Int. J. Math. And Appl., 10(1)(2022), 101–109
ISSN: 2347-1557
Available Online: http://ijmaa.in/

International Journal of *Mathematics And its Applications*

Application of Machine Learning to Analyze the Risk Factors of Stroke

Karan Keerthy[1,*]

1 Briarcliff High School, 444 Pleasantville Rd, Briarcliff Manor, NY 10510, United States

Abstract: In 2019, stroke was the second leading cause of death and disability-adjusted life years, globally. 80% of second strokes have been demonstrated to be preventable by using medication, maintaining a strict diet, and engaging in physical activity. Considering its debilitating effects, early detection of stroke is an important area of interest. Thus, this study aims to identify key risk factors for stroke, to encourage proper monitoring and lifestyle changes that can prevent stroke onset. To determine the most significant risk factors, a machine learning (ML) based artificial neural network model was derived from the Keras library in Python. With an accuracy of 92%, the model was then applied to different combinations of risk factors using the SelectKBest function. At first, a feature selection strategy using chi-squared scoring was used to select the K best features from a combination of risk factors. These prominent features were then used to train the ANN to predict presence of stroke. The accuracy of the trained model was presented in terms of area under receiver operating characteristic curve (AUC). Average glucose level, age, and BMI were determined to be the most predictive risk factors of stroke. ROC analysis yielded an AUC value of 0.73, which indicates good test performance of the model's determination of the aforementioned most significant combination of risk factors. In addition to confirming the significance of frequently reported risk factors in the existing literature such as average glucose level, age, BMI, smoking status, and hypertension, the model identifies occupation as the next most predictive risk factor for stroke, surpassing even heart disease. Thus, with information on patients, preventative measures can be given based on previously unidentified risk factors like occupation to hopefully avoid the long-term impacts of a potential stroke.

Keywords: Machine Learning, Risk Factors, Stroke.

1. Introduction

In 2019, stroke was the second leading cause of death and disability-adjusted life years, globally [7]. Stroke occurs when blood supply to a part of the brain is cut off or reduced, depriving the brain of oxygen and nutrients (Kumar, 2018). Prompt treatment is imperative as vital early action can reduce brain damage and complications like paralysis and memory loss [14]. A stroke can also occur when a blood vessel in the brain ruptures. Whether the brain is deprived of blood, or a blood vessel ruptured, parts of the brain can become damaged and die [14]. The aforementioned strokes are known as ischemic strokes and hemorrhagic strokes respectively [14]. When brain cells are damaged, the parts of the body that those brain cells control are also damaged [14]. In addition to the negative societal impact that strokes have each year on people, the frequent treatment required serves to man economic drawbacks [12].

The US healthcare system spends over 3 trillion dollars nationally on national healthcare; that is approximately 18% of the GDP [13]. A large portion of US citizen's taxes go towards healthcare, and as health care costs increasing exponentially each year, more issues will arise between taxpayers and the government [13]. As people grow older, they are more susceptible

* *E-mail: karkeerthy2445@gmail.com*

to various issues, which leads the retired demographic to use government insurance programs and sink into severe debt. In 2019, on average the US spends 11,000 dollars on each patient [12]. The healthcare system grows weaker daily; in addition to patient costs, a large portion of money is spent towards the production and maintenance of pharmaceutical. Furthermore, doctor error is also very common which also raises annual healthcare costs. It is necessary for any beneficial solution to both reduce the financial costs, while not reducing the quality of healthcare. Stroke demonstrates effects to both lifestyle and medical risk factors. Lifestyle risk factors include obesity, heavy drinking, and the use of illegal drugs like cocaine and methamphetamine [11]. Conversely, medical risk factors include high blood pressure, high cholesterol, and cardiovascular disease [11]. Thus, lifestyle risk factors can be moderated, while medical risk factors cannot be. The elderly is the most at-risk demographic as stroke most commonly affects people between 75 and 85 [11]. Although through vital research it has been determined that in the Caucasian demographic that men are more at risk than women, many demographics are under represented.

In a 2020 stroke study conducted by the Department of Neurology at Miller School of Medicine of the University of Miami on the elderly demographic, different races of people were analyzed [11]. The Black and Hispanic populations had the greatest chance of having a stroke, and it was determined that this was due to a gap in education, status of insurance, and a common low socioeconomic status [11]. This study showed that it is necessary to reduce disparities regarding socioeconomic status. There is still a need to identify and learn about different risk factors of stroke and how they affect different demographics differently. To illustrate, of the studies compiled from PubMed and Web of Science the majority predicted mortality rates using previously identified risk factors [7]. Although mortality rates are important to monitor using identified risk factors, it is important for there to be more representation for analysis of unidentified risk factors that still could have a significant effect on stroke detection: occupation, marriage status, residential status.

