. Hierarchal clustering includes adding nodes directly in parallel coordinates for hierarchical data selection. A node structured as an intuitive edge of interaction, since it represents both the coordinates and the data. In the proposed algorithm, each node consists of a collected homogeneous data clustered by hierarchical clustering. Hierarchical Clustering put each data item into a cluster and add clusters based on the shortest distance to form a new cluster [36] Agglomerative Hierarchical clustering Technique is used in this research. In this method, firstly each data point is considered as an individual cluster. At each phase, the similar clusters merge with each other until one cluster or K clusters are built. The algorithm of Agglomerative is as follows:

1- Calculate the proximity matrix
2- Let each data point be a cluster
3- Merge the two nearest clusters and update the proximity matrix, then repeat this step.
4- When only a single cluster or k (two or three) clusters remains stope the process. The Agglomerative Hierarchical clustering approach can be represented by a Dendrogram. A Dendrogram is a tree-like diagram that shows the arrangements of splits and merges.

3.4 Classification

Four machine learning algorithms are used and compared in the classification stage, they are Logisttic Regression (LR), Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF), and XGBoost.

- Logistic Regression
  As a classification technique, the logistic regression is used. It is a predictive analysis, which defines information and identifies the relationship between one dependent binary variable and at least one nominal, ordinal, proportion or interval level independent variable [26].

- Support Vector Machine
  Support Vector Machines (SVM) is a proposed approach of Machine Learning which can be used for Support Vector Regression (SVR) and Support Vector Classification (SVC). It is appropriate for both linear and nonlinear data. SVMs do predictions by automated learning from standing knowledge [14].

- Neural Network
  Neural network is a general approach of data mining. When the network output is continuous, it is doing prediction and when the output has discrete values, it is performing classification [13].

- Random Forest
  The random forest algorithm is a general-purpose classification and regression algorithm was first proposed by [37]. It aggregates and averages the prediction of several decision trees. It shows high performance when the number of variables is much larger than observations [27].

- XGBoost
  XGBoost, a scalable machine learning algorithm for end-to-end tree boosting. It is an open source package. The system has been effective in several machine learning and data mining challenges and achieve many state-of-art results [28].

3.5 Implementation Detail

Python programming language is used to simulate the proposed algorithms. Four libraries of Python are used; SMOTE, Pandas, sklearn and XGboost.

- Firstly, the dataset is loaded from [www.kaggle.com/asaumya/healthcare-datasetstrokedata].
- Then, a pre-processing includes convert the string data into numbers (encoding), then a normalization is applied, as a scaling technique. Where a new range can be discovered from a current one. Min-Max Normalization method is applied, it changes A to B which is found in the range [C, D]. It is given by the equation: $B = ((A - \text{minimum value of } A)/\text{maximum value of } A - \text{minimum value of } A) * (D - C) + C$, A=Original data point B=Normalized data point [C, D] = determined range.
- Clustering: Agglomerative Hierarchal clustering is applied. - Splitting: The dataset is divided into 80% for training and 20% for testing. Training is the process of setting the best weights on the inputs of each of the units, to use the training set to produce network output close to the desired output.
- Balancing: Synthetic is Minority Over sample technique (SMOT) is used for balancing.
- Classification: after splitting and balancing, five machine learning algorithms have been applied on the resulted data.

3.6 Performance Metrics

For evaluating the performance of algorithms, two main methods are used.

1- The confusion matrix has been used to calculate accuracy and f-measure.
2- The Area Under the Curve (AUC).

3.7 Confusion Matrix

Represents the performance of a classifier on a set of test data. Each classifier is given two types of correct predictions and two types of incorrect predictions [29]. TP is the true positive of the predicted output; TN is the true negative of the predicted output, FP is the false positive of the predicted output, and FN is the false negative of the predicted output. The accuracy, precision, recall, and f-measure (f-score) are defined as the following [10]:
Accuracy: describes the classifier performance as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: is the correctly predicted positive on the total number of the total classified positive.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall: The true positive output divided by the summation of true positive and false negative.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score shows the relationship between the Recall and Precision. F1-score is always closest to the smaller value of Recall or Precision [30]. The equation is:

$$RF1 - score = \frac{2*Recall*Precision}{Recall+Precision} \quad (4)$$

3.8 Area Under the Curve

To measure the performance of the classifiers, the AUC [31] is used because relying only on accuracy and f1 score is not a very accurate measure of performance when there is a huge data. Also, the AUC is more informative [32], especially in medical contexts.
The Receiver Operating Characteristic (ROC) curve is presented for each classifier [33]. The mean AUC of each curve is calculated for each classifier.

