ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/370478193

Using Machine Learning Algorithm as a Method for Improving Stroke Prediction

Article *in* International Journal of Advanced Computer Science and Applications · January 2023

DOI: 10.14569/IJACSA.2023.0140481

CITATIONS
0

READS
58

5 authors, including:

Nojood Alageel
University of Tabuk
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Rehab Alharbi
University of Tabuk
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Lubna A. Alharbi
University of Tabuk
18 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

A virtual companion for educational attainment View project

All content following this page was uploaded by Lubna A. Alharbi on 09 May 2023.

The user has requested enhancement of the downloaded file.

Using Machine Learning Algorithm as a Method for Improving Stroke Prediction

Nojood Alageel, Rahaf Alharbi, Rehab Alharbi, Maryam Alsayil, Lubna A. Alharbi
Faculty of Computers and Information Technology
University of Tabuk
Tabuk, Saudi Arabia

***Abstract*—Having sudden strokes has had a very negative impact on all aspects in society to the point that it attracted efforts for better improvement and management of stroke diagnosis. Technological advancement also had an impact on the medical field such that nowadays caregivers have better options for taking care of their patients by mining and archiving their medical records for ease of retrieval. Furthermore, it is quite essential to understand the risk factors that make a patient more susceptible to strokes, thus there are some factors that make stroke prediction much easier. This research offers an analysis of the factors that enhance the stroke prediction process based on electronic health records. The most important factors for stroke prediction will be identified using statistical methods and Principal Component Analysis (PCA). It has been found that the most critical factors affecting stroke prediction are the age, average glucose level, heart disease, and hypertension. A balanced dataset is used for the model evaluation which was created by sub-sampling since the dataset for stroke occurrence is already highly imbalanced. In this study, seven different machine learning algorithms are implemented: Naïve Bayes, SVM, Random Forest, KNN, Decision Tree, Stacking, and majority voting to train on the Kaggle dataset to predict occurrence of stroke in patients. After preprocessing and splitting the dataset into training and testing sub-datasets, these proposed algorithms were evaluated according to accuracy, f1 score, recall value, and precision value. The NB classifier achieved the lowest accuracy level (86%), whereas the rest of the algorithms achieved similar accuracies 96%, f1 scores 0.98, precision 0.97, and recall 1.**

***Keywords*—Stroke prediction; machine learning; PCA; decision tree; KNN; majority voting; Naïve Bayes**

I. Introduction

Strokes or cerebrovascular accidents are considered among the top three causes or morbidity and mortality in many countries all over the world [1], such that it accounts for around 10% of the world-wide deaths which makes it the second leading cause of death. As an estimation, approximately 700,000 individuals suffer from strokes each year, and by the year 2030, it is expected that this number will be greatly increased and will cause a medical cost of 240 billion dollars in the US alone [2].

The world health organization WHO defines stroke as a brain-related illness such that it leads to the dysfunction of the brain, and it could be focal, acute, or diffuse. This dysfunction is mainly a result of vessel problems and it lasts for longer than 24 hours. Ultimately, there are many types of strokes depending on the exact origin of the dysfunction, which defines four main types of strokes: ischemic stroke, subarachnoid hemorrhage, cerebral venous sinus thrombosis, and intra-cerebral hemorrhage [3].

In general, brain strokes can be classified as either ischemic or hemorrhagic. Ischemic strokes are the predominant type and they account for approximately 70% of the total stroke incidents [4]. Ischemic strokes occur as a result of clots in vessels, or hypotensive vasoconstriction, arterial tears, and sickle cell anemia [5]. On the other hand, hemorrhagic strokes account for approximately 15% of the total incidents, yet their effects are usually more detrimental as they often lead to serious morbidity and death [6]. Hemorrhagic strokes occur due to many causes among which are the vascular malfunction and uncontrolled hypertension [7].

When considering the risk factors or the reasons behind the occurrence of strokes, these can be divided into two types of factors depending on their origin, meaning that there are factors that can be changed or modified, and factors that cannot be modified [8]. Some of the modifiable (changeable) factors is hypercholesterolemia, diabetes, and hypertension. On the other hand, the non-modifiable factors include age, gender, and the genetic factors in play [9].

The traditional stroke identification methods are usually the magnetic resonance imaging MRI and Computed Tomography CT scans which are expensive and invasive [10]. However, since the stroke occurrence is a very time-sensitive issue, dealing with it in a timely efficient manner is very important because in most cases, death or permanent damage from stroke can be prevented if the diagnosis happens early on [11], [12]. Therefore, it is essential to develop medical tools and devices that allow physicians to diagnose a stroke without being invasive or uncomfortable, through relying on biomarkers for example or studying the risk factors. Machine learning poses as the perfect tool for predicting whether a stroke can occur or not based on different factors. Machine Learning is capable of diagnosing, treating, and predicting disease through analyzing clinical data.

