Early Risk Prediction of Stroke Using Machine Learning Techniques in the Context of Bangladesh

No Author Given

No Institute Given

Abstract. Stroke is a life-threatening medical condition which is one of the leading causes of death worldwide. Moreover, prediction and diagnoses of stroke are intricate tasks; require several diagnostic tests costing invaluable time and expertise of medical practitioners. In addition to that, the risk factors for stroke vary from region to region due to people's habit and lifestyle. To minimize the complexity of predicting and diagnosing of stroke, a computer-aided system can aid doctors in predicting stroke and take necessary steps by which stroke can be avoided. In this paper, various machine learning techniques have been implemented including Logistic Regression, Support Vector Machine with polynomial kernel, Linear Support Vector Classifier, Gaussian Naive Bayes, Multinomial Naive Bayes, and Decision Tree to compare their performances and finding the best classifier for the prediction of stroke on the data based in Bangladesh. After using Synthetic Minority Over-sampling Technique to balance the dataset; Logistic Regression, Linear SVC, and Decision tree performed notably well by achieving 99.52%, 99.52%, and 99.04% accuracy respectively among all the classifiers and outperforming the other work on the same dataset.

Keywords: Ischemic stroke · Hemorrhagic stroke · Stroke prediction · Machine Learning.

1 Introduction

Stroke, often referred to as a cerebrovascular accident, occurs when the blood supply to any part of the brain gets reduced or interrupted [8]. It happens if a blood vessel in the brain bursts or leaks or is blocked by a clot in an artery leading to the brain. This causes deprivation of oxygen to the brain cells, resulting in their demise [9]. There are three main types of stroke: Ischemic stroke, Hemorrhagic stroke, and Transient ischemic attack (TIA) [12]. Ischemic strokes are the most common type (87%) [15] of stroke and occurs due to the reduction of blood flow to the brain. If any blood vessel in the brain bursts or leaks blood that is called the Hemorrhagic stroke. TIA or ministroke causes a temporary (<5 minutes) decrease in blood supply to the brain [9].

Stroke is one of the leading causes of death worldwide. A stroke can occur to anyone at any time [10]. The consequences of stroke depend on which brain part is affected and how severe it is [8]. The affected part of the brain loses its functional capacity [14]. Sudden death can be caused by a severe stroke [8]. Each year about 15 million people suffer stroke worldwide and among them, 5 million dies and another 5 million are disabled permanently according to the World Health Organization (WHO) [16]. Additionally, ischemic heart disease and stroke are the top two contributors to the top 10 global causes of death in the year 2016 [1]. According to the WHO, death due to stroke in Bangladesh reached 128,190 or 16.27% of total deaths in 2017 [17].

There are several risk factors behind this disease. Among them, the most important behavioral risk factors are unhealthy diet, physical inactivity, use of illicit drugs, and consumption of tobacco and alcohol. The effects of mentioned behavioral risk factors may lead to the 'intermediate risks factors' as raised blood pressure, raised blood glucose, raised blood lipids, obstructive sleep apnea, cardiovascular diseases, overweight and obesity, etc. These are risk indicators for stroke [9,11,12]. Being of advanced age and family history of stroke also increase the chance of stroke [12,13].

In the medical domain, the disease diagnosis is often done depending on the knowledge and experience of the medical practitioner. Therefore, there are chances of errors, cognitive biases and may also take lengthier time in the proper diagnosis of disease. In the occasion of stroke, the diagnosis is not easy as it requires careful analysis of several clinical, pathological and other computer-aided technologies such as Magnetic Resonance Imaging (MRI), Computerized Tomography (CT scan), blood tests, electrocardiogram (ECG or EKG), etc. In addition, doctors need to eliminate other likely causes of the symptoms that arise in patients. Usually, strokes are diagnosed by carrying out physical tests and studying images of the brain produced during a scan (CT scan and MRI) [18]. These take time and for stroke, time is very crucial as quicker treatment can minimize brain damage [12]. Therefore, the development of a computerized automated system for the risk prediction and earlier diagnosis of stroke can play a pivotal role to minimize the damage due to stroke.

The risk factors for stroke vary from region to region. Again, countries with low and middle-income are prone to stroke. On average, the age of a stroke occurrence is 15 years lower in poorer countries compared to high-income countries [19]. Bangladesh falls between these two categories of country [4] which is why the chances of stroke are significant here. Moreover, the tests, recovery, and rehabilitation after stroke are costly and the cost is not bearable for most of the people of Bangladesh. Therefore, a risk prediction system for stroke using machine learning in the context of Bangladesh may contribute to earlier prediction and better diagnosis.

