Abstract + Introduction + Literature Review

Ashik Sathiya, Tanav Thanjavuru

*[#]College of Information Sciences & Tech.,
Pennsylvania State University, State College, PA
16801, USA*

[2]abs6607@psu.edu@psu.edu

[1]tzt5285@psu.edu

1.) ABSTRACT

Stroke has a negative impact on society, and efforts have been made to improve stroke management and diagnosis. By using technology to systematically store and analyze patients' medical records, caregivers can better manage patients. This study analyzed various risk factors in electronic health records to predict strokes. Through statistical techniques and principal component analysis, the study found that age, heart disease, average glucose level, and hypertension are the most important factors for stroke prediction. A perceptron neural network using these four factors provides the best accuracy and miss rate compared to using all available input features and other algorithms. As the dataset is imbalanced, the study used sub-sampling techniques to create a balanced dataset for analysis.

2.) INTRODUCTION

Machine learning can play a crucial role in identifying strokes in patients by analyzing various patient data such as medical history, physical symptoms, and imaging results. There are several improvements that can come out of ML in health care. Firstly is early detection. ML algorithms can analyze large volumes of patient data and detect patterns that may indicate a potential stroke. For example, an ML model could analyze a patient's blood pressure, heart rate, and other vital signs to identify warning signs of an imminent stroke. Next is risk assessment. ML algorithms can analyze patient data to determine their risk of developing a stroke. By considering various risk factors such as age, sex, medical history, and lifestyle habits, an ML model can accurately predict a patient's likelihood of suffering from a stroke. Finally is treatment planning. ML algorithms can assist doctors in determining the best treatment plan for stroke patients. By analyzing patient data such as medical history, imaging results, and other factors, an ML model can suggest appropriate treatments based on the patient's specific needs.

This paper analyzes patient records to identify key factors necessary for stroke prediction. Using a publicly available dataset, the authors use principal component analysis to reduce the dimensionality of the feature space and identify the most important risk factors. They also benchmark popular machine learning models for stroke prediction. The paper contributes to a better understanding of stroke risk factors and provides reproducible research through the availability of source code. The paper is structured into an overview of related work and the dataset, correlation analysis and feature importance analysis, principal component analysis,j data mining algorithms and their performance, and a conclusion with future work. The literature contains various studies on stroke prediction, including identifying risk factors and using machine learning models. Multiple factors can affect the results of these studies, including data collection, feature selection, data cleaning, missing value imputation, and standardization. It is important to identify how different factors in electronic health records are related to each other and their impact on stroke prediction accuracy. Identifying important features is crucial to the performance of machine learning models, and redundant or irrelevant features should be removed before using classification algorithms. Therefore, it is essential to understand the interdependence of risk factors in electronic health records and their impact on stroke prediction accuracy.

3.) RELATED WORK

The authors of the paper [1] analyzed the importance of different patient attributes in predicting the occurrence of stroke using a Learning Vector Quantization (LVQ) model. They used the varImp method from the R caret package to compute the relative feature importance, which measures the increase in the model's prediction error due to that attribute. The results show that a patient's age, presence of heart disease, patient's average glucose level, and presence of hypertension are the features with high importance in predicting the occurrence of stroke.

The authors also computed the CHADS2 score for the EHR records, which is a stroke risk score for non-valvular atrial fibrillation. The score takes into account congestive heart failure or impairment of left ventricular function, hypertension, age, diabetes, stroke or transient ischaemic attack, and history of thromboembolism. The authors found that most of the EHR observations in their dataset have a low CHADS2 score, and that the larger the CHADS2 score, the higher the occurrence of stroke cases. They also observed that most people with a CHADS2 score of 1 or 2 have a low probability of stroke, while only a small number of people with a CHADS2 score greater than 2 have a higher probability of occurrence of stroke.

The study uses Principal Component Analysis (PCA) to analyze variance in a dataset. PCA transforms a dataset into a set of linearly uncorrelated variables called principal components that extract maximum variance from the dataset. These components act as summaries of the features of the dataset, although they don't have a physical interpretation. PCA can be used for feature reduction in predictive modeling if the first few components capture most of the variance in the data. This section also explains how the results of PCA are related to predictive variables or features represented by patient attributes and individual medical health records. The section shows how the different principal components explain different underlying phenomena, which can be analyzed based on variable loadings.

The study finds that neural networks work best for stroke prediction, and using only four features, including age, heart disease, average glucose level, and hypertension, yields good accuracy of up to 80% with a low miss rate. The study also compares the results of using actual features versus principal components as inputs and finds that the former provides better accuracy and lower miss rate. Finally, the article presents the distribution of accuracy values obtained from the top four features used in the study.

4.) LITERATURE REVIEW

5.) DATASET

Analyzing electronic health records is an essential step towards building an accurate stroke prediction model. In this section, the authors use a dataset of electronic health records released by Kaggle. The dataset contains EHR records of patients, and the output response is a binary state indicating whether the patient has suffered a stroke or not. The dataset has a total of input attributes, including gender, age, hypertension, heart disease, marital status, occupation type, residence type, average glucose level, body mass index, and smoking status.

