**BRAIN STROKE PREDICTION USING DEEP LEARNING**

**T. KRUPA NANDINI, T. SAI TEJA**

**ABSTRACT:**

A brain stroke is a disruption of blood circulation to the cerebrum. As per recent analysis, adult death and disability are primarily brought over by brain stroke. The World Health Organization (WHO), reports that the primary cause of death and property damage worldwide is brain stroke. Early detection of the signs and symptoms of a stroke can help to reduce risk factor of death by up to 50%. A stroke is more likely to occur in adults over the age of 55. An increasing number of people are experiencing this crippling and frequently fatal form of stroke, which results in cerebral hemorrhage.

Various machine learning (ML) models were developed to predict the possibility that a brain stroke would occur. To predict the brain stroke, the proposed system used the CNN algorithm. The existing approaches are k-NN, Support Vector Machine (SVM), Genetic Algorithm (GA), Naïve Bayes classifier, J48 algorithm, Logistic Regression (LR) and Random Forest (RF). These requires more time to train the model and it is difficult to debug. And these are not suitable for large datasets.

The proposed system makes predictions using CNN algorithm, a deep learning technique. It includes a multilayer perceptron for the prediction task and an autoencoder for eliminating and capturing non-linear correlations between parameters. The proposed system is contrasted with existing system and it shows an enhancement in the capability to anticipate the stroke. The proposed system achieved an accuracy of 89%.

**KEYWORDS: S**troke, Machine learning, Deep learning, Convolution Neural Network and Convolution layers.

**1. INTRODUCTION:**

Artificial intelligence (AI) incorporates the machine learning (ML) as a subset which enables a model to learn from experience task and enhance the model without being explicitly programmed. The machine learning models rely on time stamped data, quantitative data, qualitative data and textual data. Supervised learning and unsupervised learning are the two fundamental types of machine learning. Deep Learning is a technique of machine learning. The deep learning models learns by finding sophisticated in the data. It can be used to solve any recognition pattern problem without human intervention.

When the brain’s blood circulation is disrupted, a brain stroke occurs. Lower blood supply may cause the death of brain cells. And when your brain isn’t getting enough oxygen, your body will send you certain warning signs. Some stroke risk factors are unaffected by medications or changes in lifestyle. Age, gender, race and a family history of stroke are one of factors. Some brain stroke risk factors have potential therapies. Obesity, heart disease, smoking and hypertension are risk factors for brain stroke.

Once the blood supply to a portion of the brain is interrupted, a dangerous medical condition known as stroke can result, which is life threating. Strokes require immediate medical attention since they constitute a medical emergency. Less damage is likely to occur the faster a person receives care for a stroke. Blood and oxygen cannot reach the brain’s tissues because of the break or obstruction.

The brain is the organ that manages our body activities, retains memories and generates the ideas, feelings and speech. In addition, the brain regulates a variety of bodily processes, including respiration and digestion. The brain need oxygen to function correctly. All areas of the brain receive oxygen rich blood through arteries. Brain cells begin to die within minutes if something blocks the blood flow because it can’t obtain enough oxygen. This leads to the brain stroke.

Convolution neural networks (CNNs) are one kind of Deep Learning neural networks (DNNs) used for image processing, recognition, classification and a model prediction. The hidden layers of CNN are used to make predictions more accurate. We used the two-activation functions that is sigmoid function and ReLu function are used to aid the model to learn complex patterns in the data.

**2. LITERATURE SURVEY:**

In **[1],** Tasfia Ismail Shoily et al were utilized four machine learning (ML) algorithms to recognize the stroke. The data is gathered from hospitals and utilized them to figure out the problem. The WEKA toolkit’s Naïve Bayes classifier, J48 algorithm, k-NN and Random Forest classifier are the four machine learning algorithms. In the beginning, importing the necessary libraries and data from the stroke database. Data preprocessing, visualization of data along with attribute selection, then the data is split into test set, train set to build and evaluate the classifier model. In terms of accuracy, the naïve bayes classifier produced better classification results for the brain stroke. As a result, the evaluation of the classification result indicates that J48, k-NN and Random Forest successfully fulfilled the obligations in the identification of stroke disorders.

