Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Machine Learning for Predicting Ischemic Stroke

Asst. Prof. Pathanjali C
Dept. of Computer Science and Engineering
B.N.M. Institute of Technology
Visvesvaraya Technological University
Bangalore, India

Priya T
Dept. of Computer Science and Engineering
B.N.M. Institute of Technology
Visvesvaraya Technological University
Bangalore, India

Monisha G
Dept. of Computer Science and Engineering
B.N.M. Institute of Technology
Visvesvaraya Technological University
Bangalore, India

Samyuktha Bhaskar
Dept. of Computer Science and Engineering
B.N.M. Institute of Technology
Visvesvaraya Technological University
Bangalore, India

Ruchita Sudarshan K
Dept. of Computer Science and Engineering
B.N.M. Institute of Technology
Visvesvaraya Technological University
Bangalore, India

***Abstract*—The stroke rate in India is much higher than in other developing countries. A small percentage of stroke patients die immediately from the initial trauma. Some of the leading causes that eventually lead to death may be initial ischemic infarction, recurrent ischemic stroke, recurrent hemorrhagic stroke, pneumonia, coronary artery disease, pulmonary embolism, and other vascular or nonvascular causes. Studies show that application of machine learning techniques to stroke, focus on predicting the risk of having a stroke or the possibility of survival given the attributes of a patient, but not so much on the likely outcomes of patients that do survive the initial stroke attack. The evaluation and treatment of Ischemic Stroke (IS) have experienced a significant advancement over the past few years, increasingly requiring the use of neuroimaging for decision-making. Therefore, the goal of the project is to apply principles of machine learning over large existing data sets to effectively predict the most probable life threatening risks that may follow the first incident. Further refinement in these algorithms could provide immense utility in clinical settings and stroke therapy and also give us an insight into the recent developments and applications of ML in neuroimaging focusing on acute ischemic stroke and apply supervised machine learning methodologies to patient profile data.**

Keywords—Machine learning; ischemic stroke; random forest; support vector machine;

I. INTRODUCTION

Ischemic stroke occurs when an artery that leads to the brain is blocked. The brain depends on the arteries to bring fresh blood from the heart and lungs. The blood carries oxygen and nutrients required to the brain, and takes away carbon dioxide and other cellular waste. If an artery is blocked, the brain cells (neurons) cannot make sufficient energy and will eventually stop working. If the artery remains blocked for a couple of minutes, the brain cells may begin to die. Thus, immediate medical treatment is necessary. Ischemic stroke predicts the risk of having a stroke, given the attributes of a patient.

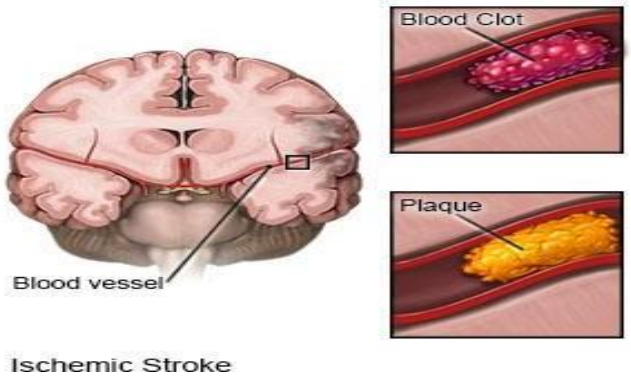

Fig. 1. Ischemic Stroke

The above Fig.1. shows when a blood vessel in the brain consists of too much plaque which clogs the brain's blood vessel. Features from different diagnoses such as high blood pressure (hypertension), heart disease, smoking, or diabetes etc., and medical interviews are used to predict the likely outcomes of fatality. Principles of machine learning are applied over large existing datasets. The prediction of outcomes in ischemic stroke patients may be useful in treatment decisions.