4. Results

The Accuracy, F1-score and AUC is measured for each classifier.

4.1 Results of using Neural Network

As shown in Figure 6, Neural network recorded 80.9% accuracy, 88.199% for F1-score and 0.845 as a mean AUC. The results are summarized in Table 5.

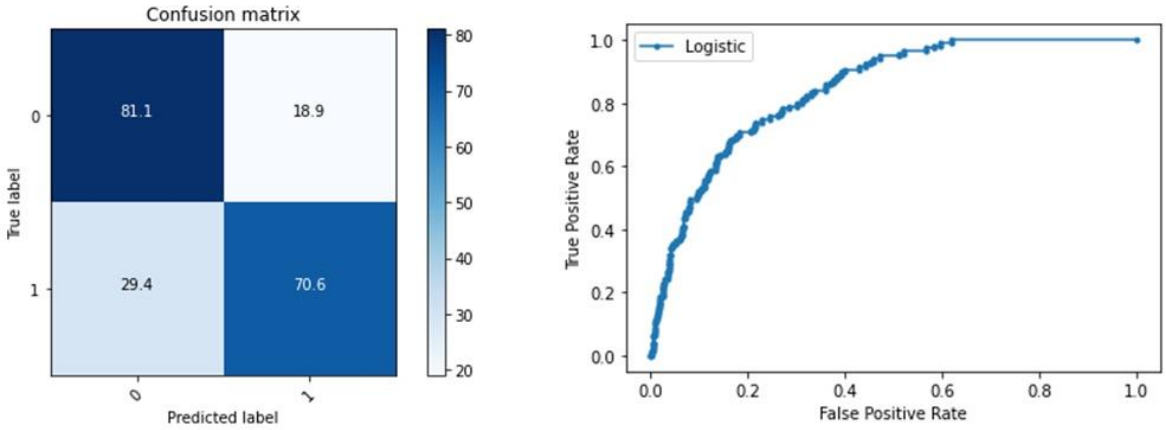

Fig. 6 (A) Confusion Matrix (B) ROC curve for NN

Table 5: Neural Network Results.

Accuracy	80.913%
F1 Score	88.199%
AUC	0.845

4.2 Logistic Regression

As shown in Figure 7, Logistic Regression recorded 74.6% accuracy, 84.18% for F1-score and 0.849 as a mean AUC. The results are summarized in Table 6.

