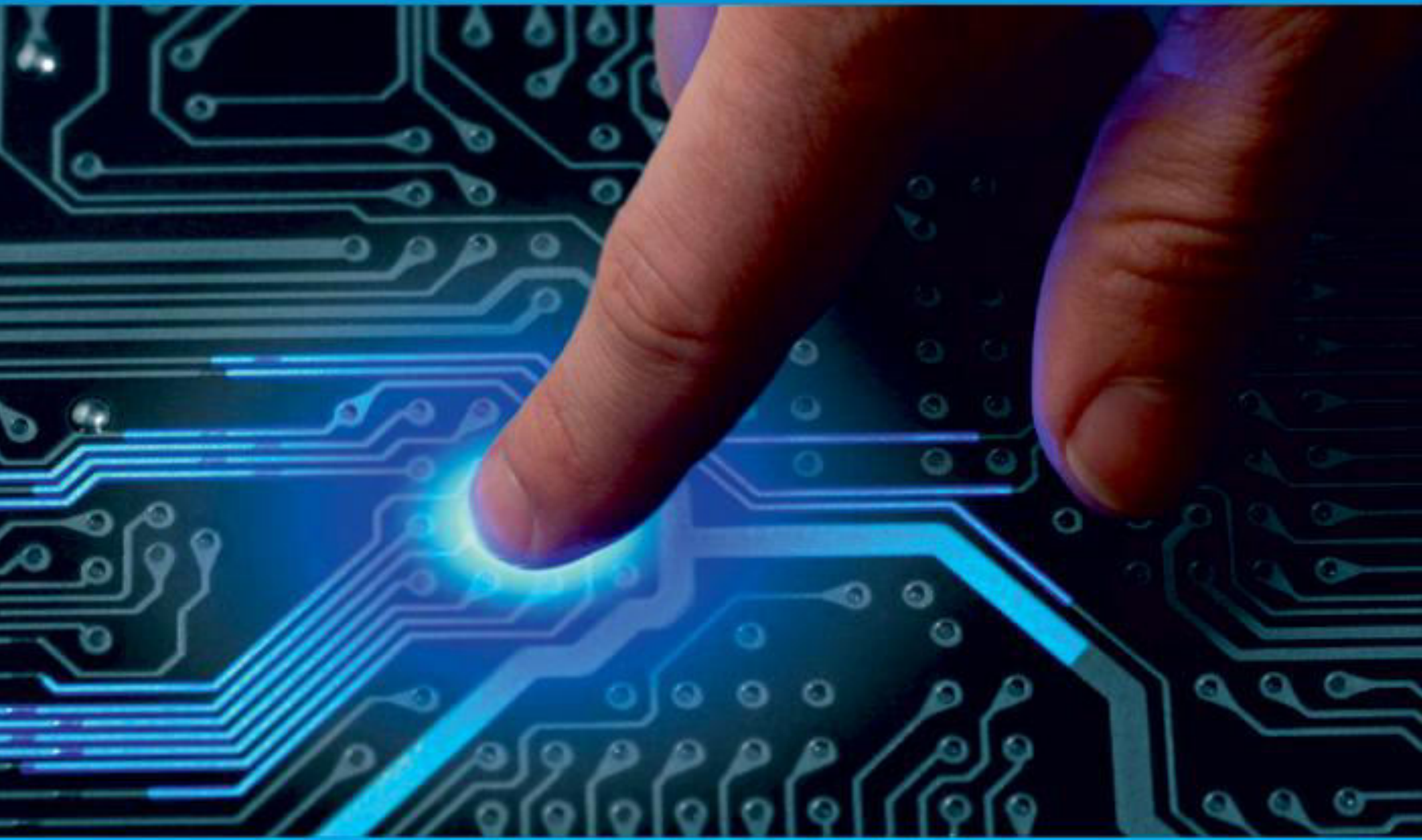
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906275 |

Detection of Strokes Using Data Mining and Machine Learning

Chintala Madhuri[1], Jujjavarapu Sumanjali[2], Avula Yasaswini Manasa[3],

Gorantla Divya Teja[4], Mr. MD Shakeel Ahmed[5]

U.G. Students, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India[1,2,3,4]

Associate Professor, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India[5]

ABSTRACT: A stroke is a medical condition in which poor blood flow to the brain results in cell death. Now a days it is a leading cause of death all over the world. Several risk factors believe to be related to the cause of stroke has been found by inspecting the affected individuals. Using these risk factors, a number of works have been carried out for predicting and classifying stroke diseases. Most of the models are based on data mining and machine learning algorithms. Machine Learning (ML) delivers an accurate and quick prediction outcome and it has become a powerful tool in health settings, offering personalized clinical care for stroke patients. In this work, we will use machine learning algorithms to detect the stroke that can possibly occur or occurred from a person's physical state and medical report data. We believe that machine learning algorithms can help better understanding of diseases and can be a good healthcare companion.The kaggle repository provided the data set that was utilized to develop and evaluate the random forest, decision tree, naïve bayes, logistic regression. The algorithm with best accuracy is chosen to design a user interface where users can input their information and get a predicted probability of having a stroke.

