Towards Stroke Detection through Feature Selection-Based Machine Learning

***Abstract*—A stroke is a health disease that causes damage to the brain by rupturing blood vessels. It can also happen when the brain's blood supply and other nutrients are cut off. Stroke is the most significant cause of death and disability worldwide, according to the World Health Organization (WHO). Most of the stroke occurs when an unexpected obstruction blocks the brain and heart. By examining the affected individuals, several risk variables believed to be related to the cause of stroke have been discovered. A lot of studies have been conducted using these risk factors to predict and diagnose stroke disorders. Early detection of stroke warning signs can minimize stroke severity. This is why we proposed a model that can successfully detect stroke from the data through machine learning. Using different machine learning approaches, we present 15 attributes to use for early stroke prediction. Six different classifiers have been trained based on the high features attributes. The classification models are Naive Bayes, Random Forest, J48, Nearest Neighbour, SGD, and Support Vector Machine. The best algorithm for performing this task is J48, which yielded an accuracy of approximately 97.87%.**

***Index Terms*—Stroke, Machine Learning, Naive Bayes, Support Vector Machine, Random Forest, Nearest Neighbour, SGD and J48.**

I. Introduction

A stroke occurs when the blood supply to part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients [1]. This injury can occur at any time and may cause loss of motion, chest pain, impotence, memory loss, or even death. Adults of all ages are affected by stroke. The main reasons for stroke in the modern day are the changes in lifestyle due to high blood sugar, diabetes, obesity, cardiac disease, and so on. There are mainly three types of stroke- Ischemic stroke, Hemorrhagic stroke, Transient ischemic stroke (also called a "mini-stroke"). Ischemic stroke is a severe condition that requires immediate medical attention. It occurs when the blood flow through the artery that delivers oxygen-rich blood to the brain becomes blocked. Hemorrhagic strokes can happen when a blood vessel bursts in the brain and blood begin leaking into it. As a result, the damaged area becomes dysfunctional. There is a possibility that it could be life-threatening. Blood vessel blockages cause most transient ischemic strokes. Of all stroke types, ischemic stroke is the most common (87%) [2].

Stroke can be avoided by leading a healthy/balanced lifestyle, including quitting harmful habits like smoking and drinking, having a healthy body mass index (BMI) and average glucose level, and maintaining good heart and kidney function. Stroke detection is essential, and it must be treated to avoid irreparable damage or death. Hypertension, BMI, heart disease, and average glucose level were used as criteria to predict stroke. Furthermore, machine learning can play an essential part in the prediction system's decision-making processes [3], [5].

According to the Center for Disease Control and Prevention(CDC), around 5% of total deaths have been caused by stroke in the past 40 years in Bangladesh [21]. Several stroke risk factors regulate each type of stroke. Machine learning techniques are undoubtedly worth researching in terms of detecting the probability of a stroke. Among the sub-fields of Artificial Intelligence, machine learning includes improvements in calculations about making forecasts based on information. Machine learning is a data analysis technology that automates the creation of logical models [6]. The iterative phase of machine learning is critical because models can evolve independently when exposed to new data samples. They learn from prior computations to make decisions that are consistent and reproducible. Using this technology, the algorithms are defined as precise representations of the data sets used by the computer to evaluate the problems. Machine learning algorithms are computer programs (math and logic) that adapt to new data and improve performance. The "learning" aspect of machine learning is that these programs modify how they analyze data over time, just like people do. Early detection and treatment are possible with these algorithms. We have used several machine learning algorithms to identify the different strokes that can occur or have already appeared in patients based on their clinical notes and statistical data.

This paper is concerned with building a model for detecting stroke by WEKA [7]machine learning. Weka is a set of data mining-related machine learning techniques. It includes data preprocessing, categorization, regression, segmentation, mining of association rules, and visual data. The fundamental *objective* of this work is to use a variety of machine learning algorithms to detect strokes. We also compare and contrast various machine learning algorithms. Our work covers: (i) Experimenting with various feature selection methods, (ii) Analyzing the impact of distinct features, and (iii) Analyzing the behavior of numerous machine learning algorithms.

The rest of the paper is organized as follows: In section II, we discussed previous research works that were done to detect stroke. In section III, the whole methodology has been explained in detail, including a description of the data processing, feature selection, and development and evaluation of the prediction model. Section IV consists of the assessment and results of the prediction model, followed by the conclusion in section V.