Of the 700,000 people that have a stroke each year 500,000 have their first stroke. This means that 71.43% of people are experiencing strokes for the first time [9]. The need for an accurate method to assess patients' pre-existing conditions and family history accurately is vital to prevent strokes before they happen for the first time. Regarding second strokes, 80% have been demonstrated to be preventable by using medication, maintaining a strict diet, and engaging in physical activity [10]. If strokes are recognized early it can lead to better outcomes with treatment [10]. Although medical professionals have been able to reduce the chance for a second stroke by a considerable margin, a way to prevent patients having their first strokes is vital. Over the past decade, several applications have been developed to help assess risk factors to determine the chances of a first stroke, and every year more of the applications implement machine learning (ML) to the data analysis process. ML is the process of developing models and predictions based on observed data to determine the best action to take. The complex process is helpful in making decisions in various scenarios. Even though the use of ML has become more common in this field of research, few ML stroke studies have led to significant and reliable real-life application of the model in clinical practice [7].

This study aims to identify novel, previously unidentified, risk factors for stroke using a ML model in hopes that identifying new risk factors will lead to an increase in preventative patient care for at-risk individuals.

2. Methods

2.1. Data Collection

The dataset analyzed was a public dataset from the coding database Kaggle, also known as the world's largest data science community. The data set, "healthcare-dataset-stroke-data.csv" was analyzed based on 5110 patients with different data for different risk factors related to stroke (Table 1). The data set focused on the features (risk-factors) 'gender', 'age',

'hypertension', 'heart disease', 'ever_married', 'work_type', 'residence_type', 'avg_glucose_level', 'bmi', and 'smoking_status' (Fedesoriano, 2021). These features are a mix of both identified and unidentified risk factors that led to stroke. The goal was to observe how much the unidentified risk factors: 'ever_married', 'work_type', 'residence_type' affected the data analysis process.

	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke	Private	Self-employed	Govt_job	formerly_smoked	never_smoked	smokes	Unknown
0	1	67.0	0	1	1	1	228.69	36.6	1	1	0	0	1	0	0	0
2	1	80.0	0	1	1	0	105.92	32.5	1	1	0	0	0	1	0	0
3	0	49.0	0	0	1	1	171.23	34.4	1	1	0	0	0	0	1	0
4	0	79.0	1	0	1	0	174.12	24.0	1	0	1	0	0	1	0	0
5	1	81.0	0	0	1	1	186.21	29.0	1	1	0	0	1	0	0	0

Table 1. healthcare-dataset-stroke-data.csv" data set, this is the data after preprocessing, one-hot encoding has been applied, there are 15 features

2.2. Patient Inclusion/Exclusion Criteria

The preprocessing method reduced the number of patients, so that there would be no discontinuities when the data is analyzed using the ML based model. The process began with all patients with any data missing being removed since that would cause unwanted problems in the data analysis stage that would lead to many errors. All patients younger than 38 were removed as well because there were only three recorded patients that had a stroke younger than the age of 38, and those outliers would skew and be negatively impactful towards the accuracy of the model.

2.3. Data Preprocessing

After basic preprocessing, the method one-hot encoding was applied to the data set. One-hot encoding is a method used to ready data for data analysis by changing categorical labels into numerical labels. () For instance, the smoking_status feature has four categorical labels associated with it: formerly_smoked, never_smoked, smokes, and unknown. In order to make this data easier to analyze, the four labels of the smoking_status feature were made into four separate features themselves. Each of the four new features were given the binary labels 0 and 1, 0 being false and 1 being true for the feature's condition. One-hot encoding is a vital step and latter step in analysis to make the data more consistent with only numerical labels (Figure 1).

