International Journal of Mechanical Engineering

Relative analysis of classification techniques for stroke prediction system

[1]SwapnilBhalerao, [2]AnkitChahar, [3]AmalMathew, [4]EshitaSohani, [5]AnishReddyBanda, [6]KaushikDaiv

[1]Department of Computer Engineering JSPM Rajarshi Shahu College Of Engineering, Tathawade, Pune, India

[2]Department Of Computer Science RV College of Engineering Bengaluru, Karnataka,India

[3]B.tech, Computer Science,Ms Ramaiah University of Applied Sciences Bangalore, Karnataka, India

[4]B.E,Department of Electronics and Telecommunications Engineering Institute Of Engineering and Technology, DAVV Indore,Madhya Pradesh, India.

[5]Department of Computer Science Engineering Mahatma Gandhi Institute of Technology Hyderabad,India

[6]Department of Computer Engineering MIT College of Engineering, Pune, Maharashtra, India

Abstract. A stroke is a life-threatening medical condition which can cause serious issues to people and their loved ones. Stroke symptoms include difficulty walking, speaking, and comprehending, as well as facial paralysis or numbness. Early therapy with TPA (clot buster) can help to prevent brain injury. Other therapies aim to reduce consequences and prevent further strokes. Then the blood in the body which flows through all the parts also flows through the portion of the brain and when this flow is stopped the brain tissues are deprived of the required amount of oxygen and nutrients which are required for the functioning of the brain properly, results in a stroke and within some time the brain cells begin to die. Stroke is considered to be the world's second-largest cause of death according to the statistics, and it is continuing to be one of the biggest health burden for both people and national healthcare system. It is considered as a big medical emergency which includes risk factors such as atrial fibrillation, diabetes, glucose metabolism dysregulation. The method proposed by the authors give an idea to the researchers about which is the best method to use from adaboost, SVM, random forest classifier, Xgboost and logistic regression.

Stroke Prediction , Random Forest classifier , Machine Learning , predictive modelling, comparative model

1 Introduction

The central and peripheral nervous systems are both affected by neurological diseases. Some of these illnesses are treatable, while others are not. Neurological disorders which pose a big problem to the human life such as Alzheimer's disease and Parkinson's disease affect people mostly beyond the age of 60, making ageing a primary role in the introduction of these diseases[2, 3]. Genetic abnormalities, infections, and lifestyle choices are among the reasons, as are other health issues that may impact the brain. There are over 600 nervous system disorders, includ-ing stroke, brain tumours, epilepsy, and many others. Around 15 million people live in the area.[4]According to the statistics around the globe it is considered that stroke is the second-greatest cause of mortality and adult impairment in the globe, with 400-800 strokes per 100,000, 15 million new acute strokes every year, 28,500,000 disability adjusted life-years, and a 28-30-day case fatality rate ranging from 17 percent to 35 percent. The impact of stroke is expected to grow, with stroke and heart disease-related fatalities expected to rise to five million in 2020, up from three million in 1998. This will be due to the ongoing health and demographic change, which will result in an increase in vascular disease risk factors and the age of the elderly people. Thus, the system proposed would be revolutionary in determining the stroke at an early stage for the patient with a much greater accuracy which would be taken into account by all the medical institutions for a better rate of avoiding strokes at an early as well as a late age.

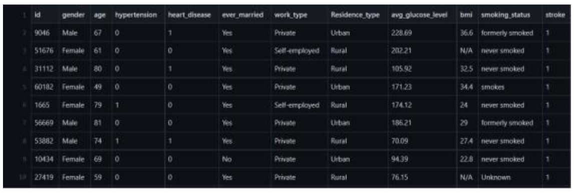
2 Related Work

Farrikh Alzami et al. (2018) developed a feature selection technique for clas-sifying epilepsy episodes. The University of Bonn in Germany contributed the EEG dataset for this study. Subsets were obtained using a hybrid feature selec-tion method. In this study, the mRMR, Fisher, Chi-Square, and Relief -F were used to identify features. Rank aggregation was used to merge the subgroups ob-tained. In addition, the aggregation subgroups were fed into a basic classifier to generate the learning model and prediction. Finally, forecasting the classification and detection tasks was done by voting, which can be used by us as an exam-ple for the detection task. Another research[6] of Alzheimer's disease includes the study by Pholpat Durongbhan et al.,(2018) which aimed to obtain biomark-ers using Quantitative Analysis of Electroencephalography through a framework consisting of data augmentation, feature extraction, KNearest Neighbour (KNN) classification, quantitative evaluation and topographic visualization. Twenty HC and 20 AD subjects participated in this research and data was collected from them. MATLAB had been used for research purposes. The proposed frame-work was able to accurately classify the records and found important features as biomarkers for proper diagnosis of disease progression.