6.) FEATURES

This dataset consists of various features that provide information about patients. Each patient is uniquely identified by an id. The dataset includes the gender of the patient which can be "Male", "Female" or "Other". The age of the patient is also included in the dataset along with whether the patient has hypertension or not. If the patient has hypertension, then the corresponding value is 1; otherwise, it is 0. Similarly, the presence or absence of heart disease is represented by 1 and 0 respectively. The dataset also includes information about the patient's marital status and work type. The Residence_type feature indicates whether the patient lives in a rural or urban area. The dataset also provides information about the patient's average glucose level in blood, body mass index (bmi), and smoking status. Finally, the stroke feature is included in the dataset, where a value of 1 represents the occurrence of a stroke, and 0 indicates the absence of a stroke.

7.) EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a process in data analysis as a part of the initial exploration and visualization of data to understand its patterns, distributions, relationships, and potential outliers. The main goal of EDA is to discover insights and relationships in data that will help with data modeling. EDA is often used to examine large datasets and identify potential trends or patterns that may not be apparent through traditional statistical analysis. For the Exploratory Data Analysis on our dataset, we focused on two main visualizations. The first visualization is a heatmap, a graphical representation of a matrix that contains correlation values of a dataset. The second visualization we focused on is a histogram, a bar chart that displays the frequencies or counts of values within predefined intervals or bins, used in our research to find potential feature imbalances

Heatmap

a.) Histogram

8.) DATA PREPROCESSING
    a.) Drop Nulls
    b.) Label Encoding
    c.) RandomOverSampler
    d.)

9.) MACHINE LEARNING
    a.) Models
        i.) Ensemble Model (XGBoost)
            (1) GridSearch CV
        ii.) Neural Network
    b.) Model Validation
        i.) Confusion Matrix
    c.) Endpoint Testingr

RESULTS

STROKE PREDICTION

In this section, the authors analyze various benchmarking algorithms for predicting strokes using patient attributes. Three popular classification approaches are benchmarked: neural network (NN), decision tree (DT), and random forest (RF). The decision tree algorithm involves building a tree-like decision process with several condition tests, which is flexible and accurate, making it suitable for medical diagnosis. The random forest algorithm is flexible and easy to use, produces good results even with minimal tuning, and can provide indicators on how it assigns significance to input variables. The authors also benchmark the performance of a 2-layer shallow neural network, which is popular and competitive. The authors implement the feed-forward multi-layer perceptron model using the nnet R package.

The dataset consists of medical records with a highly unbalanced distribution of stroke and non-stroke patients. To address this, random downsampling is used to create a balanced dataset for training three machine learning models: neural network, decision tree, and random forest. A convolutional neural network (CNN) is also implemented with four hidden layers, two convolutional and two linear, on the same dataset. The performance of these models is evaluated using three cases of features: all original features, PCA-transformed data of the first two principal components, and PCA-transformed data of the first eight components. The results are based on 100 experiments and include precision, recall, F-score, accuracy, miss rate, and fall-out rate.

CONCLUSION

This paper analyzes patients' electronic health records to predict stroke. The authors systematically analyzed different features, performed feature correlation and stepwise analyses, and found that a combination of only four features might have a good contribution to stroke prediction. Principal component analysis was also conducted, and a neural network algorithm was found to work best with a particular feature combination, achieving an accuracy of % and a miss rate of %. The authors suggest that further work could be done to collect more data and integrate the electronic records dataset with background knowledge on different diseases and drugs using Semantic Web and knowledge graph technologies, which may improve the accuracy of stroke prediction models. They also plan to collect their institutional dataset for further benchmarking and perform external validation of their proposed method. The systematic analysis of different features in electronic health records can assist clinicians in effective archival of records.

REFERENCES

Sivapalan G., Nundy K., Dev S., Cardiff B., Deepu J.
ANNet: a lightweight neural network for ECG anomaly detection in IoT edge sensors
IEEE Transactions on Biomedical Circuits and Systems (2) (2022)

Koh H.C., Tan G., *et al.*
Data mining applications in healthcare
J. Healthc. Inf. Manage., 19 (2) (2011), p. 65

Yoo I., Alafaireet P., Marinov M., Pena-Hernandez K., Gopidi R., Chang J.-F., Hua L.
Data mining in healthcare and biomedicine: a survey of the literature
J. Med. Syst., 36 (4) (2012), pp. 2431-2448

Meschia J.F., Bushnell C., Boden-Albala B., Braun L.T., Bravata D.M., Chaturvedi S., Creager M.A, Eckel R.H., Elkind M.S., Fornage M., *et al.*
Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American heart association/American stroke association
Stroke, 45 (12) (2014), pp. 3754-3832

Harmsen P., Lappas G., Rosengren A., Wilhelmsen L.
Long-term risk factors for stroke: twenty-eight years of follow-up of 7457 middle-aged men in goteborg, sweden
Stroke, 37 (7) (2006), pp. 1663-1667

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

Pathan M.S., Jianbiao Z., John D., Nag A., Dev S.
Identifying stroke indicators using rough sets
IEEE Access, 8 (2020), pp. 210318-210327

Jeena R.S., Kumar S.
Stroke prediction using SVM
Proc. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (2016), pp. 600-602

Hanifa S.-M., Raja-S K.
Stroke risk prediction through non-linear support vector classification models
Int. J. Adv. Res. Comput. Sci., 1 (3) (2010)