2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579763

STROKE PREDICTION USING SMOTE-TOMEK AND NEURAL NETWORK

Chirag Rana*, Nikita Chitre†, Bhargavi Poyekar‡, and Pramod Bide¶
Department of Computer Engineering, Sardar Patel Institute Of Technology
Mumbai, India
Email: *chirag.rana@spit.ac.in, †nikita.chitre@spit.ac.in, ‡bhargavi.poyekar@spit.ac.in, §pramod_bide@spit.ac.in

***Abstract*—Stroke is a medical condition that occurs due to inadequate blood supply to the brain causing the death of brain cells. If a stroke is not diagnosed correctly can lead to brain injury, paralysis or even death. As the symptoms of Stroke are very instantaneous and can be triggered by unforeseen conditions also, it makes prevention of this situation very difficult to predict. The pandemic has made this worse as depression is the next stage of it which can be seen by the suffering of the people all across the globe. In our paper, we have taken a step to highlight this important topic of discussion by using advanced Machine learning and Deep Learning techniques to predict any possibility of stroke. We have proposed our final model using Artificial Neural Network which gives the best roc score of 0.84 and given a comparative analysis of how well do other Machine Learning Algorithms like ensemble-based, tree-based and Naive Bayes-based Algorithms perform in predicting Stroke.**

***Index Terms*—Stroke, Deep Neural Network, Machine Learning, SMOTE-TOMEK Sampling method, Class Imbalance.**

I. Introduction

A stroke occurs due to an insufficient supply of blood to the brain causing a lack of oxygen and nutrients to the brain tissue. Blockage of arteries and bursting of vessels are the two main causes of stroke. Stroke requires emergent treatment [1]. Being obese, a smoker, diabetic patient, having heart-related disease are some of the potential risk factors of stroke [2].

In 2015, causing 6.5 million deaths, stroke became the second most frequent cause of death [3]. Around 7,95,000 people experience stroke every year [4]. In India, stroke is one of the leading causes of death and the rate of stroke ranges approximately from 84-262/100,000 in rural areas and 334-424/ 100,000 in urban areas [5].

Early recognition of stroke is very important as the early treatment reduces the risk of brain damage. Diagnosis of stroke is done by several techniques including MRI and CT scans. CT scans have a sensitivity of 16% and specificity of 96% for diagnosing ischemic stroke. MRI scans have 83% sensitivity and 98% specificity [3]. Misdiagnosed stroke-causing delayed or no treatment may cause brain injury, paralysis, seizures, memory problems, and the worst of it all, death [6]. Hence, early and correct diagnosis of stroke is really important.

In this paper, we have proposed a new methodology for the prediction of stroke. We first performed data cleaning by handling missing values and then pre-processed the data by label encoding and one-hot encoding. As the data was imbalanced, we performed SMOTE-Tomek to overcome the problem of imbalance.

SMOTE-Tomek performs both oversampling and undersampling. First SMOTE oversamples the minority class and Tomek Links removes samples from the majority class with overlapping values. So the ratio of samples becomes 1:1 [7]. Then we classified the data using models like Logistic Regression, SVM, KNN, GaussianNB, BernoulliNB, Random Forest, LightGBM, XGBoost, Bagging, AdaBoost, Decision Tree and Neural Network. For better results, we did parameter tuning and later evaluated the results using accuracy, ROC-AUC scores precision, recall and f1-score. We found that Neural Network gave us the best roc score which fulfills the requirement. The results we achieved outperformed the results as compared to the existing work.

When analyzing the related work, we realized that most of them misunderstood the SMOTE technique and have used it on the entire Dataset. Whereas the test set is supposed to be untouched. While conducting this research we could understand the problem statement in a much better way, which led us to achieve better results. As the data is highly imbalanced, the models result in True Positive values being much greater than True Negative, False Positive, False Negative values, which gives high accuracy value even if only the non-stroke samples are predicted correctly, and the stroke samples have a low precision value which led to accuracy Paradox as shown in [8]. Hence roc score should be the metric of comparison as it justifies the result by considering the specificity and sensitivity. But most of the existing work has deemed accuracy as the perfect fit.