In **[2],** Hyuna Lee et al identifies the patients with severe ischemic stroke within 4.5 hours after the onset of symptoms using three machine learning (ML) models. On the basis of a reasonable request, the information was collected from the appropriate author and is used to support the authors of the study conclusions. The MRI Protocol, Image Processing, Infract Segmentation, and Co-registration are used to manipulate the collected data, which is then divided into test and train sets for feature extraction and feature selection to train the machine learning model. The best machine learning is then placed to the test, and later, it is assessed by every top machine learning algorithm to spot test subjects who are examined within 4.5 hours of the commencement of symptoms. Random Forest, Support Vector Machine, and Logistic Regression were the three machine learning techniques that were utilized. Each patient in the independent test set was identified using one of these three ideal machine learning models. Finally, ML approaches may be practical and helpful in finding medication prospects among patients with a stroke onset that is ambiguous. Korean Health Technology R&D Project, through the Korea Health Industry & Welfare, provided funding for the author's research.

In **[3],** Jeena R S et al were worked with Support Vector Machine (SVM) models, which are well-known for their capacity to simulate complicated systems, are frequently employed in classification tasks. The International Stroke Trail Database was used to get the data for this project. Information about the patient, the hospital, risk factors, and symptoms are all included in the database. The 300 data samples are under goes for training set and 50 data samples for testing set. The classification accuracy of different kernel functions had been compared. MATLAB was utilized to implement SVM. It is used to determine the parameters sensitivity, specificity, accuracy, precision, and F1 score to evaluate the effectiveness of various SVM classifier kernel functions. The outcome was assessed across a range of patients in various age groups.

In **[4],** Grant C et al applied the Genetic Algorithm (GA) and k-NN algorithm to find a pattern of gene expression in peripheral circulation that might be enhanced to recognize Androgen Insensitivity Syndrome (AIS) during the early phase of treatment. 24 neurologically asymptomatic controls and 39 AIS patients make up the data. The genetic algorithm, k-nearest neighbors (GA/k-NN) machine learning method was used to find a pattern of gene expression that can clarified the groups. It was evaluated for its ability to identify between 20 acute stroke models and an additional 39 AIS patients, 30 neurologically asymptomatic controls.

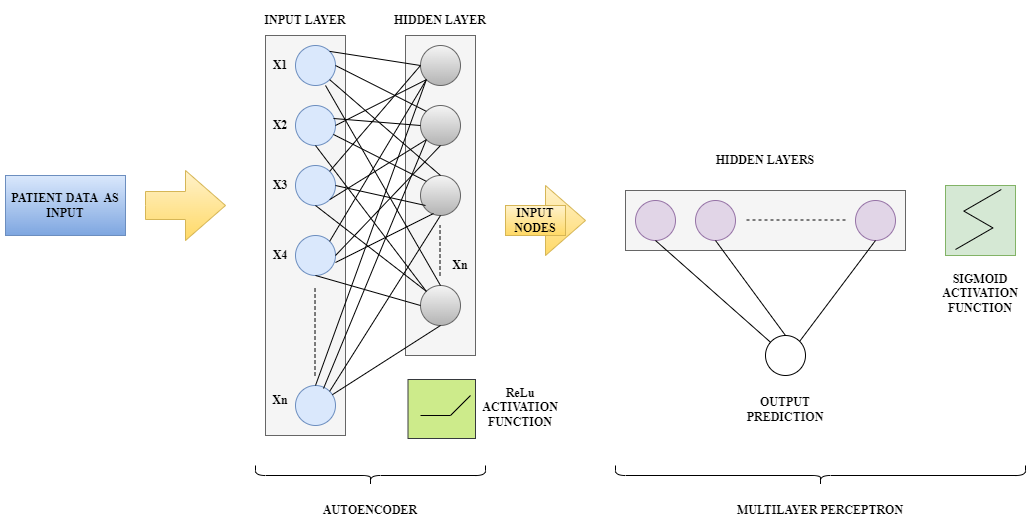
In **[5],** Minhaz Uddin Emon et al employed the Base Algorithm, which consists of the base 10 methods. LR, SGD, DTC, AdaBoost, Gaussian, QDA, MLP, k-NN, GBC, and XGB are the 10 base algorithms. The vulnerability predictors are built using these well-known classifiers, and these models are evaluated by using the confusion matrix. The performance of the stroke is predicted using these ten classifiers.