2 Related Works

Few works have been done for improving the detection of stroke to reduce the mortality rate.

A study has been done in [7] by Asadi et al. to predict potential outcomes of 107 consecutive acute anterior circulation ischaemic stroke patients by comparing generalized linear model (GLM), generalized additive model, support vector machine, adaptive boosting, and random forest. Though their dataset was small, prediction the outcome is approaching to 70%.

Khosla et al. [5] proposed an integrated machine learning approach combining the elements of data imputation, feature selection and prediction on the Cardiovascular Health Study (CHS) dataset. Their Margin-based Censored Regression prediction model and feature selection approach combined with SVM significantly outperformed the Cox proportional hazards model on the dataset for stroke risk prediction.

In [2], Farzana Islam et al. proposed a system based on Fuzzy Logic to analyze the potential risk factors of stroke and predict it on a dataset of Bangladesh context. Fuzzy C-means classifier and Fuzzy Inference System (FIS) has been used to classify between stroke and non-stroke patients. For better prediction they used Adaptive neuro-fuzzy inference system (ANFIS). An accuracy of 95.1% was achieved by their proposed system.

In [3], Li et al. offered an integrated machine learning and data mining approaches for building 2-year thromboembolism (TE) prediction models for Atrial Fibrillation (AF) from Chinese Atrial Fibrillation Registry data. In terms of area under the curve (AUC) and area under the precision recall curve (AUPR), wrapper selection method achieved the best prediction performance though the time complexity is very high. Additionally, generalized linear model (GLM) and Naive Bayes did not work well, but it can be improved by better feature selection.

Rahma et al. in [6], worked on an automatic Acute Ischemic Stroke (AIS) severity classifier built on electroencephalogram (EEG) signals using Wavelet transform and feedforward type of neural network with ELM algorithm. Their proposed systems achieved performance is above 72% in terms of test accuracy, sensitivity, and specificity.

Hung et al. in [21], compared deep neural network (DNN) with three other machine learning approaches to predict 5-year stroke occurrence on a large population-based electronic medical claims (EMC) database of around 800,000 patients. The DNN and gradient boosting decision tree (GBDT) perform very well achieving an AUC of 92%; better performance than logistic regression and SVM.

3 Methodology

3.1 Dataset

The dataset was originally collected by Farzana et al. [2] from Dhaka Medical College, Dhaka, Bangladesh for their research work. This dataset consists of 500 patients' data among them 232 are female and 268 are male. Additionally, 350 patients are diagnosed with stroke and 150 are diagnosed as normal or non-stroke. Data was collected from patients' case history, pathological test and

Lipid profile test. In their work, they used 15 parameters for their work. However, after studying about the risk factors of stroke [20] and consulting with a domain expert it has been concluded that 17 parameters (including class) in Table 1 are convenient for predicting stroke.

Table 1: Selected parameters list

S/N	Parameters
1	Age (years)
2	Sex (Male- 1, Female- 0)
3	Heredity (1°) (Yes- 1, No- 0)
4	Systolic Pressure (SP) (mmHg)
5	Diastolic Pressure (DP) (mmHg)
6	Pulse (bpm)
7	Diabetes Mellitus (DM) (Yes- years, No- 0)
8	Total cholesterol (mg/dL)
9	Tri-glyceride (TGs) (mg/dL)
10	High-Density Cholesterol (HDL) (mg/dL)
11	Low-Density Cholesterol (LDL) (mg/dL)
12	Myocardial infarction (Yes- 1, No- 0)
13	Stroke history (Yes- times, No- 0)
14	Tobacco intake (Yes- 1, No- 0)
15	Pain killer intake (Yes- 1, No- 0)
16	Hours of physical work
17	Class (Stroke- 1 or Normal- 0)

3.2 Oversampling Technique

Oversampling techniques are used to balance the class distribution of a dataset. We used Synthetic Minority Over-sampling Technique (SMOTE) [22] which is one of the most popular oversampling techniques in terms of classification problems. SMOTE is an over-sampling approach where the minority class is over-sampled by creating 'synthetic' examples. It generates synthetic examples by operating in 'feature space' rather than 'data space'. SMOTE can under-sample the majority class too. In our paper, we used SMOTE only to over-sample the minority class to remove bias.