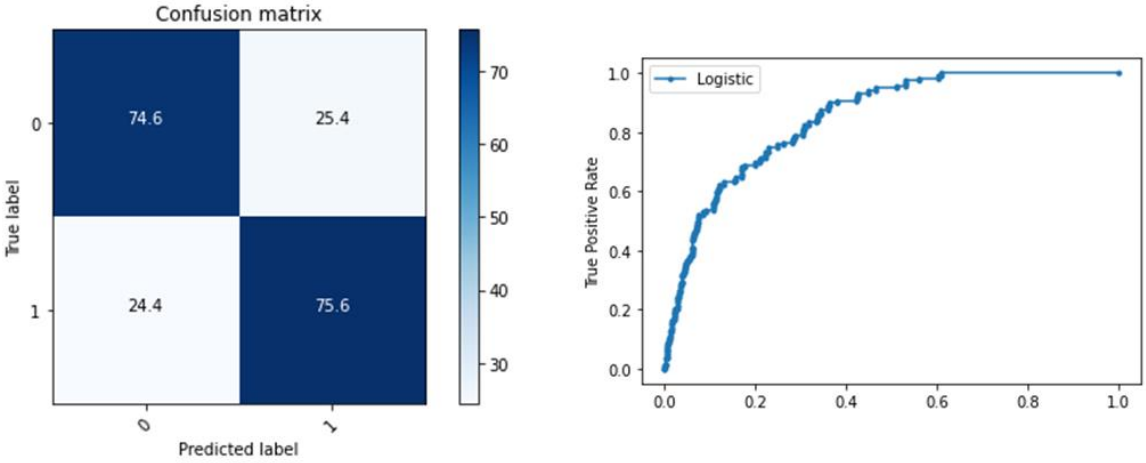

Fig. 7. (A) Confusion Matrix and (B) ROC curve for Logistic Regression

Table 6. Logistic Regression Results

Accuracy	74.619%
F1 Score	84.184%
AUC	0.849

4.3 Results of XGBoost

As shown in Figure 8, XGBoost recorded 75.29% accuracy, 84.6% for F1-score and 0.840 as a mean AUC. The results are summarized in Table 7.

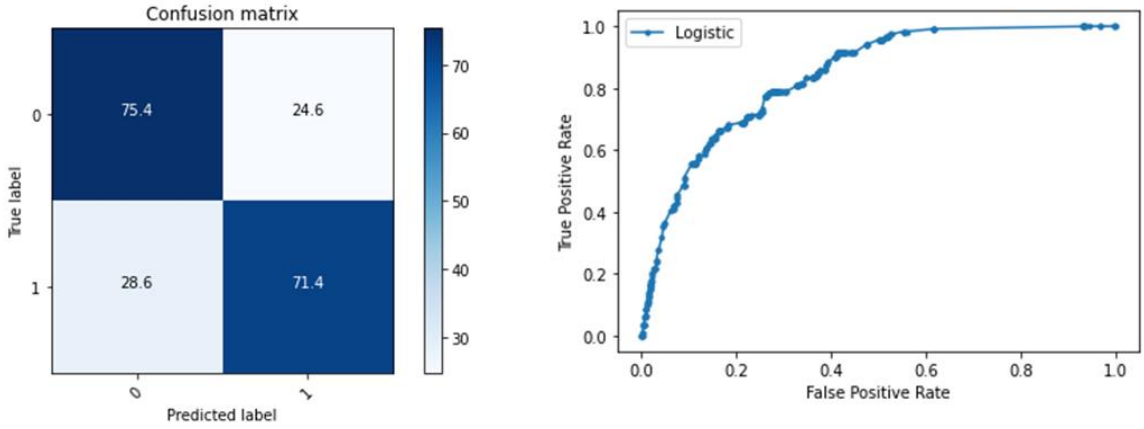

Fig. 8. (A) Confusion Matrix and (B) ROC curve for XGBoost

Table 7. Xgboost Results

Accuracy	72.298%
F1 Score	84.635%
AUC	0.840

4.4 Results of Support Vector Machine

As shown in Figure 9, SVM recorded 80.2% accuracy, 87.77% for F1-score and 0.800 as a mean AUC. The results are summarized in Table 8.