In this research, the aim is to develop and implement a machine learning-based system for the accurate prediction of future occurrence of stroke in patients based on several features including age, gender, BMI, and medical history. The primary objective is to get this system to predict the occurrence of stroke by 100% accuracy so that lives can be saved. The

contributions that are provided in this report can be listed as follows:

- Predictive analytics approach to predict stroke recurrence is suggested.
- Machine learning and neural network algorithms are implemented.
- A publicly available dataset of electronic health records is used.
- The subsampling techniques for balancing the dataset is followed.
- Dimensionality reduction techniques are implemented in analyzing the attributes.
- The most impactful features for predicting strokes are picked out and shown.

Thus, after mentioning the contributions, it can be said that the added value of this paper lies in the fact that it uses simple algorithms to achieve high accuracies with explainable results, instead of using complex algorithms. More precisely, the majority of the chosen algorithms were able to score similarly high results.

The rest of the paper is distributed as follows: Section II is the literature review where some studies are mentioned with their relative results. Section III is for describing the details of the methodology followed in this study. Section IV shows the results that were obtained by the proposed model. Finally, the paper is concluded with Section V as a conclusion.

II. Literature Review

Since technologies like machine learning and deep learning can greatly benefit the medical sector by increasing the accuracy of stroke prediction, many studies were conducted to explore how exactly machine learning models can be used in predicting strokes. In this section, a group of similar studies that relied on freely available datasets such as Kaggle and datasets from local hospitals or labs were selected.

Dritsas and Trigka [13] gathered data from Kaggle such that the participants were 3254. The dataset consists of 10 independent features such as age, BMI value, glucose level, smoking status, hypertension, and whether the individual had contracted a stroke before. Data preprocessing was performed on the dataset, and class balancing was implemented through a resampling method known as SMOTE. Machine learning models namely Stacking, Decision Tree, Random Forest, Majority Voting, Naïve Bayes, Multilayer Percepton, KNN, Stochastic Gradient Descent, and logistic regression were used for predicting stroke or no-stroke. It appears from the results that the stacking classifier performed best and achieves 0.989 AUC value, with 0.974 precision and 0.974 recall. The other high performing models were Random Forest, KNN, and Majority Voting.

Rakshit [14] also relied on the Kaggle dataset and some of the algorithms as [13] namely Decision Tree, Naïve bayes, Support Vector Machine, Random Forest, K-Nearest Neighbor and Logistic Regression. According to their results, the best performance was recorded by Decision Tree followed by KNN (96.3%).

Using Kaggle dataset, Sailasya and [15] discussed the prediction of stroke based on machine learning algorithm namely Logistic Regression, K-Nearest Neighbour, Random Forest, Support Vector Machine, Naïve Bayes, and Decision Tree algorithms. Undersampling method was used to handle the imbalanced data. The results showed that among these algorithms, Naïve Bayes had the best performance with 82% overall accuracy compared to 80 % for both K-NN and support vector machine, and 78% for logistic regression.

Emon et al. [16] collected information for 5110 patients were taken from Bangladesh's medical clinic. Then, ten different machine learning classifiers, which are ANN, MLP, K Neighbours algorithm, SGD, QDA, AdaBoost, Gaussian, QDA, GBC, and XGB were used. The weighted voting classifier offered the highest accuracy of about 97%, GBC and XGB classifiers achieving 96% accuracy, right before AdaBoost classifier that scored 94% accuracy. On the other hand, the lowest accuracy was recorded by the SGD classifier with a value of 65%.

Shoily et al. [17] used KNN, Naïve Bayes, J48, and Random Forest classifiers. They gathered data from multiple sources to create their dataset of 1058 individuals overall and took a total of 28 features. The authors performed integer encoding to make the machine learning algorithms suitable for WEKA processing. After that, feature selection took place, and the models were trained and tested then evaluated according to f1 score, accuracy, precision, and recall. In terms of accuracy, Random Forest as well as KNN and J48 achieved the same results: 0.998 accuracy, 0.998 f1 score, 0.998 precision and 0.98 recall, whereas Naïve Bayes achieved 0.856 accuracy and 0.861 f1 score.