3.3 Feature Dependency

Dependency of features plays a very important role in machine learning as compressing the dataset from higher dimensions to lower dimensions, and finding a representation that is more informative can boost the efficiency of the model. To find out the important features or more contributing parameters in stroke in our dataset, we implemented correlation heatmap.

Correlation Heatmap The correlation among the features of the dataset is calculated by a correlation matrix shown in Fig. 1 as a heatmap. If the correlation index is high then the two features are very closely related. Therefore, one of them can be eliminated from the feature list. From the correlation heatmap, systolic and diastolic blood pressure has the highest correlation index of 0.81 and the other features don't have any notable correlation index. So, eliminating a feature from the dataset is challenging and we proceeded without discarding any feature.

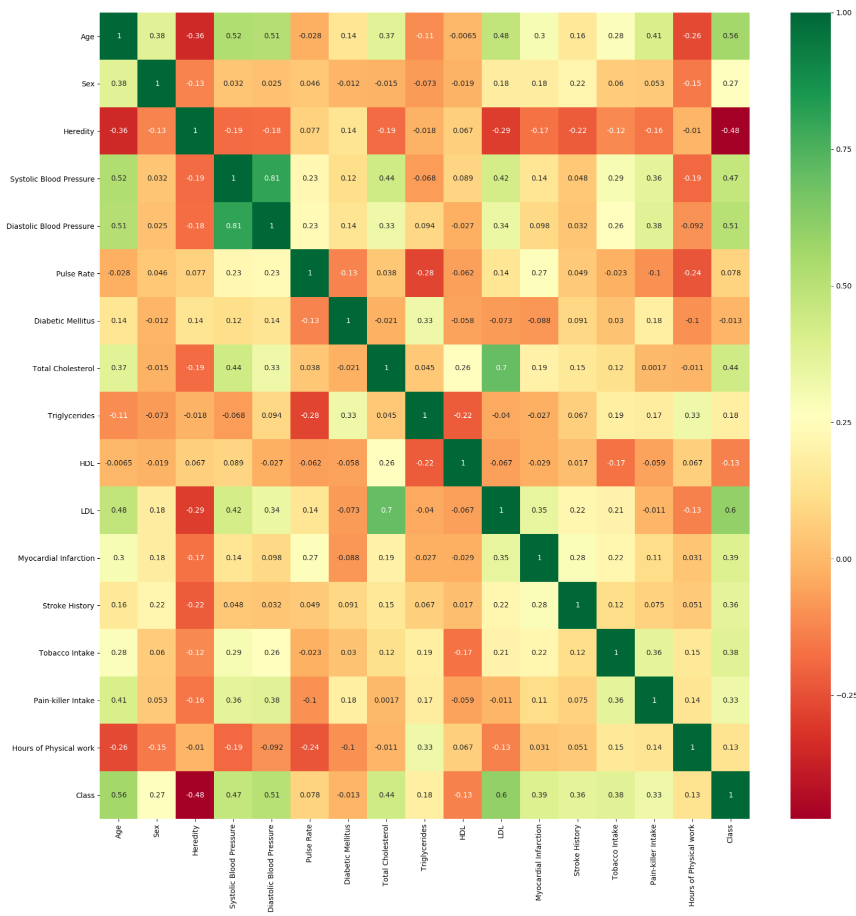


Fig. 1: Correlation Heatmap

3.4 Machine learning models

The proposed prediction system is a supervised classification model. As the purpose is to classify between stroke and normal, we implemented several algorithms which is well known for binary classification problems. Parameter tuning is done by 5-Fold cross validation using Grid Search.

Logistic Regression Logistic Regression is one of the simplest and efficient algorithms for classification problems. It tries to find the best hyperplane which is obtained by minimizing the cost function (hyperplane has $n-1$ dimensions if the data has n features) to separate the classes. It calculates the probability using a hypothesis function which uses a sigmoid function.

Support Vector Machine Support Vector Machine (SVM) is a much prevailing classification algorithm, defined by a separating hyperplane. It produces support vectors and tries to maximize Euclidean distance (margin) between the data points and the decision boundary. It is a non-probabilistic classifier and has the hinge loss function which measures the number of misclassified data examples. In our work, we used Support Vector Classifier (SVC) with kernel trick and Linear Support Vector Classifier (Linear SVC). SVM and Linear SVC both belong to the SVM family, while Linear SVC uses the squared hinge loss function as the loss calculation.

