2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386474

Brain Stroke Prediction Using Machine Learning Techniques

Ibrahim Almubark

Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia

imbark@qu.edu.sa

***Abstract*—Machine learning (ML) techniques have gained prominence in recent years for their potential to improve healthcare outcomes, including the prediction and prevention of stroke. The primary objective of this study is to develop and validate a robust ML model for the prediction and early detection of stroke in the brain. This study aimed to address some of the limitations of previous studies by utilizing a representative dataset and applying proper feature engineering techniques. The results showed that the Artificial Neural Network (ANN) model with synthetic minority oversampling technique (SMOTE) (ratio = 0.3) and hyperparameter turning performs the best in terms of average precision, which is crucial for making accurate predictions in clinical settings. By adjusting the threshold for classification, the model can be tailored to achieve the desired balance between precision and recall, depending on the specific requirements of a clinical workflow.**

***Keywords*—brain stroke, machine learning, artificial neural network**

I. Introduction

Globally, brain stroke is a significant public health concern. The disease incurs a high mortality rate and lifelong disability is high among those who experience it [1]. Stroke is reported to be the 2nd leading cause of death, with approximately 5.5 million deaths attributed to it each year [2]. A brain stroke is a medical condition that occurs due to a restriction of the flow of blood to the brain leading to the death of brain cells [2]. Frequent indicators of a stroke include a loss of ability to move or feel sensations on one side of the body, confusion, difficulty in speaking, dizziness, or loss of vision on one side [3]. These symptoms usually appear with the onset of the stroke. A brain stroke usually happens suddenly, so it is important to investigate and monitor certain risk factors. Several risk factors have been found to be associated with an increase in the risk of an individual experiencing a stroke. These include living with high blood pressure, high blood cholesterol, smoking, obesity, and diabetes [4]. Of these factors, high blood pressure is the leading cause of stroke. It is also important to note that many of these risk factors share underlying risk factors themselves [5], with the result that individuals are often living with more than one. Typically, a physical exam, medical imaging (such as a CT or MRI scan), alongside blood tests are used to identify risk factors and eliminate the condition.

Studies have shown that variables such as age, gender, family history, and lifestyle choices can affect the risk of a brain stroke [6]. Machine learning (ML) techniques can be applied to analyze and classify the risk factors associated with brain stroke, it can also be used for the early detection/prevention of stroke [7] and predict mortality from the disease [8]. A great advantage of such techniques is that the developed models become more accurate with the generation of high-quality data over time, thus lowering computational costs combined with increasingly accurate diagnoses.

There is a multitude of studies that investigate the performance of different prediction models aiming to detect brain strokes from different types of data including demographic, lifestyle, and clinical data expressed in numeric values or presented as images [1, 8, 9]. One of the biggest limitations in previous brain stroke studies, that frequently occurs and impacts the reliability of the results, is biased and limited data due to the small sample size or imbalanced classes. Some datasets have low qualities, including missing labels/data and imbalanced data without sufficient discussion/actions taken to prevent potential model bias. In addition, some studies lack careful discussion about feature selection and hyperparameter tuning. Moreover, many studies focus on developing ML models and their performance and ignore how to convert this to meaningful clinical workflow. Some of them lack good model interpretability, which makes it even harder to integrate ML into the clinical setting [10-14].

The aim of our current study is to improve brain stroke prediction and early detection through the use of advanced ML techniques. The dataset will be carefully examined through exploratory data analysis, and then ML models will be employed to test its accuracy.

The paper is organized as follows. Section II presents the dataset and the methodology used. Section III presents the results of the modeling. The final section derives conclusions from the study and proposes directions for future work.

II. Materials and Methods

A. Dataset

The brain stroke dataset is a publicly available dataset that can be found on Kaggle (https://www.kaggle.com/datasets/fedesoriano/stroke prediction-dataset). The dataset was collected from patients who were admitted to a hospital in India and contains information about their medical history, demographics, and lifestyle factors. The dataset consists of 5,110 samples from clinical brain stroke covering patients from different groups. The dataset includes some of the most important risk factors for predicting a brain stroke, including gender, age, presence of heart disease, hypertension, diabetes, high body mass index (BMI), and smoking status. The variables are categorical and numerical. The output variable is binary. Proper cleaning on the raw dataset was applied before using it for our ML model training.