Abedi et al. [18] created a dataset termed “GNSIS”, which is a collection of electronic health records from 2003 to 2019. Data preprocessing was performed, and the individuals within the dataset were classified into six groups totaling 2091 individuals, 1 group consists of those who didn’t contract stroke in the last 5 years, and the other 5 groups are of stroke patients. After that, the dataset was split into training and testing by 80 to 20 ratio, where data imputation was also done. From the dataset, 53 features existed including BMI, diastolic blood pressure, creatinine, and smoking status. Then, four feature selection sets were created with exclusion of some features at times, and six machine algorithm models were used each in all of the 5 recurrence prediction window, which makes 24 models in total. For 1year prediction window, Random Forest achieved the better results with 90% accuracy, whereas the average accuracy of all models was 88%. The average accuracy achieved in the 5 years prediction window was 78%, thus the wider prediction window results in less accurate performances.

Relying on electronic health records, Nwosu et al. [19] used a dataset published by McKinsey & Company, containing 11 different attributes including body mass index, heart disease, marital status, age, average blood glucose, and smoking status. In the dataset, 548 patients suffered from stroke whereas 28524 patients didn’t suffer from any previous strokes, thus the

dataset needed downsizing. In fact, 1000 downsizing experiments were done to avoid sampling bias. After that, 70% of the dataset was selected for training and 30% for testing. Over the 1000 experiments, the Neural Network model achieved the best accuracy of 75.02%, followed by Random Forest at 74.53% accuracy and Decision Tree at 74.31%.

In [13], the dataset was large and their study was able to score very similar results to ours, even though at times our metrics were better. However, they did not mention the scored accuracy. Similarly, oir proposed model acheived better performances than [14].

It's notworthy that the proposed method in this study acheived 96.7% accuracy, which is significantly higher than the accuracy of [15] (80%).

In [16] the authors chose complex algorithms such as ADABoost and XGB and were only able to acheive similar results to ours, whereas we acheived the high performances using much simpler algorithms, which is more desirable.

In [17] the study relied on 28 inputs to predict stroke occurance, which is usually difficult to obtain from patients for a quick prediction. Conversely, the proposed salgorithms in the proposed system in this paper relied on 9 factors only as an input. In addition, [17] used a much smaller dataset.

Similarly, [18] used a very high number of input, which is not desireable for ML algorithms.

III. Methodology

Machine learning permits the advancement of a system by making it capable of learning and improving from past experiences without the need of constant continuous programming. These systems learn through machine learning how to analyze data to identify patterns that help them make decisions in the future without the help of humans.

The real influence of machine learning becomes crystal clear in the fields that deal with a huge amount of data such as retail, health, government, finance, and transportation. This is mainly due to the decision-making capabilities of machine learning since it can understand the data and fit them into the different models such that human can rely on them for decisions. Machine learning models are efficiently used for identifying diseases and computing risk satisfaction in the healthcare sector. The previous are only a few examples of the capabilities of machine learning.

Nonetheless, real-life data cannot be simply directly processed by the selected machine learning algorithms which is why data preprocessing is an essential step before applying the ML models. After that, the available dataset must be divided into training and testing datasets. The training step is performed in order to teach the algorithm about the data. In addition, unknown data can be predicted through ML algorithms, yet the prediction results are checked against each other.

This study is dedicated to implementing machine learning algorithms for stroke prediction, since it is a dangerous and common disease. Machine learning is often suitable for datasets due to its simplicity, structure, and compatibility with a wide range of machine learning platforms and tools. For this reason, machine learning algorithms were chosen in this study. However, the limitation of this method is that it requires many inputs for the model to be able to make predictions. It is possible that when predicting a person's status, not all inputs are available, and then the model will not be able to predict. This issue was removed since the chosen dataset was large.

In general, a wide set of attributes are used to predict strokes such as gender, age, and blood pressure data among many others. Additionally, the performance of a number of machine learning algorithms was examined to see which one is best suited for predicting stroke incidence based on the dataset. Ultimately, the chosen ML algorithm must give the predictions with the highest accuracy.

A. Implementation

In this section, the machine learning algorithms that will be implemented and put to the test are presented and described.

1) Naive bayes: In the cases when the features are highly independent, the Naïve Bayes NB algorithm can lead to probability maximization [20]. There is a feature vector fi for every subject i at that class c such that $P(c|fi1, \ldots, fin)$ is maximized. The formula that defines the conditional probability is as in (1):

$$P(c|fi1, \ldots, fin) = Pfi1, \ldots, fin|c/P(c)Pfi1, \ldots, fi4 \quad (1)$$

In (1), $Pfi1, \ldots, fin|c = \prod n\, j = 1\, P(fij|c)$ resembles the features probability given class, whereas the previous feature probability is resembled by $Pfi1, \ldots, fin$, and previous class probability is resembled by P(c). Through maximizing the numerator of 1, its number is also maximized, and the optimization becomes as in (2):

$$\hat{c} = arg\ max\ P(c)\ n \prod j = 1\, P(fij|c) \quad (2)$$

2) Random forest: There are multiple decision trees in a Random Forest (RF) classifier [21]. When these independent trees are combines in an ensemble through resampling, the results become subsets of instances that are used for classification and regression. In a random forest, the final output is a result of majority voting, since each independent tree generates its own classification outcome.

