Predicting Stroke Risk: A Machine Learning Approach for Identifying Influential Factors and Early Detection

Sabbir Hussain Meraj[a,1], Md. Nafis Faisal[a], Dr. B. M. Mainul Hossain[a]

[a]*Institute of Information Technology, University of Dhaka,*

***Abstract* - Stroke, a leading cause of mortality and morbidity worldwide, demands effective risk assessment and early detection to mitigate its impact on public health. In this research, we employ five distinct machine learning algorithms—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, Random Forest, and Logistic Regression—to predict stroke risk. Through rigorous experimentation and analysis, we identify the Random Forest model as the standout performer, boasting an impressive Area Under the Curve (AUC) of 0.979 and an Accuracy of 0.964. Furthermore, the SVM model emerges as a robust alternative, with a commendable AUC of 0.973 and an F1-Score of 0.920. Our investigation extends beyond predictive accuracy to elucidate the pivotal factors influencing stroke risk. Notably, our analysis underscores age, heart disease, and hypertension as significant contributors to stroke risk prediction. These findings provide crucial insights for healthcare practitioners, enabling them to focus their efforts on the most salient factors when assessing and managing stroke risk in patients.**

***Keywords* - Stroke, Machine Learning, Healtcare, Area Under Curve (AUC), Feature Importance, Classification**

1. Introduction

Stroke, a critical medical emergency, arises from the disruption or reduction of blood flow to specific regions of the brain, resulting in a deprivation of essential nutrients and oxygen to brain cells, ultimately leading to their demise. Stroke manifests in two primary forms: ischemic and hemorrhagic. Ischemic strokes are characterized by the occlusion of blood vessels due to clots, while hemorrhagic strokes result from the rupture of weakened blood vessels within the brain, leading to bleeding. A significant worldwide burden of stroke, impacting nearly 5.5 million individuals each year, leading to a distressing loss of lives at an alarming rate of every 4-5 minutes and chronically disabling up to 50% of survivors [5]. Timely identification of individuals at risk of stroke is pivotal for effective preventive measures and better patient outcomes. Machine learning, a subset of artificial intelligence, has emerged as a promising tool in the field of medical research for early disease prediction and risk assessment.

This research paper presents an in-depth investigation into the development and evaluation of machine learning-based predictive models for stroke risk assessment. We have used 5 machine learning models - Logistics Regression, Naive Bayes, Random Forest, Support Vector Machine & KNN.

Our research aims to address several critical questions in the context of stroke prediction using machine learning:

1. Which features and variables are the most relevant for accurate stroke risk prediction?
2. What machine learning algorithms and techniques are best suited for building robust and accurate predictive models for stroke risk assessment?

The rest of the paper is organized as following. Section 2 discusses some literature review on the existing research. Research methodologies are stated in section 3 and it is separated as three parts: dataset description, feature selection techniques and classification technique are discussed. In section 4, findings and result analysis are discussed. Finally, the conclusion is discussed in section 5.

2. Related Work

Numerous researchers have already applied machine learning methodologies to predict the occurrence and outcomes of strokes. Cheng et al. [3] focused on estimating the prognosis of ischemic strokes, drawing data from 82 patients. They deployed two distinct Artificial Neural Network (ANN) models, achieving precision rates of 79% and 95%.

Singh et al. [12] conducted a study on stroke prediction, utilizing a diverse range of methodologies on the Cardiovascular Health Study (CHS) dataset. They attained a remarkable 97% accuracy by employing the decision tree algorithm for feature extraction, followed by neural network classification.

Monteiro et al. [9] focused on predicting the functional outcomes of ischemic stroke patients using machine learning, achieving an impressive AUC value exceeding 90%. Kansadub et al. [20] conducted a study to predict stroke risk, employing Naive Bayes, Decision Trees, and Neural Networks for data analysis.

Govindarajan et al. [6] conducted a study involving 507 patients, employing a combination of text mining and machine learning to categorize stroke disorders. They achieved an impressive accuracy rate of 95% using the Stochastic Gradient Descent (SGD) algorithm.

Amini et al. [1] collected data from 807 subjects and categorized 50 risk factors for strokes, including diabetes, car-

diovascular disease, smoking, hyperlipidemia, and alcohol use. Their approach utilized the C4.5 decision tree algorithm and K-nearest neighbor, yielding accuracies of 95% and 94%, respectively.

Sung et al. [13] aimed to create a stroke severity index, examining data from 3,577 patients with acute ischemic strokes. Their most effective model employed the k-nearest neighbor approach, producing a promising result of 95%.

