International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

A Hybrid Feature Selection Method for Stroke Detection Using Machine Learning Approaches

Priyanka S R [1], Meera M [2]

P.G. Student, Department of Computer Science and Engineering, Government Engineering College, Idukki, Kerala, India[1]

Associate Professor, Department of Computer Science and Engineering, Government Engineering College, Idukki, Kerala, India[2]

ABSTRACT: Strokes are brain attacks that occur when the blood supply to the brain becomes blocked. A stroke is a medical emergency that wants immediate medical attention. Machine learning (ML) is an application of artificial intelligence that permits the systems the ability to learn and enhance from experience without being explicitly programmed. This work proposes a supervised hybrid model for stroke detection using hybrid feature selection that is applicable for various machine learning approaches. Feature selection is based on wrapper and filter models. Wrapper models evaluate the accuracy produced by the use of the selected features in classification and filter model uses a particular feature evaluation indices to rank the features. A correlation based feature ranking is used. Then a subset of features is selected and given as input to four classifiers to evaluate its performance. From the experimental results, it is found that the Support Vector Machine (SVM) classifier achieves an accuracy of 83% with a reduced set of features.

KEYWORDS: Stroke; Machine Learning; Feature Selection, Correlation.

I. INTRODUCTION

The stroke disease occurs due to the rapid loss of brain function caused by the disruption in the blood supply to the brain. This can happen due to ischemia (lack of blood flow) caused by blockage, or a haemorrhage (leakage of blood).As a result, the affected area of the brain cannot work properly. Symptoms are: hemiplegia (an inability to move one or more limbs on one side of the body), aphasia (inability to understand or use language), or an inability to see one side of the visual field. It can cause permanent damage. If it's not quickly treated, it's going to cause death. It is the third most common cause of death and the most common cause of disability for adults in the United States and Europe. Thus it has become a need to detect the stroke as early as possible with the symptoms that the patients are showing. Early detection of stroke helps to minimize brain damage by medication like tpa.

This paper focuses on detection of stroke by finding the most appropriate features that will lead to the disease. Techniques such as machine learning and feature selection were used for the proper detection of stroke. Feature selection is is a famous data-pre-processing method in machine learning. It is a dimensionality reduction method used basically for the removal of unenviable and useless features from any dataset. That is, selecting subsets from original features. There are four approaches in feature selection they are filter approach, wrapper approach, embedded approach, and hybrid approach [2]. This paper focuses on hybrid feature selection approach that combines both filter and wrapper approaches. It first selects possible optimal feature set and then it is tested by wrapper approach.

Machine learning allows computers the capability to learn without being explicitly programmed. Idea of machine learning came from artificial intelligence. Machine learning gains knowledge and make predictions on data. Machine learning includes collection of data, exploration of data, training the model, evaluation of model performance, and performance improvisation. It attains greater detection accuracy with effective algorithms, applications and frameworks. Machine learning allows computers the capability to learn without being explicitly programmed. This paper uses the concept of machine learning to detect the stroke more accurately.

The main goal of this study is to present a hybrid machine learning framework by ranking of features using correlation and by subset generation. To validate the model, different classifiers are used which are evaluated on the

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

basis of performance metrics that is accuracy. This paper is organized as follows II Related work . III Materials and methods. IV Proposed Framework. V Result Analysis and VI Conclusion.

II. RELATED WORK

Dr. V. Ilango, et al. [1] shows various prediction techniques and tools for Machine Learning has been discussed in this paper. A.K.Shafreen Banu et al. [2] presents several feature selection approaches has been discussed in this survey. Basic method, merits and demerits of feature selection available at recent innovation are explained. N. Leibowitz et al. [3] proposes a novel quantitative approach to the assessment of proprioception deficits in stroke patients. Here an automated protocol is designed and implemented where a magnetic motion tracking system and a sensor attached to each of the patient's hands. The system provides a reliable quantitative measure of upper limb proprioception, and considerable advantage over the conventional means applied within the clinic. G.E. Wang et al. [4] To address feature selection the comparison of hybrid methods based on SVM is presented in this paper. It specialize in the foremost relevant features for use in representing information so as to delete those features considered as irrelevant. Guixiong Liu et al. [5]: This paper proposes a multi-class classification method of support vector machines based on double binary tree (DBT-SVM) to solve the problems of 'irreversibility', 'error accumulation' and randomicity of classification order in multi-class classification of support vector machines based on binary tree (BT-SVM). Sancho Salcedo-Sanz et al. [6]: This paper uses Support Vector Machine (SVM) as classifier. And proposes two approaches for feature selection and ranking based on Simulated Annealing (SA) and Walsh analysis.

