INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 476-482

e-ISSN : 2583-1062

Impact Factor : 5.725

PROGNOSIS OF STROKES USING MACHINE LEARNING

Divya K[1], Evanjeline Oswald E[2], Gomathy K[3], Jancy Rani K[4]

[1,2,3]Student, Computer Science and Engineering, Agni College of Technology

[4]Assistant Professor, Computer Science and Engineering Department, Agni College of Technology

ABSTRACT

The interruption or reduction of the blood supply to the brain causes a stroke. A stroke is a condition where there is insufficient blood supply to the brain, which causes cell death. Today, it is the main cause of death in the entire world. Examining the affected individuals has shown a number of risk variables that are thought to be connected to the stroke's origin. Numerous studies have been conducted for the prediction and classification of stroke disorders using these risk variables. Machine learning and data mining methods are the foundation of the majority of the models. The stroke deprives the brain of oxygen and nutrients, perhaps leading to the death of brain cells. Numerous studies have been conducted to compare the effectiveness of predictive data mining methods in the prediction of various diseases. On the Cardiovascular Health Study dataset, we evaluate various approaches for predicting stroke with our algorithm in this article. Here, the principal component analysis algorithm is used to reduce the dimension, the decision tree algorithm is used to pick the features, and random forest algorithm is used to build a classification model. Our work offers the best predictive model for the stroke disease with 94.7% accuracy after analyzing and comparing classification efficiency with other approaches and variation models.

Key words: CSV Dataset, Artificial Intelligence, and Stroke.

1. INTRODUCTION

Stroke is one of the most dangerous diseases for anyone over the age of 65. It injures the brain in the same way as a "heart attack" injures the heart and is the third greatest cause of death in the United States and developing countries. When a stroke happens, it not only results in significant medical costs and chronic impairment, but it can also result in death. A stroke kills someone every 4 minutes, however up to 80% of strokes can be avoided if we can identify or forecast the incidence of stroke in its early stages. In general, data mining plays an important role in disease prediction in the health care industry. To extract usable knowledge from massive amounts of medical data, powerful data analysis methods are required. Medical data, in the form of patient records, are an ever-expanding stream of information generated by hospitals. However, the information included in these data represents a massive resource bank for medical research. The goal of our work is to make a medical decision, which is a highly specialized and difficult task owing to a variety of circumstances, particularly in the case of diseases with similar symptoms or unusual conditions. It is a significant area of Artificial Intelligence (AI) in medicine. An AI system would analyze the patient's data and recommend a set of suitable predictions. The system can extract hidden knowledge from a history clinical database and predict individuals with disease using medical profiles such as age, blood pressure, blood sugar, and so on. Classification algorithms with a large number of attributes are used to predict disease. The AI system could answer complicated queries, each with its own strengths in model interpretation, access to extensive information, and accuracy. Following a review of various related and unrelated research studies, it is discovered that neural networks provide superior classification accuracy than other classification algorithms.Our system employs the following modules: (i) data collecting (ii) preprocessing (iii) feature extraction and selection (iv) dimension reduction (v)classification (vi) result analysis. Published scientific work is also a huge help in recognizing one's own potential. Now, we'll go over the tried-and-true methods for publishing a research paper in a journal. We obtained 5,110 samples in this initial module, 2,115 of which are male and 2,994 of which are female. The data was then put into a Jupyter notebook by connecting the colab with drive. The dataset has two sections: training and testing. Once the input dataset containing training and testing sets has been loaded, Preprocessing is the next step that is conducted to remove undesirable noise and distortions from the data. The Pandas library includes a complete collection of data preparation functions. Handling missing values, handling duplicates, normalization, scaling, encoding categorical variables, and feature engineering are some of the most significant strategies. The preprocessed data will then be returned to the following module, which is feature extraction and selection.The preprocessed data will be fed into another module in the third module where the model will be trained. The correlation coefficient is used for feature extraction and selection. The Random Forest technique will be used as the classifier since it is simple to develop, resilient to noisy training data, and effective when training data is huge. As a result, it is regarded as one of the finest classifiers for classifying strokes. Once the classification module has been completed, the next stage is to analyze the results, which entails using several morphological techniques to analyze the data.The final model analyses the results using a

confusion matrix. It produces better results than other procedures. Because one class has more data instances than the others, a model may predict the majority class in all situations and have a high accuracy score even when it is not predicting the minority classes. Confusion matrices come very handy here. A confusion matrix is a table pattern that helps visualize the different outcomes of a classification problem's forecast and results. It generates a table containing all of a classifier's predicted and actual values.