KEYWORDS: stroke detection, machine learning, data mining, random forest, decision tree, naïve bayes, logistic regression

I. INTRODUCTION

Performance of an algorithm means predicting the resources which are required to an algorithm to perform its task. We compare algorithms with each other which are solving the same problem, to select the best algorithm. To compare algorithms, we use a set of parameters or set of elements like memory required by that algorithm, the execution speed of that algorithm, easy to understand, easy to implement, etc. That means when we have multiple algorithms to solve a problem, we need to select a suitable algorithm to solve that problem.

Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Just like any other computer science domain machine learning also have different algorithms for performing a task. But machine learning differs in one way that it has to deal with accuracy. So, some difference performance metrics like accuracy, F1score, time taken for training, time taken for testing and memory used. Machine learning is a branch of artificial intelligence that aims at solving real life engineering problems. It provides the opportunity to learn without being explicitly programmed and it is based on the concept of learning from data. It is so much ubiquitously used dozen a times a day that we may not even know it.One of the important problems in multivariate techniques is to select relevant features from the available set of attributes. The common feature selection techniques include wrapper subset evaluation, filtering and embedded models. Embedded models use classifiers to construct ensembles, the wrapper subset evaluation method provides ranks to features based on their importance and filter methods rank the features based on statistical measurements.

Supervised learning is the most common form of machine learning scheme used in solving the engineering problems. It can be thought as the most appropriate way of mapping a set of input variables with a set of output variables. The system learns to infer a function from a collection of labeled training data. The training dataset contains a set of 12 input features and several instance values for respective features. The predictive performance accuracy of a machine learning algorithm depends on the supervised learning scheme. The aim of the inferred function may be to solve a regression or classification problem. There are several metrics used in the measurement of the learning task like accuracy, precision, f-score, recall, area under the curve etc. In this work, we will use machine learning algorithms to detect the type of stroke that can possibly occur or occurred from a person's physical state and medical report data. Before solving any engineering problem, it is vital that it is necessary to choose a suitable algorithm for the training purpose based on the type of the data. The selection of a method depends primarily on the type of the data as the field of machine learning is data driven. The next important aspect is the optimization of the chosen machine learning algorithms

II. LITERATURE SURVEY

Govindarajan et al. used Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Tree, Logistic Regression, and ensemble methods (Bagging and Boosting) to classify the stroke disease. They have collected the data from Sugam Multi-specialty Hospital, India which contains information about 507 stroke patients ranging from 35 to 90 years of age. The novelty of their work is in the data processing phase, where an algorithm called novel stemmer was used to attain the dataset. In their collected dataset, 91.52% of patients were affected by ischemic stroke and only 8.48% of patients were affected by hemorrhagic stroke. Among the mentioned algorithms, Artificial Neural Networks with stochastic gradient descent learning algorithm have the highest accuracy with 92.3% for classifying stroke.

Jeena and Kumar proposed a model based on Support Vector Machine for stroke prediction they have collected data from International Stroke trial Database. The dataset contains 12 risks factors (attributes). They have used 350 samples for their work. For training purpose 300 samples and for testing 50 samples were used. Different kernel functions like polynomial, quadratic, radial basis function and linear functions were applied. The highest accuracy of 91% was found with the linear kernel which gives the balance measure F1-score F-measure 91.7

Adam et al. have been developed a classification model for ischemic stroke using decision tree algorithm and k-nearest neighbor (k-NN). Their dataset was collected from several hospitals and medical centers in Sudan which is the first dataset for ischemic disease in Sudan. It contains 15 features and information about 400 patients. The results of the experiment show that the performance of decision tree classification is higher than the performance of k-NN algorithm. Sudha et al. used the Decision Tree, Bayesian Classifier, and Neural Network for stroke classification. Their dataset contains 1000 records. PCA algorithm was used for dimensionality reduction. In ten rounds of each algorithm, they have got the highest accuracy as 92%, 91%, and 94% in Neural Network, Naive Bayes classifier, and Decision tree algorithm respectively.