**Table 1: Analytical Existing approach methodologies**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.NO** | **AUTHOR** | **ALGORITHM** | **MERITS** | **DEMERITS** | **ACCURACY** |
| 1. | Tasfia Ismail Shoily | Naïve Bayes classifier,  J48 Algorithm,  k-NN,  Random Forest | It works quickly & save a lot of time.  It is used to make accurate predictions from the data.  Easy to implement.  RF can automatically handle missing values. | Evaluation can be wrong in some cases.  A small change in the data can cause large change.  Sensitive to noisy and missing data.  It requires more time to train a model. | 85%  79%  69%  80% |
| 2. | Hyuna Lee | Logistic Regression,  SVM  Random Forest | It is very fast at classifying unknown records.  It is more productive in high dimensional spaces.  RF algorithm is less prone to overfit  . | It constructs linear boundaries.  It is not suitable for large data sets.  RF computations may go far more complex compared to other algorithms. | 83%  82%  85% |
| 3. | Jeena R S | SVM | It has more stability. | Difficult to interrupt. | 75% |
| 4. | Grant C | Genetic Algorithm,  k-NN | It provides multiple optimal solutions.  It is simple to implement. | It is difficult to debug.  It does not work well with large datasets. | 85%  80% |
| 5. | Minhaz Uddin Emon | Base Algorithm | It does not require as much training data. | It is a zero-probability problem. | 87% |

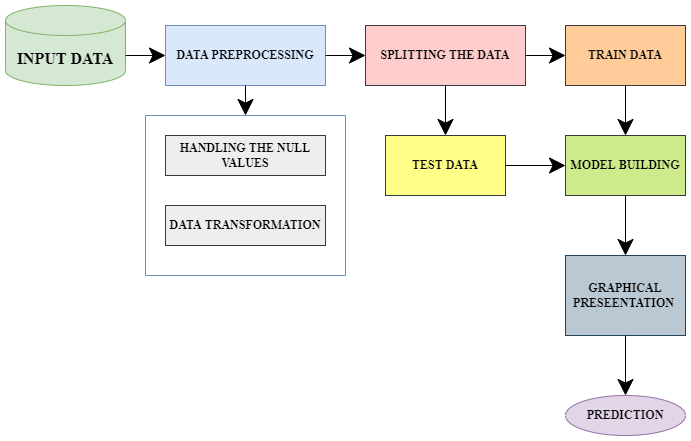
**3. PROPOSED METHODOLOGY:**

The Convolution Neural Network (CNN) of our proposed model. CNN is a type of Deep Learning (DL) algorithm to predict the brain stroke for the particular person according to accumulated data. An autoencoder network and a multilayer perceptron network are the two networks that constitute the architecture.



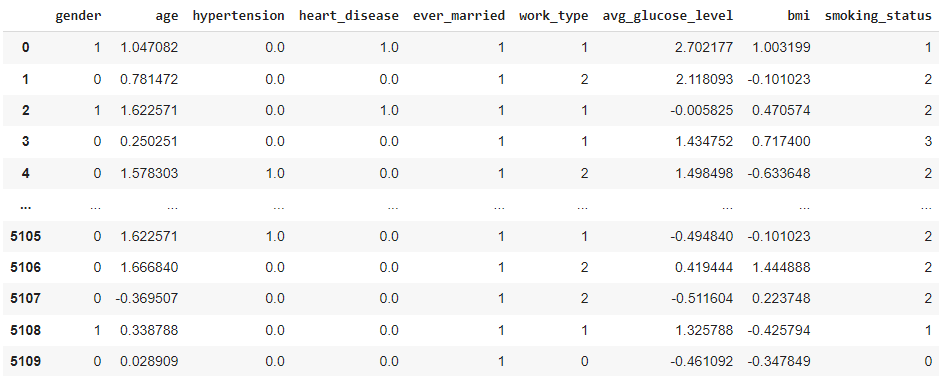
**Figure 1: Architecture of proposed CNN model**

Figure 1 presents the architecture of CNN using autoencoder and multilayer perceptron. It demonstrates how the autoencoder connects to the multi-layer perceptron (MLP) network and describes the input and output layers in detail.