B. *Data Pre-processing*

Analyses were implemented using Python and several libraries, including *Scikit-learn*, *Pandas*, and *Numpy*. Our first step was exploratory data analysis which included descriptive analysis of each variable and data visualization. This delivered some insights on the nature of the data, that helped us to build a highly accurate prediction model.

Table I shows the descriptive statistics of the continuous variables age, average glucose level, and BMI. The 'Count' row shows that there were 5,110 observations for each variable. The 'Missing values' row indicates that, the 'BMI' column has 201 null values. The 'Mean' row shows the average value of each variable, which is 43.42 years for age, 105.94 mg/dL for average glucose level, and 28.50 for body mass index (BMI: weight in kilograms (kg) by height in metres (m) squared). The 'Standard deviation' row indicates the amount of variability or dispersion in the data, with higher values indicating greater variability. The 'Min' and 'Max' rows show the range of values observed for each variable. The '25%', '50%', and '75%' rows represent the quartiles of the data distribution, with the median (50 percentile) of the data being the value at the '50%' row.

TABLE I. DESCRIPTIVE STATISTICS FOR CONTINUOUS VARIABLES

Statistic	Age	Average glucose level	BMI
Count	5,110	5,110	5,110
Missing values	0	0	201
Mean	43.42	105.94	28.50
Standard deviation	22.66	45.08	6.79
Min	0.08	55.12	14.00
25%	25.00	77.23	23.70
50%	45.00	91.85	28.10
75%	61.00	113.86	32.60
Max	82.00	271.74	48.90

Fig. 1 shows the distribution of the stroke and no stroke samples for each categorical feature. Note, because we have very imbalanced labels between stroke and non-stroke samples, the stroke samples are multiplied by 10 to be more visible. Among all the features, related medical conditions, including hypertension and heart disease seem to have very important effects. The work type, gender, and residence type variables have lower impacts. Interestingly, smoking status is not an important factor, the smoking group seem to have a lower rate for having strokes compared with non-smoking/formerly-smoked groups. Marital status seems to be crucial, with the married group having a significantly higher chance of having a stroke, but this might partially be explained by this group also being older and/or having comorbidities.

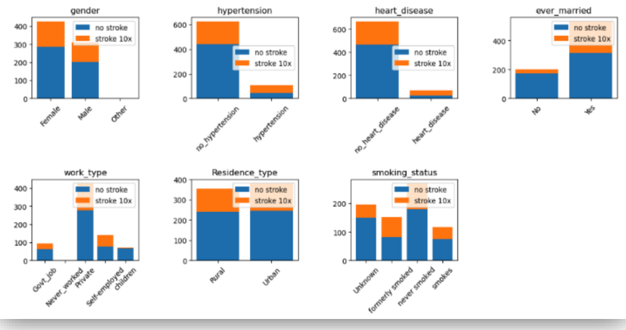

Fig. 1. Bar plots showing the distribution of stroke vs. non-stroke samples on different classes for the 7 categorical features.

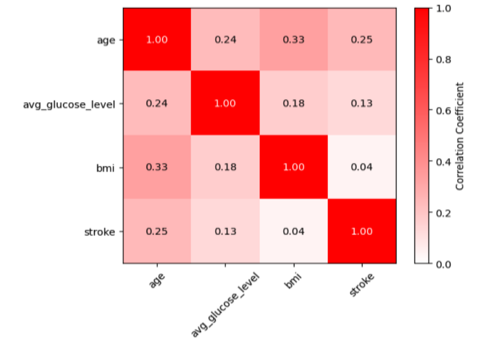

Fig. 2. Correlation coefficient matrix plot of the 3 numerical features and the label column.