III. PROPOSED METHODOLOGY

Machine Learning (ML) delivers an accurate and quick prediction outcome and it has become a powerful tool in health settings, offering personalized clinical care for stroke patients.We used the Random Forest, Decision Tree, Naïve Bayes and Logistic Regression methods to detect the stroke disease. By feature selection process we select automatically those features which contribute most to our work and trained our models with those features. We have achieved 94.5% as highest accuracy for decision tree classifier after feature selection.

The method which gives best accuracy is chosen to build the user interface. As Decision Tree classifier gives us best performance, we train our model using the Decision Tree classifier. The user interface which is a desktop application that detects the stroke and gives the probability of a person getting the stroke. The application takes the person medical details and personal details as input and gives the probability of getting the stroke to that person.

Design Architecture

Kaggle repository provided the data set that was utilized to develop and evaluate models. Our design architecture involves various steps among them first step is extracting data from the data set, next step is data preprocessing and then integer encoding after that splitting data set in to the test and train sets, training the model and then evaluating the model performance.

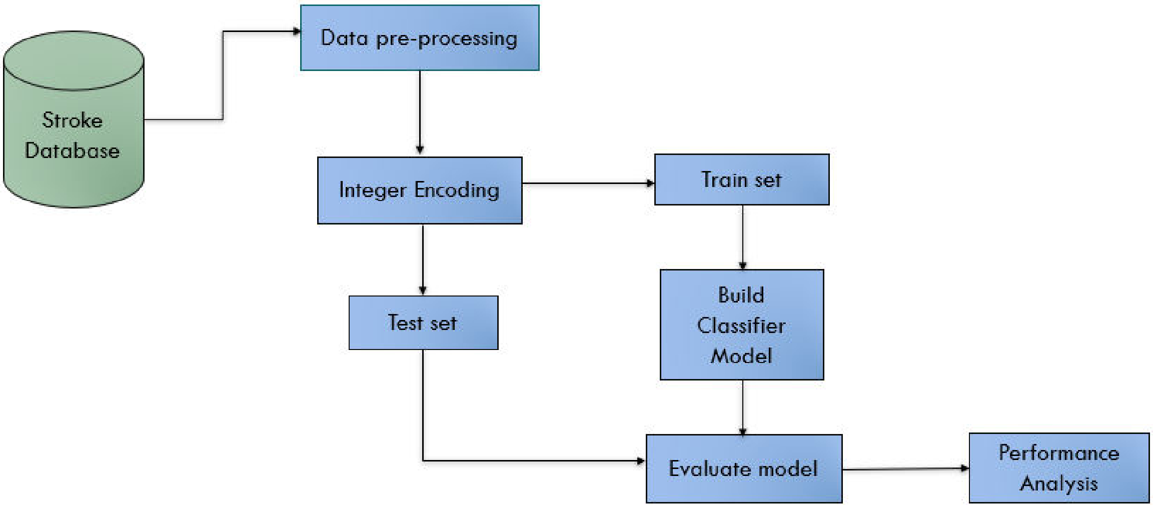


Figure 1: Architecture Diagram

Evaluating Model Performance

Here we build the model using three algorithms and the algorithm that builds a model with best accuracy is chosen to predict the final output. The accuracy of decision tree classifier is 94.5%, naïve bayes classifier is 86.5%, random forest classifier is 93.9% and logistic regression is 93%. As Decision Tree classifier gives us best performance, we train our model using the decision tree classifier. The performances of different models are as follows.