3) K-Nearest neighbors: K-nearest neighbors (KNNs) classifier depends on Manhattan or Euclidean distances to evaluate similarities or differences between instances in the dataset [22]. More often than not, the Euclidean distance is the metric of choice in KNN classifiers. In stroke prediction, the features vector of the new samples would be fnew. The closest K vectors (neighbors) to fnew is determined through KNN. After that, the class where most neighbors belong is given the fnew value.

4) Decision tree: In the proposed Decision Tree model [23], J48 resembling the single classifier, and RepTree [24]resembling the base classifier were chosen. The classes are denoted by the leaf nodes, whereas the features are denoted by the internal nodes. The Gini index technique is employed by the J48 classifier in order to split a single feature

at each node. Gini index is a fast and simple decision learner that is capable of building a DT through the gained information as an impurity measure and pruning via reduced-error pruning.

5) Majority voting: Soft or hard voting is implemented through simple majority voting, assuming an ensemble of K basis models. This method allows the prediction of the class label associated to an instance [25]. The hard voting collects the votes related to each class label and choses the one with most votes as an output, that is the candidate class. On the other hand, the predicted probabilities for every class label are collected by soft voting, and the class label with the largest probability is predicted. In the proposed model, the hard voting is adopted. Its general function of hard voting is represented by (3):

$$\max \sum_{K=1}^{K} 1\ Pk, c \qquad (3)$$

Such that Pk,c is the prediction or probability of k-th model in class c, and c = {Stroke, Non − Stroke}.

6) Stacking: One of the ensemble learning techniques is the Stacking, where the predictions of multiple heterogeneous classifiers are integrated within a meta-classifier. Usually, the training set is used for training the base models whereas the outputs of the base models are used to train the meta-classifier. Here, J48, RF, NB, and RepTree were chosen to be included in the stacking ensemble classifier. The predictions of these collective classifiers are used for training a logistic regression meta-classifier.

The influence of machine learning parameters on the performance of a model can vary depending on the specific algorithm used, the dataset being analyzed, and the problem being solved. However, in general, adjusting the values of these parameters can have a significant impact on the accuracy and speed of a machine learning model. In this study, several parameters for the different algorithms were modified to make sure better results are achieved. The modifications to the parameters of each algorithm are shown in Table I.

TABLE I. THE CHANGED PARAMETERS FOR EACH ALGORITHM

Algorithm	Specific Parameters
KNN	- Number of neighbors (k): value is 6 - Distance metric: (Euclidean distance)
SVM	- Kernel type: default kernel is radial basis function (RBF) - Regularization parameter (C): default value is 1.0
DT	- Tree depth: 3
NB	No modifications
RF	- Number of trees in the forest: default value is 100 - Maximum depth of each tree: 9

B. Pre-Processing

1) Dataset description: For this study, the dataset of choice was adopted from Kaggle. The dataset comprises a large number of participants of which only those above 18 years old are chosen, making the total of the participants 3254. The 9 input attributes (most of which are nominal) as well as the target class are briefly described in Table II.

TABLE II. DESCRIPTION OF THE ATTRIBUTES/FEATURES IN THE DATASET

Risk factor	Description	Details
Age (year)	The actual age of participants	All of the participants are older than 18
Gender	Whether the participants is male or female	In the dataset, 1260 participants are males, and 1994 participants are female
Hypertension	The participant suffering from hypertension or not	12.54% of the participants in the dataset are hypertensive
Heart disease	The participants suffering from heart diseases in general or not	6.33% of the participants in the dataset suffer from heart diseases
Marital status	The participant is married or not	In the dataset, 79.84% of the participants are married
Work type	The work status of the participants	65.02% of them work in the private sector, 19.21% are self-employed, 15.67% have a job, while 0.1% have never worked
Residence type	Whether the participant lives in an urban or rural place	51.14% of the participants in the dataset live in urban place whereas the rest live in rural places
Avg glucose level (mg/dL)	The average level of a participant's blood glucose	Numerical values for each patient
BMI (Kg/m2)	Participant's body mass index of the participants	Numerical values for each patient
Smoking status	Whether a participant currently smokes or not	22.37% of the participants smoke, 24.99% of them have smoked in the past, and 52.64% of them have never smoked
Stroke history	Whether the participant has had a stroke previously or not	5.53% of the participants have previously had a stroke

2) Data pre-processing: If the data were kept in their raw form, it might negatively affect the quality of the predictions, which is why data preprocessing is essential. In the raw data, there might be some missing values and redundancy as well as noisy data, so tasks like data discretization and reduction of redundant values are performed. Furthermore, one of the data pre-processing tasks is to balance the classes through selecting one of the available resampling techniques. In the proposed workflow, the SMOTE technique will be used so that the participants can be distributed over the stroke and non-stroke classes in a balanced way. In more details, the minor class which belongs to the stroke participants, oversampling was done to increase the number of participants in this class. In addition, there were not missing or null values, so neither dropping nor data imputation was applied.