Figure 1. Demonstrates the process to ready the data for analysis so there are no idiosyncrasies with it and it does not cause any problems

2.4. Split into Training and Validation Sets

The process of splitting the data set into a training and validation set is to observe how effective the trained model will be in determining the likelihood of a stroke occurring. The training set learns from applying the model to the original data received in order to predict the correct outcome, while the validation set is used to fine-tune the parameters when training the model (). In order to ensure the accuracy of the outcomes the model, the number of patients that did have a stroke and those who did not, had to be equal. The validation set was 75% the size of the training set. If the data had been split 50%-50% only half of the data would be trained, and it is likely that important id left out when training the model. With the use of the Synthetic Minority Over-sampling Technique (SMOTE) as a stratifier the data was able to be balanced. SMOTE is an over sampling technique that helps balance the number of people who had strokes compared to the number of people who did not. The use of a training and validation set will also allow for if over fitting is occurring in the model. Over fitting occurs when the model has memorized each of the points data in the training data set instead of generalizing across the test data, this is not desirable.

2.5. Scaling the Data Set

Scaling the data set is an important part of data preprocessing that helps prevent any outliers in the data set; thus, it normalizes the data by making the data more generalized to reduce the numerical difference between different data points (). This process is vital in ensuring that the model runs smoothly on the data and so the outliers do not have an impact on skewing the data, which can lead to an undesirable and inaccurate model. To measure the effectiveness of the scaling process, the standard deviation of the data set was taken next, which is an effective method in determining the amount variation between data points (). The standard deviation was 1.44 for this data set, which represents an accurate representation that the scaling process is working, and that the data is ready to be applied to a model. The scalar (.StandardScaler()) was applied to both the training set (X_train_res) and the validation set (X_val_new) (Figure 2).

```
scaler = preprocessing.StandardScaler()
scaler.fit(X_train_res)
X_train_scaled = scaler.transform(X_train_res)
X_val_scaled = scaler.transform(X_val_new)
```

Figure 2. **Code for applying the scalar to the training and validation set, to ensure that the data was more generalized**

2.6. Artificial Neural Network (ANN) model

After the processed data was split into the training and validation set, the data was ready for ANN to be applied to it. ANN is the most common process used to analyze the accuracy and validity of ML models. The ANN model is the easiest way to apply supervised machine learning algorithms and it shows how accurately a model can analyze the data. The ANN model utilizes a logistic regression curve to evaluate if a model is trained properly. It is ideal for the curve to go to zero to represent that the data has been completely understood. Logistic regression assumes the predictions to represent this functional form

$$\hat{y} = (w_1 x_1 + w_2 x_2 + \cdots + w_k x_k + b),$$

where the numbers $w_1, w_2, \ldots, w_k, b$ are trainable parameters. Logistic regression on the most basic level, models the optimal line to describe data in two different classes. That line becomes more accurate with each run by utilizing the binary cross entropy (BCE) error to minimize the error. Each time it calculates the function the BCE changes the line parameters in $y = mx + b$ to represent a better line that represents the two different classes. The loss or logistic regression model used 2 hidden layers and epochs 1000. Hidden layers are the number intermediate steps that the data takes until it is completely analyzed, and epochs are the number of times the training data goes through the model (Figure 3).

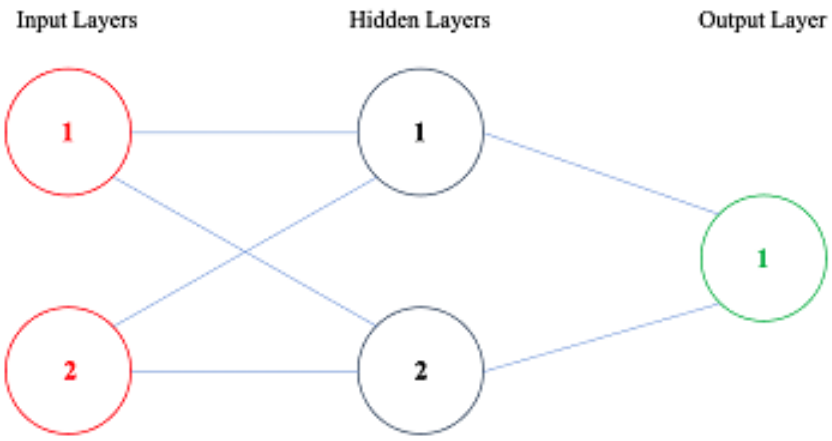

Figure 3. The structure of the artificial neural network (ANN) model and what the data was put through to receive optimal results. Two hidden layers was found to be the most effective process to give the model the most accuracy