2. RELATED WORK

In this Section, we looked at a few studies that demonstrate how deep learning is linked to the Strokes Analysis system.

A. The author develops a model for predicting thromboembolic stroke, using an Artificial Neural Network to enhance existing diagnosis methodologies. The ANN design was trained using the back-propagation technique, which was then tested for various types of stroke sickness. According to this study, ANN-based prediction of stroke disease improves diagnosis accuracy by 89%.

B. Govindarajan et al. classified the stroke disease using Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, Logistic Regression, and ensemble approaches (Bagging and Boosting). In their gathered dataset, 91.52% of patients suffered from ischemic stroke, whereas just 8.48% suffered from hemorrhagic stroke. Among the algorithms discussed, Artificial Neural Networks with stochastic gradient descent learning algorithm had the highest accuracy for classifying stroke, with 95.3%.

C. Singh and Choudhary created a stroke prediction model using an Artificial Neural Network (ANN). The completed dataset has 357 properties and 1824 entities, with 212 stroke occurrences. The C4.5 decision tree approach was utilised for feature selection, and Principle Component Analysis (PCA) was applied for dimension reduction. They used the Back Propagation learning approach in the ANN implementation. They obtained 95%, 95.2%, and 97.7% accuracy for the three datasets, respectively.

D. Adam et al. created a decision tree method and knearest neighbour (k-NN) classification model for ischemic stroke. Their dataset, which is the first for ischemia illness in Sudan, was gathered from different hospitals and medical centres in Sudan. It includes 15 features as well as information on 400 patients. The experiment findings reveal that the performance of the decision tree classification algorithm is superior to the performance of the k-NN algorithm.

E. For stroke classification, Sudha et al. used the Decision Tree, Bayesian Classifier, and Neural Network [8]. Their dataset is made up of 1000 records. Dimensionality was reduced using the PCA technique. In ten rounds of each algorithm, they achieved 92%, 91%, and 94% accuracy in the Neural Network, Naive Bayes classifier, and Decision tree algorithms, respectively.

F. The author developed a categorization method to assist surgeons in determining where to send patients after surgery. This is referred to as post-operative decision-making. The author advocated support vector machines and Artificial Neural Networks as two categorization methods to assist clinicians in making critical decisions. Their proposed strategies achieved SVM and ANN classifier accuracy of 88.5417% and 82.8125%, respectively.

G. The author proposed a methodology for predicting stroke risk based on demographic data. First, demographic data was established and collected from Thailand's Faculty of Physical Therapy at Mahidol University. Detecting stroke takes time and is challenging for medical personnel. As a result, an automated technique based on patient demographic data for predicting stroke symptoms is required. Demographic information on patients, such as gender, age, and education.

3. DATASETS

According to the World Health Organization (WHO), stroke is the world's second biggest cause of death, accounting for around 11% of all deaths. Based on input variables like gender, age, a variety of diseases, and smoking status, this dataset forecasts the likelihood that a patient will experience a stroke. Each row of data contains essential information about the patient. The collection contains 5,110 samples, 2,115 of which are male and 2,994 of which are female.

Attribute Details

- id: a distinct identification
- Gender: "Male", "Female", "Other"
- Age: the patient's age.
- Hypertension: If the patient has it, they receive a score of 1, else they receive a score of 0.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 04, April 2023, pp : 476-482

e-ISSN : 2583-1062

Impact Factor : 5.725

- heart_disease: 0 if the patient does not have any heart illnesses, 1 if the patient does.
- married_ever: "No" or "Yes"
- job_type: "children", "Govt_jov", "Never_worked", "Private", or "Self-employed"
- Type of residence: "Rural" or "Urban"
- avg_glucose_level: blood glucose level average
- BMI stands for body mass index.
- smoking_status: "used to smoke," "don't smoke," "smokes," or "Unknown"
- stroke: If the patient suffered a stroke, the value is 1, otherwise it is 0.

4. PROPOSED SYSTEM

The three main modules of the proposed stroke analysis system are preprocessing, classification, and result analysis. There are 5,110 samples total in the collection, 2,115 of which are male and 2,994 of which are female. To pick features, one uses the correlation coefficient. to collect a collection of fundamental variables in order to create better classification features. The goal is to develop a machine that can learn and behave like the human brain. In contrast to traditional approaches, we used a Decision Tree classifier with one-hot encoding. To increase forecast accuracy, this approach can be used in hospitals. The most crucial variables in the training dataset are the top few nodes on which the decision tree is split, and feature selection is carried out automatically. For problems involving categorization and forecasting, this algorithm is frequently employed. The proposed strategy performs better in this instance in terms of accuracy than the other two methods. Finally, 94% validation accuracy is discovered. The proposed system concentrates on increasing the system's effectiveness and accuracy in comparison to other systems and aids in the development of a resilient system.