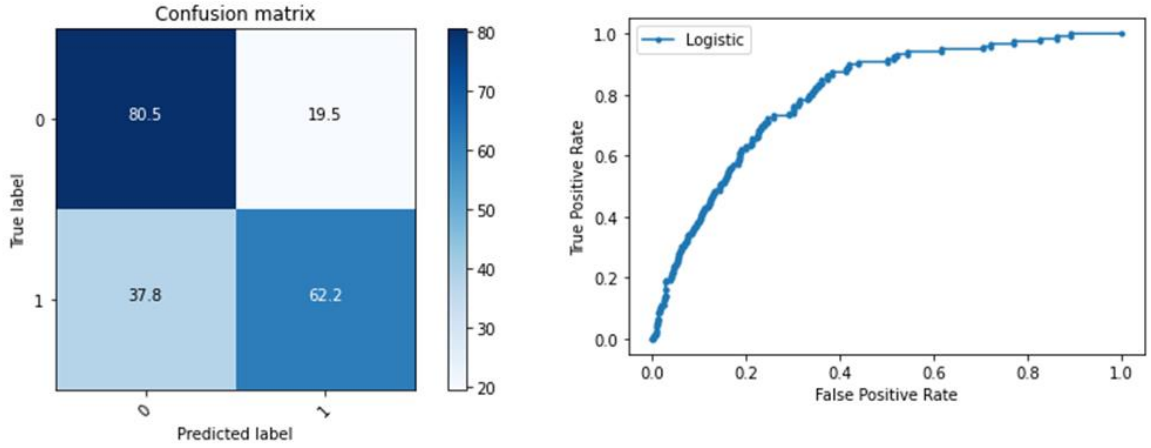

Fig. 9. (A) Confusion Matrix and (B) ROC curve for SVM

Table 8. SVM Results

Accuracy	80.222%
F1 Score	87.772%
AUC	0.800

4.5 Results of using Random Forest

As shown in Figure 10, RF recorded 97.616 % accuracy, 97.616 % for F1-score and 0.806 as a mean AUC. The results are summarized in Table 9.

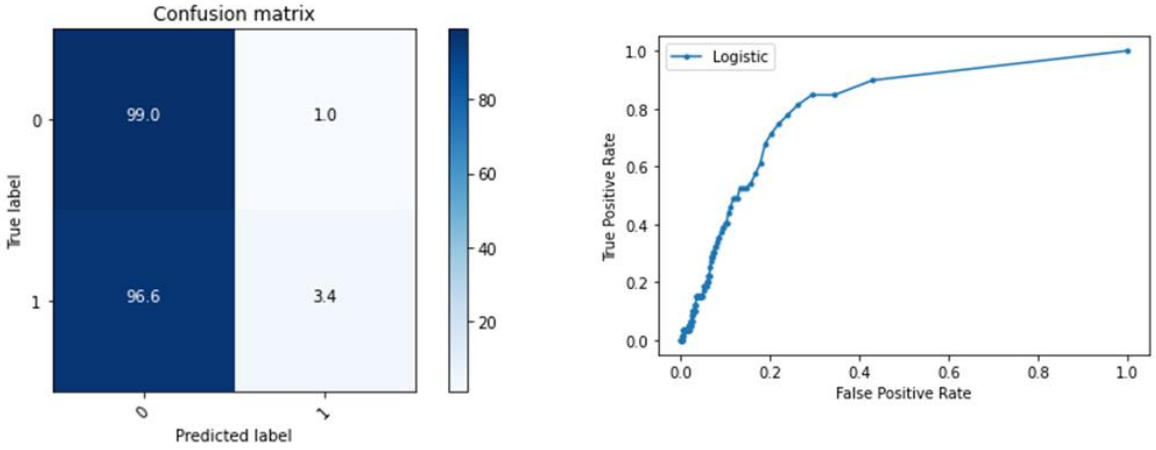

Fig. 10. (A) Confusion Matrix and (B) ROC curve for the Random Forest classifier.

Table 9. SVM Results

Accuracy	97.616 %
F1 Score	97.616 %
AUC	0.806

5. Conclusion

Stroke disease is one of the major problems in now a day which can lead to death. Prediction of stroke diseases is probable by the consideration of attributes, by the employing data mining techniques. Data mining includes using machine learning algorithms such as neural networks, naive Bayes, clustering mechanisms, etc. To predict stroke disease in a big data environment, several steps of implementation should be done. The prediction analysis is the technique in where user predicts the future based on current conditions. The proposed prediction analysis involves two main steps. The first step is clustering, which clusters the similar and dissimilar type of data. The second step is classification which classifies the clustered data for the prediction analysis. In this research, hierarchal clustering is used for the clustering. Five classifiers are evaluated for classifying and predicting the complex data. The Agglomerative hierarchal clustering consists of two key steps. In the first step, each data point represents one cluster. In the second step, the closest clusters are merged with each other until one of k clusters remains. The

accuracy of classification may reduce when many points are wrongly clustered or unclustered. Normalization has been applied in this work to improve the accuracy of clustering by enhancing the Euclidean distance to obtain the maximum accuracy, normalization works better with large dataset. The proposed improvement leads to increase the accuracy of classification. The proposed algorithms are being implemented in Python and the accuracy, f-score and AUC is measured for each classifier. All classifiers achieved good results, while Random Forest has the best overall results among other classifiers.

References

[1] Veerbeek, J. M., Kwakkel, G., van Wegen, E. E., Ket, J. C., & Heymans, M. W. (2011). Early prediction of outcome of activities of daily living after stroke: a systematic review. Stroke, 42(5), 1482-1488.