2.7. Analysis of Dataset and Each Feature Separately

After the effective use of the ANN function and minimization of the BCE, the accuracy of the model will be possible to determine. Following a successful analysis of the data set using the ANN model, a final accuracy of above 80%, the data is ready to for each feature to be analyzed separately, and the goal of a positive result to the research question is within reach. To evaluate the accuracy of each feature separately, instead an AUC curve will be utilized, an effective analysis for accuracy method that returns quantitative data. The Select-K-Best function from skikit learn, will be applied to determine, out of the number of features chosen to be analyzed, which features are having the greatest impact on a positive stroke result. Furthermore, when the no_features (number of features) variable is set equal to three, the Select-K-Best function returns age, bmi, and avg_glucose_level to be the most effective risk factors leading to a stroke. This is understandable as these three risk factors have been previously proven as significant risk factors that need to monitor and reduced, if possible, to reduce the chance of stroke. In addition to reaffirming previous studies, this function will also allow for the effectiveness of the previously unidentified features, ever_married, work_type, and residence_type, to be measured according to this specific data set, in hopes to obtain a breakthrough that can be applied to a larger sample source and reaffirmed on a wide scale level.

2.8. AUC curve

In addition to the logistic regression model the subsets of data were also analyzed using the area under the receiver operating characteristic curve (AUC) curve. With a successful logistic regression model the application of an AUC curve was possible. It is a graph that measures the performance of the model based on sensitivity (true-positive) and specificity (true-negative) rate. Sensitivity measures the amount of people that were correctly diagnosed as positive stroke patients and looks after the number of false-negative results, and specificity represents the number of patients correctly identified as negative and refers

to the number of false-positive results. The AUC curve is used to observe how well the model has represented the data set by plotting the sensitivity vs 1 - specificity rate and comparing the area under that curve to the set threshold value. The threshold value monitors the projected probability into a label (Deepchecks, 2021). The default threshold value is 0.50, so as long as the AUC value is greater than 0.50 the model is predicting accurately and is better than a random guess.

3. Results

3.1. Analysis of Entire Data Set

Following the pre-processing of the data the original number of patients was reduced from 5110 to 2934. After analysis the data's training and validation error were 0.547 and 0.572 respectively **(Figure 4).** The errors are close in value which is desired because it denotes that no over fitting occurred in the data analysis process. Using a basic analysis model, the accuracy, precision, recall, and other values were determined after analyzing the entire data set. In the bottom left corner of the table, it can be observed that the weighted average for the entire model based on the validation set was 0.92 or 92% (Figure 4), which was beneficial, as it allowed for the next phase of the analysis process on each feature separately to develop a ranking on the extent of each features' impact on the chance of a stroke.

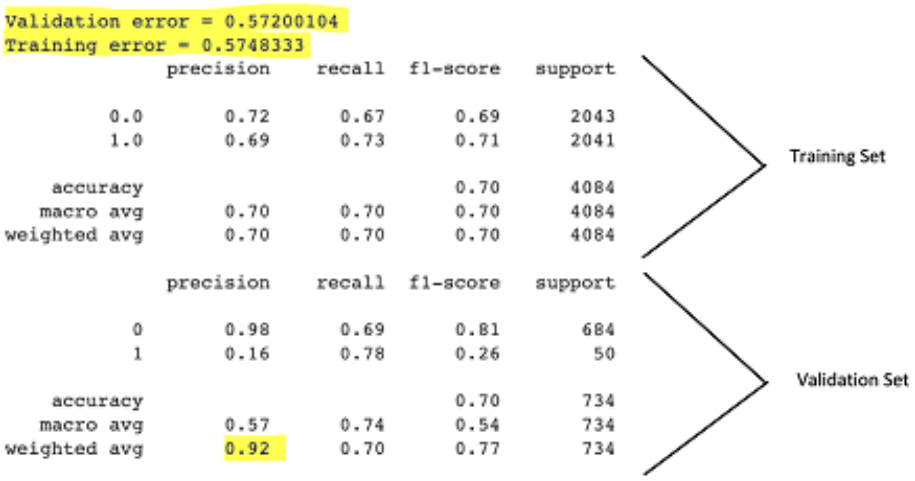

Figure 4. The basic analysis of the entire model. Validation and training set error calculated, and validation set weighted accuracy determined to be 92% (0.92 in the table)