3 Methodology

3.1 Data preparation

Data has been taken from Health care dataset of strokes, which has played an important part in many researches. As the data is pretty tough to collect, the only action possible to perform the model to do the prediction was to take a data from a good source .so the dataset has been taken from Health care dataset of strokes. Based on input factors including gender, age, and numerous illnesses and smoking status, this dataset is used to predict whether a patient is likely to have a stroke. For Machine Learning and Data Visualization purposes, a subset of the original train data is selected using the filtering approach. The glimpse of the dataset can be seen in the figure 1 with 11 clinical features in it. The dataset has approximate 5000 entries in it.



id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1

Fig. 1. Dataset overview

3.2 Data Preprocessing:

Data preprocessing is referred to manipulating the data or dropping the data before it is used for it's main purpose to enhance the performance of the model.So now starting with improving the dataset quality the first step would be Encoding Categorical Features which means that the dataset contains categorical values and encoding should be done before one can fit and evaluate the model. The next step would be to scale the variance in the features, as feature scaling is a method to standardize the independent features present in a data in a fixed range, in this case BMI, age and avg-glucose-level is scaled. And finally, the last step would be to drop the id feature because the number of entries isn't required and finding the null value as null creates a problem as it hinders with the accuracy of the model.

3.3 Feature Selection

All the variables in a dataset are unlikely to be helpful for developing a machine learning model in the actual world. Duplicate variables reduce the model's gen-eralization ability and can reduce a classifier's overall accuracy. Adding more variables to a model raises the overall complexity of the model. The optimum solution to a problem is one that requires the fewest assumptions. As a result, feature selection in the building of machine learning models is becoming increasingly crucial. In this case, the author has 11 characteristics that will assist in achieving adequate accuracy and improving the model's performance.Figure 2 shows that the majority of patients who have had a stroke do not have heart dis-ease, but this does not rule out the possibility that it was a contributing cause. Smoking has been linked to an increased risk of stroke, as seen in Figure 3. The relationship between employment type and stroke is peculiar, since it has been noticed that persons who work in private enterprises have a higher risk of stroke. Age is a factor; the older you become, the more likely you are to have a stroke. Finally, because stroke is nothing more than a blockage in our heart, which may arise as a result of obesity, the box plot in displayed most persons undergoing treatment for stroke have a high BMI

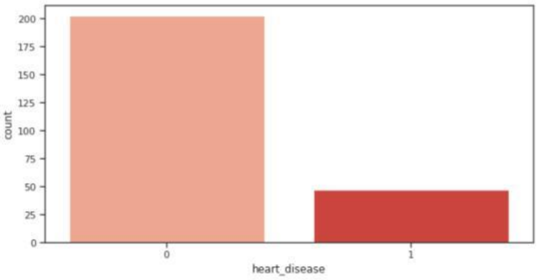


Fig. 2. heart disease correlation

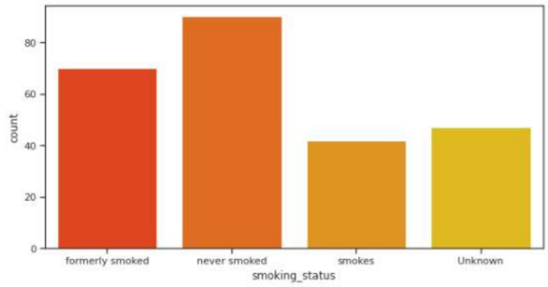


Fig. 3. smoking feature correlation

3.4 Model Architecture

Adaboost Adaptive boosting also known as adaptive boosting which is a boost-ing technique mainly used in an ensemble method in machine learning. The name adaptive boosting is there because the weights are re-assigned to every possible instance, with higher weights assigned to incorrectly classified instances. The main aim of boosting is to reduce the bias as well as the variance for the su-pervised machine learning method. It follows the principle of learners growing sequentially. So, weak learners are finally then converted into strong ones.