**Figure 2: Flow chart of proposed methodology**

**3.1 INPUT DATA:** The information on the 5110 patients' various medical conditions, including the occurrence of stroke disease, is gathered. The information is gathered from the Kaggle website.The dataset is used to predict a patient’s likelihood of having a stroke based on input factors such as gender, age, numerous diseases and smoking status.

**Figure 3: INPUT DATA**

**3.2 DATA PREPROCESSING:** Data preprocessing is the process to improve the quality of data. Preprocessing helps to make the data consistent by removing duplicates and anomalies, increasing accuracy and simplifying the data for comparison.

**3.2.1 HANDLING THE MISSING VALUES:** Based on the kind of missing data, the mean/median or mode are used to replace the null values in the dataset.

**NUMERICAL DATA:**  It is a type of data that can be expressed as a numerical value. When a numeric value is missing, it is replaced with a NAN value and the mean or median.

**CATEGORICAL DATA:** Categorical is a type of data which is used to gather information with similar characteristics. When the categorical data is missing, then it was replaced with a value which was most occurring.

**3.2.2 DATA TRANSFORMATION:** The process of transforming raw data into structured data that may be used for building the model. Data transformation is used standard-scaler function to transform into the standard form of mean.

**3.3 SPLITTING THE DATA:** It was applied to prevent data overfitting. The majority of the data is divided into train data and the leftover into test data.

**3.4 MODEL BUILDING:** The model is developed using a convolution neural network (CNN). The simplest and easiest method to construct a model is sequential modelling. A model is constructed using it layer by layer. Our model layers are added using the method add (). Our model uses the 7 dense layers and 10 dropout layers. The dense layers are used to connect each layer in the neural network and the dropout layers are used to ignore a set of neurons and it was preventing the overfitting in the neural network randomly. The rectified linear activation (ReLu) function is used for hidden layers and the sigmoid is a type of activation function which is used to take any values between 0 and 1.

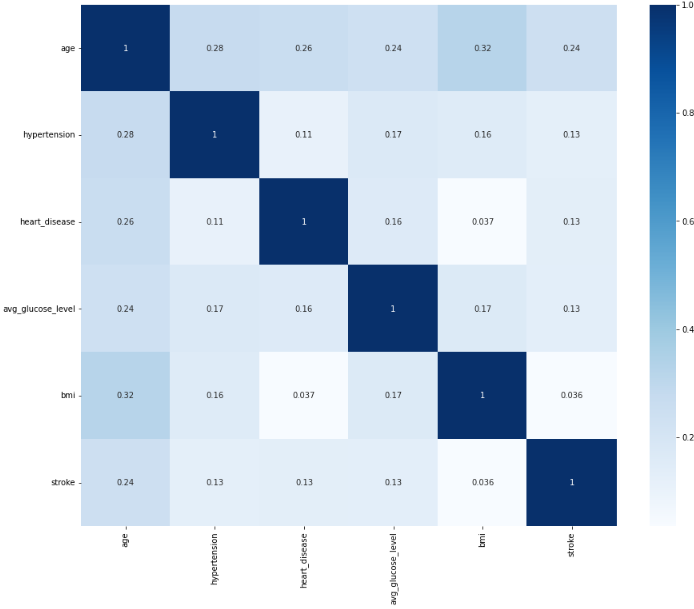
**3.5 TRAIN DATA:** It is used to train the model. 70% of the dataset is used for training the model.