3.2. Analysis of Each Feature Separately

Following the successful analysis of the entire model and determining that the overall accuracy of the model's stroke prediction was 92%, the use of Scikit-learn's SelectkBest function was used to analyze each feature separately, and eventually develop a ranking system for the features based on the chance of a stroke occurrence. The Selectkbest function is a pre-programmed function that determines which features cause the most impact on the return of a positive value in response to the model (Pedregosa, 2011). The function utilizes a variable, k, to represent the number of features being analyzed. The k can be changed with each run; for clarity the k variable was renamed to "no_features" (Figure 5). After the function has been applied, a ranking is created for the features analyzed based on their effect on the model returning a positive result of a stroke occurring.

```
from sklearn.datasets import load_iris
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

print(X.shape)

X_train, X_val, y_train, y_val = train_test_split(X,y,test_size=0.75,random_state=4,stratify=y)

no_features = 4

kbest = SelectKBest(chi2, k=no_features)
X_train_new = kbest.fit_transform(X_train, y_train)
X_val_new = kbest.transform(X_val)

cols = kbest.get_support(indices=True)

print(X_train.shape,X_val.shape)

print((y_train == 0).sum())
print((y_train == 1).sum())
print((y_val == 0).sum())
print((y_val == 1).sum())

print(df.columns[cols])
```

```
(2934, 15)
(733, 15) (2201, 15)
681
52
2046
155
Index(['age', 'avg_glucose_level', 'bmi', 'smokes'], dtype='object')
```

Figure 5. Application of the SelectKBest function on the data set to develop a ranking of the features. E.g. no_features = 4, according to the function the four most impactful features contributing to the chances of having a stroke are (1) 'avg_glucose_level', (2) 'age', (3) 'bmi', (4) 'smokes' in order from most to least impact

3.3. Testing the ANN Model

To ensure that the ANN model was effective on this specific data set, a basic test was run by comparing two ANN models on the same data but changing the number of epochs trained on the two models, the first with 100 epochs and the second with 1000 epochs (Figure 6). If the ANN model is effective, additional epochs, runs through the model, will increase the percent of the model trained. Only when approaching 800-1000 epochs the slope of the regression graph is more horizontal or closer to zero meaning the entire model has been trained. This observation validates the fact the ANN model is affective on the "healthcare-dataset-stroke-data.csv" data set because when more epochs are added the model trains better.

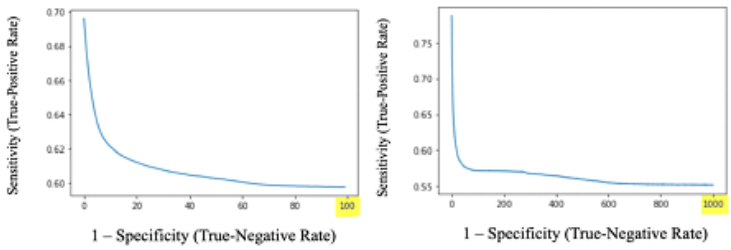

Figure 6. Comparison between 100 and 1000 epoch ANN models to assess the ANN models effectiveness on the stroke data set

3.4. Application the ANN Model to find the AUC

The AUC curves were determined for all the different number of features and compared to the original AUC value obtained when no_features = 1. This process allows for the ranking to be validated using a more quantitative method. The |Δ AUC| showed that there is fluctuation throughout the AUC curve when different numbers of features are analyzed, the values of 0.003, 0.080, and 0.062, for government job (occupation), residence type (living status), and ever married (marriage status) respectively **(Table 2).** These values represent major fluctuations compared to the control value AUC of 0.617, which shows that they have an impact on the chance of stroke. Furthermore, in this dataset occupation caused more chance of stroke than heart disease; these previously unidentified risk factors do have an effect on chance of stroke occurrence. The "–" in the table represent data that is associated with another risk factor that was broken down during one-hot encoding. There are four smoking features, but only values for the feature "smokes" was recorded because it was ranked higher by SelectKBest and the AUC for the other three smoking features would be similar.

# Features	Features Ranking	AUC Value	|Δ AUC|
1	Average Glucose Level	0.617	0
2	Age	0.693	0.076
3	BMI	0.707	0.090
4	Smokes	0.700	0.083
5	Unknown	–	–
6	Never Smoked	–	–
7	Formerly Smoked	–	–
8	Hypertension	0.606	0.011
9	Government Job	0.620	0.003
10	Heart Disease	0.529	0.088
11	Private Job	–	–
12	Residence Type	0.537	0.080
13	Gender	0.525	0.092
14	Self Employed Job	–	–
15	Ever Married	0.555	0.062

Table 2. finding the AUC and change in AUC from no_features = 1; no_features is increased by one in each calculation and compared to the original AUC to measure its effectiveness and accuracy in more quantitative manner

4. Discussion/Conclusions

With the availability of a larger data set analyzing more risk factors it will be helpful in this research. The analysis showed that this method worked to analyze stroke. These positive results will allow for further application to obtain more accurate results to expand the data analysis process. With additional and more accurate data an interface will be applied to the data to add a more interactive way for the patient to connect with their symptoms and get better treatment.