In the following paper, Section II consists of the literature survey we carried to get a better idea about this topic. Section III consists of a Dataset Description of the data we used for building our model, Section IV presents our proposed methodology, Section V shows the experimental results and Section VI concludes the paper along with the Future Scope.

II. Related Work

In [9], they collected demographic data of patients like gender, age, education from the Faculty of Physical Therapy, Mahidol University, Thailand from 2012 to 2015. Because of very large data, consulting specialists performed attribute and data selection. Later they performed resampling because of an imbalance in data to get 250 strokes and 500 non-stroke patients. Once the data was ready, models like Decision Tree, Naive Bayes, and Artificial Neural Networks were used to

get the best accuracy by Decision Tree of 75%. They did not consider the symptoms of the patient, only demographic data was used.

M. S. Singh and P. Choudhary have used the Cardiovascular Health Study dataset with 5,888 samples (3,228 males, 2,660 females) and more than 600 attributes [10]. They performed data preprocessing to get a dataset with 1824 samples and 357 features. Later Feature Selection was done using a Decision Tree with C4.5 and dimension reduction using PCA. Back-propagation neural network classification algorithm was used for classification to get an accuracy of 97.7%.

UCI's Heart Disease Dataset with 899 rows and 76 attributes was used in [11] by omitting the missing values and predicting the stroke by Deep Neural Network with a feedforward multilayer artificial neural network. The results showed a Mean Square Error of 0.2596. They could have used more stroke-related risk factors.

In [12], Kaggle Healthcare Problem Dataset with information of about 29072 patients with 12 attributes. Variance in the dataset was analyzed by PCA using 10 variable attributes. Random Down-sampling technique was used because of imbalance in the dataset and a balanced dataset with 548 strokes and 548 non-stroke patients were created. Decision Tree, Random Forest and Neural Network were applied for classification to achieve the best accuracy of 75.02% by Neural Network. The effect of the use of a subset of features on the accuracy of the model is not studied.

Chun-Cheng Peng [13] used Kaggle's Healthcare Problem: Prediction Stroke Patients dataset with 43400 records further divided into 70% training, 15% test and 15% validation data. They trained the ANN by 2 methods, i.e., the Levenberg Marquardt (LM) Algorithm and Scaled Conjugate Gradient (SCG) Algorithm. The accuracy rate of 98% was achieved by both methods using 1000-fold cross-validation. This accuracy is achieved on the entire data and not the testing data. They could have pre-processed the data more efficiently to handle the class imbalance problem.

Exploratory data analysis by univariate and bivariate plots was performed in [14] for analyzing the correlation between attributes. Data cleaning was done by removing the duplicate and filling the missing values with the mean. To handle the imbalanced data, they used the SMOTE technique and then the data were classified using 5 models namely, Naive Bayes, Logistic Regression, Random Forest, Decision Tree and Gradient Boosting Algorithms. After tuning, gradient boosting outperformed the others with a ROC score of 0.95. But the SMOTE technique was used inaccurately.

Collection of 100 standard 12-leads ECG traces sampled at 360 Hz was done in [15]. In total 98 images were collected. A Convolution Neural Network with 12 layers Dense Net architecture was used to predict stroke. They used Batch Normalization with ReLU and Convolution. They achieved an accuracy of 85.82% with a learning rate of 0.001.

[16] also used Kaggle's Healthcare Problem Dataset and created a balanced dataset with 548 stroke and non-stroke patients. The data was split into 70-30% train and test. They used a rough set theory to find favorable features for stroke prediction. R scores (found by rough set theory) for age, hypertension, average glucose level, and heart disease made these attributes favorable candidates for stroke prediction in comparison with other attributes. The highest correlation coefficient (r=0.675) was achieved by their proposed methodology.

Most of the papers have misused the SMOTE technique for over-sampling by applying it on the entire dataset rather than the training dataset. Furthermore, the majority of them have employed accuracy as an evaluation metric, which is incorrect in the case of an imbalanced dataset. In our paper, we have used SMOTE-TOMEK for overcoming the class imbalance problem. In addition, ROC curve is used as an evaluation metric in place of accuracy.

