2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) | 978-1-6654-2829-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAECA52838.2021.9675652

Applying Machine Learning Techniques for Stroke Prediction in Patients

Dr. R.K. Kavitha
kavitha.rk.mca@kct.ac.in
Assistant Professor [SRG]
Department of Computer Applications
Kumaraguru College of Technology

Dr. W. Jaisingh
jaisingh.w.mca@kct.ac.in
Assistant Professor [SRG]
Department of Computer Applications
Kumaraguru College of Technology

Ms. S.R. Sujithra
sujithra.20mca@kct.ac.in
PG Scholar
Department of Computer Applications,
Kumaraguru College of Technology

***Abstract*- Nowadays, an increase in stroke rate is observed in India when equated to other countries. Also, a minor percentage of fatality was observed after early trauma in patients suffering from stroke. Several studies have been conducted throughout the world to predict the disease at an early stage with the help of machine learning algorithms. The objective of the research is to build a machine learning prototype which has the ability to predict the occurrence of stroke based on the observed characteristics and symptoms. Also, this work aims to find the best features of stroke disease by using feature selection approach. The stroke dataset which is publicly available was utilized in this work. Application of various machine learning algorithms on the data set was done and the prediction accuracy of the algorithms were compared. Out of various techniques, it was observed that the features selection methods like Chi-square, principal component analysis and removing data outliers when applied on the Decision Tree classifier yielded a good accuracy of 98% thus helping in early identification, medication and therapy for the disease.**

Keywords- Machine learning, Predictive techniques, Stroke, feature selection, classification.

I. INTRODUCTION

Machine Learning (ML) is getting focus and attention nowadays in the field of healthcare. Chronic diseases are rapidly increasing among human population and many countries are shifting their focus in healthcare systems to prevent diseases and increase the wellness of the people [5]. A considerable volume of medical data has been made available and when these data are analysed in depth, solutions for medical problems can be recommended which leads to prevention of several diseases at the early stage. As per the reports of World Health Organization (WHO), illness due to cancer and heart defects are the main life-threatening diseases. The third critical disease was identified as stroke accounting for higher death rate in recent times. Stroke happens when there is a disturbance to the blood flow to brain causing damage of the cells in the brain [6][8]. Generally, stroke often occurs among elderly people and this disease causes cerebral dysfunction leading to unconsciousness, hemiplegia etc. which may sometimes lead to fatality in adults [2]. Thus, if there is a possibility of detecting this disease early, the severity of it can be lowered. The occurrence of stroke can be decreased when the risk factors leading to stroke like smoking, hypertension, body mass index and heart disease are appropriately managed. To improve the situation, there is a pressing necessity to carry out research and apply measures to overcome the disease.

Machine learning entails computers realizing how to execute tasks without explicitly programmed to carry out the task. Machine learning concentrates on implementing computer programs which can gain access to data and make use of it to learn for themselves [9][12]. This field has turn out to be an important tool for researchers and medical experts to predict and treat diseases at the appropriate time. Main challenges of healthcare data are the huge, diverse and complicated nature of it [15]. The requirement for the processes which provide highly accurate results is necessary in medical diagnosis since it is deemed as a quite considerable task which must be done correctly and effectively. Popular techniques which can be applied to generate predictive models on the data set are support vector machines, naïve Bayesian classifier and decision trees. Bayesian classifier is considered to be the easiest and a reasonably precise predictive data mining approach [13].

This work aims to propose a novel stroke prediction system based on a publicly available data set. Classification algorithms namely Decision Trees (DT), Support Vector Machines (SVM) and Naïve Bayes (NB) were utilized for forecasting the existence of stroke disease with a variety of associated attributes. A commonly used technique for lowering the dimensions of data is the Principal Component Analysis (PCA) [11]. It helps in deciding more appropriate attributes to forecast stroke disease. Chi square test is one of the feature selection technique used in this work. Normally chi square is employed in statistics to check the independence of two events. An outlier is an observation point which is far away from other observations. Identifying and removing outliers improves the accuracy of data classification [14]. Feature selection was accomplished by employing Chi square, PCA, a combination of Chi square and PCA and finally outlier removal combined with Chi square and PCA in this work. The performance of the classifiers was compared by applying the above-mentioned feature selection techniques.