II. EXISITING METHODOLOGIES

A. Machine Learning for Predicting Delayed Onset Trauma Following Ischemic Stroke

In this paper, with the combination of unsupervised learning algorithms such as K-Means and supervised outcome data builds canonical profiles of the most common patients doctors are likely to encounter following initial stroke attack. Features from differential diagnoses and medical interviews are used to predict the likely outcomes of fatality. Crucial features were identified through Principal component analysis (PCA). Applying K-Means to group individuals into canonical

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

patient profiles. It will categorize the items to K-group of similarity. Cluster patient profile represents each patient with profile information. Support vector machine (SVM) used for both classification and regression challenges. Regression is used when the output variable happens to be a real or continuous value. SVM looks at data and sorts it into one of two categories and is applied to predict patient fatality outcome. New patients can be fitted into the most representative profile and plan of action will be taken to minimize chances of the most likely risks. The combination of both supervised and unsupervised we build the canonical profile of patients who are likely to encounter stroke attack. New patients can be fitted into the most representative profile and plan of action will be taken to minimize the chances of the most likely risks. Different predictions result for different datasets and techniques. Less accurate and less effective prediction result. PCA is used to reduce a large set of variables to a small set that still contains most of the information in a large set hence there is loss of data.

B. Use of Combination of PCA and ANFIS in Infarction Volume Growth Rate Prediction in Ischemic Stroke

Treatment of stroke using a procedure called Decompressive Hemicraniectomy requires the patient to undergo multiple CT scans in order to determine the size of the stroke affected area, also known as the infarction volume. A system was proposed that is able to predict the infarction volume growth rate based on only one CT scan and several clinical measurements. The main problem addressed in this is related to the prediction of infarction volume at specific time. The proposed technique applies a combination of Principal Component Analysis (PCA) and Adaptive Neuro-Fuzzy 0% Plagiarized 100% Unique Inference System (ANFIS) and has proven to perform better in predicting the infarction volume. Adaptive Neuro-Fuzzy Inference System (ANFIS), an algorithm that combines Artificial Neural Network (ANN) and Fuzzy Inference System (FIS) to investigate infarction growth pattern and use that to predict infarction growth rate and infarction volume at a particular instance for stroke patients that had large vessel occlusion in their anterior circulation. Principal Component Analysis (PCA) is a feature reduction process that reduces the dimensionality of a dataset with minimal loss of information. The proposed solution performs feature reduction to predict the second infarction growth rate from the reduced data set which shows better results. It has low accuracy and efficiency.

C. Fuzzy Data to Crisp Estimates: Helping the Neurosurgeon Making Better Treatment Choices for Stroke Patients

A strategy has been presented that utilizes the Infraction growth rate (IGR) as the key element in defining the infraction volume reaching critical levels such that a surgery is inevitable within 48 hours. This volume defines some of the sensitive treatment decisions that the neurosurgeon has to make. ML platform is used for mapping of infarction volume which is based on Adaptive Neuro-Fuzzy Inference System (ANFIS). ANFIS is composed of the artificial neural network in the first stage that inputs the data and combines them in clusters based on their similarities. These clusters are connected with the help of fuzzy rules to generate specific decision surfaces which calculate specific output. Framework for predicting the IGR from clinical data obtained from the monitoring devices that are CT scan using open source. It identifies the factors that affect the behavior of the infarction volume and gives 90% accuracy. It's a less efficient model because the dataset is comprised of 130 patients.

D. Stroke Prediction using Artificial Intelligence

The aim was to take a Medical decision which is a highly specialized and challenging job due to various factors, especially in the case of diseases that show similar symptoms, or regarding rare diseases. Different methods are compared with the approach for stroke prediction on the Cardiovascular Health Study (CHS) dataset. It is a major topic of Artificial Intelligence (AI) in medicine. An AI system would take the patients data and propose a set of appropriate predictions. Here, decision tree algorithm is used for feature selection process, the principle component analysis algorithm is used for reducing the dimension and it determines the attributes involving more towards the prediction of stroke disease and also adopted back propagation neural network classification algorithm is used to construct a classification model. Neural network classification is a crude electronic network of neurons based on the neural structure of the brain. Neural Network gives better classification accuracy than decision tree and Naive Bayes classification algorithms. CHS dataset is very challenging to use effectively due to a significant fraction of missing values and a large number of features in the dataset.