C. Proposed Workflow

The details of the proposed approach and methodology can be summed up in a workflow chart presented in Fig. 1.

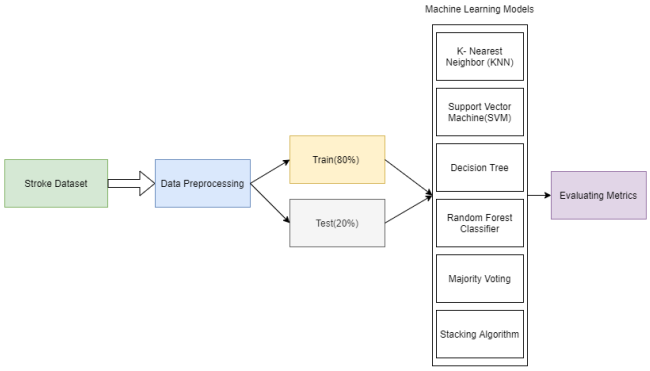

Fig. 1. Workflow of the proposed model.

Initially, the Kaggle dataset with 3254 participants is acquired. Then, the data is visualized to determine the specifics such as visualization of column and the relevant attributes. In this stage, the distribution of the participants can be visualized over the different features such as the age and gender distribution. After that, data preprocessing takes place where the data is being prepared through reduction of redundant information or resampling. In this approach, the SMOTE technique is selected. Later, the data is split into 80% for training and 20% for testing. Six different algorithms were selected to perform the predictions: Naive Bayes, Random Forest, K-Nearest Neighbors, Decision Tree, Majority Voting, and Stacking. These algorithms are then evaluated according to the evaluation metrics.

D. Evaluation

A group of performance metrics were chosen to evaluate the performance of the chosen machine learning methods. The most commonly used metrics in general will also be used in this study [see (4), (5), (6) and (7)]. Sensitivity, which is also termed Recall, represents the true positive results where participants who have had a stroke were successfully classified into the stroke class from the collective totality of the participants. Precision on the other hand specifies how many of those who had a stroke actually belong to this class. Whereas, Recall shows how many of those who had a stroke are correctly predicted. F-measure is the harmonic mean of the precision and recall and sums up the predictive performance of a model.

$$\text{Ρεχαλλ} = \text{ΤΠ} / (\text{ΤΠ} + \text{ΦΝ}) \quad (4)$$

$$\text{Πρεχισιον} = \text{ΤΠ} / (\text{ΤΠ} + \text{ΦΠ}) \quad (5)$$

$$\text{Φ–Μεασυρε} = 2\ \text{Πρεχισιον} \cdot \text{Ρεχαλλ}\ \text{Πρεχισιον} + \text{Ρεχαλλ} \quad (6)$$

$$\text{Αχχυραχψ} = (\text{ΤΝ} + \text{ΤΠ}) / (\text{ΤΝ} + \text{ΤΠ} + \text{ΦΝ} + \text{ΦΠ}) \quad (7)$$

Where, true positive is designated by TP and false negative is designated by FN, false positive is designated by FP and true negative is designated by TN.

On the other hand, Area under curve (AUC) is also a beneficial metric, where the values must be between 0 and 1, such that the higher the AUC value, the better the performance. If the model can discriminate between the instances of two classes perfectly, then AUC would be 1. Conversely, if the model fails to distinguish between any instances, the AUC would be 0.

IV. Results

A. Data Visualization

The dataset can be visualized where each of the features or attributes are analyzed separately and against each other. Fig. 2, for instance, illustrates how the participants from the dataset are distributed according to age and gender. It can be seen that the patients have an average of 41 years old, and that there are slightly more females than males, specifically, 56% of the participants are female.

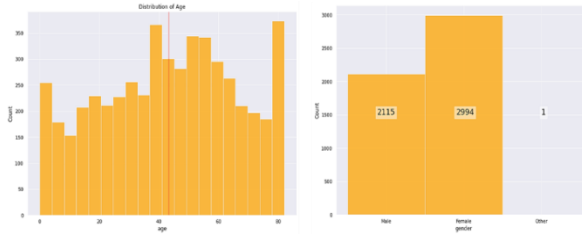

Fig. 2. Distribution of data by age and gender.