Raid Alzubi et al. [7]: The proposed approach consists of using the CMIM filter feature selection method for selecting a subset of the most informative SNPs. Affected samples and healthy samples of SNP data are distinguished between supervised classification steps. Yvan Saeys et al. [8]: This paper reviewed the main contributions of feature selection research in a set of well-known bioinformatics applications. Two main issues in the bioinformatics domain are: the large input dimensionality, and the small sample sizes. T. Kalaiselvi et al. [9]: This paper describes the algorithm of GSO and reviewed GSO based on the several field like clustering, optimization problem, multicast routing problem (MQMR) problem and multi-robot based problems. The result confirms that the outcomes of GSO are better when compared to the other optimization methods namely, PSO, ACO and GA but it has high computational complexity. Debasish Ghose et al. [10]: This paper presents multimodal function optimization, using a GSO algorithm, with the applications to collective robotics. Divya Jain et al [11]: For the diagnosis of breast cancer and diabetes a hybrid machine learning framework is used using efficient feature selection and classification technique. It identifies significant risk factors related to the chronic disease datasets by applying different feature selection techniques. ReliefF Feature Ranking with Principal Component Analysis (PCA) method is used in this paper. Yonglai zhang et al. [12]: This paper proposed a feature selection model for detecting the risk of stroke. Standard deviation is used as a parameter to rank the features, and SVM is used for classification, which was effective in feature selection of stroke. Mazaher Ghorbani [13]: In this paper, the mining of frequent item sets along with their temporal patterns. New notion of TCs is presented to consider time hierarchies in data mining process. It enables us to find different kinds of temporal patterns. A new threshold, called density, was proposed to mine valid patterns and solve the problem of overestimating the time periods.

III. MATERIALS AND METHODS

A. Stroke Dataset

The stroke dataset is downloaded from the kaggle site. Dataset consist of 12 attributes. They are id, age, hypertension, gender, heart disease, ever married, work type, residence type, bmi, smoking status, and the label. It contains 1860 rows and 12 columns. The dataset Description in given in Table 1.

B. Hybrid Feature Selection and Subset Generation

Feature selection process is a data pre-processing method that eliminates the undesirable features from a dataset and selects relevant and non-redundant subset of original attributes. It is often used prior to applying any machine learning task and eliminates superfluous attributes from a dataset which do not have any significance for model construction in a classification process. This article applies hybrid feature selection approach method which significantly increases the performance of the classifier as well as in reduces the computational complexity. The hybrid approach consist of feature ranking using correlation method and generating subsets. After the filter stage features are reduced and then

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

from the reduced feature set subset of features is generated. And the generated subsets is given to SVM for finding the best subset that gives the highest accuracy. Then the highest accurate subset is given for classification

S.NO	ATTRIBUTE	FORMAT
1	ID	NUMERIC
2	AGE	NUMERIC
3	HYPERTENSION	NUMERIC
4	EVER-MARRIED	NUMERIC
5	WORK-TYPE	CATEGORICAL
6	RESIDENCE TYPE	NUMERIC
7	BMI	NUMERIC
8	SMOKING STATUS	CATEGORICAL
9	GENDER	CATEGORICAL
10	HEART DISEASE	NUMERIC
11	AVG-GLUCOSE-LEVE;	NUMERIC
12	STROKE	NUMERIC

Table 2. Attribute details of stroke dataset

C. Classfication

Classification is a widely used method by researchers to construct a classification model from a particular dataset. During classification, the dataset is divided into train data and test data. The train data is used identify the relationship between each feature and class label to create model for disease detection. The test data is used to evaluate the performance of model created.

For example, when filtering emails spam or not spam, when watching at transaction data, fraudulent, or authorized. In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and therefor the values (class labels) in classifying attributes and uses it in classifying new data. There are a number of classification models. Classification models include SVM logistic regression (LR), decision tree (DT), Random Forest (RF), gradient-boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes. In this paper for classification we use four different classifiers they are LR, SVM, Naive Base (NB), RF.

IV. PROPOSED FRAMEWORK

In the proposed model, first the original dataset is loaded into the system for data pre-processing. Data Pre-processing is a technique that is used to change the raw data into a clean data set. In other words, the data that is collected from different sources is gathered in raw format which is not feasible for the analysis. Forms of data pre-processing are data cleaning, data integration, data transformation and data reduction. Data from the real world tend to be incomplete, noisy, and inconsistent.