III. Dataset Description

The Dataset used in this study is provided by the author fedesoriano which is available on Kaggle [17]. The dataset consists of 5110 entries of patients with 12 attributes which is shown with description in Table - I. The Dataset is highly imbalanced with a total of 249 out of 5110 having a stroke as shown in Fig. 1. The Dataset even has 201 missing values of bmi and in almost 30% of the Dataset, the person's smoking status is unknown, which further makes the model difficult to fit.

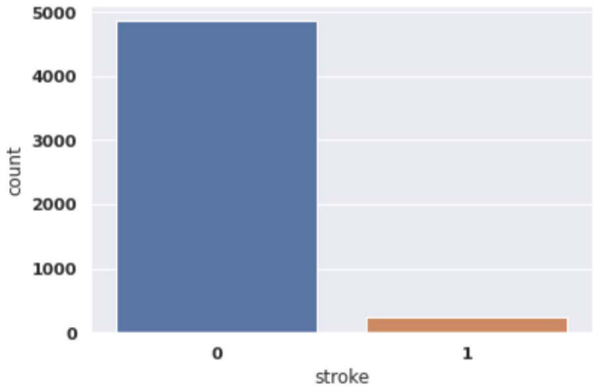

Fig. 1. Class Imbalance

IV. Methodology

For the proposed methodology, the selection of our final model is done on the comparative analysis of all types of models based on Tree, Naive Bayes, and ensemble-based Learning. A Deep neural network gave the best results. The System Diagram of the working of our model is shown in Fig. 2 and the results of all the models are provided in Table - II.

a) Data Preprocessing: is the most important part of the entire Methodology as the dataset suffers a lot from Missing values and Class Imbalance problems. Throughout the process Data Exploration and Data wrangling was done simultaneously.

The analysis of Data and steps taken for making the Dataset ready for modeling are given below:

TABLE I
DESCRIPTION OF ATTRIBUTES IN DATASET

Attribute	Description
Id	unique identifier for indexing
gender	includes ”Male”, ”Female” or ”Other”.
age	age of the patient (Integer)
hypertension	Boolean value (0,1) representing suffering from Hypertension.
heart_disease	Boolean value (0,1) representing suffering from Heart related disorder.
ever_married	Marital Status of the Patient as ”Yes” or “No”
work_type	Status Job Type as ”children”, ”Govt_job”, ”Never_worked”, ”Private” or ”Selfemployed”
Residence_type	”Rural” or ”Urban”
avg_glucose_level	average glucose level in blood
bmi	Body Mass Index (Decimal Value)
smoking_status	”formerly smoked”, ”never smoked”, ”smokes” or ”Unknown”
stroke	“1” if the patient had a stroke or “0” if not

1) Dropping the single data point with ’Other’ in gender.
2) Filling the 201 missing values of ’bmi’ using KNNImputer.
3) Label Encoded ’gender’, ’ever_married’,’Residence_type’.
4) One Hot Encoded ’work_type’ and ’smoking_status’.
5) Scaling the Dataset using StandardScaler.

As the size of the dataset is small a split of the Dataset to training and testing further decreased the size. We did a 70% split of the dataset. To Balance the values of the Dataset we first applied SMOTE for oversampling and Tomek Links to undersample the dataset. Here SMOTE could have over-generalized the dataset as the size of the entire dataset is small and to avoid the overlapping values generated Tomek-Links is one of the best undersampling algorithms.

b) Modelling: is the process of fitting the data to the algorithm, to identify the hidden patterns and use them to further apply in real-world problems.

c) Machine Learning and Deep Learning Algorithms: A list of machine learning algorithms and Deep Neural Network has been used to find out which one is the best-suited algorithm for the detection of Stroke. The list with their results are shown in Table II.

The Neural Network model used consists of two hidden layers having 50 and 20 neurons respectively as shown in Fig 3. Relu activation function is used in the hidden layers as it adds non-linearity to the model and, hence can easily back-propagate the errors. Also, Relu is computationally less expensive than Sigmoid and Tan-h activation functions due to simpler mathematical operations. The Sigmoid function is used in the output layer of the neural network as its value lies between 0 and 1 and hence is suitable for binary classification. The model was compiled with Stochastic Gradient Descent optimizer and Binary Cross-Entropy loss function as parameters and was trained in a batch size of 32 and for 100 epochs.