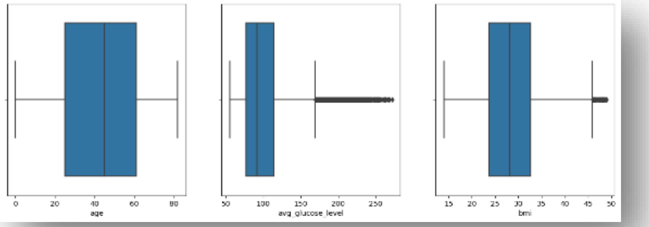

Fig. 3. Box plots.

Fig. 2 displays the correlation coefficient matrix of the three numerical features, age, average glucose level, and BMI, as well as the label column "stroke". All of the features have positive correlations. Among these numerical features, prior research tells us that age has the largest correlation with having a stroke, with a correlation coefficient of 0.25. BMI has little impact, with only a 0.04 correlation. Glucose level also has a positive correlation of 0.13 with having a stroke.

Finally, we checked if there were any outliers in our continuous data. The dataset was found to have outliers in BMI and average glucose level. They could represent valid data

Identify applicable funding agency here. If none, delete this text box.

points, such as individuals with unique medical conditions or lifestyle factors that lead to high or low values. (Fig. 3). These outliers were not removed as they might represent important meanings and removing these outliers, without a valid justification may bias the analysis and affect the accuracy of the results, hence we decided to keep all these values.

After performing descriptive and preliminary statistical analysis, we then undertook cleaning steps. We noticed that the BMI column had missing values, we used the k-nearest neighbors (kNN) imputer from the "fancyimpute" package (https://pypi.org/project/fancyimpute/) to impute the missing values. We checked the mean and standard deviation after the imputation, and they stayed relatively steady, being 28.89 and 7.85 before and 28.94 and 7.75 after the imputation, respectively. There is only one row from the gender column with value "other", we removed it because we don't have enough information for analysis or predictions. The dataset has no duplicates to remove. The numerical columns ('age', 'avg_glucose_level', 'bmi') are preprocessed by the 'MinMaxScalar' to scale into the range between 0 and 1 linearly. All the categorical columns are processed by 'OneHotEncoder'. After this, we had a total of 17 features.

The original dataset had 4,860 negative samples and 249 positive samples. To address the issue of class imbalance, we explored techniques including Synthetic Minority Oversampling Technique (SMOTE) and applying class weight.

- **Train validation test splitting:** first, we split the data into 80% train + validation set and 20% for the test set. The 20% test set has 968 negative samples and 54 positive samples.
- **SMOTE oversampling:** the ratio between the minority class (stroke) and the majority class (non-stroke) was around 0.05 in our original dataset. We generated 10 different datasets based on the ratio, from 0.1 to 1.0 with a step size of 0.1. These oversampling datasets will be used to train different versions of the model, and all will be evaluated on the same test dataset for performance comparison.
- **Class weights:** class weights on the original dataset were treated as a tuning parameter, similarly to the SMOTE oversampling ratio. We trained different versions of the model using class weights ratios between the positive samples and the negative samples on a continuous integer grid of one through 10.

C. Feature selection

SelectKBest was applied to select the best features of the dataset. Univariate feature importance was determined by calculating the F-value of the ANOVA tests between each feature and the target. From Fig. 4, we can see that there are seven features (out of 17) having close to zero F-values. These features are mostly related to 'work_type' and 'smoking_status'. The negligible F-values suggest that these features will not be helpful for making brain stroke predictions. These results align with our observations in the feature exploration section. As a result, we removed these 7 features from the list by running a SelectKBest with k = 10.

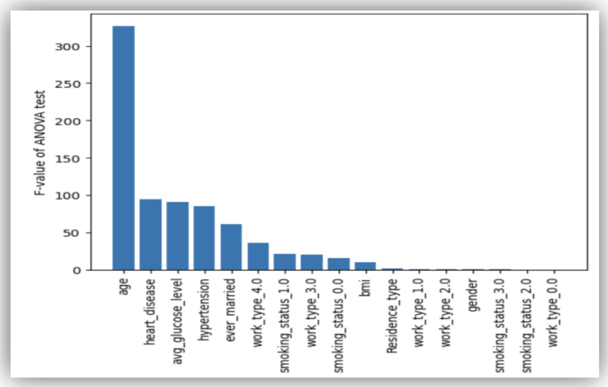

Fig. 4. ANOVA test F-value of the 17 features on the training dataset.