II. EXISTING WORK

Throughout the world, many researchers have conducted studies on the initial analysis of neurological conditions. Ane Alberdi et al. has gathered the behavioural statistics of patients at their residence to identify the indications for

978-1-6654-2829-3/21/$31.00 ©2021 IEEE

Alzheimer's disease [1]. The symptoms of such patients were related to the feelings, awareness, and movement of patients. Yonglai Zhang et al. performed a study on stroke patients to discover the threat of stroke [2]. The dataset comprised of archived data of 792 patients at a hospital located in Beijing. The data was analysed in several stages namely filter, voting and wrapper stage. Data filtering was performed on the standard deviation variable. Also, Support Vector Machine technique was utilized for finding the classification correctness of feature subsets. The fusion of SVM with glow-worm swarm optimization algorithm was used to determine essential features which helped to identify the key risk factors for occurrence stroke disease among patients. Farrikh Alzami et al. presented a technique which used feature selection to categorize seizures triggered by epilepsy. The research dataset was offered by a German University [3]. Techniques like mRMR, Relief - F, Fisher, and Chi-Square were used for feature selection. Another researcher has employed classifiers like DNN, SVM, LR and GBDT to analyse a patient database with the intention of stroke happening among patients [8]. Another study conducted by Rajkomar et al. utilized deep learning technique which was applied on EMR data of several patients [5].

A stroke prediction model was proposed in a research which helped determining the stroke risk centred on health examination records obtained by National Health Insurance Service (NHIS), with attributes namely age, total cholesterol, diabetes, high blood pressure, drinking volume, smoking, workout and body mass index [4]. Khosla et al. [8] has stated the prediction and verification results after changing parameters of the model by making use of the kernel functions of classifier SVM for threats of stroke disease. In specific, the function like RBF kernel was utilized to achieve a better classifier accuracy. Though, these SVM-based findings concentrated on forecasting seriousness and diagnosis following an outbreak, as contradicted to initial detection of early indications. Revanth et al. has applied machine learning to detect stroke among patients using various classifiers and SVM has proved to provide a higher accuracy of 98% accuracy [9]. Soodamani Ashokan et al. has proposed an effective method for detecting stroke and the model provides better results for Decision Tree model and it was concluded that persons above sixty years of age are more prone to the stroke disease [10]. Also, people in the age group of sixty-one to seventy displayed less risk of getting stroke if they never have smoked. The research concluded that people above the age of sixty are most likely to get affected by stroke. One More research on Alzheimer's disease was done by Pholpat Durongbhan et al. and the work aimed to obtain biomarkers utilizing Quantitative Analysis of Electroencephalography via a framework comprising of data reinforcement, quantitative evaluation, feature extraction, K Nearest Neighbour (KNN) classification and topographic visualisation [7]. The recommended framework was capable to precisely classify records and discovered vital features as biomarkers for appropriate analysis of disease progression.

III. PROPOSED WORK

This research uses a publicly available healthcare dataset with the intention of detecting the possibility of stroke among people by analysing various attributes. It was decided to use different techniques chi square, principal component analysis and outlier detection for feature selection. Also, a combination of two or more of these feature selection methods were tried. Data set was classified with the help of three classifiers Decision Trees, Support Vector Machines and Naïve Bayes and their performances were compared.

The dataset used in this study consisted of 5110 records. The dataset consisted of 10 features which included patients' demographic data namely type of work, gender, age, marital status and type of residence and based on health records like hypertension, average glucose level which is measured after meal consumption, heart disease, Body Mass Index (BMI), status about smoking and earlier experience of stroke. Out of the 5110 records, 3577 (70%) was used as training data set and 1533 (30%) was used for testing the model. In the total data set, female was more than the male as shown in Fig.1. Also, it can be understood from Fig.2 that age of most of the people in the dataset was between 30 to 50 years.