The results show significant reason to broaden the search for new risk factors as ever_married, work_type, and residence_type were significant features in determining the likelihood of stroke (Table 2). Although measuring the change in AUC is an innovative method to measure the accuracy of different features, using the AUC is not the best method because AUC does not create a bias on the size of the test data, while accuracy is dependent on the size of the data. It would be more effective to apply a new and more complex method to accurately identify features. Scikit-Learn's lasso feature selection explains this process. Lasso feature selection is a quantitative ranking approach that aims to push its values towards 0. The values after lasso feature selection that are not 0 are important to the model and should be observed; this will make identifying previously unidentified risk factors much more explicit and straightforward. The lasso function has a variable alpha that must be determined depending on the size and complexity of the data set. Like the ANN model, lasso features selection utilizes logistic regression in the form of a cost function, which is the summation of all the BCE between the all the samples. This method would be much more effective and present better qualitative results to allow for a seamless segway into future research that delves deeper

For instance, with the opportunity to work on a larger data set, the analysis process will be much more accurate and positive results will allow for more opportunities in helping the field and helping people with current and future stroke issues and complications. This ML model provides an accurate representation that analyzes risk factors for stroke, so that future preventative measure based on unidentified risk factors can be put into place and instituted for different populations.

References

[1] E. Alpaydin, *Introduction to machine learning*, 2nd Edition, (2010).

[2] A. Geron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to*

build intelligent systems, O'Reilly Media, (2019).

[3] G. Zaccone and M. R. Karim, *Deep learning with tensorFlow: Explore neural networks and build intelligent systems with python*, Packt Publishing Ltd, (2018).

[4] A. C. Muller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*, O'Reilly Media, (2016).

[5] L. P. Chen, Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, *Foundations of machine learning*, (2019).

[6] K. P. Murphy, *Machine learning: a probabilistic perspective*, MIT Press, (2012).

[7] W. Wang, M. Kiik, N. Peek, V. Curcin, I. J. Marshall, A. G. Rudd and B. Bray, *A systematic review of machine learning models for predicting outcomes of stroke with structured data*, PloS one, 15(6)(2020), e0234722.

[8] A. Khosla, Y. Cao, C. C. Y. Lin, H. K. Chiu, J. Hu and H. Lee, *An integrated machine learning approach to stroke prediction*, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, (2010), 183-192.

[9] H. Kamal, V. Lopez and S. A. Sheth, *Machine learning in acute ischemic stroke neuroimaging*, Frontiers in neurology, 9(2018), 945.

[10] R. Feng, M. Badgeley, J. Mocco and E. K. Oermann, *Deep learning guided stroke management: a review of clinical applications*, Journal of Neurointerventional Surgery, 10(4)(2018), 358-362.

[11] E. J. Lee, Y. H. Kim, N. Kim and D. W. Kang, *Deep into the brain: artificial intelligence in stroke imaging*, Journal of Stroke, 19(3)(2017), 277.

[12] H. Gardener, R. L. Sacco, T. Rundek, V. Battistella, Y. K. Cheung and M. S. Elkind, *Race and ethnic disparities in stroke incidence in the Northern Manhattan Study*, Stroke, 51(4)(2020), 1064-1069.

[13] S. Sealy-Jefferson, J. J. Wing, B. N. Sanchez, D. L. Brown, W. J. Meurer, M. A. Smith and L. D. Lisabeth, *Age-and ethnic-specific sex differences in stroke risk*, Gender Medicine, 9(2)(2012), 121-128.

[14] G. Kumar and R. Patnaik, *Inhibition of Gelatinases (MMP-2 and MMP-9) by Withania somnifera Phytochemicals Confers Neuroprotection in Stroke: An In Silico Analysis*, Interdiscip Sci Comput Life Sci., 10(2018), 722-733.

[15] Fedesoriano, *Healthcare-dataset-stroke-data.csv*, version 1, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/version/1

[16] Pedregosa, *Scikit-learn: Machine Learning in Python*, JMLR, 12(2011), 2825-2830.