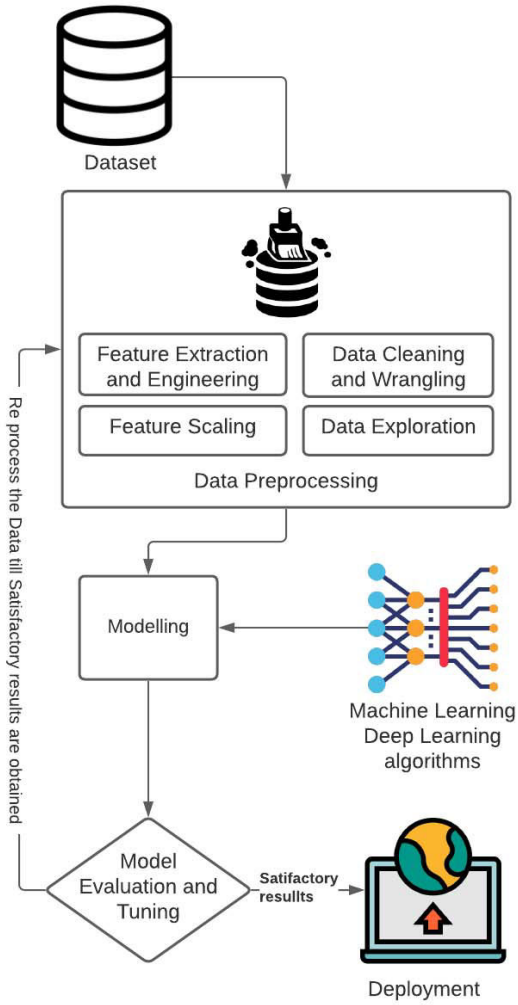

Fig. 2. Proposed Methodology System Diagram

d) Model Evaluation and Tuning: is another important phase of the process to check whether the model is ready for deployment. We have used GridSearch to find the best parameters for all the models we have used for comparative analysis. In our problem, the most important evaluation metric is ROC-AUC Score and f1-score as they avoid the accuracy paradox and measured on the testing dataset.

e) Model Deployment: is the step where the model is to be deployed on a cloud platform in the form of an API which could be further used by other authorized applications and websites.

The tools used for the methodology are:
1. numpy, pandas, matplotlib, seaborn for Data Handling and preprocessing.
2. imblearn for Oversampling and Undersampling.
3. sklearn, tensorflow, xgboost and lgbm for classification

TABLE II
PERFORMANCE COMPARISION USING SMOTE AND TOMEK-LINKS SAMPLING METHOD

Model	Accuracy (percentage)	ROC-score	Precision (0,1)	Recall (0,1)	F1 score (0,1)	Training Time (seconds)	Testing Time (seconds)
Neural Network	86	0.84	(0.99, 18)	(0.87, 0.55)	(0.92, 0.27)	43.91	0.07
Logistic Regression	91.68	0.63	(0.96,0.23)	(0.95,0.30)	(0.96,0.24)	0.24	0.00
SVM	94.22	0.61	(0.96,0.30)	(0.98,0.14)	(0.97,0.19)	11.88	0.04
KNeighbors	80.43	0.67	(0.97,0.12)	(0.82,0.50)	(0.89,0.20)	0.00	0.25
GaussianNB	55.87	0.63	(0.97,0.07)	(0.56,0.62)	(0.71,0.12)	0.0	0.00
BernoulliNB	79.25	0.51	(0.95,0.06)	(0.82,0.22)	(0.88,0.09)	0.01	0.01
Decision Tree	90.99	0.54	(0.96,0.12)	(0.95,0.14)	(0.95,0.13)	0.03	0.00
Random Forest	93.64	0.50	(0.95, 0.06)	(0.98, 0.02)	(0.97, 0.03)	1.35	0.05
XGBoost	93.05	0.55	(0.95,0.11)	(0.98,0.06)	(0.96,0.08)	0.64	0.00
AdaBoost	88.06	0.66	(0.97,0.18)	(0.91,0.40)	(0.94,0.25)	1.35	0.65
Bagging	92.66	0.54	(0.95, 0.07)	(0.97, 0.04)	(0.96,0.05)	3.45	0.18
LGBM	93.44	0.53	(0.95,0.16)	(0.98,0.08)	(0.97,0.11)	1.11	0.01