D. Machine learning models

We tested the performance of multiple supervised ML algorithms, including Random Forests (RF), Support Vector Machine (SVM), k-nearest neighbors (KNN), Logistic Regression (LR), and Artificial neural network (ANN). The model performance was evaluated by the metrics discussed in the next section. The final choice of the final model was based on a balance between model performance and interpretability.

The hyperparameter turning was done by using BayesSearchCV method from the *Scikit-learn* optimization package (https://scikit-optimize.github.io/stable/index.html). In contrast to the traditional GridSearchCV, not all parameter values are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions, which makes it runs significantly faster. The use of BayesSearchCV enabled us to explore a large search space, including multiple models, datasets, and hyperparameter grids.

E. Performance Evaluation

Given that our class label was highly imbalanced, we used average precision as the main metric for model selection and hyperparameter tuning instead of accuracy. The average precision is essentially the area under the curve (AUC) of the precision-recall curve, which has been used as the main metric for imbalance datasets such as fraud detection models. Other standard evaluation metrics, including accuracy, precision, recall, F1-score, were also reviewed and are detailed in the results section.

III. Results

We plotted the average precision vs. the SMOTE oversampling ratio between the minority class (stroke) and the majority class (non-stroke) (Fig. 5). We noticed that all the models prefer a more balanced dataset (higher ratio) in general. However, when the ratio between the minority class and the majority class is above 0.3, all the models seemed to benefit less from an increased oversampling ratio. Among all the algorithms, LR benefited from SMOTE the most, and RF benefited from SMOTE the least.

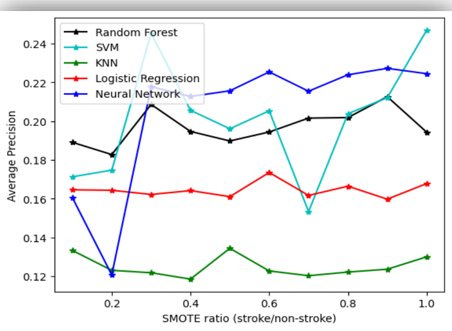

Fig. 5. Average precision vs. SMOTE ratio (stroke/non-stroke) for all 5 different models.

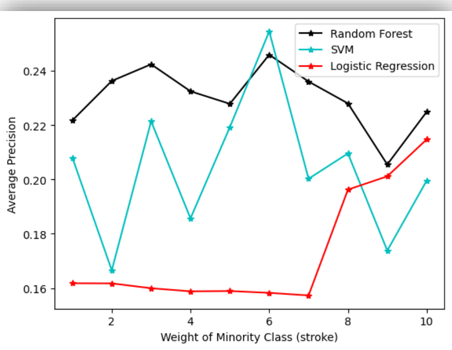

Fig. 6. Average precision vs. class weight of the minority class (stroke) for LR, RF, and SVM models.

Due to the limitation of the *Scikit-learn* package, kNN and ANN do not support the choice of class weights. We plotted the average precision vs. the class weight on the stroke class for the other 3 models (Fig. 6). We found that the SVM model is not dependent on the choice of class weight. The LR/RF model had a strong dependency on the class weight, but the effect is slow after class weight is greater than six.

The hyperparameter tuning was done by the BayesSearchCV method discuss in Section II. We listed the hyperparameters and the treatment for the imbalanced dataset of the best performing models for all the five algorithms (Table II). For most of the algorithms, having a 0.3 ratio on the oversampling training dataset gave us the best performing model. kNN preferred a higher oversampling ratio and LR performed the best when given a large class weight on the minority class.

We measured the performance metrics on the out of sample test dataset. Note that the models selected for each algorithm were determined by the average precision on the validation dataset. The average precision, F1-score, accuracy, precision, and recall metrics are shown in Fig. 7. Here, the threshold is determined by having the largest gap between the True Positive Rate (TPR) and the False Positive Rate (FPR). The Neural Network model performs the best in terms of average precision (0.296). Having a higher average precision means we had good model performance across different choices of the threshold.