II. Related Works

In the last few years, there have been several publications on machine learning algorithms. Several of them are discussed here:

A. Research works

As a method for classifying stroke disease, Govindarajan et al. [8] used Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, Logistic Regression, and ensemble methods (Bagging and Boosting). Their data comes from the Sugam Multispeciality Hospital in India, which contains information about 507 stroke patients in the range of 35 to 90 years old. An algorithm called novel stemmer was used to obtain the dataset, which is a novelty of their work. They found that 91.52 percent of stroke patients had an ischemic stroke, while only 8.48% had a hemorrhagic stroke. When comparing the three algorithms, the stochastic gradient descent algorithm with artificial neural networks has the highest accuracy with 95.3% for classifying strokes.

Amini et al. [9] studied 807 healthy and unhealthy individuals, and 50 different risk factors were determined, such as diabetes, heart disease, smoking, hyperlipidemia, and alcohol use. Based on the C4.5 decision tree algorithm, the accuracy was 95%, and for K-nearest neighbors, it was 94%.

Shanthi et al. [10] proposes a model for the prediction of thromboembolic strokes using an Artificial Neural Network to assist existing diagnosis methods. The dataset contains 50 patients with stroke symptoms. After examining all fifty cases with the assistance of a physician, 25 parameters are selected. Additionally, a backward stepwise method is used to examine the 25 parameters on the list and finalize 20 parameters. ANNs were trained using back-propagation algorithms, and they were tested against the different stroke categories. The results of this study show that ANN-based stroke prediction improves diagnostic accuracy by 89%.

The Cardiovascular Health Study dataset (CHS) was used to predict stroke using five machine learning techniques in [11]. A combination of Decision Trees and the C4.5 algorithm was used with Principal Component Analysis, Artificial Neural Networks, and Support Vector Machines, as an optimal solution. But the CHS Dataset used for this study contained fewer parameters.

Chin et al. [12] researched to see if they might detect an automated early ischemic stroke. The primary goal of their research was to create a system that could automate prior ischemic stroke using CNN. To train and test the CNN model, they gathered 256 pictures. They utilized the dataprolongation approach to increase the assembled picture in their system image prepossessing to eliminate the im-possible area that cannot arise of stroke. Their CNNmethod has a 90% accuracy rate.

Sung et al. [13] created a stroke severity index. A total of 3577 patients with acute ischemic stroke had their data collected. They employed data mining methods and linear regression to create their forecasting models. From the k-nearest neighbor model, their prediction feature produced the best results (95 percent CI)

A stroke prediction has been made using social media posts in [14]. Using the DRFS method, the authors identify various symptoms associated with stroke. As a result of using Natural Language Processing to extract the texts from social media posts, the model has a longer execution time than is desirable.

B. Different Dataset

- An Artificial Neural Network (ANN) model for stroke prediction has been developed by Singh and Choudhary [15]. These datasets come from the Cardiovascular Health Study (CHS). The data are divided into three datasets that each contain 212 strokes (all three) and 52, 69, and 79 non strokes.In the feature selection process, the C4.5 decision tree algorithm and Principal Component Analysis (PCA) were used. To implement ANNs, they use a back propagation approach.Their accuracy for the three datasets was 95%, 95.2%, and 97.7%, respectively. The decision tree algorithm and k-nearest neighbor (k-NN) are used by Adam et al. to develop a classification model for ischemic stroke.It is the first dataset for ischemic disease to be collected from Sudanese hospitals and medical centers.It includes 15 features and information about 400 patient. Finally the experiment shows that decision tree classification offers better performance than the k-NN algorithm.
- The stroke prediction dataset comes from Kaggle [16]. There are 5110 rows and 12 columns in this dataset. The main attributes of the columns are 'id,'gender,'age,'hypertension,heart disease,ever married,'work type,'Residence type,'avg glucose level,'BMI','smoking status,'stroke'. The value of the output column 'stroke' is either '1' or '0.' The value '0' indicates that there is no risk of stroke, but the value '1' suggests that there is a chance of stroke. The probability of a '0' in the output column ('stroke') exceeds the possibility of a '1' in the same column, resulting in a severely imbalanced dataset. Only 249 rows in the stroke column have the value '1', while 4861 rows have the value '0'.
- Support Vector Machine modeling has been proposed for stroke prediction by Jeena and Kumar [17].The data have been collected from the International Stroke Trial Database. There are 12 risk factors (attributes) in the dataset. For their study, they used 350 samples. To train and test, 300 samples were used. There were several kernel functions that were applied, including polynomial, quadratic, radial basis function, linear and others. In testing, the linear kernel was found to be the most accurate at 91%, which is the equivalent to 91.7 balance measure F1-scores.