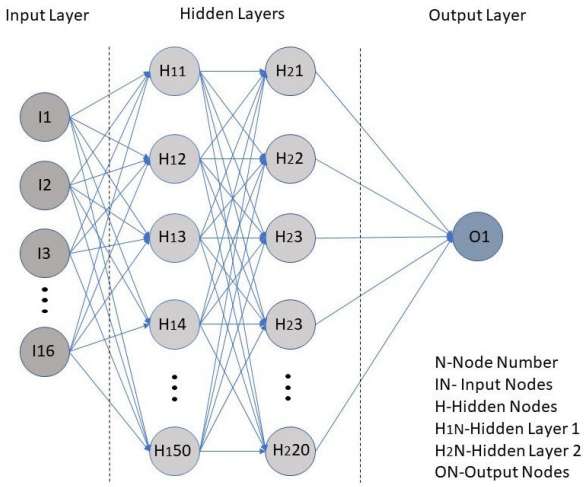

Fig. 3. Deep Neural Network

models and model evaluation.
4. Use flask and Heroku cloud app for model deployment.

V. RESULTS AND ANALYSIS

The results and model comparison is given in Table II. Accuracy is given in terms of percentage and (0,1) mentioned in Precision, Recall and f1-score corresponds to (No-Stroke, Stroke) and the training and testing time is given in seconds.

From the table, it can be inferred that Support Vector Machines have done the best job in terms of accuracy but fails in the ROC-AUC score and f1-score which is the perfect example of accuracy Paradox. This issue has occurred in several other algorithms in which the accuracy is high but recall, f1-score, and ROC-AUC are very low. This is because both the classes have a very high amount of overlapping values which becomes difficult to separate and even if the False Positive values over exceed the True Negative, it won't make a difference due to the large size of True Positives. A model with such in-efficiency cannot be used as it does not fit the practical applications of the Stroke Prediction.

The best results in terms of ROC-AUC score of 0.84 were obtained by Neural Network as shown in Fig. 4 and maintaining recall and f1-score relatively high. It did struggle in the accuracy due to an increase in the number of False Negatives but when we consider the size of the dataset, Missing values, imbalance nature, and Practical application, the main aim of alerting a person from a stroke can be achieved more as compared to the other Models.

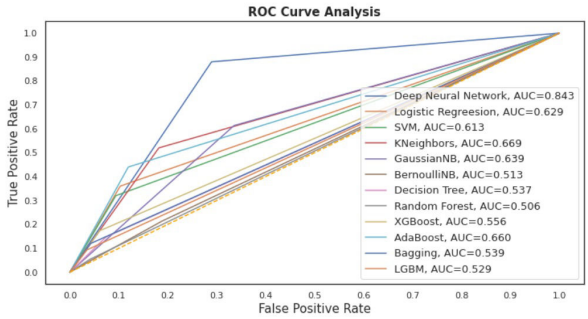

Fig. 4. ROC-AUC curve of all models

VI. CONCLUSION AND FUTURE SCOPE

The aim was to build a system that is cost-efficient and could be deployed on a large scale. Though the best treatment can always be provided by doctors and experts in this field, our system could predict the possibility of stroke through very general information and thus reduces the cost. Our final Model is Neural Network which gave the best ROC-score of 0.843 as the aim is to prevent the possibility of stroke.

Limitation: The dataset used in this research is imbalanced and has a lot of missing values which makes it uncertain.

The Proposed system has a very high Future Scope owing to the following reasons:
Fitness tracking applications used by several users not only can provide more data for more analysis and generalization of

causes of Stroke but also can suggest preventive measures to their users accordingly. Due to the Pandemic, many people are suffering from anxiety and a huge working crowd has moved online which might lead to obesity and several other reasons leading to stroke. As stroke is an emergent dysfunction of the brain which is fatal it is important to have the preventive measures ready.