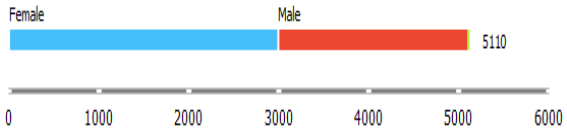

Fig. 1. Gender statistics

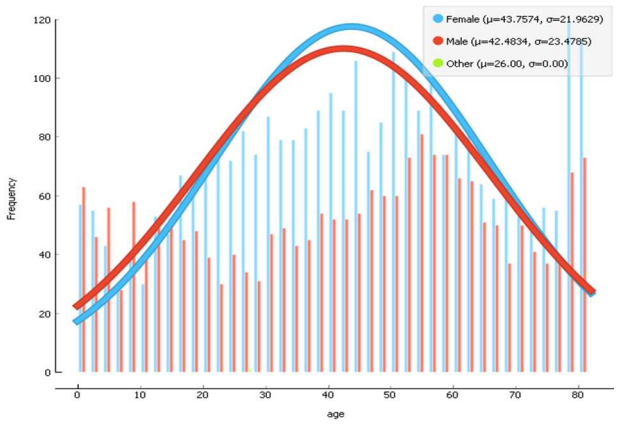

Fig. 2. Age and gender relation in the dataset

The model used in this study is shown in Fig. 3. The dataset was first pre-processed. Normally, feature selection methods are meant to lessen the quantity of input variables and are understood to be highly helpful for a model to foresee the target variable. Feature selection is mainly centered on eliminating non-informative or unnecessary

predictors from the model. In this work, it is decided to use feature selection methods like Principal Component Analysis (PCA), Chi Square, combination of Chi Square and PCA, removing outliers combined with Chi Square and PCA. Classification is a data mining method which defines objects or items in a collection to target classes. Main objective of classification is to correctly forecast the target class for every single case in the data set. Following the feature selection, the data is classified by applying methods namely Decision Trees (DT), Support Vector Machine (SVM) and Naïve bayes (NB). Finally, the performance of the classifiers on application of various feature selection techniques were studied.

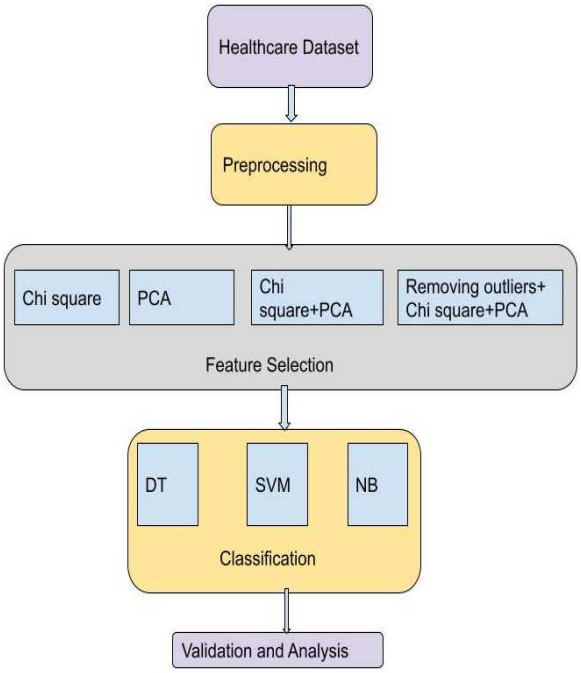

Fig. 3. Proposed model

IV. RESULTS AND DISCUSSIONS

From the Fig.4, it can be inferred that the possibility of stroke occurrence is more in employees who worked in private organizations when compared to people who worked in government organizations or those who were self-employed. Also, from the Fig.5, it is evident that the people residing in urban areas were affected more by stroke disease when compared to people residing in rural areas.