III. Methodology

In this section, we'll go over the processes that must be followed in order to construct a suitable model. We use the

International Stroke Trial Database (IST) [18]. This database has information of 19,435patients including stroke, non-stroke and undiagnosed patients. We filter our data in a number of built-in filter techniques to find the key characteristics. We generate a model in WEKA to apply our machine learning algorithms.

For the algorithms, we take six basic classifiers which are Naive Bayes, Random Forest,J48, Support Vector Machine, SGD and Nearest Neighbour. The classifier from which we get the highest accuracy is the best model for this study. While discussing this section, we illustrated the workflow in Fig. 1 and thoroughly followed the stages.

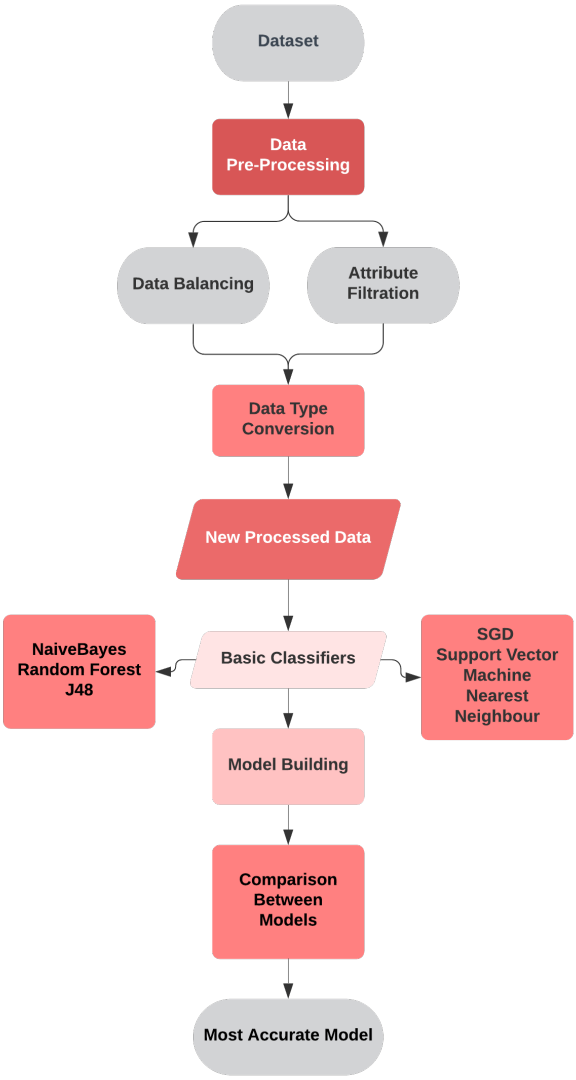

Fig. 1. Flowchart of working process

A. Dataset

We used an international trial dataset [19]. Age, sex, time from onset to randomized, presence or absence of atrial fibrillation (AF), aspirin administration within three days previous to randomized, systolic blood pressure at randomized, degree of awareness, and neurological impairment are baseline variables included in the dataset. The impairments were categorized as complete anterior circulation syndrome (TACS), partial anterior circulation syndrome (PACS), posterior circulation syndrome (POCS), or lacunar syndrome by the Oxford Community Stroke Project (OCSP) (LACS). Recurrent stroke, pulmonary embolism, and mortality were all retrieved within 14 days (date and cause of death). Table I represents all the attributes and their values.