Cheon et al. [4] sought to predict stroke patient mortality using a deep neural network approach coupled with Principal Component Analysis (PCA) on data from 15,099 patients. Their study yielded a commendable Area Under the Curve (AUC) value of 83%.

3. Methodology

This section is divided into three part, these are: Dataset Description, Feature Selection Technique, Classification approach.

3.1. Dataset Description

Within the scope of this research study, our dataset originated from medical clinics, where practical considerations sometimes led to certain attributes having missing values. To maintain data quality and analysis integrity, instances with missing attribute values (a total of 201 instances) were methodically excluded from the dataset. This step was taken to ensure that our dataset remained robust and reliable for the specific task of stroke prediction. In the dataset, there were originally 4909 instances, with the majority class having 4700 instances and the minority class having 209 instances. To address the issue of class imbalance, we employed the SMOTE technique, which resulted in a dataset of 7154 instances (4700 belonging to the majority class and 2454 belonging to the minority class)

Moreover, we conducted a thorough feature selection process to focus on attributes directly pertinent to stroke prediction. As part of this process, one specific attribute, referred to as "ID," was identified as non-contributory and subsequently removed from the dataset.

As a result of these meticulous data curation and feature selection procedures, our final dataset comprises 7,154 instances, each characterized by 11 carefully chosen features.

According to selection, the features are as follows:

1. Age: It refers to a person's age. It is numerical data.
2. Gender: It refers to a person's gender, categorized as categorical data.
3. Hypertension: This attribute denotes whether an individual is hypertensive or not. It has value of either 1 or 0.
4. Work Type: This attribute characterizes an individual's occupational scenario and falls under categorical data.
5. Resindece Type: This attribute delineates the living arrangement of an individual, categorized as categorical data. These categories are: rural & urban.
6. Heart Disease: This attribute denotes whether an individual has heart disease or not. It has value of either 1 or 0.
7. Average Glucose Level: This attribute refers to a person's glucose level, represented as numerical data.
8. Body Mass Index: BMI, reflecting an individual's body mass index, is conveyed as numerical data.
9. Ever Married: It refers to a person's marital status. It has value of either yes or no.
10. Smoking Status: This attribute encapsulates a person's smoking habits and falls under categorical data. These categories are: unknown, smokes, formerly smoked & never smokes.
11. Stroke: It refers to whether an individual has previously experienced a stroke or not. It is encoded as numerical data.

Here, target class is Stroke.

3.2. Feature Selection Technique

Within this study, we employed the Information Gain Attribute Evaluator for feature selection and assessing the relative importance of features. Information Gain (IG) serves as an entropy-based feature evaluation technique, quantifying the knowledge acquired from specific features. It evaluates the worth of an attribute by measuring the information gain with respect to the class. Initially, IG computes the Entropy, which serves as a measure of "impurity" within the dataset. For effective classification, it is crucial to reduce entropy, as higher entropy levels introduce increased uncertainty. The entropy of a given set, denoted as E(T), is defined as follows:

$$E(T) = \sum_{i=1}^{C} (P_i * \log_2 P_i) \tag{1}$$

Where, C is the set of classes in T, and Pi is the proportion of the number of elements in class i to the number of elements in set T.

According to the Entropy, the Information Gain Attribute Evaluator calculates the Information Gain (IG). The Information Gain of T for attribute a, IG (T, a) is defined as follows:

$$IG(T, a) = E(T) - E(T|a) \tag{2}$$

where, $E(T|a)$ is the conditional entropy of T given the value of attribute a.

3.3. Classification Technique

According to features, five machine learning model are trained for predicting stroke risk. These models are:

Logistic Regression: Logistic Regression, in its fundamental form, is a statistical model that utilizes a logistic function to effectively model a binary dependent variable. This method finds extensive application, particularly when dealing with categorical dependent variables. In the medical domain, logistic regression has been employed to create various assessment scales aimed at quantifying the severity of a patient's condition [10].

Naive Bayes: The Naive Bayes classification method is a family of algorithms grounded in Bayes' theorem. It assumes that each pair of features under consideration for classification

is independent of one another. This approach finds practical utility in the realm of automated medical diagnosis [11].

Random Forest: Random Forests represent an ensemble learning technique that operates by constructing a multitude of decision trees during the training process. The core objective of this method is to generate multiple trees within randomly selected subspaces of the feature set. These trees, situated in diverse subspaces, contribute to classification in unique ways, collectively enhancing the effectiveness of their combined classification [14].