**3.6 TEST DATA:** It is used to test the trained model. 30% of the dataset is used for testing the model.

**3.7 GRAPHICAL REPRESENTATION:** It represents the values of training and validation loss, accuracy graphs on epochs.

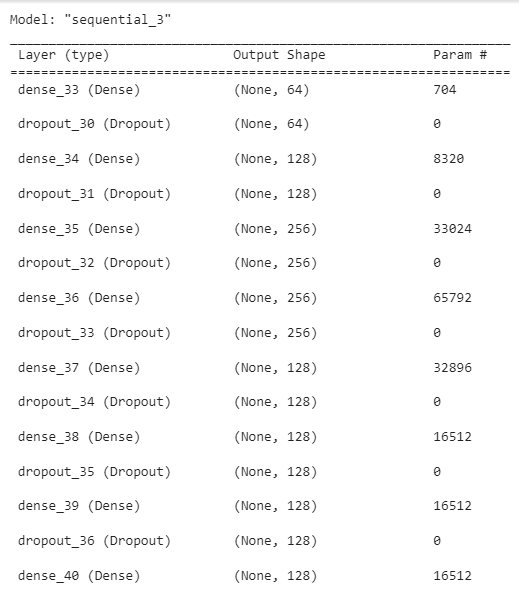
**3.8 PREDICTION:** Optimizer (Adam), loss and metrics are the three parameters to compile our model and accuracy is a metric, utilized to evaluate the correct output from the input.

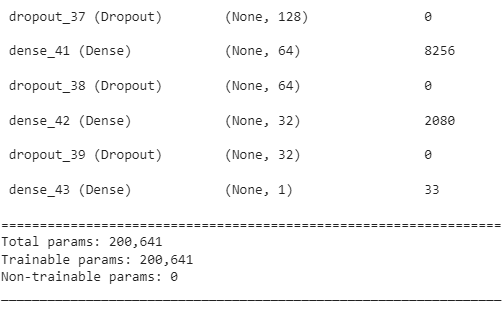
**4. RESULTS AND DISCUSSIONS:**

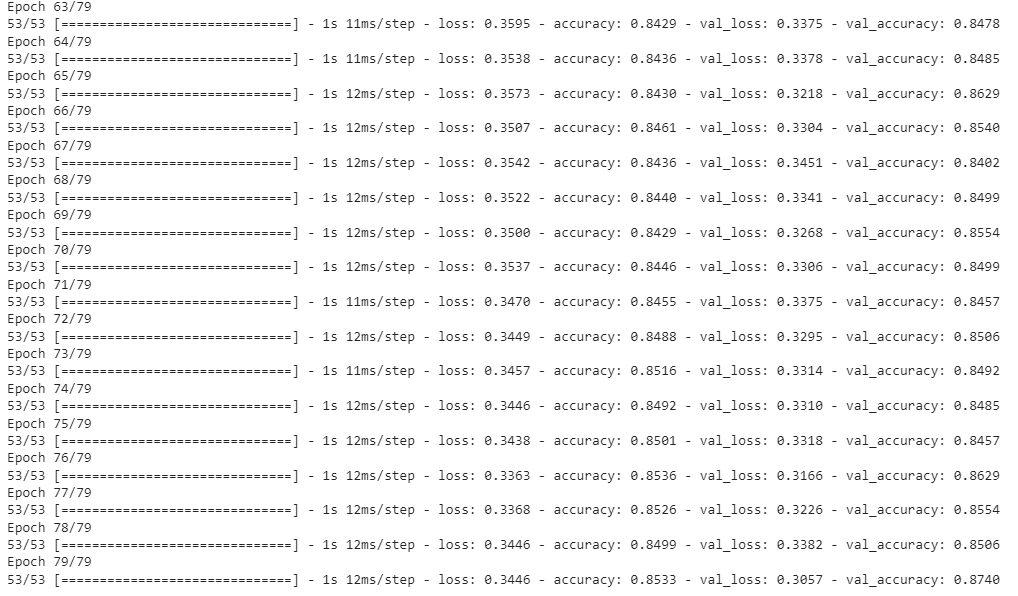


**Figure 4: Heat map**

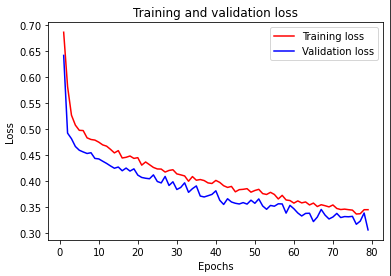
In contrast to the heat map, it visualize the amount of correlation among variables. It assisted in the discovery of attributes that are ideal for construction of machine learning models. The correlation matrix is color coded with cmap (colormap) using “Blues”.