On the other hand, Fig. 3 shows how the patients who had suffered from a previous stroke are distributed according to age, where it becomes clear that approximately all of them were older than 40 years old, and the largest number of stroke patients was 80 years old. While the patients who didn't suffer from stroke were distributed among the different age groups.

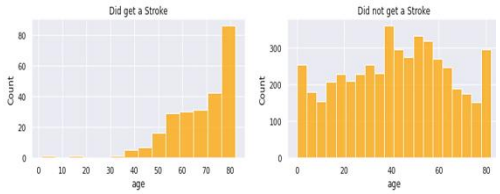

Fig. 3. Distribution of patients who suffered from a stroke and those who didn't according to age.

In addition, Fig. 4 shows that the majority of the participants didn't suffer from any heart diseases, nor did they suffer from Hypertension.

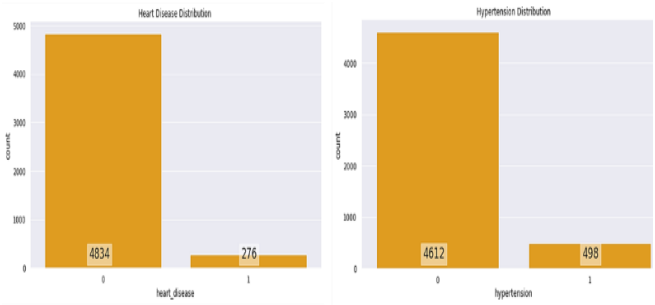

Fig. 4. Distribution of data over heart disease and hypertension cases.

Moreover, 25% of the patients were obese, and 18% of the participants were overweight according to Fig. 5.

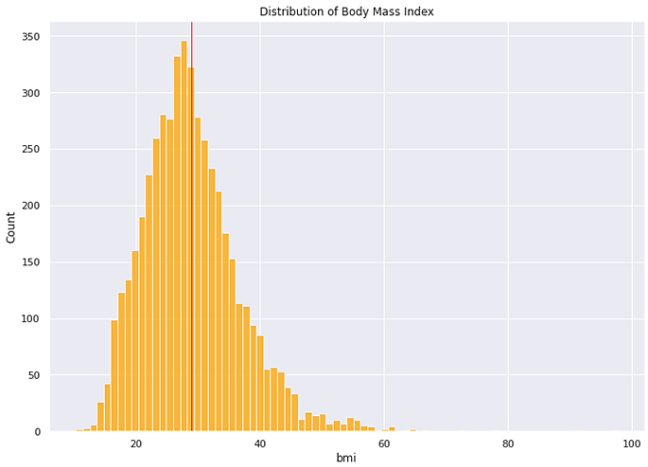

Fig. 5. Distribution of data according to BMI.

In Fig. 6 depicts that the majority of the patients were smokers, followed by a large group of participants with unknown smoking status (1544 participants).

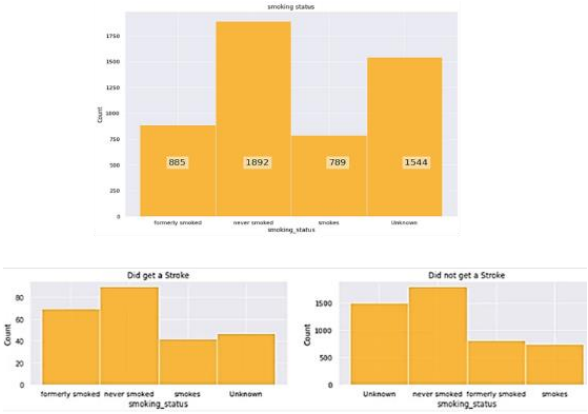

Fig. 6. Distribution of data according to smoking status and relation to stroke.

Additionally, the majority of the participants are employed in the private sector. Meanwhile, the data was almost equally distributed between living in rural and urban areas, as depicted in Fig. 7.

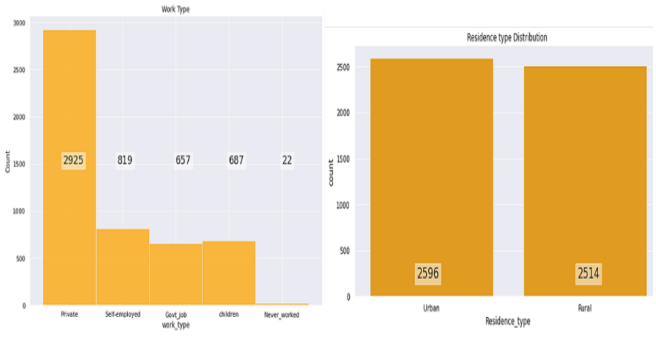

Fig. 7. Distribution of data according to work type and residency.