Data cleaning routines try to fill in missing values, smooth noise while identifying outliers and correct inconsistencies in the data. The missing values are handled during the transformation stage. Data in the real world are rarely clean and homogeneous. Typically, they have a tendency to be incomplete, noisy and inconsistent. It is foremost to be handled as they might cause wrong prediction or classification for any given model getting used. Many different methods are used to handle missing values which are deleting the row, replacing value by mean, median and mode values and predicting missing values by machine learning algorithms. In this work the missing values are handled during data pre-processing stage. That is, the features which have more than 35 percent of missing values are replaced by mean value. ID is also removed. This stage perform checking for text data, duplicate columns and redundant records. It also performs encoding of string values or categorical values in the dataset. Then after data-pre-processing feature selection is applied on dataset for reducing dimensionality of data. In this filter feature selection method

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

features are ranked based on the correlation. That is the correlation between the feature and the label is calculated. Correlation is a measure of the linear correlation between two variables. It has a value between +1 and -1, where 1 is total positive linear correlation and -1 is total negative correlation. We select the top n features having high correlation between the feature and label. Then the features are ranked in descending order of their correlation value. Those features whose correlation value below 0.035 is eliminated. After the removal of irrelevant features, the dataset with reduced set of features is passed to Subset generation stage. In this stage we generate subsets with 2 features, 3 features etc. and each subset is given to SVM then calculates the accuracy. For generating subsets idea of candidate generation gorithm is taken [13]. Here, subset with seven features is taken as final set of subsets. From that subset who is getting highest accuracy is taken as final subset. This final subset is given fo classification. For classification we use 4 different algorithms. That is SVM, RF, LR, and NB. Figure 1 illustrates process of the hybrid model for the stroke detection.

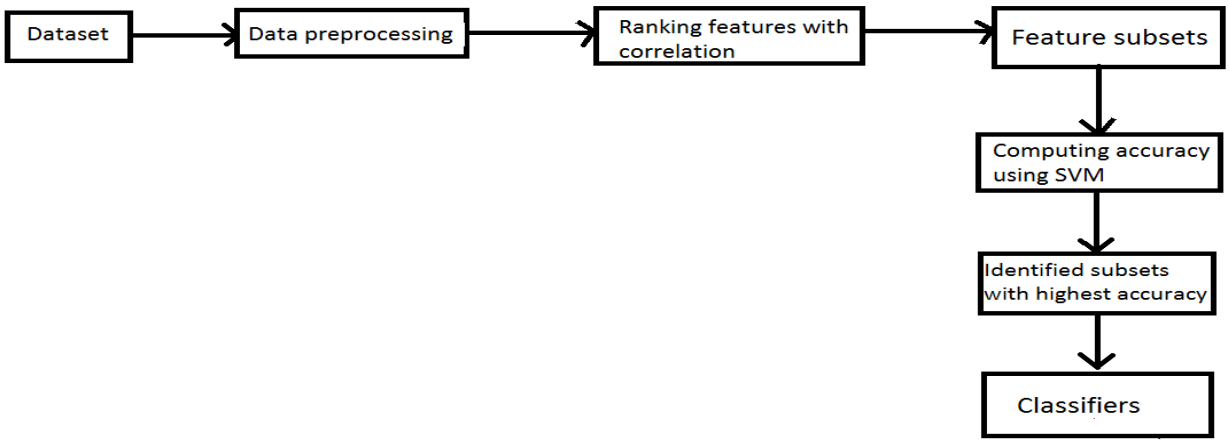

Fig. 1. Proposed Framework model

V. RESULT ANALYSIS

The main focus of this work is to demonstrate that machine learning models, which are widely used and are feasible for the stroke disease diagnosis and find a standard model with good diagnosis performance from multiple machine learning models. In order to get the real test results, I have taken cross-validation in the experiment to improve the accuracy of the model. To evaluate the performance of the proposed hybrid feature selection method based on Correlation and subset selection a stroke data set with 12 features were used.

For this experimental analysis, the data before subset selection is applied to four different classifiers to evaluate their performance. Before Subset selection Support Vector Machine and Logistic Regression provides better accuracy than the other classification models. The data after Subset selection is also applied to these classification models to compare their results. After the data is processed, the SVM Still keeps a high accuracy.

By comparing the data before and after the implementation of the hybrid feature selection method, it is found that the accuracy of SVM, RF and LR has been improved as shown in figure2.The accuracy of RF is improved from 75% to 79%, the accuracy of LR is improved from 77% to 79% and the accuracy of SVM is improved from 80% to 83%. Whereas NB classification model provides the same classification accuracy before and after Subset generation. From these four classification algorithm SVM achieves the highest classification accuracy and thus SVM can be chosen as the final classification algorithm for classifying the stroke patients. Table 2 shows the accuracy comparison before and after subset generation