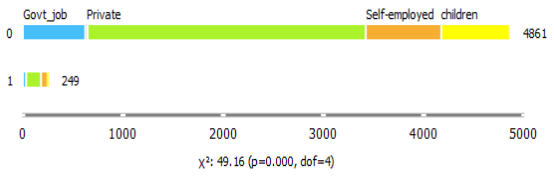

Fig. 4. Job wise stroke

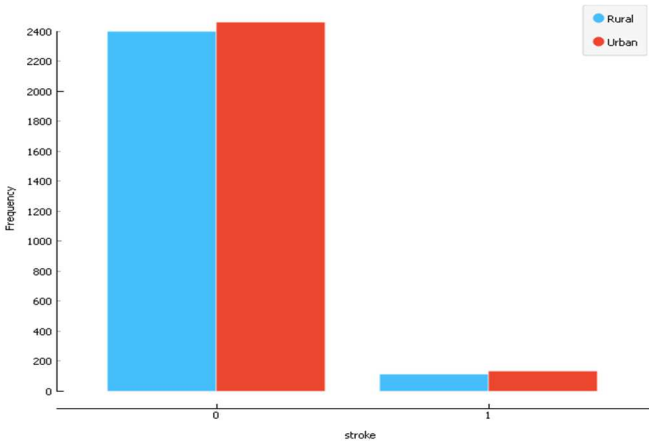

Fig. 5. Stroke and Residence type

After constructing a classification model, next step is to assess the performance of it, that is demonstrating how great the model is in forecasting the results of the test data. Also, it is essential to assess the model prediction correctness and prediction errors with help of the test data set. The results of data set classification performance based on metrics like Classifier Accuracy (CA), F1, Precision and Recall are exhibited in Table I. Techniques namely chi square, PCA, combination of chis quare and PCA, removing outliers combined with chi square and PCA were utilized to decrease the quantity of features considered for classification. The feature reduction techniques were used for classifying the dataset by applying three algorithms namely DT, SVM and NB. From table 1, it could be understood that Decision Tree (DT) classifier displays a improved classification precision when related to all other classifiers used in this study. Also, from Fig.6, it can be concluded that a better classifier accuracy of 98 percent is obtained when the data set features were reduced by removing outliers and applying chi square and PCA.

TABLE I. HEALTHCARE DATASET – CLASSIFIER PERFORMANCE

Data set	Features	classifier	CA	F1	Precision	Recall
Health care	chi	DT	0.951	0.928	0.905	0.951
		SVM	0.891	0.9	0.909	0.891
		NB	0.923	0.922	0.921	0.923
	PCA	DT	0.962	0.951	0.905	0.951
		SVM	0.894	0.903	0.913	0.894
		NB	0.945	0.927	0.914	0.945
	Chi+ PCA	DT	0.971	0.968	0.905	0.951
		SVM	0.917	0.912	0.907	0.917
		NB	0.870	0.894	0.924	0.870
	Removing outliers +chi+ PCA	DT	0.985	0.993	0.913	0.955
		SVM	0.919	0.917	0.915	0.919
		NB	0.890	0.907	0.927	0.890

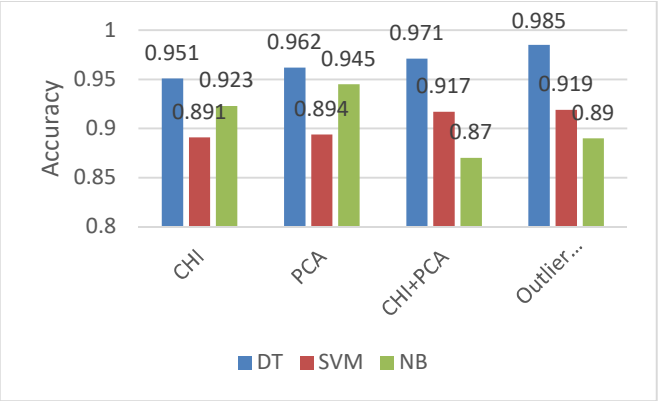

Fig. 6. Comparison of ML classifiers for Healthcare dataset using accuracy

The classifier performance based on the metric F1 Score is shown in Fig.7 for various feature selection techniques. Again, for DT classifier, a better classification accuracy of 99 percent is obtained when the data set features were reduced by removing outliers and applying chi square and PCA.

