Review on Stroke Prediction Mobile App using Machine Learning Model

Abhishek Khot[1], Prof. Megha Patil[2]
[1]*Author, BE final year student at Bharati Vidyapeeth Collage of Engineering, Lavale, Pune*
[2]*Author, Assistant Prof. at Bharati Vidyapeeth Collage of Engineering, Lavale, Pune*

Abstract -Stroke is the third largest leading cause of death and the principal cause of serious long-term disability in the United States. According to the world health organization (WHO) stroke is the second largest cause of death globally responsible for approximately 11% of the total death. Accurate prediction of stroke is highly valuable for early intervention and treatment.

A stroke is caused when blood flow to part of the brain is stopped abruptly. Without the blood supply, the brain cells gradually die, and disability occurs depending on the aera of the brain affected. Early recognition of the symptoms can significantly carry valuable information for the prediction of stroke and promoting a healthy life. Stroke is third largest leading cause of death and the principal cause of serious long-term disability in the United States. According to the world health organization (WHO) stroke is the second largest cause of death globally responsible for approximately 11% of the total deaths. Accurate prediction of the stroke is highly valuable for early intervention and treatment. In research work with help of the machine learning, several models are developed and evaluated to design a robust framework for the long-term risk prediction of stroke occurrence further this machine learning models can be used in backend for developing mobile applications so that real time monitoring of the health could be done using some sensors and input data from user.

There are numbers of mobile application exits which detect chances of any person getting stroke using some health-related information. Most of the mobile application uses machine learning algorithm in backend to predict the chances of getting stroke by proving the health-related data as input to the algorithms.

In this paper we are going to look at some of the previous mobile application and machine learning algorithm which are used for predicting chances of getting stroke. We will look at some important factors that matters the most for accurate prediction of stroke and how can we improve machine learning algorithm to make accurate prediction of stroke also some further work related to using some external monitoring devices to get real-time input data such as wearable devices sensors to get real-time heart rate and blood oxygen level so that we can improve machine learning model accuracy and efficiency.

LITERATURE REVIEW

Year : 2021
Author: Gangavarapu Sailasya, Gorli L Aruna Kumari.
Source: International journal for advanced computer Science and application
Title: Analyze the performance of stroke Prediction machine learning model.
Abstract : Paper suggests the implementation of Various Machine learning algorithm on the dataset taken and applied Random Forest algorithm for predictions.

Year : 2021
Author: Soumyabrata Dev, Nishtha Jain
Title : A predictive analytical approach for stroke Prediction using machine learning And Neural Networks.
Abstract : Authors collected patient dataset and Did Analysis of Dataset attribute in electronic record for prediction.
How to choose relevant features from dataset according to the algorithm to be developed. Also, how can we use Artificial Neural Network for stroke Prediction in more generalized way.

Year : 2022
Author : Dr. Harish B.G, Noorul Huda Khanum
Source: International journal of creative research Thoughts.
Title : Stroke prediction using (machine learning) Logistic Regression Model
Abstract : Focused on using Logistic Regression Model for stroke risk prediction and Comparing accuracy with Decision Tree Models accuracy. How different Factors affects the accuracy of model.

And features from dataset which are Important for generalized model for Stroke risk prediction using machine Learning.

Year : 2022
Author: Maria Trigka, Elias Dritsas
Source : Sensors Journal
Title : Stroke Risk Prediction with Machine Learning Techniques.
Abstract : Shown two models for stroke risk Prediction and their evaluation factors comparison.
Model 1: Logistic Regression Model.
Model 2: Random Forest Model.

INTRODUCTION

In past mobile application for stroke prediction using machine learning algorithm were using Random Forest and Decision Tree Algorithms and logistic regression and K-nearest neighbors' algorithms to make predictions. Among all of these algorithms for making more precise and accurate prediction we will use logistic regression algorithm gives most accuracy upon providing accurate data.

Creating mobile application which will use a machine learning algorithm based on logistic regression to make prediction of stroke and provide user data as input to the algorithm is what we are going to do in the near future. Gathering data from user related to his personal health and converting that textual data into binary format like yes (0) or no (1) so that model can make more accurate prediction and also efficiency of the model would be more than what we used in the past. For real-time monitoring of the user health, we can do some changes in the mobile application to be able to work and take input from the sensors of the wearable devices take input such as heart rate and blood oxygen level if there could be uneven changes in this input app should take some emergency majors such a mobile application can be very efficient and will do early prediction of the getting stroke.

PROBLEMS AND SOLUTIONS

Developing a mobile application for stroke risk prediction using machine learning model.

Following is some machine learning model that are created in the past for stroke risk prediction and their accuracy rates are given below. Machine learning model for stroke prediction using naïve Bayes with accuracy of 85.6%, J48 algorithm with accuracy of 99.8% also k-nearest neighbor and random forest both with accuracy of 99.8% but all of the above models are trained and test for limited no data with lots of features considered.

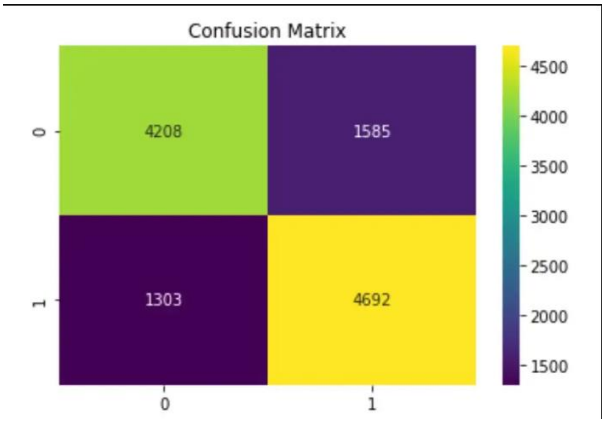
There are also some studies where researchers were trying to figure out the important factors/features that affects most on high chances of getting stroke using clustering method to classify some features in a particular group and some less important in another group basically forming different clusters according the importance of the features according to the dataset available.