Naive Bayes Naive Bayes (NB) is a probabilistic classifier method of supervised learning algorithms. It is based on applying Bayes theorem, strongly assuming that every pair of features is independent. Among the three classifiers of Naive Bayes, we used Gaussian Naive Bayes (GNB) and Multinomial Naive Bayes (MNB). Bernoulli Naive Bayes is omitted as all feature vectors of our dataset is not binary. In Gaussian Naive Bayes, it is assumed that the features follow a simple Gaussian distribution. On the other hand, in Multinomial Naive Bayes it is assumed that the features are generated by a simple multinomial distribution.

Decision Tree A decision tree is a non-parametric supervised learning method which predicts the value of a target variable. At the root of the tree the input enters and then at each node it traverses down the tree according to the split decision. In our work, the degree of randomness of elements (entropy) has been used to measure the quality of the split.

4 Experimental Setup and Result Analysis

We worked on the dataset in 2 setups as the dataset is imbalanced because the number of stroke patients (350 samples) is significantly larger than non-stroke patients (150 samples). Setup 1 consists of the original dataset while in Setup 2, synthetic data was added using SMOTE for both the train and test sets to make them evenly distributed. The details of both setups are given in Table 2. All the models were trained and evaluated using k-Fold cross validation where $k = 10$ which has been found to be low biased.

Table 2: Description of the setups

	Train Set			**Test Set**			**Subtotal**
	Stroke	Normal	Total	Stroke	Normal	Total	
Setup 1	275	75	350 (70%)	75	75	150 (30%)	500
Setup 2	245	245	490 (70%)	105	105	210 (30%)	700

For classification performance evaluation of the proposed methods, different evaluation techniques are used in this research such as Precision, Recall, F-measure and Receiver Operating Characteristic (ROC) curve along with training and testing accuracy. Among these evaluation metrics, testing accuracy, precision, recall and F-measure can be obtained from the confusion matrix. Precision is a measurement of result relevancy whereas recall is a measure of how many truly relevant results are returned. So, both the precision and recall are important for evaluating the models; where achieving high precision and high recall is the goal. F-measure is also an important measurement for evaluating the models since its the Harmonic Mean of precision and recall.

We started working with Setup 1 which has a uniform test set. Confusion Matrices of the classifiers are given in Table 3.

Table 3: Confusion Matrices of Different Classifiers (Setup 1)

		Predicted	
		0	1
Actual	0	74	1
	1	29	46

(a) Logistic Regression

		Predicted	
		0	1
Actual	0	69	6
	1	13	62

(b) SVM with poly. kernel

		Predicted	
		0	1
Actual	0	74	1
	1	32	43

(c) Linear SVC

		Predicted	
		0	1
Actual	0	75	0
	1	24	51

(d) Gaussian NB

		Predicted	
		0	1
Actual	0	55	20
	1	4	71

(e) Mulitnomial NB

		Predicted	
		0	1
Actual	0	70	5
	1	35	40

(f) Decision Tree

From the confusion matrices, it can be observed that the rate of missclassification is high for all the models. However, it is severe for the stroke class. The comparison of different classifiers in terms of precision, recall, F-measure, and accuracy are shown in Table 4.

Table 4: Comparison of different classifiers in terms of precision, recall, F-measure and accuracy (Setup 1)

Classifier	Class	Precision	Recall	F-measure	Train Acc. (%)	Test Acc. (%)
Logistic Regression	0	0.72	0.99	0.83	99.70	80.0
	1	0.98	0.61	0.75		
SVM Poly kernel	0	0.84	0.92	0.88	99.13	**87.33**
	1	0.91	0.83	0.87		
Linear SVC	0	0.70	0.99	0.82	99.41	78.0
	1	0.98	0.57	0.72		
Gaussian NB	0	0.76	1.0	0.86	87.69	84.0
	1	1.0	0.68	0.81		
Multinomial NB	0	0.93	0.73	0.82	91.69	84.0
	1	0.78	0.95	0.86		
Decision Tree	0	0.67	0.93	0.78	97.43	73.33
	1	0.89	0.53	0.67		

The outcomes from the classifiers are not satisfactory. This result is not really surprising because the models are over-fitted due to the skewedness of the dataset as the train set contains 275 stroke patients data where the number of normal data is only 75. Hence, Setup 2 has been introduced with a uniform train and test set. In Setup 2, we created synthetic data to build uniformly distributed train and test set. SMOTE has been used to oversample the dataset where minor class (Non-stroke or Normal) was increased to 350 from 150. Confusion Matrices of the classifiers for Setup 2 are given in Table 5.