TABLE I
LIST OF ATTRIBUTES OF THE DATASET

SL No.	Attributes	Explanation
1	Age	Age in years
2	Sex	M = Male; F = Female
3	RXHEP	Trial heparin allocated(M/L/N)
4	DMAJNCH	Major non-cerebral haemorrhage (Y/N/U = unknown)
5	DMAJNCHD	Date of Above (days elapsed from randomisation)
6	DMAJNCHX	Comment on above
7	DSIDE	Other side effect (Y/N/U=unknown)
8	DSIDED	Date of above (days elapsed from randomisation)
9	DSIDEDX	Comment on above
10	FDEADX	Comment on death
11	CMPLASP	Compliant for heparin (N/Y)
12	CMPLHEP	Compliant for heparin (N/Y)
13	DEAD7	Other vascular or unknown (1 = Yes, 0 = No)
14	TRAN14	Indicator of major non-cerebral bleed within 14 days (1 = Yes, 0 = No)
15	NCB14	Indicator of any non-cerebral bleed within 14 days (1 = Yes, 0 = No)

After six months, they collected data on the degree of recovery, residence, and current antiplatelet or anticoagulant medication usage, as well as mortality (date and cause of death). Initial stroke, recurrent ischaemic stroke, recurrent hemorrhagic stroke, pneumonia, coronary artery disease, pulmonary embolism, other vascular cause, or non-vascular causes were all included as causes of death. Patients were categorized into one of six groups based on where they lived six months after a stroke: their own home, relatives' homes, residential care, nursing homes, other hospital departments, or unknown.

The final diagnosis is somewhat inaccurate, given that 5569 people were scanned for the first time after selection and 846 were not examined at all. However, because the study was conducted to treat, all people were included in the study, regardless of the ultimate diagnosis. Table II shows the number of patients with each final diagnosis.

TABLE II
THE NUMBER OF PATIENTS AFTER FINAL DIAGNOSIS

Type of Strokes	Number of patients
Ischemic Stroke	17,398
Hemorrhagic Stroke	599
Definite Stroke (pathological type unknown)	992
Non-Stroke	420
Uncertain Diagnosis	26
Total	**19,435**

B. Data Collection

We utilized the International Stroke Trial Database (IST), one of the biggest randomised studies in acute stroke ever undertaken, which is open to the public for trial preparation and secondary analysis. They have extracted data on the factors assessed at randomized, the early outcome point (14 days after randomized or prior discharge), and the 6-months outcome point for each randomized patient and provided them as an analyzed database. This database comprises information on 19,435 individuals who had an acute stroke and were followed for 99 percent of the time. At the time of study entrance, about 26.4 percent of patients were above the age of 80.

C. Data Pre-Processing

For developing a model, data preprocessing is essential to eliminate undesirable noise and outliers from the dataset, which can cause a deviation from good training. This stage takes care of everything that prevents the model from performing as efficiently as possible. In our dataset, among 19,435 patients 400 were non-stroke patients, and the rest had strokes of varying kinds. In order to balance the non-stroke and stroke data, we took 400 stroke patients and 400 non-stroke patients. As a result we got best accuracy from our dataset.

D. Feature Selection

Feature selection aims to improve predictor performance, resulting in faster and more cost-effective predictors as well as a better understanding of the data's underlying process. The University of Waikato in New Zealand created and maintained the Waikato Environment for Knowledge Analysis (WEKA), a machine-learning toolbox. The built-in algorithms of WEKA were employed to detect stroke disease. WEKA has proven to be a very reliable machine learning suite in previous tests. After collecting the appropriate dataset, the next step lies in cleaning the data and ensuring it is ready for model building. The dataset taken had 109 attributes. To sort out the key features exclusively, we used numerous built-in filter methods such as Information Gain Value, Correlation attribute evaluator, Numerical Cleaner, principal component analysis, StringtoNumeric, RemoveWithValues, and the stages are as follows:

Step 1: Information Gain, or IG for short, is a technique for calculating the reduction in entropy or surprise in a dataset by dividing it into segments based on the value of a random variable. It can figure out how much information each attribute in the output variable adds up to. The entry values range from 0 (no information) to 1 (a lot of information).Those traits that contribute more information have a higher information gain value and can be chosen, while those that do not give much information have a lower score and can be deleted. We used the Information Gain Value to delete a couple columns that didn't make much of a difference in model building.

Step 2: The value of an attribute is measured by the correlation between it and the class by the correlation attribute evaluator. Each value is treated as an indicator when considering nominal qualities on a value-by-value basis. We reduced the features to 72 using the Correlation attribute evaluator.