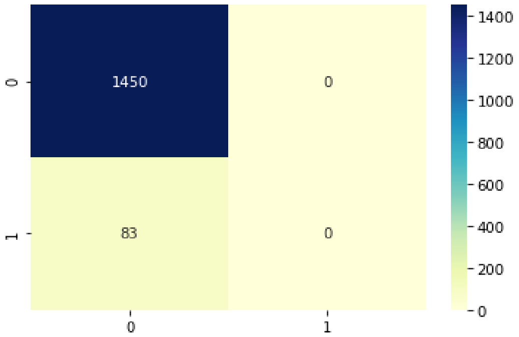


Fig 2: Confusion matrix for Decision tree classifier

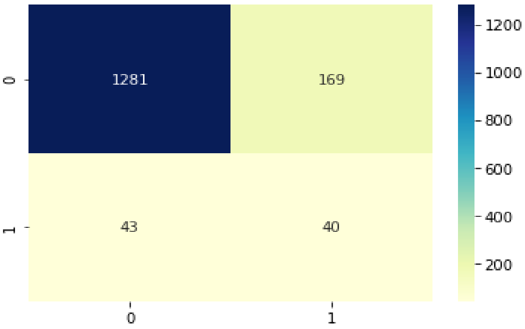


Fig 3: Confusion matrix for Naïve Bayes classifier

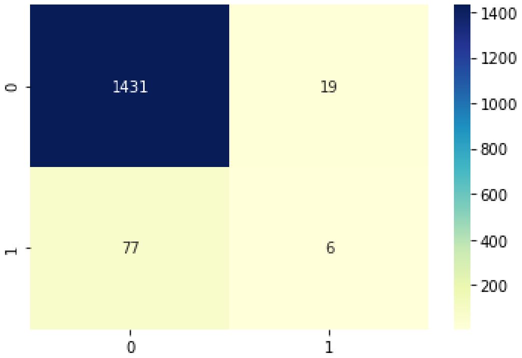


Fig 4: Confusion matrix for Random Forest

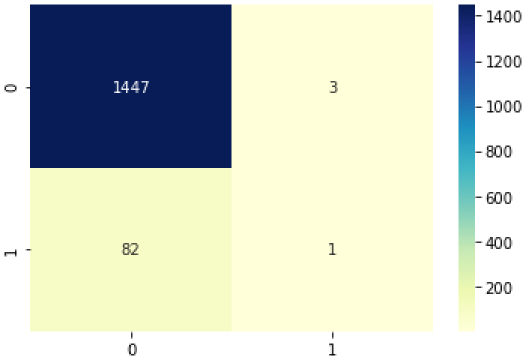
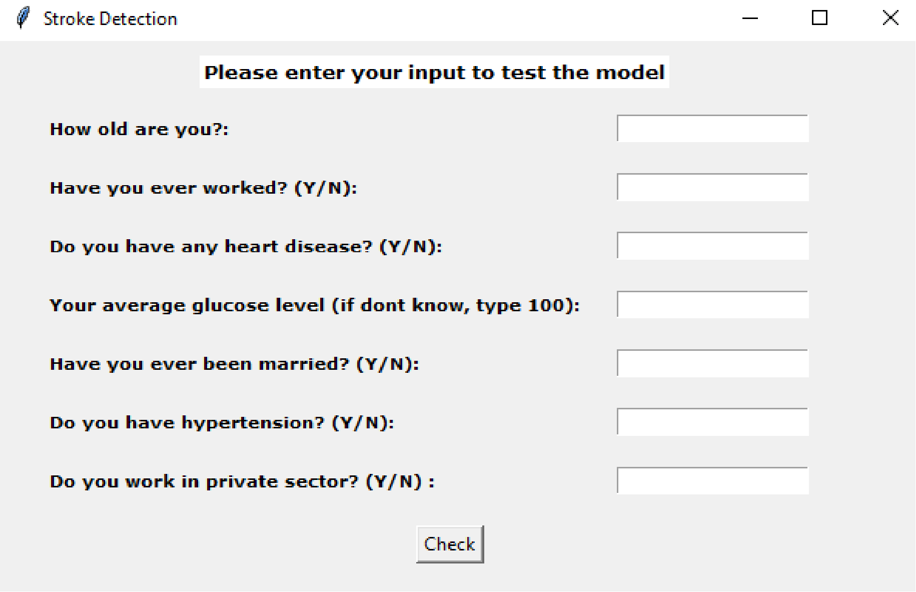


Fig 5: Confusion matrix for Logistic Regression

IV. RESULTS

The final output is a simple GUI built using Tkinter to interact with the user. It mainly performs two operations accepting the inputfrom the User and gives the probability of getting a stroke.



Please enter your input to test the model

How old are you?:

Have you ever worked? (Y/N):

Do you have any heart disease? (Y/N):

Your average glucose level (if dont know, type 100):