Table 5: Confusion Matrices of Different Classifiers (Setup 2)

		Predicted 0	Predicted 1
Actual	0	104	1
	1	0	105

(a) Logistic Regression

		Predicted 0	Predicted 1
Actual	0	104	1
	1	6	99

(b) SVM with poly. kernel

		Predicted 0	Predicted 1
Actual	0	104	1
	1	0	105

(c) Linear SVC

		Predicted 0	Predicted 1
Actual	0	105	0
	1	19	86

(d) Gaussian NB

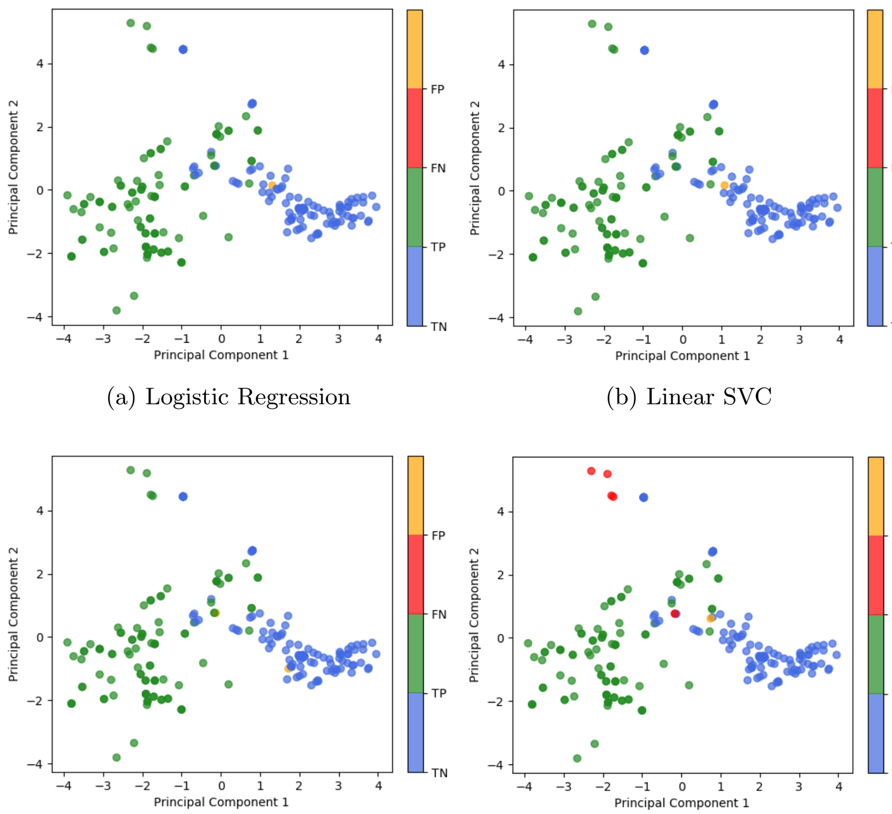
		Predicted 0	Predicted 1
Actual	0	87	18
	1	8	97

(e) Mulitnomial NB

		Predicted 0	Predicted 1
Actual	0	103	2
	1	0	105

(f) Decision Tree

According to Table 5, the missclassification rate is reduced by a great margin for every classifier. Fig. 2 contains the scatter plots of correctly classified and missclassified samples of better performing classifiers. It is showed by decreasing the dimensions to 2 by Principal Component Analysis where x-axis represents principal component 1 and y-axis represents principal component 2. The data samples are plotted according to their labels in the confusion matrix such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Here, Fig. 2a and Fig. 2b have only one missclassified (False Positive) sample. Fig. 2c has two missclassified (False Positive) samples. On the other hand, Fig. 2d classified 7 samples (both False Positive and False Negative) wrongly.