Support Vector Machine (SVM): SVM is a classification technique primarily designed for two-group classification problems. In this method, input vectors undergo a non-linear transformation, projecting them into a considerably high-dimensional feature space. Within this feature space, a linear decision boundary is established. The distinctive attributes of this decision boundary contribute to endowing the learning machine with robust generalization capabilities [2].

K-Nearest Neighbors (KNN): KNN is a classification and regression algorithm that relies on the proximity of data points in feature space. It determines the class of an instance by considering the classes of its nearest neighbors, where 'K' denotes the number of neighbors to be considered. This approach allows KNN to make predictions based on the local characteristics of data, making it particularly useful in situations where data exhibits spatial clustering or patterns [7].

To evaluate the models, we used five different metrics. Among the metrics used for model evaluation, AUC is a preferred measure when comparing algorithms due to its distinct advantages [8]. The metrics used for comparing machine learning models are:

1. AUC (Area Under the ROC Curve): AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots True Positive Rate (Recall) against False Positive Rate (FPR) at various threshold values.

2. Precision : Precision is the ratio of true positive predictions (TP) to the total number of positive predictions (TP+FP).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

3. Recall : Recall is the ratio of true positive predictions (TP) to the total number of actual positive instances (TP + FN).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

4. Accuracy : Accuracy is the ratio of correct predictions (TP + TN) to the total number of predictions (TP + TN + FP + FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

5. F1-Score : The F1 Score is a metric used to balance precision and recall in binary classification. It provides a single value that combines the precision (the accuracy of positive predictions) and recall (the ability to identify all positive instances) into a single measure.

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6}$$

4. Findings & Result Analysis

4.1. Feature Ranking

The feature ranking based on Information Gain provides valuable insights into the relative importance of various attributes in predicting cardiovascular risk. The Ranked Features are given in Table 1:

Ranking	Features
1	Age
2	Heart Disease
3	Hypertension
4	Marital Status
5	Occupation
6	Gender
7	Residence Type
8	Smoking Status
9	BMI
10	Average Glucose Level

Table 1: Ranking of Features based on Information Gain

4.2. Classification Performance

For testing purposes, we used a subset of 1431 instances. The AUC (Area Under the Curve) is the primary metric utilized to determine whether the model assigns higher probabilities to individuals who have had a stroke as opposed to those who have not. Additionally, various other metrics for model evaluation, including accuracy, precision, recall, and F1-score, are computed in this research. This study assesses the performance of five distinct machine learning models, namely SVM, Random Forest, Logistic Regression, KNN and Naïve Bayes.

The results regarding False Positive Rate and False Negative Rate are visually presented in Figure-1.

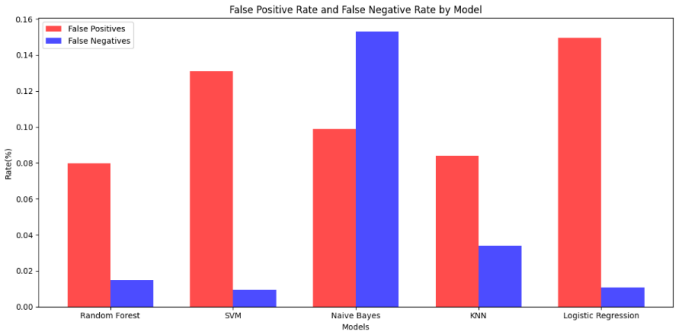

Figure 1: FPR & FNR of Evaluated Machine Learning Models

The outcomes of machine learning model evaluation are presented in Table-2.

The Random Forest model exhibits a remarkable AUC score of 0.979, indicating excellent discriminative ability. It maintains a strong balance between Precision (0.969) and Recall (0.924), resulting in a high F1-Score of 0.946. The Accuracy

	AUC	Precision	Recall	F-1 Score	Accuracy
Random Forest	**0.979**	0.969	0.924	**0.964**	**0.963**
SVM	0.973	**0.979**	0.869	0.920	0.949
Naive Bayes	0.861	0.935	**0.951**	0.943	0.895
Logistic Regression	0.943	0.976	0.856	0.908	0.942
KNN	0.923	0.933	0.916	0.924	0.949

Table 2: Classification Performance

of 0.964 suggests that it correctly classifies instances most of the time. This model stands out as a top performer.

The SVM model also demonstrates strong performance with an AUC of 0.973. It excels in Precision (0.979), which suggests a low rate of false positives. However, its Recall (0.869) is relatively lower, indicating some missed positive cases. The F1-Score of 0.920 shows a good balance between Precision and Recall, and the high Accuracy of 0.949 underscores its effectiveness.