TABLE II. THE HYPERPARAMETERS OF THE BEST MODEL ALONG WITH THE CHOICE OF DEALING WITH IMBALANCED DATASET, THE NUMBER AFTER SMOTE OR CLASS WEIGHT INDICATES THE OVERSAMPLING RATE (N_STROKE/N_NON-STROKE) AND EXTRA CLASS WEIGHT ON THE STROKE CLASS, RESPECTIVELY

Algorithm	**Hyperparameter of the best model**	**Best imbalance data treatment**
RF	('criterion', 'gini'), ('max_depth', 8), ('max_features', 4), ('min_samples_split', 5)	SMOTE, 0.3
SVM	('C', 0.001), ('degree', 2), ('gamma', 7.374757919230442), ('kernel', 'rbf')	SMOTE, 0.1
kNN	('n_neighbors', 22), ('weights', 'distance')	SMOTE, 0.6
LR	('C', 0.374502627266191), ('l1_ratio', 0.012168749309768193)	class weight, 0:1, 1:7
ANN	('layer1', 45), ('layer2', 48), ('layer3', 16)	SMOTE, 0.3

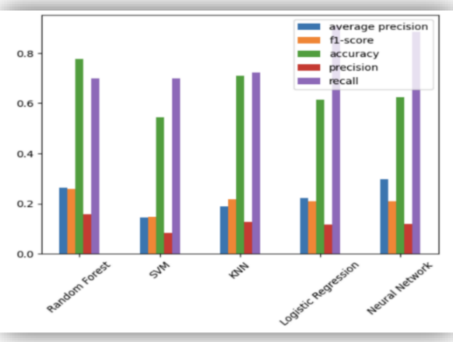

Fig. 7. Average precision, F1-score, accuracy, precision, and recall on the test dataset of the best performing model for each algorithm.

The accuracy, precision, and recall for all the algorithms with four sets of experiments are summarized in Table III. In the first two experiments, we used the raw dataset with and without hyperparameter tuning and normalization. As shown in the table, almost all the models predicted all the test dataset as non-stroke because the models were biased toward the majority class. The strong imbalance in the labels created the 0 recall scores, which made these models essentially useless. We considered the treatment of the imbalance in the labels a must, which is why we carefully checked the SMOTE and class weight treatments and found the best ways and parameters (Table II). After applying the correct treatment for the label imbalance, we received significantly better results in experiments three and four. Both experiments had hyperparameter tuning, the only difference was with (experiment four) or without (experiment three) feature selection (with SelectKBest). The model performance was not significantly different, as feature selection contributed significantly to model interoperability than model performance.

TABLE III. THE ACCURACY (ACC.), PRECISION (PREC.), AND RECALL (REC.) ON THE TEST DATASET OF THE BEST-PERFORMING MODEL FOR EACH ALGORITHM WITH RAW DATA, DATA AFTER NORMALIZATION, DATA WITH SMOTE, AND DATA WITH FEATURE SELECTION.

Algorithm	Raw data without normalization and default model parameters			Data after normalization and hyperparameters tuning with imbalanced labels			Data after normalization and hyperparameters tuning with SMOTE/class weight applied			Data after normalization and hyperparameters tuning with SMOTE/class weight and feature selection		
	Acc.	**Prec.**	**Rec.**	**Acc.**	**Prec.**	**Rec.**	**Acc.**	**Prec.**	**Rec.**	**Acc.**	**Prec.**	**Rec.**
RF	0.942	0	0	0.942	0	0	0.895	0.218	0.318	0.775	0.158	0.698
SVM	0.942	0	0	0.942	0	0	0.713	0.117	0.613	0.545	0.082	0.698
KNN	0.942	0.500	0.026	0.942	0	0	0.749	0.144	0.681	0.708	0.128	0.721
LR	0.942	0	0	0.941	0	0	0.732	0.160	0.863	0.614	0.117	0.907
ANN	0.942	0	0	0.942	0	0	0.649	0.123	0.840	0.623	0.118	0.883

The current research has some limitations. First, the dataset we are using is very small, which limits the model performance of our ML models. Additionally, the current data we are using only has physical data which is not the best input for clinical predictions. With the help of recent developments in advanced deep learning models and foundation models, we can make the best use of historical clinical data, including images, and measurements/sensor data.