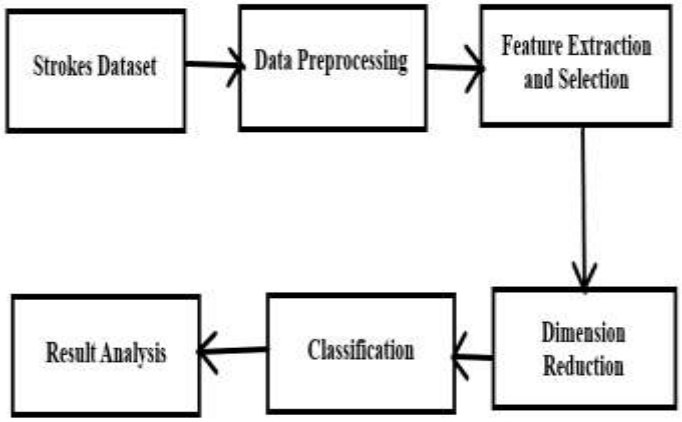

Fig1: FLOW DIAGRAM OF THE PROPOSED SYSTEM

5. MODULES-

Modules include:

a. Input

b. Preprocessing

c. Feature extraction and selection

d. Dimension reduction

e. Classification

f. Output

a. INPUT MODULE-

Based on input variables like gender, age, a variety of diseases, and smoking status, this dataset forecasts the likelihood that a patient will experience a stroke. Important details about the patient are included in each row of data. There are 5,110 samples total in the collection, 2,115 of which are male and 2,994 of which are female.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 476-482

e-ISSN : 2583-1062

Impact Factor : 5.725

b. PREPROCESSING MODULE

The input image needed to be obtained, and then it needs to be preprocessed to make it bigger and remove noise. The following phase is preprocessing, which is finished to remove unwelcome noise and distortions from the data. A well-liked Python package for analysing and manipulating data is called Pandas. It offers tools for data cleansing, aggregation, and transformation as well as data structures for effectively storing and handling massive datasets. the procedures for deleting or adding missing data pieces. For instance, the fillna() method fills in any missing data with a provided value whereas the dropna() method drops any rows or columns that contain them.

```
# Filling NaN Values in BMI feature using mean:
dataset['bmi'] = dataset['bmi'].fillna(dataset['bmi'].median())
```

```
# After filling Missing (NaN) Values in BMI feature:
dataset.isnull().sum()
```

Fig2: PREPROCESSING DATA

c. FEATURE EXTRACTION AND SELECTION MODULE

The preprocessing step, when the undesired sounds and distortions from the data are removed, is followed by feature extraction and selection. Now, only the necessary features and components have been removed from the data to make classification easier. Following feature extraction, the extracted features required to be chosen according to a set of standards. The feature selection process employs the correlation technique. We can forecast one variable based on another through correlation. Good variables have a strong correlation with the aim, which is the justification for utilizing correlation in feature selection. Variables should also be uncorrelated among themselves but correlated with the aim.If two features are associated, the model only needs one, as the second does not offer extra information. Once selection is complete, it is then transferred to the following module for classification and processing.

Fig3: CORRELATION USING HEATMAP

d. DIMENSION REDUCTIOND

imensionality refers to how many input features, variables, or columns are present in a given dataset, while dimensionality reduction refers to the process of reducing these features. In many circumstances, a dataset has a vast number of input features, which complicates the process of predictive modelling. For training datasets with a large number of features, it is extremely challenging to visualize or predict the future; hence, dimensionality reduction techniques must be used. Because the random forest algorithm only accepts numerical variables, we must use hot encoding to transform the input data into numeric data.

```
# One Hot Encoding:
X = pd.get_dummies(X, drop_first=True)
```