[2] Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019, July). Predicting Stroke from Electronic Health Records. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5704-5707). IEEE.

[3] Meschia, J. F., Bushnell, C., Boden-Albala, B., Braun, L. T., Bravata, D. M., Chaturvedi, S., ... & Goldstein, L. B. (2014). Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, 45(12), 3754-3832.

[4] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), 2431-2448.

[5] Abedi, V., Goyal, N., Tsivgoulis, G., Hosseinichimeh, N., Hontecillas, R., Bassaganya-Riera, J.,& Alexandrov, A. V. (2017). Novel screening tool for stroke using artificial neural network. Stroke, 48(6), 1678-1681.

[6] Chantamit-O-Pas, P., & Goyal, M. (2018). A case-based reasoning framework for prediction of stroke. Advances in Intelligent Systems and Computing.

[7] Rosado, J. T., & Hernandez, A. A. (2019). Developing a Predictive Model of Stroke using Support Vector Machine. In 2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA) (pp. 35-40). IEEE.

[8] McKinley, R., H¨ani, L., Gralla, J., El-Koussy, M., Bauer, S., Arnold, M., ... & Wiest, R. (2017). Fully automated stroke tissue estimation using random forest classifiers (FASTER). Journal of Cerebral Blood Flow & Metabolism, 37(8), 2728-2741.

[9] Al-Talqani, H. M. (2017). Dyslipidemia and Cataract in Adult Iraqi Patients. EC Ophthalmology, 5, 162-171.

[10] Harmsen, R., Helms-Lorenz, M., Maulana, R., van Veen, K., & van Veldhoven, M. (2019). Measuring general and specific stress causes and stress responses among beginning secondary school teachers in the Netherlands. International Journal of Research & Method in Education, 42(1), 91-108.

[11] Schneider, S., Kornejeva, A., Vibo, R., & K˜orv, J. (2017). Risk factors and etiology of young ischemic stroke patients in Estonia. Stroke research and treatment, 2017.

[12] Koton, S., Sang, Y., Schneider, A. L., Rosamond, W. D., Gottesman, R. F., & Coresh, J. (2020). Trends in Stroke incidence rates in older us adults: an update from the Atherosclerosis Risk in Communities (ARIC) Cohort Study. JAMA neurology, 77(1), 109-113.

[13] Shanthi, D., Sahoo, G., & Saravanan, N. (2009). Designing an artificial neural network model for the prediction of thrombo-embolic stroke. International Journals of Biometric and Bioinformatics (IJBB), 3(1), 10-18.

[14] Hanifa, S. M., & Raja-S, K. (2010). Stroke risk prediction through nonlinear support vector classification models. International Journal of Advanced Research in Computer Science, 1(3).

[15] Kansadub, T., Thammaboosadee, S., Kiattisin, S., & Jalayondeja, C. (2015, November). Stroke risk prediction model based on demographic data. In 2015 8th Biomedical Engineering International Conference (BMEiCON) (pp. 1-3). IEEE.

[16] Jeena, R. S., & Kumar, S. (2016, December). Stroke prediction using SVM. In 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 600-602). IEEE.

[17] Cheng, C. A., Lin, Y. C., & Chiu, H. W. (2014, July). Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. In ICIMTH (pp. 115-118).

[18] Sung, S. F., Hsieh, C. Y., Yang, Y. H. K., Lin, H. J., Chen, C. H., Chen, Y. W., & Hu, Y. H. (2015). Developing a stroke severity index based on administrative data was feasible using data mining techniques. Journal of clinical epidemiology, 68(11), 1292-1300.

[19] Lee, H. C., Chang, K. C., Huang, Y. C., Lan, C. F., Chen, J. J., & Wei, S. H. (2010). Inpatient rehabilitation utilization for acute stroke under a universal health insurance system. The American journal of managed care, 16(3), e67-e74.

[20] Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010, July). An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 183-192).

[21] Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches-based prediction of ischemic stroke. Computer methods and programs in biomedicine, 130, 87-92.

[22] Adam, S. Y., Yousif, A., & Bashir, M. B. (2016). Classification of ischemic stroke using machine learning algorithms. Int J Comput Appl, 149(10), 2631.