Random Forest It's a technique which belongs on the ensemble model cate-gory. It may be used to develop a good prediction model by combining classi-fication and regression techniques. In this work, decision trees are used as the foundation estimators. On their own, decision trees are a poor predictor, but when combined with other decision trees, they improve. Decision trees vote on how to categorize a specific instance of input data in classification tasks, and they output the class that is the mode of the classes or the mean of predictions in regression tasks. In this manner, we may prevent parameter tinkering and reduce overfitting.

Support Vector Machine To implement nonlinear class borders, Support Vector Machines use a linear model. To separate the target classes, support vectors (lines or hyperplanes) are created. To handle a nonlinear problem, the model uses a mapping function to apply numerous transformations to the data and then trains a linear SVM model to classify the data in a higher-dimensional feature space.

Logistic regression The method of modelling the probability of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary outcome, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be used to model situations with more than two discrete outcomes. Logistic regression is a valuable approach of analysis.

XGBoost XGBoost mainly known as the decision tree ensemble machine learn-ing algorithm especially designed for speed and performance. It uses the gradient boosting framework, when it comes to prediction of problems involving unstruc-tured data the artificial neural networks tend to outperform all the other de-signed algorithms and frameworks. But, when it comes to small and medium structured data, the tree based algorithms are considered best.

4 Experimental results

Finally, we've progressed through data gathering through preprocessing to fea-ture selection to model architecture and application, the algorithms used by the

6 F. Author et al.

authors(SVM, Adaboost, XGBoost, random forest classifier and logistic regres-sion) have been compared based on the accuracy provided by each of them after performing the algorithms individually and not as an ensemble method. The confusion matrix for the XGboost can be seen in the figure 5. Similarly, for the SVM algorithm we would be getting the matrix which gives us a clear picture for the comparison. All the algorithms with their confusion matrix have been analysed and compared.

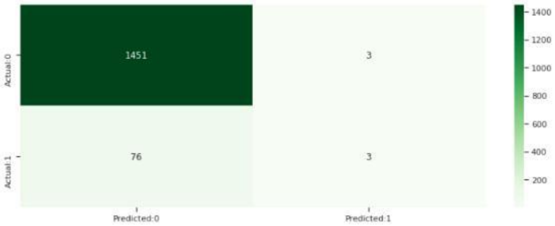
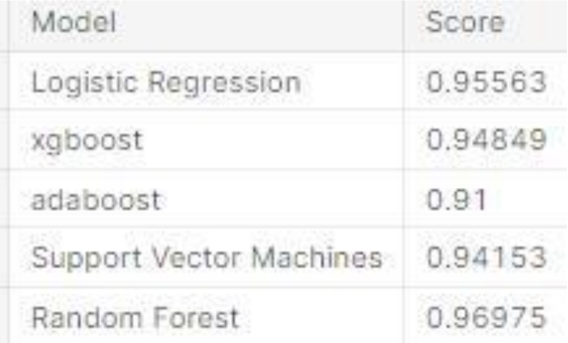


Fig. 4. XGBoost confusion matrix



Model	Score
Logistic Regression	0.95563
xgboost	0.94849
adaboost	0.91
Support Vector Machines	0.94153
Random Forest	0.96975

5 Conclusion

The broad comparison between the machine learning models used provide us with the best accuracy out of all which is achieved by random forest classifier and helps the people working in the similar field and topic to use it up and explore more using the provided algorithm and along with it assists physicians in making the best, most accurate, and quickest decisions possible, as well as lowering total treatment costs. By predicting strokes at an early stage, the suggested method significantly lowers treatment costs and enhances the quality of life. By using the Artificial neural networks, we have been able to get a whooping accuracy of approximate 96 percent on a given dataset, which is a great outcome in terms of science and invention and will help us to have fewer patients dying with strokes. The below table states the different accuracies retrieved after performing each algorithm.