The Naive Bayes model achieves a decent AUC score of 0.861 and excels in Recall (0.951), suggesting a low rate of false negatives. However, its Precision (0.935) is relatively lower, resulting in a lower F1-Score of 0.943. The Accuracy of 0.895 indicates reasonably accurate classification.

The Logistic Regression model achieves a solid AUC score of 0.943. It maintains a good balance between Precision (0.976) and Recall (0.856), resulting in a competitive F1-Score of 0.908. The Accuracy of 0.942 suggests reliable classification performance.

The K-Nearest Neighbors (KNN) model shows respectable AUC performance at 0.923. It maintains a good balance between Precision (0.933) and Recall (0.916), resulting in a commendable F1-Score of 0.924. The model's high Accuracy of 0.949 indicates reliable classification.

The Random Forest model stands out as the best choice with an AUC of 0.979 and an Accuracy of 0.964. The SVM model also offers a strong option with a high AUC of 0.973 and a good F1-Score of 0.920.

5. Conclusion

In this study, we harnessed the capabilities of machine learning to predict stroke risk and identify influential factors within a medium range dataset. The practical implications of this research are substantial. Clinicians can utilize these machine learning tools to identify individuals at risk of stroke, enabling timely interventions and tailored preventive strategies. Our findings reveal promising results that can significantly impact healthcare practice.

Among the models examined, Random Forest stood out with an AUC of 0.979 and an impressive accuracy of 96.4%. The Support Vector Machine (SVM) model also performed well, emphasizing precision. These models hold the potential for precise stroke risk assessment.

Our analysis pinpointed key factors affecting stroke risk. Notably, age, heart disease, and hypertension emerged as significant contributors. These insights equip healthcare providers with valuable knowledge for early intervention.

Future works will involve refining our models further, incorporate additional features, and expand our dataset for enhanced predictive accuracy. Prospective clinical studies will validate the real-world applicability of our findings.

In summary, this study bridges the gap between machine learning and stroke risk assessment within a medium-range dataset, offering substantial benefits to healthcare practices and emphasizing the importance of early stroke prevention.

References

[1] L. Amini, M. T. Farzadfar R. Azarpazhouh, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar. Prediction and control of stroke by data mining. *International Journal of Preventive Medicine*, 4(Suppl 2):245–249, May 2013.

[2] Cortes C and Vapnik V. Support-vector networks. *Machine Learning*, 20(3):273–97, Sep. 1995.

[3] C.-A. Cheng, Y.-C. Lin, and H.-W. Chiu. Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. *Studies in Health Technology and Informatics*, 202:115–118, 2014.

[4] S. Cheon, J. Kim, , and J. Lim. "the use of deep learning to predict stroke patient mortality. *International Journal of Environmental Research and Public Health*, 6(11), 2019.

[5] Donkor and Eric. Stroke in the 2 1 s t century: A snapshot of the burden, epidemiology, and quality of life. *Stroke Research and Treatment*, 2018:1–10, 11 2018.

[6] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomia, R. Patan, P. Jayaraman, and R. Manikandan. Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications*, 32(3):817–828, Feb. 2020.

[7] Guo, Gongde, Wang, Hui, Bell, David, Bi, and Yaxin. Knn model-based approach in classification. 08 2004.

[8] Ling, Charles, Huang, Jin, Zhang, and Harry. Auc: A better measure than accuracy in comparing learning algorithms. pages 329–341, 01 2003.

[9] M. Monteiro, A. C. Fonseca, A. T. Freitas, T. Pinho e Melo, A. P. Francisco, J. M. Ferro, and A. L. Oliveira. Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15:1953–1959, Nov. 2018.

[10] Peng, Joanne, Lee, Kuk, Ingersoll, and Gary. An introduction to logistic regression analysis and reporting. *Journal of Educational Research - J EDUC RES*, 96:3–14, 09 2002.

[11] Rish and Irina. An empirical study of the naïve bayes classifier. *IJCAI 2001 Work Empir Methods Artif Intell*, 3, 01 2001.

[12] M. S. Singh and P. Choudhary. Stroke prediction using artificial intelligence. *8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pages 158–161, Aug. 2017.

[13] S.-F. Sung, C.-Y. Hsieh, Y.-H. Kao Yang, H.-J. Lin, C.-H. Chen, Y.-W. Chen, and Y.-H. Hu. Developing a stroke severity index based on administrative data was feasible using data mining techniques. *Journal of Clinical Epidemiology*, 68(11):1292–1300, Nov. 2015.

[14] Ho TK. Random decision forests. *3rd international conference on document analysis and recognition*, 1:278–282, Aug. 1995.