[23] U.S. National Library of Medicine (2020). Blood sugar test. Retrieved from: https://medlineplus.gov/ency/article/003482.htm

[24] National Institute of diabetes and Digestive and Kidney diseases. (2020). Low Blood Glucose (Hypoglycemia). Retrieved from: https ://www .niddk .nih .gov /health - information/diabetes/overview/preventing-problems/lowblood-glucose-hypoglycemia

[25] American Diabetes Association (2020). Diagnosis and Classification of Diabetes Mellitus. Retrieved from: https://care.diabetesjournals.org/content/ 37/Supplement 1/S81

[26] Singh, R., & Rajesh, E. (2019). Prediction of Heart Disease by Clustering and Classification Techniques Prediction of Heart Disease by Clustering and Classification Techniques.
[27] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.
[28] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
[29] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.
[30] Chai, K. M. A. (2005, August). Expectation of F-measures: Tractable exact computation and some empirical observations of its properties. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 593-594).
[31] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of mathematical psychology, 12(4), 387-415.
[32] Meares, C., Badran, A., & Dewar, D. (2019). Prediction of survival after surgical management of femoral metastatic bone disease–A comparison of prognostic models. Journal of bone oncology, 15.
[33] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.
[34] Healthcare dataset stroke data. [Cited 2020; Available from: https ://www.kaggle.com /asaumya / healthcare-dataset-stroke-data.
[35] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863-14868.
[36] Huang, S., Kang, Z., Tsang, I. W., & Xu, Z. (2019). Auto-weighted multi-view clustering via kernelized graph learning. Pattern Recognition, 88, 174-184.
[37] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Hosam Alhakami received his B.Sc. degree in Computer Science from King Abdulaziz University, Saudi Arabia in 2004. From 2004 to 2007, he worked in software development industry, where he implemented several systems and solutions for a national academic institution. Following that, he started his postgraduate studies in UK, where he received his MSc degree in Internet Software Systems from Birmingham University, Birmingham, UK in 2009. Then he successfully acquired his PhD in Software Engineering from De Montfort University in 2015. His research interests include algorithms, semantic web and optimization techniques. He focuses on enhancing real-world matching systems using machine learning and data analytics in a context of supporting decision-making.

Shouq Alraddadi received her B.Sc. degree from Umm Al-Qura University in 2019. Currently, she is Master of Science candidate at Umm Al-Qura University, College of Computer and Information Systems, Computer Science Department.

Shurug Alseady received her B.Sc. degree from Umm Al-Qura University in 2019. Currently, she is Master of Science candidate at Umm Al-Qura University, College of Computer and Information Systems, Computer Science Department.

Abdullah Baz received the B.Sc. degree in electrical and computer engineering from UQU, in 2002, the M.Sc. degree in electrical and computer engineering from KAU, in 2007, and the M.Sc. degree in communication and signal processing and the Ph.D. degree in computer system design from Newcastle University, in 2009 and 2014, respectively. He was a Vice-Dean, and then the Dean of the Deanship of Scientific Research with UQU, from 2014 to 2020. He is currently an Assistant Professor with the Computer Engineering Department, a Vice-Dean of DFMEA, the General Director of the Decision Support Center, and the Consultant of the University Vice Chancellor with UQU. His research interests include VLSI design, EDA/CAD tools, coding and modulation schemes, image and vision computing, computer system and architecture, and digital signal processing. Since 2015, he has been served as a Review Committee Member of the IEEE International Symposium on Circuits and Systems (ISCAS) and a member of the Technical Committee of the IEEE VLSI Systems and Applications. In 2017, IEEE has elevated him to the grade of IEEE Senior Member. He served as a Reviewer in a number of journals, including the IEEE Internet of Things, the IET Computer Vision, the Artificial Intelligence Review, and the IET Circuits, Devices and Systems.

Tahani Alsubait is a faculty member of College of Computer and Information Systems. She earned her PhD in AI and instruction from the University of Manchester. She hold a Bachelor's in Computer Science from King Saud University and a Master's from King Abdulaziz University. Her research interests include knowledge representation and reasoning, data analytics and HCI.