Fig4: ONE HOT ENCODING

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 476-482

e-ISSN : 2583-1062

Impact Factor : 5.725

e. CLASSIFICATION MODULE

The preprocessed data will be fed into a different module in the fourth module, where we will train the model. The Random Forest algorithm is the classifier that we're going to utilize because it's easy to develop, resilient to noisy training data, and efficient with plenty of training data. As a result, it is regarded as one of the best classifiers for predicting strokes. Some decision trees may predict the correct output, while others may not, because the random forest combines numerous trees to forecast the class of the dataset. However, when all of the trees are joined, they predict the correct outcome. As a result, the following two assumptions for a better Random forest classifier: There should be some actual values in the dataset for the dataset's feature variable to predict true outcomes rather than an imagined consequence. The forecasts of each tree must have extremely low correlations. In comparison to other algorithms, it requires less training time. Even with the enormous dataset, it operates effectively and predicts the outcome with a high degree of accuracy. When a significant amount of the data is absent, accuracy can still be maintained.

```
# RandomForestClassifier:
from sklearn.ensemble import RandomForestClassifier
RandomForest = RandomForestClassifier()
RandomForest = RandomForest.fit(X_train,y_train)

# Predictions:
y_pred = RandomForest.predict(X_test)

# Performance:
print('Accuracy:', accuracy_score(y_test,y_pred))
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

```
Accuracy: 0.9461839530332681
[[966   2]
 [ 53   1]]
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       968
           1       0.33      0.02      0.04        54

    accuracy                           0.95      1022
   macro avg       0.64      0.51      0.50      1022
weighted avg       0.92      0.95      0.92      1022
```

Fig5: CLASSIFICATION

f. RESULT ANALYSIS MODULE

The final model uses a confusion matrix to examine the data. Compared to other methods, it yields better results. A model may predict the majority class in all circumstances and have a high accuracy score even when it is not predicting the minority classes because one class has more data instances than the others. Confusion matrices are quite helpful in this case. A confusion matrix is a table design that aids in visualizing the many outcomes of the forecast and outcomes of a classification task. It creates a table with all of the predicted and actual values from a classifier. Here, the correctness of training and validation will be taken into account.

Fig6: CONFUSION MATRIX