(a) Logistic Regression (b) Linear SVC

(c) Decision Tree (d) SVM with polynomial kernel

Fig. 2: Illustration of Confusion Matrices

The comparison of different classifiers in terms of precision, recall, F-measure, and accuracy are shown in Table 6. The results are significantly better than Setup

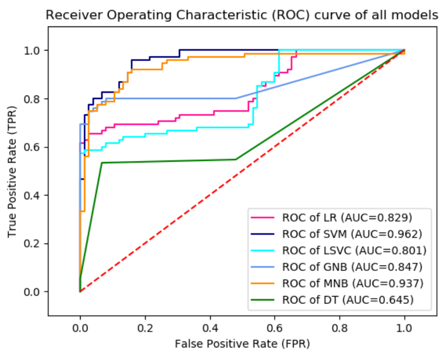
1 as the data is uniformly distributed in Setup 2. Overfitting issue is also resolved in this setup. Logistic Regression and Linear SVC performed very well with an accuracy as high as 99.52% where for Setup 1 it was 80% and 78% respectively. These two classifiers achieved the highest accuracy among the other classifiers. However, the models with low complexity such as Linear models, Bayesian models perform well with small amount of data. For this reason, Linear SVC and Logistic Regression worked well. But, the Bayesian models didn't perform accordingly as bayesian models need the features to be independent. Though the features of the dataset have less mutual dependencies, some of them are related to each other. This impeded the performance of the classifier. On the other hand, the kernel SVM is not a less complex model but it performs well for high dimensional feature set. Besides, Decision Tree can handle numerical and categorical data very well. Thus, it performed well with an accuracy of 99.04%.

Table 6: Comparison of different classifiers in terms of precision, recall, F-measure and accuracy (Setup 2)

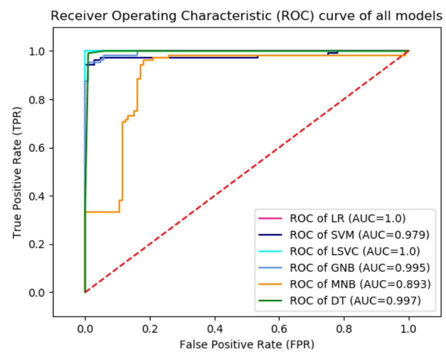
Classifier	**Class**	**Precision**	**Recall**	**F-measure**	**Train Acc. (%)**	**Test Acc. (%)**
Logistic Regression	0	1.0	0.99	1.0	99.58	**99.52**
	1	0.99	1.0	1.0		
SVM Poly kernel	0	0.95	0.99	0.97	97.76	96.66
	1	0.99	0.94	0.97		
Linear SVC	0	1.0	0.99	1.0	98.57	**99.52**
	1	0.99	1.0	1.0		
Gaussian NB	0	0.85	1.0	0.92	89.02	90.95
	1	1.0	0.82	0.90		
Multinomial NB	0	0.92	0.83	0.87	86.80	87.61
	1	0.84	0.92	0.88		
Decision Tree	0	1.0	0.98	0.99	96.55	99.04
	1	0.98	1.0	0.99		

Precision, recall, and accuracy all could be affected slightly negative if the dataset is imbalanced like Setup 1. On the contrary, Receiver Operating Characteristic (ROC) curve gives more credibility to the performance of a model as it is not affected by any imbalanced class. Therefore, the ROC curve is plotted and the score of Area Under the Curve (AUC) is calculated for different classifiers for both setups are shown in Fig. 3.

For Setup 1, though Gaussian Naive Bayes seems better than Multinomial Naive Bayes in terms of confusion matrices and accuracy, Multinomial Naive Bayes has greater AUC score according to Fig. 3a. On the other hand, in Setup 2 SVM has better accuracy than Gaussian Naive Bayes, but Gaussian Naive Bayes has greater AUC score than SVM according to Fig. 3b. However, the ROC curve of Logistic Regression and Linear SVC is at the top left which indicates it outperforms all the other models with an AUC score of 1.0.



(a) Setup 1



(b) Setup 2

Fig. 3: ROC curve and AUC of all models

5 Conclusion

Few linear, non-linear, and Bayesian models have been applied to the dataset collected from a Bangladeshi hospital before and after using SMOTE. Several evaluation techniques including confusion matrix, ROC curve, AUC score have been used to check the model's performance and whether the proposed models are over-fitted or not. While analyzing the results, it is found that when the dataset is not uniformly distributed the models are overfitted. Unlike that, after generating synthetic data by SMOTE and making the dataset uniform, none of the classifiers are over-fitted and the performances become exceedingly commendable. The accuracy achieved by Logistic Regression and Linear SVC both found to be 99.52%. The performance of Decision Tree is also impressive. However, the performance of the Bayesian models is not as good as the previous ones due to some degree of feature dependencies. Additionally, few of our models exceed the accuracy achieved (95.1%) by Farzana Islam et al.'s [2] proposed method using ANFIS on the same dataset. Due to the promising result in prognosticating stroke, this proposed method could be used in the real world to predict strokes earlier to depreciate the loss and also might help in more reliable diagnosis of the disease.