Furthermore, the participants in the dataset scored mostly healthy levels of blood glucose (below 100) as shown in Fig. 8.

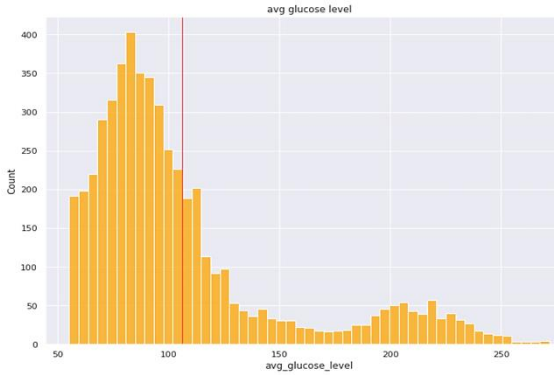

Fig. 8. Distribution of data according to average glucose level.

B. Model Evaluation

After acquiring the data, preprocessing it, and visualizing it, it was used to train and test several classifiers whose role was to predict whether a stroke occurs to a patient or not. The evaluation results for each classifier are presented in Table III.

TABLE III. EVALUATION OF THE DIFFERENT CLASSIFIERS IN TERMS OF ACCURACY, F1 SCORE, RECALL, AND PRECISION

Algorithm	Accuracy	F-1 Score	Recall	Precision
KNN	0.9633	0.98	1.00	0.97
SVM	0.9674	0.98	1.00	0.97
Decision Tree	0.9674	0.98	1.00	0.97
Gaussian NB	0.8655	0.93	0.89	0.97
Random Forest	0.96741	0.98	1.00	0.97
Voting Classifier	0.9674	0.98	1.00	0.97
Stacking Classifier	0.96741	0.98	1.00	0.97

In addition, these evaluation metrics can be seen in Fig. 9 which clearly illustrates that in fact, all of the proposed algorithms in this study have similar results in predicting the stroke occurrence in patients, except for Naïve Bayes which clearly has the worst performance among these classifiers.

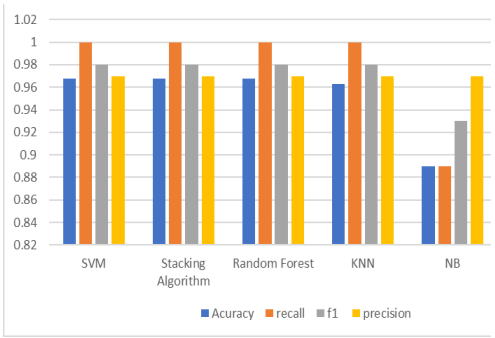

Fig. 9. Evaluation of different algorithms according to several evaluation metrics.

However, taking into consideration that the stacking classifier is an ensemble model, it can be said that choosing stacking algorithm might enhance the prediction results in case of stroke.

V. CONCLUSION

Stroke is among the top medical accidents that lead to death but even in the case of survival, stroke leaves serious implications on the lives of its patients. A patient who has previously suffered from brain stroke, shall he remain alive, might suffer the consequences in what seems like paralysis, among many other life-long complications. Since there are several risk factors that enhance the chances of strokes, its prediction beforehand is possible. Machine learning algorithms have been employed for this purpose promising fast and efficient prediction results.

In this study, the aim was to develop the optimal system that can predict stroke occurrence with high accuracy based on several risk factors collected about the patients. Here, multiple machine learning algorithms are implemented such as Naïve Bayes, SVM, Random Forest, KNN, Decision Tree, Stacking, and majority voting to check the results provided by each algorithm. After that, the choice of the optimal algorithm will be made depending on the evaluation results.

After appropriate preparation of the data, it was divided into training and testing parts such that all of the proposed algorithms are tested for their ability of predicting stroke occurrence. The evaluation metrics of choice were accuracy, f1 score, recall value, and precision value. Ultimately, the results showed that the selected algorithms perform quite well in predicting the strokes, such that SVM, DT, RF, KNN, Voting, and stacking classifier almost scored the same values. The algorithms scored 96% accuracy, 0.98 f1 score, 1 recall value, and 0.97 precision value. However, the achieved results suggest that the Naïve Bayes algorithm might not be the best choice for creating a stroke prediction model since it scored less accuracy levels (86%), less f1 score (0.93), less Recall (0.89), but the same precision value (0.97).

REFERENCES

[1] V. L. Feigin, C. M. M. Lawes, D. A. Bennett, and C. S. Anderson, "Stroke epidemiology: A review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century," The Lancet Neurology, vol. 2, pp. 43-53, 2003.