E. Stroke Prediction Context-Aware Health Care System

Ubiquitous computing connects a user to his environment, being present everywhere and in real-time using many kinds of devices such as smart phones and sensors. It is an emergent technology of both knowledge and information dissemination. Its main purpose is to adapt intelligently to the user's context. The user can query this intelligent system which instantly responds according to several parameters (context).It proposes a prediction framework based on ontology and Bayesian Belief Networks (BBN) to support medical teams in everyday life. A stroke Prediction System (SPS) is proposed. SPS is a new software component that handles the uncertainty of having a stroke disease by determining the risk score level. SPS senses, collects, and analyses data of a patient, then uses wearable sensors and the mobile application to interact with the patient and staff. When the risk reaches a certain critical limit, SPS notifies all concerned parties; the patient, the doctor, and the emergency department. Bayesian networks are used to deal with inference and to handle uncertainty. A Bayesian model is designed and implemented using the Netica tool for better efficiency i) by handling patient context remotely and verifying its changes locally and ii) on predicting the missing probabilities and calculates the probability of high risk level for emergency cases. Context aware systems provide a clear improvement in medical monitoring and decision making systems. Medical context-aware systems cannot always identify the current context precisely.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

F. Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy

The objective is to predict the outcome of endovascular intervention in acute anterior circulation ischemic stroke. It predicts from a set of variables. Implemented using supervised learning algorithms like Artificial Neural Network (ANN) and Support vector machine (SVM). ANN is inter-connected group of nodes inspired by simplification of neurons in a brain. SVM is used for both classification and regression challenges. Network is reduced to a certain value of error that the training has been completed. The value is generated. Logistic regression model allows for the identification and validation of predictive variables. Regression problem is applied when the output variable is a real or continuous value. Artificial neural network and support vector algorithms are applied to design a supervised machine capable of classifying these predictors into potential good and poor outcomes. The algorithms were trained, validated and tested using randomly divided data. A set of criteria identifying those patients who may benefit from intervention, whilst avoiding potential unwanted catastrophic treatment related complications. It proves the effectiveness of the invasive stroke treatments. Having fault tolerance corruption of one or more cells does not prevent it from generating output. Appropriate network structure is determined through experience and trial and error. Network is reduced to a certain value of error that the training has been completed. This value does not give optimal results. Requires large training datasets to improve their performance.

III. PROPOSED METHODOLOGY

The proposed strategy focuses on a novel machine learning procedures for Ischemic Stroke prediction, thus overcoming the existing problem. Different machine learning methods may not perform equally on the same feature set. Therefore, optimal feature sets for each machine learning methods were defined systematically. By utilizing Random Forest (RF) and Support Vector Machine (SVM) algorithms, the model can be used in order to increase the performance and accuracy. MATLAB was used to develop the proposed system. The proposed model describes the approach taken to develop the proposed solution which entails pre-processing, feature reduction and the final classification. The collected dataset is given as the input which is pre-processed to remove unwanted rows and columns to produce modified dataset. The data modified is analyzed and is compared to people with stroke and without stroke and the dataset is split into train and test data. The test data is used to get the predictions to produce accuracy report as shown in the figure below.