Step 3: Large datasets are becoming more frequent, and they might be challenging to decipher. PCA is a method for lowering the dimensionality of such datasets, boosting interpretability while minimizing information loss. It accomplishes this by generating new uncorrelated variables that optimize variance in a sequential manner. We found our ultimate key 15 attributes that provide the best overall accuracy after using the principle component analysis on our dataset.

Step 4: The dataset is then checked for null values, missing values, and filled using "NumericalCleaner," a filter that 'cleans' numeric data by replacing all missing values for nominal and numeric attributes in a dataset with the modes and means, and "RemoveWithValues [20]," a filter that replaces all missing values for nominal and numeric attributes in a dataset with the modes and means.

Step 5: The next duty is Label Encoding, which comes after removing the null values from the dataset. We transformed string values to numeric values using this filter method in WEKA StringtoNominal Converts a range of string attributes to nominal.

The correlation matrix is a crucial data analysis measure that is used to summarize data in order to better understand the link between various variables and make informed decisions. When dimensionality reduction on high-dimension data is needed, it is also a crucial pre-processing step in Machine Learning pipelines to calculate and evaluate the correlation matrix. The correlation matrix is generated only using the integer values. Fig.2 shows the correlation matrix for the attributes we have got.

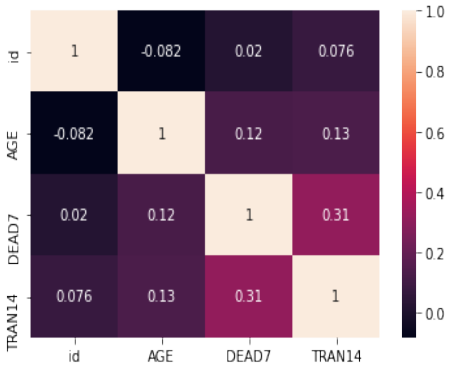

Fig. 2. Correlation Matrix of selected features

IV. Result & Analysis

The result is analyzed by applying machine learning models in WEKA. The following Table III shows the accuracy of machine learning algorithms. After applying some basic classifiers we got different accuracies. For Naive Byes we got 93.21%, J48 97.87%, Random Forest 97.69%, SGD 97.55%, Support Vector Machine 97.80% and Nearest Neighbour 95.56%. J48 has the best accuracy of all, with an accuracy of 97.87%.

TABLE III
Percentage of Accuracy

Classifier	Accuracy	Precision	Recall	F-measure
Naive Bayes	93.21	0.927	0.928	0.926
Random Forest	97.69	0.997	0.998	0.977
J48	97.87	1.000	1.000	0.979
SGD	97.55	0.995	0.978	0.978
Support Vector Machine	97.80	1.000	1.000	0.980
Nearest Neighbour	95.56	0.957	0.958	0.954

Initially, we used 800 samples (400 strokes and 400 non-stroke patients) and 108 features in our dataset. After processing the dataset we found 15 attributes. Then, we applied same classifiers on both dataset. Using 108 features, we got the accuracy of 88.63 from Naive Bayes, 84.87 from Random Forest and so on. After using 15 features, we got the accuracy of 93.21 from Naive Bayes, 97.69 from Random Forest and so on. We treated our dataset, comprises of 15 features, to improve the accuracy and get a lot more refined result.

Fig.3 shows a comparison of balanced and imbalanced data. We can't accurately determine stroke or non-stroke patients based on the imbalanced data. After balancing our dataset we got the best accuracy.

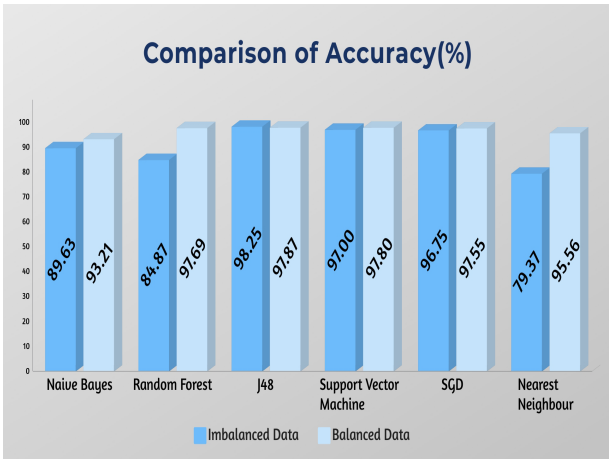

Fig. 3. Comparison of accuracy between imbalanced and balanced data.