Moreover, the Kaggle dataset [1] is applied in [2]. This research work suggests the implementation of various machine learning algorithms, such as logistic regression, decision tree, random forest, K-nearest neighbor, support vector machine and naive Bayes. The naive Bayes, compared to the other algorithms, achieved a better accuracy, with 82% for the prediction of stroke. The categories of support vector machine and ensemble (bagged) provided 91% accuracy, while an artificial neural network trained with the stochastic gradient descent algorithm outperformed other algorithms, with a higher classification accuracy greater than 95%. In addition, an analysis of patients' electronic health records in order to identify the impact of risk factors on stroke prediction was performed in [5]. The classification accuracy of the neural network, decision tree and random forest over 1000 experiments on the dataset of electronic health records was 75.02%, 74.31% and 74.53%, respectively.

Now lest know about the features of dataset that are begin used in the training and testing of stroke prediction machine learning model.

Age(years): This feature tells age of particular user who are over 18-year-old.
Gender: Refers to the participant's gender
Hypertension: Refers to the whether this participant has a hypertension or no
Heart disease: This feature tells us whether participants have a any kind of heart disease
Ever Married: Participants is ever married or not
Work type: Refers to the participant work status and has 4 categories (private work, government job, self-employed and never worked)
Residence Type: Refers to the residence type of participant has 2 categories (urban and rural aera)
Average Glucose level: Glucose level of participant

BMI (kg/m2): This feature refers to the body mass index of the participant
Smoking status: this feature tells participants smoking profile and has 3 categories (smokes, Never smoked, Formerly smokes)
Stoke: This feature tells us past history of participant whether they had stroke if yes how many in the past.
Let's now learn about stroke prediction using logistics regression algorithm based above mentioned features of the same data set.
What is Logistic regression is a statistical method to predict a binary outcome, such as yes or no, based on prior observations of a data set.
A logistic regression model predicts a dependent variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular collage.
Using logistic regression model provided the data and following are confusion matrix of model which tells the evaluation metric of the model.



The confusion matrix shows 4692+4208 = 8900 correct predictions and 1585+1303= 2888 incorrect ones
True positive = 4692
True Negative = 4208
False positive = 1585
False Negative = 1303
The confusion matrix shows 2360+2048 = 4408 correct predictions and 310+541= 851 incorrect ones
True positive = 2360
True Negative = 2048
False positive = 310
False Negative = 541

Following are the metrics which tells the efficiency of the machine learning model developed using logistic regression algorithm.

	Precision	Recall	f1-score	Support
0	0.76	0.73	0.74	5793
1	0.75	0.78	0.76	5995
accuracy			0.76	11788

Precision, Recall and F1 score values

From the above table we conclude following are the important features of the database.
Age, marital status, heart disease, hypertension and BMI are the important features for the prediction of stroke for given model and database consideration
Smoking is the partition parameter of the dataset we partition the dataset based on participant smoking status in 3 categories
Accuracy of this model is 84%, for getting more accuracy by this model providing the converted data as data of participant is in textual form converting at the time of training the model is what make loss in mode so converting that dataset values beforehand in binary format of yes (0) or no (1) and then provide those data values to the

CONCLUSION

As stroke is the responsible for huge number deaths which happens in single year, by creating mobile application which can predict early chances of getting stroke could reduce of amount of people which die due to untreated stroke. Machine learning application are becoming more widely used in the health care sector and there is major contribution of this models in this sector. As in case of the stroke prediction using machine learning based mobile application there are several such mobile application is already developed in the past but getting accuracy and efficiency out of them is the future need. As seen earlier in the Logistic Regression based machine learning model which accuracy of about 84% of limited no of dataset and limited no features into consideration. For getting more accuracy out of the same model, providing binary data to the model is very important also training of the model on large no of dataset which give more exposed to the model will understand better how different types of datasets containing different features are going to come in near future.

Another best machine learning algorithm that can be used for risk of stroke prediction is random forest which internally uses decision tree to make prediction which can make more accurate prediction for large no dataset for handling large no of user of different profile.

REFERENCES

[1] Stroke Prediction Dataset. Available online: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed on 25 May 2022).
[2] Sailasya, G.; Kumari, G.L.A. Analyzing the performance of stroke prediction using ML classification algorithms. Int. J. Adv. Comput. Sci. Appl. 2021, 12, 539–545.
[3] Logistic Regression https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression
[4] Stroke prediction using machine learning by international journal of creative research thoughts
[5] Nwosu, C.S.; Dev, S.; Bhardwaj, P.; Veeravalli, B.; John, D. Predicting stroke from electronic health records. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5704–5707.
[6] Medium article about logistic regression in stroke prediction machine learning algorithm https://medium.datadriveninvestor.com/stroke-prediction-using-logistic-regression-72f62edf4792
[7] Stroke prediction using machine learning methods – by Saumya Gupta, Supriya Raheja IEEE journal, 27-28 january 2022
[8] Predicting Risk of Stroke from Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models by National Library Medicine
[9] Analyzing the Performance of Stroke Prediction using ML Classification Algorithms by (IJACSA) International Journal of Advanced Computer Science and Applications, Gangavarapu Sailasya1, Gorli L Aruna Kumari2 Department of Computer Science and Engineering GITAM Institute of Technology, GITAM (Deemed to be University) Visakhapatnam, Andhra Pradesh – 530045