REFERENCES

[1] ”Stroke-symptoms and causes”, on mayo clinic[online document], Available: https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113[Accessed: May 16, 2021]

[2] ”Stroke risk factors and prevention”, on better-health [online document], Available: https://www.betterhealth.vic.gov.au/health/ConditionsAndTreatments/str-oke-risk-factors-and-prevention[Accessed: May 16, 2021]

[3] ”Stroke”, on Wikipedia[online document], Available: https://en.wikipedia.org/wiki/Stroke#cite_note-GBD2015De-11[Accessed: May 1, 2021]

[4] Salim Virani, Alvaro Alonso, Emelia Benjamin, Márcio Bittencourt, Clifton Callaway, April Carson, Alanna Chamberlain, Alex Chang, Susan Cheng, Francesca Delling, Luc Djoussé, Mitchell Elkind, Jane Ferguson, Myriam Fornage, Sadiya Khan, Brett Kissela, Kristen Knutson, Tak Kwan , Daniel Lackland, Connie Tsao, ”Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association”, 2020 Circulation. 141. 10.1161/CIR.0000000000000757.

[5] Pandian, Jeyaraj Durai, and Paulin Sudhan, “Stroke epidemiology and stroke care services in India.” Journal of stroke vol. 15,3 (2013): 128-34. doi:10.5853/jos.2013.15.3.128.

[6] ”Stroke Misdiagnosis’, on Weiss and Paarz[online document], Available: https://www.weisspaarz.com/emergency-room-malpractice/stroke-misdiagnosis/[Accessed: May 1, 2021]

[7] Jason Brownlee, ”How to Combine Oversampling and Undersampling for Imbalanced Classification”, on Machine Learning Mastery[online document], Available: https://machinelearningmastery.com/combine-oversampling-and-undersampling-for-imbalanced-classification/[Accessed: May 1, 2021]

[8] Tejumade Afonja, ”Accuracy Paradox”, on toward data science[online document], Available: https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b[Accessed: May 1, 2021]

[9] T. Kansadub, S. Thammaboosadee, S. Kiattisin and C. Jalayondeja, ”Stroke risk prediction model based on demographic data,” 2015 8th Biomedical Engineering International Conference (BMEiCON), 2015, pp. 1-3, doi: 10.1109/BMEiCON.2015.7399556.

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

[11] Chantamit-o-pas P., Goyal M. ”Prediction of Stroke Using Deep Learning Model” In: Liu D., Xie S., Li Y., Zhao D., El-Alfy ES. (eds) Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science, vol 10638. Springer, Cham. 2017, doi: 10.1007/978-3-319-70139-4_78

[12] C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli and D. John, ”Predicting Stroke from Electronic Health Records,” 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 5704-5707, doi: 10.1109/EMBC.2019.8857234.

[13] C. -C. Peng, S. -H. Wang, S. -J. Liu, Y. -K. Yang and B. -H. Liao, ”Artificial Neural Network Application to the Stroke Prediction,” 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2020, pp. 130-133, doi: 10.1109/ECBIOS50299.2020.9203638.

[14] Rajora M., Rathod M., Naik N.S.”Stroke Prediction Using Machine Learning in a Distributed Environment.” In: Goswami D., Hoang T.A. (eds) Distributed Computing and Internet Technology. ICDCIT 2021. Lecture Notes in Computer Science, vol 12582. Springer, Cham., doi: 10.1007/978-3-030-65621-8_15.

[15] Xie, Y., Yang, H., Yuan, X. et al. ”Stroke prediction from electrocardiograms by deep neural network.” Multimed Tools Appl 80, 17291–17297 (2021), doi: 10.1007/s11042-020-10043-z

[16] M. S. Pathan, Z. Jianbiao, D. John, A. Nag and S. Dev, ”Identifying Stroke Indicators Using Rough Sets,” in IEEE Access, vol. 8, pp. 210318-210327, 2020, doi: 10.1109/ACCESS.2020.3039439.

[17] Fedesoriano, ”Stroke Prediction Dataset”, Kaggle[online dataset], Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset[Accessed: April 20, 2021]