IV. CONCLUSIONS AND FUTURE WORK

This study addressed some of the limitations of previous researches by employing ML techniques for brain stroke prediction using a carefully cleaned and pre-processed dataset. Various ML algorithms were investigated and the optimal model for predicting brain stroke was determined through a comprehensive evaluation of model performance and interpretability. We found that the ANN model with SMOTE oversampling and hyperparameter turning performed the best in terms of average precision. This has important implications for the use of ML in clinical settings.

Further, the study demonstrated the importance of handling class imbalance, feature selection, and hyperparameter tuning in order to achieve high performance in predicting brain stroke. We also found that, by adjusting the threshold for classification, the model can be tailored to achieve the desired balance between precision and recall, depending on the specific requirements of a clinical workflow.

This research contributes to the ongoing efforts to utilize ML for the early detection and prevention of brain stroke. The developed model can potentially be integrated into clinical decision support systems to assist healthcare professionals in the identification and management of high-risk patients.

Future work can focus on refining the model using larger and more diverse datasets, as well as exploring the integration of additional features or data sources, such as medical imaging or genomic data, to further improve the prediction accuracy and utility of the model in real-world clinical settings.

ACKNOWLEDGMENT

The author would like to thank Qassim University and the Deanship of Scientific Research for their support and funding the publication.

REFERENCES

[1] Someeh, N., Mirfeizi, M., Asghari-Jafarabadi, M. *et al.* "Predicting mortality in brain stroke patients using neural networks: outcomes analysis in a longitudinal study." Sci Rep 13, 18530 (2023). https://doi.org/10.1038/s41598-023-45877-8.

[2] Donkor, E. S. "Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life." Stroke Research and Treatment, 2018. (2018): 3238165.

[3] Soto-Cámara, Raúl, *et al.* "Knowledge on signs and risk factors in stroke patients." Journal of clinical medicine 9.8 (2020): 2557.

[4] Boehme, Amelia K., Charles Esenwa, and Mitchell SV Elkind. "Stroke risk factors, genetics, and prevention." Circulation research 120.3 (2017): 472-495.

[5] Flora, Gagan D., and Manasa K. Nayak. "A brief review of cardiovascular diseases, associated risk factors and current treatment regimes." Current pharmaceutical design 25.38 (2019): 4063-4084.

[6] Rodgers, Jennifer L., et al. "Cardiovascular risks associated with gender and aging." Journal of cardiovascular development and disease 6.2 (2019): 19.

[7] Lee, Hyunna, *et al.* "Machine learning approach to identify stroke within 4.5 hours." Stroke 51.3 (2020): 860-866.

[8] Tazin, Tahia, *et al.* "Stroke disease detection and prediction using robust learning approaches." Journal of healthcare engineering 2021 (2021).

[9] Sirsat, Manisha Sanjay, Eduardo Fermé, and Joana Camara. "Machine learning for brain stroke: a review." Journal of Stroke and Cerebrovascular Diseases 29.10 (2020): 105162.

[10] Liu, Tianyu, Wenhui Fan, and Cheng Wu. "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset." Artificial intelligence in medicine 101 (2019): 101723.

[11] Wu, Yafei, and Ya Fang. "Stroke prediction with machine learning methods among older Chinese." International journal of environmental research and public health 17.6 (2020): 1828.

[12] Emon, Minhaz Uddin, *et al.* "Performance analysis of machine learning approaches in stroke prediction." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.

[13] Sailasya, Gangavarapu, and Gorli L. Aruna Kumari. "Analyzing the performance of stroke prediction using ML classification algorithms." International Journal of Advanced Computer Science and Applications 12.6 (2021).

[14] Ozaltin, Oznur, *et al.* "A Deep Learning Approach for Detecting Stroke from Brain CT Images Using OzNet." Bioengineering 9.12 (2022): 783.