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

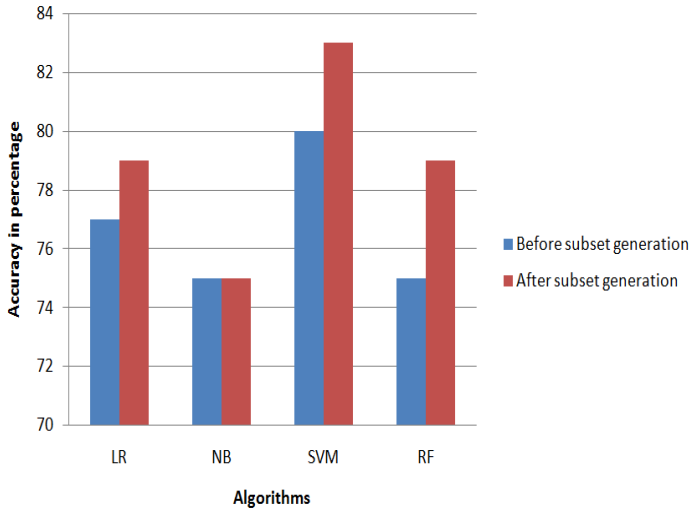

Fig. 2. Accuracy Comparison

Classifier	Before subset generation	After subset generation
SVM	80	83
NB	75	75
LR	77	79
RF	75	79

Table 2. Accuracy Comparison

VI. CONCLUSION

Stroke occurs when a blood clot forms in another part of the body, often the heart or arteries in the upper chest and neck and moves through the bloodstream to the brain. The clot gets stuck in the brain's arteries, where it stops the flow of blood and causes a **stroke**. Medical diagnosis is always essential for stroke patients. And also lab tests and imaging is required. Treatment can help but this condition can't be healed. So before the disease occurs if we can detect the symptoms we can cure it. Thus , a hybrid feature selection model for stroke detection is proposed. This method is more accurate and computational complexity is less compared to other methods that detects the stroke. Among 12 features, five features are removed. From the analysis of experimental results it is found that the most suitable machine learning model for stroke disease is SVM, LR and RF. SVM achieves an accuracy of 83% and both LR and RF achieves an accuracy of 79%.

REFERENCES

[1] Dr. V. Ilango et al." Predictive Analytics in Health Care Using Machine Learning Tools and Techniques" International Conference on Intelligent Computing and Control Systems'17.

[2] A.K.Shafreen Banu et. al. "A Study 0f Feature Selection Approaches for Classification" IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems'15.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

[3] N. Leibowitz et al." Automated measurement of proprioception following stroke" Research gate article on Disability and Rehabilitation'14.
[4] G.E. Wang et al. "Comparison of Feature Selection Approaches based on the SVM Classification" IEEE conference on Industrial Engineering and Engineering Management'2008.
[5] Guixiong Liu et al. "Multi-Class Classification of Support Vector Machines Based on Double Binary Tree" IEEE International Conference on natural computation'08
[6] Sancho Salcedo-Sanz et al. "Feature Selection Methods Involving SVMs for Prediction of Insolvency in Non-life Insurance Companies" International conference on Financial Economy
[7] Raid Alzubi et al. " SNPs-based Hypertension Disease Detection via Machine Learning Techniques" IEEE conference on neural networks and learning systems
[8] Yvan Saeys "A review of feature selection techniques in bioinformatics" journal on bioinformatics'07
[9] Kalaiselvi Thiruvenkadam "A Review on Glow worm Swarm Optimization" International Journal of Information Technology'17
[10] K.N. Krishnanand et al. "Glow-worm swarm based optimization algorithm for multimodal functions with collective robotics applications" International Journal on Multiagent and Grid Systems'06
[11] Vijendra Singh et al. "Diagnosis of Breast Cancer and Diabetes using Hybrid Feature Selection Method" 5th IEEE International Conference on Parallel, Distributed and Grid Computing'18
[12] Yonglai zhang et al. "Risk Detection of Stroke Using a Feature Selection and Classification Method" IEEE Translations and content mining'18
[13] Mazaher Ghorbani"A New Methodology for Mining Frequent Itemsets on Temporal Data" IEEE transactions on engineering management.
[14] Shutao Li," Extremely High-Dimensional Feature Selection via Feature Generating Samplings" IEEE transactions on cybernetics'14
[15] Yuanning Liu, "An Improved Particle Swarm Optimization for Feature Selection " Journal of Bionic Engineering' 11
[16] Dr. A.Anushya et al. "An Analysis of Particle Swarm Optimization for Feature Selection on Medical Data" International Conference on Energy, Communication, Data Analytics and Soft Computing'17
[17] Xiaojun Chen et al. "Local Adaptive Projection Framework for Feature Selection of Labeled and Unlabeled Data" IEEE Transactions on neural networks and learning systems'18
[18] Song Zhang" Prevalence of stroke and associated risk Factors among middle-aged and older Farmers in western China" Environmental Health and Preventive Medicine'17
[19] Xiaomei Wu" Prevalence, Incidence, and Mortality of Stroke in the Chinese Island Populations: A Systematic Review" International Journal of Computer Applications'14
[20] XIN-SHE YANG"Nature-Inspired Optimization Algorithms"