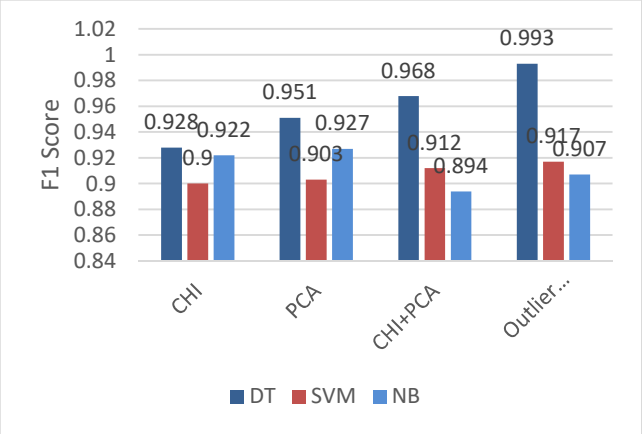

Fig. 7. Comparison of ML classifiers for Healthcare dataset using F1

V. CONCLUSION

In this work, it was suggested to utilize feature reduction techniques like chi-square (CHI), PCA, combination of CHI and PCA, outlier removal combined with CHI and PCA in order to enhance the prediction of machine learning models. The objective of a classifier was to guess whether a patient has the possibility of getting stroke disease or not. Utilization of complete features is not possible while considering system resources. However, dimensionality reduction techniques were successfully applied to improve the results in this work. Also, it was observed that amongst the various classifiers, combination of CHI and PCA when applied after outlier removal exhibited maximum performance for Decision Tree classifier with 98% accuracy for the healthcare stroke dataset. The main intention of this work was to discover the best dimensionality reduction method for predicting stroke in terms of performance. The experimental findings demonstrated that by combining outlier removal along with CHI and PCA displayed better performance in most classifiers and hence found to be the most reliable and desirable method.

REFERENCES

[1] A Alberdi, A Weakley et al., "Smart home-based prediction of multi-domain symptoms related to Alzheimer's Disease," IEEE Journal of Biomedical and Health Informatics, vol.6, pp.1720-1731,2018.
[2] Yonglai Zhang, Wenai Song et al., "Risk Detection of Stroke Using a Feature Selection and Classification Method," IEEE Access, vol. 6, pp. 31899-31907, 2018.
[3] 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.
[4] Chen-Ying, H Wei-Chen, C et al., "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database", Annu Int Conf IEEE Eng Med Biol Soc, 2017, pp. 3110–3113.
[5] Rajkomar, A, Oren, E et al., "Scalable and accurate deep learning with electronic health records", NPJ Digit. Med. 2018, pp. 1- 18.
[6] Lee, J.S, Park, J.M et.al, "Development of a stroke prediction model for Korean", Korean Neurol. Assoc. 2010, vol. 28, pp. 13–21.
[7] Pholpat Durongbhan, Yifan Zhao et al., "A Dementia Classification Framework using Frequency and Time frequency Features based on EEG signals," IEEE Transactions on Neural Systems and Rehabilitation Engineering, pp. 1-10, 2018.
[8] Khosla, A, Cao, Y et al. "An integrated machine learning approach to stroke Prediction", Proceedings of the Sixteenth ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining pp. 183–192, 2010.
[9] Revanth S, Sanjay S, Sanjay N, Vijayaganth V, "Stroke Prediction using Machine Learning Algorithms", International Journal of Disaster Recovery and Business Continuity Vol.11, No. 1, (2020), pp. 3081–3086.
[10] Soodamani Ashokan, Suriya G.S Narayanan, Mandresh S, Vidhyasagar B, Paavai Anand G, "An Effective Stroke Prediction System using Predictive Models", International Research Journal of Engineering and Technology (IRJET), Volume 07 Issue 03 , Mar 2020.
[11] Chen R, Sun N, Chen X, Yang M, Wu Q, "Supervised feature selection with a stratified feature weighting method", IEEE Access 2018;6:15087–98.
[12] Parthiban G, Srivatsa SK. "Applying machine learning methods in diagnosing heart disease for diabetic patients" Int J Appl Inf Syst 2012;3(7):25–30.
[13] Domingos P, "A few useful things to know about machine learning", Communications of ACM, 2012;55(10):78.
[14] Anna Karen Garate-Escamila, Amir Hajjam El Hassani, Emmanuel Andres, "Classification models for heart disease prediction using feature selection and PCA", Informatics in Medicine Unlocked 19 (2020), pages 1-11.
[15] Zhang D, Zou L, Zhou X, He F, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer", IEEE Access 2018;6:28936–44.