V. Conclusion And Future Work

Stroke is a critical medical condition that needs to be treated on time to avoid complications. A machine learning model can help in the early detection and prediction of strokes, reducing the severity of their long-term consequences. In this paper, multiple variables are evaluated in order to find out how well different machine learning algorithms can detect stroke. The stroke detection method, which employs machine learning, can be used to determine whether or not a patient is suffering from a stroke. In J48, the detection provides an accurate result of 97.87%. We can detect the patient's stroke and provide the optimum treatment utilizing this technology.

Weka has some drawbacks, such as the inability to handle large datasets. Whenever a dataset is more significant than a few megabytes, an OutOfMemory error occurs. As a result, we intend to expand the research in the future by incorporating neural networking and various machine learning algorithms.

References

[1] "Stroke,"2021 (accessed September 14,2021).[Online]. Available: https://www.cdc.gov/stroke/types_of_stroke.htm
[2] "Hemorrhagic Stroke,"2021 (accessed September 14,2021).[Online]. Available:https://www.mayoclinic.org/diseasesconditions/stroke/symptoms-causes/syc-2
[3] M. Mahmud et al., "A brain-inspired trust management model to assure security in a cloud based iot framework for neuroscience applications,"Cognitive Computation, vol. 10, no. 5, pp. 864–873, 2018.
[4] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. Al Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia," Brain Informatics, vol. 7, no. 1, pp. 1–21, 2020
[5] M. Mahmud, M. S. Kaiser, and A. Hussain, "Deep learning in mining biological data," arXiv preprint arXiv:2003.00108, 2020.
[6] "Machine Learning Algorithms,"2021 (accessed September 12,2021).[Online]. Available:https://www.analyticsvidhya.com/blog/2017/09/commonmachine-learning-algorithms/
[7] "weka," 2021 (accessed September 12, 2021). [Online]. Available: https://en.wikipedia.org/wiki/Weka(machinelearning)
[8] P. Govindarajan, R. Soundarapandian, A. Gandomi, Rizwan Patan, Premaladha Jayaraman and R. Manikandan. "Classification of stroke disease using machine learning algorithms." Neural Computing and Applications 32 (2019): 817-828.

[9] L. Amini, R. Azarpazhouh, M. T. Farzadfar, 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.

[10] D.Shanthi, Dr.G.Sahoo, and Dr.N.Saravanan, “Designing an artificial neural network model for the prediction of thrombo-embolic stroke,”2008.

[11] Singh, M.S., Choudhary, P., Thongam, K.: A comparative analysis for various stroke prediction techniques. In: Springer, Singapore (2020).

[12] C. Chin, B. Lin, G. Wu, T. Weng, C. Yang, R. Su, and Y. Pan, “An automated early ischemic stroke detection system using CNN deep learning algorithm,” in 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), Nov. 2017, iSSN: 2325-5994.

[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, vol. 68, no. 11, pp. 1292–1300, Nov. 2015.

[14] Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach, A. K.: DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques. In Springer (2020).

[15] M. S. Singh and P. Choudhary, ”Stroke prediction using artificial intelligence,” 2017 8th Annual Industrial Automation and Electrome chanical Engineering Conference (IEMECON), 2017, pp. 158-161, doi: 10.1109/IEMECON.2017.8079581.

[16] Gangavarapu Sailasya and Gorli L Aruna Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms” International Journal of Advanced Computer Science and Applications(IJACSA), 12(6), 2021.

[17] Jeena, R. and Sukesh Kumar. “Stroke prediction using SVM.” 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (2016): 600-602.

[18] “IST,” 2021 (accessed September 14, 2021). [Online]. Available: https://datashare.ed.ac.uk/handle/10283/124(IST)

[19] P. A. Sandercock, M. Niewada, and A. Członkowska, “The international stroke trial database,” Trials, vol. 13, no. 1, pp. 1–1, 2012.

[20] ”Remove with values,”2021 (accessed September 14,2021).[Online]. Available: http://www.cs.tufts.edu/ ablumer/weka/doc/weka. fil.ters.ReplaceMissingValuesFilter.html

[21] ”Global Health Bangladesh,”2021 (accessed September 14,2021).[Online]. Available:https://www.cdc.gov/globalhealth/countries/bangladesh/default.htm