Acknowledgement. Authors would like to show their gratitude to Farzana Islam, North South University for great help by providing the dataset without which we wouldn't be able to conduct the research.

References

1. 'The top 10 causes of death', 2018. [online]. Available at: http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [Accessed 5- July- 2018].
2. F. Islam, S. B. A. Shoilee, M. Shams, and R. M. Rahman, "Potential risk factor analysis and risk prediction system for stroke using fuzzy logic", In *6th proceedings of Computer Science On-line Conference 2017 (CSOC2017)*, vol. 1, pp. 262–272, 2017.
3. X. Li, H. Liu, X. Du, P. Zhang, G. Hu et al. "Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation", In *AMIA Annual Symposium Proceedings*, pp. 799-807, 2016.
4. 'Bangladesh to be declared lower-middle income country in March', 2018. [online]. Available at: http://www.daily-sun.com/post/285331/Bangladesh-to-be-declared-lowermiddle-income-country-in-March [Accessed 3- Aug- 2018].
5. A. Khosla, Y. Cao, C. Lin, H. Chiu, J. Hu and H. Lee, "An integrated machine learning approach to stroke prediction", In *16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*, pp.183–192.
6. O. N. Rahma, S. K. Wijaya, Prawito and C. Badri, "Electroencephalogram analysis with extreme learning machine as a supporting tool for classifying acute ischemic stroke severity", In *2017 International Seminar on Sensors, Instrumentation, Measurement and Metrology (ISSIMM)*, pp. 180–186, 2017.
7. H. Asadi, R. Dowling, B. Yan and P. Mitchell, "Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy", *PLOS ONE 9*, no. 2, pp. 1-11, 2014.
8. 'Stroke, Cerebrovascular accident' [online]. Available at: http://www.who.int/topics/cerebrovascular_accident/en/ [Accessed 10- Jul- 2018].
9. 'Stroke', 2018. [online]. Available at: https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113 [Accessed 10- Jul- 2018].
10. 'What is stroke?' [online]. Available at: http://www.stroke.org/understand-stroke/what-stroke [Accessed 11- Jul- 2018].
11. S. C. Johnston, S. Mendis, C. D. Mathers, "Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling", *The Lancet Neurology*, vol. 8, no. 4, pp. 345–354, 2009.
12. J. McIntosh, 'Everything you need to know about stroke', 2017. [online]. Available at: https://www.medicalnewstoday.com/articles/7624.php [Accessed 12- Jul- 2018].
13. 'Stroke facts', [online]. Available at: http://www.stroke.org/understand-stroke/what-stroke/stroke-facts [Accessed 12- Jul- 2018].
14. B. Wedro, 'Stroke Signs, Symptoms, Warning Signs, and Treatments', [online]. Available at: https://www.medicinenet.com/stroke_symptoms_and_treatment/article.htm#stroke_definition_and_facts [Accessed 13- Aug- 2018].
15. E. J. Benjamin, M. J. Blaha, S. E. Chiuve et al., "Heart disease and stroke statistics-2017 update: a report from the American Heart Association", *American Heart Association Statistics Committee and Stroke Statistics Subcommittee*, vol. 135, no. 10, 2017.
16. 'Stroke Statistics', [online], Available at: http://www.strokecenter.org/patients/about-stroke/stroke-statistics/ [Accessed 14- Aug- 2018].

17. 'Bangladesh Stroke', [online] Available at: http://www.worldlifeexpectancy.com/bangladesh-stroke [Accessed 14- Aug- 2018].
18. 'Diagnosis Stroke', [online]. Available at: https://www.nhs.uk/conditions/stroke/diagnosis/ [Accessed 16- Aug- 2018].
19. 'Stroke: a global response is needed.' [online]. Available at: http://www.who.int/bulletin/volumes/94/9/16-181636/en/ [Accessed 16- August-2018].
20. N. R. Colledge, B. R. Walker, S. Ralston and S. Davidson, *Davidson's principles and practice of medicine*. Edinburgh: Churchill Livingstone/Elsevier.
21. C. Hung, W. Chen, P. Lai, C. Lin and C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database", In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3110–3113, 2017.
22. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", In *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.