```
[[966   2]
 [ 53   1]]
```

Fig6: CONFUSION MATRIX

6. CONCLUSION AND FUTURE WORKS

The significant aspect of our system is that we used the Random Forest model, which is used not only for the classification process but also for the dimension reduction process. In this paper, we suggested an automated strokes analysis method that will be beneficial in many places. The potential benefits of the future work are its versatility, cost effectiveness, and speed of execution. Research and analysis show that the suggested approach is a useful method for doctors to diagnose strokes. Future development should consider adding more feature specifics. Building more adaptive models for other forms will be interesting to pursue along the same lines of research as those presented here. The identification of mild stroke symptoms is another prospective study direction that can be explored.

7. REFERENCES

[1] Feigin V. L., Forouzanfar M. H., Krishnamurthi R., et al., "Global and Regional Burden of Stroke During 1990-2010: Results from the Global Burden of Disease Study 2010," doi: 10.1016/S0140-6736(13)61953-4. [No cost PMC paper] [PubMed] [CrossRef] Google Scholar usage

[2] C. O. Johnson, G. A. Roth, G. A. Mensah, et al. An update on the global burden of cardiovascular illnesses and risk factors from 1990 to 2019 according to the GBD 2019 research. Journal of the American College of Cardiology, 2020;76(25):2982–2921. Cite this article as 10.1016/j.jacc.2020.11.010. [No cost PMC paper] [PubMed] [CrossRef] Google Scholar usage

[3] A thorough investigation of the techniques for using exhaled breath (EB) to identify disorders in the human body was conducted by Kaur M., Patil D. D., and Mule N. M. 2021;26, paper 100715 10.1016/j.imu.2021.100715, kindly. Unlocking Medical Informatics. [CrossRef] Google Scholar is used.

[4] Wang Y., Zhang L., Sun W., et al. Clinical outcome and mortality of stroke patients with coronavirus infection in Wuhan, China in 2019. China Stroke 2020;51(9):2674-2682. 10.1161/STROKEAHA.120.030642 is the URL to cite. [No cost PMC paper] [PubMed] [CrossRef] Google Scholar is used.

[5] Automated epileptic seizure waveform detection method based on the characteristic of the mean slope of wavelet coefficient counts utilising a hidden Markov model and EEG signals. Lee M., Ryu J., Kim D. The ETRI Journal, 2020;42(2):217-229. Please use this citation: 10.4218/etrij.2018-0118. [CrossRef] Google Scholar is used.

[6] Using deep convolutional neural networks and EEG, Adeli H., Subha D. P., Tan J. H., Acharya U. R., Oh S. L., and Hagiwara Y. automated depression identification. 2018;161:103–113, Biomedical Computer Methods and Programmes. Citation needed: 10.1016/j.cmpb.2018.04.012. [CrossRef] and [PubMed] Google Scholar is used.

[7] Shin S. B., Kwon Y. H., and Kim S. D. an electroencephalography-based two-dimensional (2D) convolution neural network (CNN) system for mood recognition. Sensors: 2018;18(5):p. 1383. Doi: 10.3390/s18051383. [No cost PMC paper] [PubMed] [CrossRef] Google Scholar is used.

[8] Leahy R. M., Haldar J. P., Kim B., Schweighofer N. Corticospinal tract microstructure in chronic stroke predicts improvements in distal limb motor function. Journal of Neurologic Physical Therapy 2021;45(4):273-281. 10.1097/NPT.0000000000000363 is the citation. [No cost PMC paper] Cross Reference and PubMed Google Scholar is used.

[9] Rezal, M., Badri, C., Wijaya, S., and Adhi, H. A. Scaling exponent electroencephalograms can be used to automatically diagnose ischemic stroke using an extreme learning machine. 2017;820:12005-12013. Conference Series in the Journal of Physics, DOI: 10.1088/1742-6596/820/1/012005.

[10] Ramadhan R. I., Mandasari M. I., Djamal E. C., and Djajasasmita D. The post-stroke EEG data is recognised using wavelet and convolutional neural networks. Bulletin of Electrical Engineering and Informatics, 2020;9(5):1890–1898; 10.11591/eei.v9i5.2005 is the identifier. [CrossRef] Google Scholar is used.

[11] Thrombophilia testing in young patients with ischemic stroke," Thrombosis research, vol. 137, pp. 108–112, 2016. S. H. Pahus, A. T. Hansen, and A.-M. Hvas.

[12] Classification of stroke disease using machine learning algorithms," Neural Computing and Applications, pp. 1–12; P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan.

[13] To err is human: establishing a better health system, vol. 6. National academy press Washington, DC, 2000. L. T. Kohn, J. Corrigan, M. S. Donaldson, et al.

[14] Stroke prediction using svm," in 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 600–602, IEEE.

[15] P. A. Sandercock, M. Niewada, and A. Czonkowska, "The international stroke trial database," Trials, vol. 13, no. 1, pp. 1-1, 2012.

[16] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in IEEE, 2017, 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), pp. 158–161.

[17] S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of ischemic stroke using machine learning algorithms," International Journal of Computing Applications, vol. 149, no. 10, pp. 26–31, 2016.

[18] [Treatment options for the suppression of inflammation in ischemic and hemorrhagic stroke, Current Neurological Opinion. 2009;22(3):294–301. Kleinig T. J., Vink R. The doi is 10.1097/WCO.0b013e32832b4db3.

[19] Laura A. Boyd, Kristopher S. Hayward, Ward, et al. The fundamental suggestions of the roundtable on stroke recovery and rehabilitation are based on consensus and pertain to biomarkers of stroke recovery. 2017;12(5):480-493; Stroke International Journal. Doi: 10.1177/1747493017714176. [Free article on PMC] .

[20] Using an image-based deep recurrent convolutional neural network, the motor imagery EEG data may be categorised into multiple classes. Ulbert I., Ibrahim Y., Wahdow M., Kollod C., and Fadel W. In 2020, Gangwon, South Korea will host the eighth international conference on brain-computer interface (BCI).

[21] EEG signal classification for BCI systems using integrated spatial and temporal dimensions and convolutional neural networks, 2020 Montreal, Quebec, Canada: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); pp. 434–437. Anwar A. M. and Eldeib A. M.

[22] A multivariate, multimodal time series classification architecture for the classification of temporal sleep stages was developed by Gramfort A., Wainrib G., Galtier M. N., Arnal P. J., and Chambon S. 2018;26(4):758-769 Neural Systems and Rehabilitation Engineering, an IEEE Transaction.

[23] Stroke biomarkers: problems and developments in diagnosis, prognosis, differentiation, and therapy Clinical Chemistry 2010;56(1):21–33 Saenger A. K., Christenson R. H. 10.1373/clinchem.2009.133801.

[24] [Jia B., Huo X., Sun D., et al. Information on endovascular therapy for acute ischemic stroke with substantial vessel obstruction according to different subtypes of stroke from the ANGEL-ACT registry. 2022;11(1):151-165. Therapy and neurology. Doi: 10.1007/s40120-021-00301-z.

[25] Y. M. Zhang, F. P. Jiang, N. H. Chen, and others. After endovascular thrombectomy, FLAIR vascular hyperintensity in individuals with acute ischemic stroke indicates early neurological damage. In the year 2022, the following DOI will be used in neurological sciences: 10.1007/s10072-021-05853-4.

.