[2] B. Ovbiagele, L. B. Goldstein, R. T. Higashida, V. J. Howard, S. C. Johnston, O. A. Khavjou, *et al.*, "Forecasting the future of stroke in the United States," Stroke, vol. 44, pp. 2361-2375, 2013.

[3] World Health Organization. Noncommunicable Diseases and Mental Health Cluster, "WHO STEPS stroke manual : the WHO STEPwise approach to stroke surveillance," World Health Organization, Geneva, 2005.

[4] B. C. V. Campbell, D. A. De Silva, M. R. Macleod, S. B. Coutts, L. H. Schwamm, S. M. Davis, *et al.*, "Ischaemic stroke," Nature Reviews Disease Primers, vol. 5, p. 70, 2019.

[5] R. V. Krishnamurthi, V. L. Feigin, M. H. Forouzanfar, G. A. Mensah, M. Connor, D. A. Bennett, et al., "Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990–2010: findings from the Global Burden of Disease Study 2010," The Lancet Global Health, vol. 1, pp. e259-e281, 2013.

[6] S. Kamalakannan, A. S. V. Gudlavalleti, V. S. M. Gudlavalleti, S. Goenka, and H. Kuper, "Incidence & prevalence of stroke in India: A systematic review," Indian Journal of Medical Research, vol. 146, pp. 175-185, 2017.

[7] E. S. Donkor, "Stroke in the 21st century: A snapshot of the burden, epidemiology, and quality of life," Stroke Research and Treatment, vol. 2018, p. 3238165, 2018.

[8] V. L. Feigin, B. Norrving, and G. A. Mensah, "Global burden of stroke," Circulation Research, vol. 120, pp. 439-448, 2017.

[9] M. J. O'Donnell, D. Xavier, L. Liu, H. Zhang, S. L. Chin, P. Rao-Melacini, *et al.*, "Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): A case-control study," The Lancet, vol. 376, pp. 112-123, 2010.

[10] M. Kaur, S. R. Sakhare, K. Wanjale, and F. Akter, "Early stroke prediction methods for prevention of strokes," Behavioural Neurology, vol. 2022, p. 7725597, 2022.

[11] M. Lee, J. Ryu, and D.-H. Kim, "Automated epileptic seizure waveform detection method based on the feature of the mean slope of wavelet coefficient counts using a hidden Markov model and EEG signals," ETRI Journal, vol. 42, pp. 217-229, 2020.

[12] B. Kim, N. Schweighofer, J. P. Haldar, R. M. Leahy, and C. J. Winstein, "Corticospinal tract microstructure predicts distal arm motor improvements in chronic stroke," Journal of Neurologic Physical Therapy, vol. 45, pp. 273-281, 2021.

[13] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," Sensors, vol. 22, p. 4670, 2022.

[14] T. Rakshit and A. Shrestha, "Comparative analysis and implementation of heart stroke prediction using various machine learning techniques," International Journal of Engineering Research & Technology, vol. 10, pp. 886-890, 2021.

[15] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," International Journal of Advanced Computer Science and Applications, vol. 12, pp. 539-545, 2021.

[16] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, "Performance analysis of machine learning approaches in stroke prediction," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2020, pp. 1464-1469.

[17] T. I. Shoily, T. Islam, S. Jannat, S. A. Tanna, T. M. Alif, and R. R. Ema, "Detection of stroke disease using machine learning algorithms," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, 2019, pp. 1-6.

[18] V. Abedi, V. Avula, D. Chaudhary, S. Shahjouei, A. Khan, C. J. Griessenauer, *et al.*, "Prediction of long-term stroke recurrence using machine learning models," Journal of Clinical Medicine, vol. 10, p. 1286, 2021.

[19] C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John, "Predicting stroke from electronic health records," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 5704-5707.

[20] D. Berrar, "Bayes' theorem and naive bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. ed Oxford: Academic Press, 2019, pp. 403-412.

[21] S. Alexiou, E. Dritsas, O. Kocsis, K. Moustakas, and N. Fakotakis, "An approach for personalized continuous glucose prediction with regression trees," in *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, Preveza, Greece, 2021, pp. 1-6.

[22] P. Cunningham and S. J. Delany, "K-Nearest neighbour classifiers - a tutorial," ACM Computing Surveys, vol. 54, p. Article 128, 2021.

[23] M. B. A. Snousy, H. M. El-Deeb, K. Badran, and I. A. A. Khlil, "Suite of decision tree-based classification algorithms on cancer gene expression data," Egyptian Informatics Journal, vol. 12, pp. 73-82, 2011.

[24] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, 2018, pp. 1-7.

[25] A. Dogan and D. Birant, "A weighted majority voting ensemble approach for classification," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, Samsun, Turkey, 2019, pp. 1-6.

View publication stats