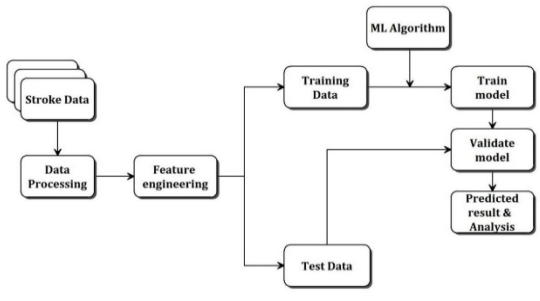

Fig. 2. Proposed Model

Firstly, Random Forest can be used to solve both classification as well as regression problems and tasks. It is designed with a combination of multiple decision trees, and it enables random selection for variables in model construction, which makes it more powerful for dealing with over-fitting in prediction. Larger the number of trees, more accurate is the results. Hence, we are providing a large dataset for producing accurate predictions. It can automatically handle missing values and can be modeled for categorical value. No feature scaling (standardization and normalization) is required in case of Random Forest as it uses rule based approach instead of distance calculation. Features such as age, sex, blood pressure becomes important in determining patient outcome. For Random Forest, over the minimum number of samples per leaf, the minimum number of samples per split and the number of estimators was searched. Another algorithm applied is Support Vector Machine (SVM) which is used for both classification and regression challenges. Support vector machine is a maximum interval classification method. It can obtain global optimal solutions with the assurance of its perfect theoretical basis and structural risk minimization criteria. In addition, SVM can handle complicated non-linear classification issues through kernel functions efficiently, so it is also a widely used classical disease prediction method. It looks at the data and sorts it into one of the two categories i.e., patients with stroke and patient without stroke. It is applied to predict patient fatality outcome. Support vector algorithm is applied to design a supervised machine capable of classifying the predictors into potential good and poor outcomes. The algorithm is trained, validated and tested using randomly divided data. Conceptually, SVM optimizes the margin between the positive and negative examples. It can formulate the stroke prediction problem as predicting the occurrence of stroke over a pre-defined time frame, which makes it a binary classification problem that fits into the framework of SVM. 10-fold cross-validation was applied for model derivation and validation. Among the features, age, hypertension, heart disease, residence type, average glucose level, BMI and smoking status comes as a significant variable. A few of them are intuitive as well, but gender, marriage status and work status are some which can be ignored. Finally, the average value of the results is measured to obtain a more stable performance. The two algorithms are applied to the train data which trains the model. This train data evaluates and builds the model. The validated model helps in predicting the result and gives an analysis of the report. The prediction models for the risk of stroke have been helpful to guide screening and interventions and to predict stroke event. Early treatment with medication can minimize the brain damage. The proposed model helps to demonstrate that using random forest and support vector machine gives the prediction classification report. This system uses the advantages of the SVM and RF to create an efficient and an accurate report. The proposed model removes the redundancy of providing the same datasets over and over again for prediction. It provides a simple and accurate process. The design of the system can be described by using the Data Flow Diagram (DFD). This diagram can be explained at various levels. For instance the level 0 diagram would represent the high level view of the system, which is the input and output.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

A drill down to the next level would describe the system in a more detail view. The diagrams explain the flow of data in the system. The proposed model removes the redundancy of providing the same datasets over and over again for prediction. It provides a simple and accurate process. Support Vector Machines and Random Forest is applied to predict the accuracy outcome respectively in classifying different death outcomes following initial ischemic trauma, using crucial features. The accuracy report carries immense clinical utility in improving a patient's chance of survival and quality of life.

IV. RESULTS COMPARISION

Supervised machine learning algorithms were applied to the developed predictive models, which are SVM and RF. 10-fold cross-validation with machine learning algorithms was applied to improve results. Using 10-fold cross-validation, 30% is used for the testing data, and 70% is used for the training Data. Accuracy performance measure was used to evaluate the performance of classification models. For evaluating the performance of the model, confusion matrix is used to calculate the accuracy. Confusion matrix describes the performance of a model on a set of test data. It gives two types of correct predictions and two types of incorrect predictions for the classifier. Early signs of potential stroke are important to be noticed because it is life-threatening. It could improve the patient's life expectancy and health condition. A supervised algorithm known as Support Vector Machine (SVM) along with random forest algorithm was used to develop the model of stroke prediction. The result is analyzed using confusion matrix and the accuracy is calculated using the following rule:

Accuracy = (TP+TN) / (TP+TN+FP+FN) Where, TP = Actual results is true and obtained results is also true.

FP = Actual result is false and obtained result is also false.

TN = Actual result is true and obtained result is false.

FN = Actual result is false and obtained result is true.

Below table shows the experimental result of different methods and proposed method applied on our selected dataset.

TABLE I. RESULT COMPARISION

Dataset	Support Vector Machine	Random Forest Classifier
Stroke Dataset	98.450167	98.442219

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization," or "Magnetization, M," not just "M." If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization (A (m(1)," not just "A/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K."

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g." Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.