References

1. Akash, Kunder Shashank, H .S, Srikanth A.M, Thejas. (2020). Prediction of Stroke Using Machine Learning.
2. "What is Alzheimer's Disease? Symptoms Causes — alz.org," Alzheimer's Associ-ation.[Online].
3. "Parkinson'sdisease-Symptomsandcauses,"Mayo Clinic,2018.[Online].Available:https://www.mayoclinic.org/diseases-conditions/parkinsons disease/symptoms-causes/syc-20376055.
4. "Stroke Statistics," The Internet Stroke Center. [On-line].Available:http://www.strokecenter.org/patients/about-stroke/stroke- statistics/.
5. Farrikh Alzami, Juan tang et al.,"Adaptive hybrid feature selection-based classifier ensemble for epileptic seizure classification," IEEE Access, vol. 6, pp. 29132-29145, 2018.
6. Pholpat Durongbhan, Yifan Zhao et al., "A Dementia Classification Framework using Frequency and Timefrequency Features based on EEG signals," IEEE Trans-actions on Neural Systems and Rehabilitation Engineering, pp. 1-10, 2018.
7. "Development of an Algorithm forStroke Prediction: A National HealthInsurance-Database Study" - Min SN, Park SJ, KimDJ, Subramaniyam M, Lee KS
8. "Stroke prediction using artificialintelligence"- M. Sheetal Singh,PrakashChoudhary
9. Y. Zang, H. Lu, Y. Zhang, E. Alghannam, Z. Guo and L. Li, "A Straightness Con-trol System for Motor Shaft Straightening with the Stroke Prediction Algorithm," 2019 6th International Conference on Systems and Informatics (ICSAI), 2019, pp. 57-62, doi: 10.1109/ICSAI48974.2019.9010553.
10. B. A. Kobrinskii and V. V. Donitova, "Building a Knowledge Base of an Expert System for Personalized Stroke Risk Prognosis," 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), 2021, pp. 2815-2817, doi: 10.1109/ElConRus51938.2021.9396306.
11. P. A, N. G, V. K. R, P. P and S. R. R.V.T, "Stroke Prediction System Us-ing Artificial Neural Network," 2021 6th International Conference on Commu-nication and Electronics Systems (ICCES), 2021, pp. 1898-1902, doi: 10.1109/IC-CES51350.2021.9489055.
12. S. J. Park, I. Hussain, S. Hong, D. Kim, H. Park and H. C. M. Benjamin, "Real-time Gait Monitoring System for Consumer Stroke Prediction Service," 2020 IEEE International Conference on Consumer Electronics (ICCE), 2020, pp. 1-4, doi: 10.1109/ICCE46568.2020.9043098.

13. T. Kansadub, S. Thammaboosadee, S. Kiattisin and C. Jalayondeja, ”Stroke risk prediction model based on demographic data,” 2015 8th Biomedical Engi-neering International Conference (BMEiCON), 2015, pp. 1-3, doi: 10.1109/BME-iCON.2015.7399556.

14. Indarto, E. Utami and S. Raharjo, ”Mortality Prediction Using Data Mining Clas-sification Techniques in Patients With Hemorrhagic Stroke,” 2020 8th International F. Author et al.Conference on Cyber and IT Service Management (CITSM), 2020, pp. 1-5, doi:10.1109/CITSM50537.2020.9268802.

15. S. Modi and M. H. Bohara, ”Facial Emotion Recognition using Convolution Neural Network,” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1339-1344, doi: 10.1109/ICICCS51141.2021.9432156.

16. C. -H. Lien, F. -H. Wu, P. -C. Chan, C. -M. Tseng, H. -H. Lin and Y. -F. Chen, ”Readmission Prediction for Patients with Ischemic Stroke after Discharge,” 2020 International Symposium on Computer, Consumer and Control (IS3C), 2020, pp. 45-48, doi: 10.1109/IS3C50286.2020.00019.

17. J. Chen, Y. Chen, J. Li, J. Wang, Z. Lin and A. K. Nandi, ”Stroke Risk Prediction with Hybrid Deep Transfer Learning Framework,” in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2021.3088750.

18. J. Cho, Z. Hu and M. Sartipi, ”Post-stroke discharge disposition pre-diction using deep learning,” SoutheastCon 2017, 2017, pp. 1-2, doi: 10.1109/SECON.2017.7925299.

19. E. Zamsa, ”Medical software user interfaces, stroke MD application design,” 2015 E-Health and Bioengineering Conference (EHB), 2015, pp. 1-4, doi: 10.1109/EHB.2015.7391403.

20. Le Zheng et al., ”Risk prediction of stroke: A prospective statewide study on patients in Maine,” 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 853-855, doi: 10.1109/BIBM.2015.7359796.