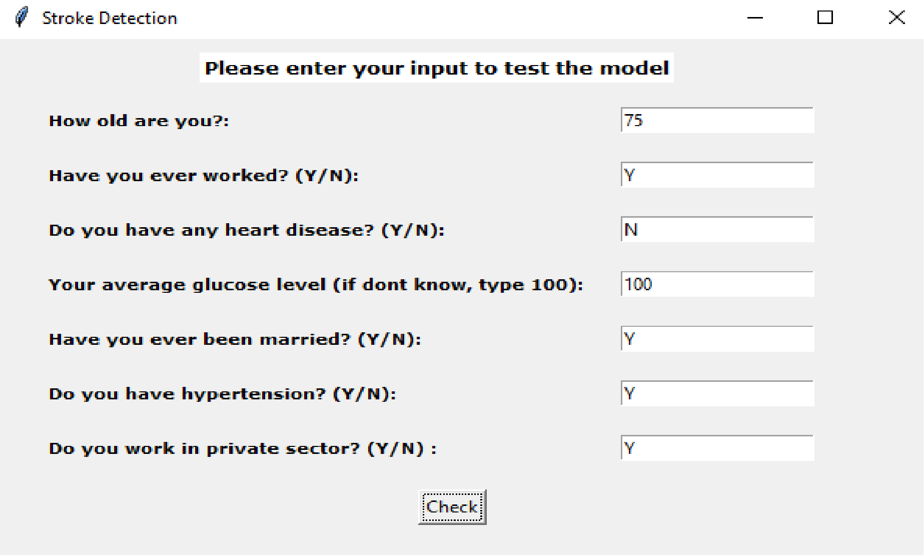
Have you ever been married? (Y/N):

Do you have hypertension? (Y/N):

Do you work in private sector? (Y/N) :

Fig 6: Requesting for the input details

The input details are to be entered in the text box available as in the above figure and the check button is to be clicked in order to detect the stroke.



Please enter your input to test the model

How old are you?: 75

Have you ever worked? (Y/N): Y

Do you have any heart disease? (Y/N): N

Your average glucose level (if dont know, type 100): 100

Have you ever been married? (Y/N): Y

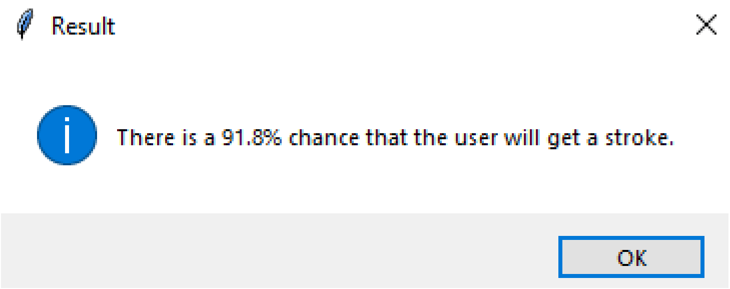
Do you have hypertension? (Y/N): Y

Do you work in private sector? (Y/N) : Y

Fig 7: Submitting the details for detection

After submitting the details, the system module comes into the action in order to get probability of stroke that the user has.

| DOI: 10.15680/IJIRCCE.2021.0906275 |



There is a 91.8% chance that the user will get a stroke.

Fig 8: Output Screen

V. CONCLUSION

According to WHO, stroke is the second leading cause of death and major cause of disability worldwide. So, it will be helpful to people to monitor their risk of getting the stroke and it is the purpose of our work. The work indicates how the stroke is affecting the people and provides a method to detect the stroke of the individual. As you can see, the process of the work is training the model, testing, predicting, checking the accuracy and executed in Jupyter Notebook.The work has the potential to monitor the risk of getting stroke for significant number of people.

REFERENCES

[1] R. Jeena and S. Kumar, “Stroke prediction using svm,” in 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 600–602, IEEE, 2016.
[2] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, “Classification of stroke disease using machine learning algorithms,” Neural Computing and Applications, pp. 1–12.
[3] A. Sudha, P. Gayathri, and N. Jaisankar, “Effective analysis and predictive model of stroke disease using classification methods,” International Journal of Computer Applications, vol. 43, no. 14, pp. 26– 31, 2012.
[4] S. Y. Adam, A. Yousif, and M. B. Bashir, “Classification of ischemic stroke using machine learning algorithms,” Int J Comput Appl, vol. 149, no. 10, pp. 26–31, 2016.
[5] Vu MT, Adali T, Ba D, Buzsáki G, Carlson D, Heller K. A shared vision for machine learning in neuroscience. *J Neurosci.* (2018) 38:1601–7. 10.1523/JNEUROSCI.0508-17.2018
[6] Kim, Y.D.; Jung, Y.H.; Saposnik, G. Traditional risk factors for stroke in East Asia. J. Stroke 2016, 18, 273–285.
[7] Johnson, C.O.; Nguyen, M.; Roth, G.A.; Nichols, E.; Alam, T. Global, regional, and national burden of stroke, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019, 18, 439–458.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details