**Figure :5 Sequential model using CNN**

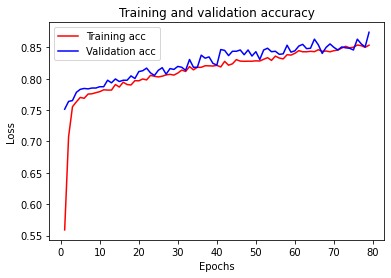
Figures 5 represent a description of the model, in which each neuron in the dense layer sends one output to the next layer while also receiving all of the outputs from the previous layer. A dropout layer is a technique to remove the unwanted data or noise to improve the processing and results in time.**Figure 6: Epochs in CNN**

In contrast to the epochs, a neural network is built for one cycle or an epoch which was used all the training data. Every piece of information is used exactly once during an epoch.



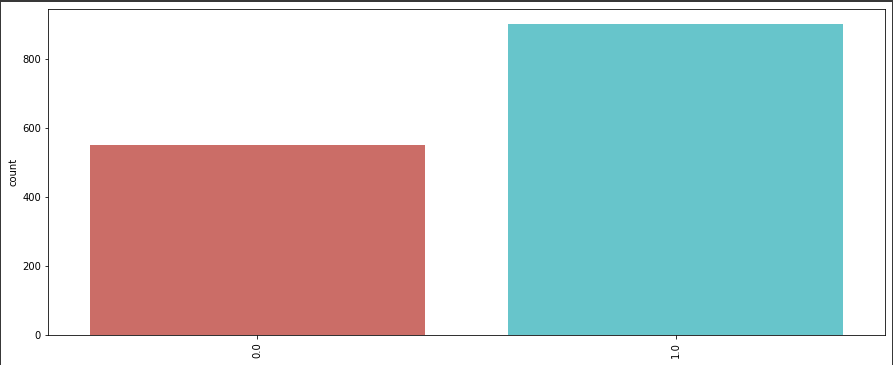
**Figure 7: Graphical representation of Training Loss and Validation Loss**

In contrast to the training loss, which shows how well the model fits with training data, the validation loss illustrates how well the model fits with new data.



**Figure 8: Graphical representation of Training Accuracy and Validation Accuracy**

In contrast to the training accuracy, which uses the validation set to train the model, the validation accuracy uses validation set to evaluate the performance of the model.



**Figure 9: Output of Brain Stroke Prediction**

In contrast to the brain stroke prediction model’s binary output, which ranges from 0 to 1, indicates that the sigmoid function changes the output between 0 and 1. A brain stroke is not present when the value is 0, but it is predicted when the value is 1.

**7. CONCLUSION:**

The proposed approach proposed a deep learning strategy based on neural networks and encoders to predict the stroke. This approach identifies status of patient through input parameters (dataset). The proposed system has been compared with another systems. The outcome of proposed system outperforms with another approaches. Our goal is to enhance the proposed methodology for light strokes in further work and plan to test additional deep learning architectures as well. Furthermore, also consider adding additional input features.

**8. REFERENCES:**

[1] T. I. Shoaly, T. Islam, S. Jannat, S. A. Tana, T. M. Alif and R. R. Emma, “Detection of Stroke Disease using Machine Learning Algorithms,” 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-6, doi:10.1109/ICCCNT45670.2019.8944689.

[2] Lee, Hyuna; Lee, Eun-Jae; Ham, Sungwon; Lee, Han-Bin; Lee, Ji Sung Kwon, Sun U.; Kim, Jong S.; Kim Namkung; Kang, Dong-Wha (2020). Machine Learning Approach to Identify Stroke Within 4.5 Hours, Stroke, (), -, doi:10.1161/STROKEAHA.119.027611.

[3] Jeena, R S; Kumar, Sukesh (2016). [IEEE 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) - Kumara coil, India (2016.12.16-2016.12.17)] 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) - Stroke prediction using SVM, (), 600-602, doi:10.1109/ICCICCT.2016.7988020.

[4] O’Connell, Grant C; Petrone, Ashley B; Treadway, Madison B; Tennant, Connie S; Lucke-Wold, Noelle; Chantler, Paul D; Barr, Taura L (2016). Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischemic stroke, npj Genomic Medicine, 1(1), 16038-. Doi:10.1038/npjgenmed.2016.38.

[5] Emon, M, U; Keya, M. S; Meghla, T. I; Rahman, M. M; Mamun, M. S. A; & Kaiser, M.S. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). Doi:10.1109/iceca49313.2020.9297525.