IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 11, November. 2019, Series -I, PP 15-23

Hybrid Multimodal Evolution Dense Support Vector Machine based classification for Parkinson's disease diagnosis

Sherina T.M[1], Mr. Senthil kumar .B[2]
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [1]
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[2]
Corresponding Author: Sherina T.M

Received 19 October 2019; Accepted 5 November 2019

Abstract-Parkinson's disease research work is carried out in a systematic manner with the help of data mining techniques. Machine learning algorithms are used in the proposed work to diagnose the Parkinson's disease. Several researches made for disease classification. The proposed work aims to concentrate on the above issues by developing a new hybrid algorithm. For that, the proposed work utilizes a new hybrid classification algorithm named as Hybrid Multimodal Evolution Dense Support Vector Machine (HMEDSVM) classification algorithm to check the ability. The fitness is measured in this case. The HMEDSVM selects optimal feature in case of high dimensional incomplete datasets. The Parkinson's disease has several set of datasets which are complicated to process. The proposed work utilizes the dataset from the UCI repository and performs the classification algorithm. The experiments are carried out with real time dataset using Weka and dot net platforms. The result of the proposed work is compared with existing classification algorithms. The proposed work is better in different parameters such as accuracy, detection time and reduced error rate. The results show that the proposed hybrid classifier algorithm achieves remarkable dimensionality reduction on the Parkinson's disease medical data and after stabilization; the data are classified as disease and normal classes. The results are evaluated based on different evaluation metrics and the accuracy of classification using existing SVM model is 69.89%, the accuracy of Naïve bayes model is 96.56% and the accuracy of proposed work is achieved to be 99.89%.

Key Words: HMEDSVM, Disease Classification, Data Mining.

I. INTRODUCTION

Parkinson's disease is a disorder of the central nervous system. This disease caused by a complaint in the central nervous system which affects the human movement. This affects the human brain. The indications of the disease increase due to the death of dopamine producing cells in the mid brain. This disease does not have any known cause and the signs and symptoms can be different for everyone. The symptoms differ in different stages of the disease but generally they all involve cognitive and behavioral problems. Parkinson's disease is the second most common neurodegenerative. Alzheimer's disease affects many people worldwide. Due to the death of nigral neurons, patients experience both motor and non-motor symptoms, affecting their quality of life. The reasons for the cell death are still poorly understood, and there is currently no cure for Parkinson's disease. Physicians try to manage patients' symptoms by introducing medications therapies, using antiparkinson medications. Physicians need to carefully prescribe medications therapies since the prolonged intake—in particular of higher dosages of antiparkinson medications—can have significant side-effects. Changes of the status of Parkinson's disease patients through time is a result of the natural progression of the disease and the medications that the patients are prescribed in order to keep their status stable as long as possible.

II. RELATED WORK

In this paper [1], the author Describes Classification is the process of supervised learning process, which is an important task in pattern recognition. Imbalanced class distribution is always a trouble in achieving high accurate classification. In this paper [2], the author Describes Personal satisfaction of patients with Parkinson's ailment debases fundamentally with malady movement. This paper introduces a stage towards customized the executives of Parkinson's sickness patients, in light of finding gatherings of comparative patients. [3], the author Describes Parkinson's sickness is a neurodegenerative issue that influences individuals around the world. Cautious administration of patient's condition is critical to guarantee the patient's autonomy and personal satisfaction. [4], Personal satisfaction of patients with Parkinson's ailment corrupts fundamentally with illness movement. This paper displays a stage towards customized medication the executives of Parkinson's malady patients, in light of finding gatherings of comparable patients. [5], Arrangement principles and guidelines depicting intriguing subgroups are significant parts of expressive AI.

In this paper [6], the author Describes Electronic wellbeing records contain a lot of longitudinal information that are significant for biomedical informatics inquire about. The use of AI is a promising option in contrast to manual examination of such information. [7], the author proposed a study with some features based on entropy, energy and intrinsic measures of the handwriting skills of an individual. The main objective of this paper is to present the PaHaW Parkinson’s disease handwriting database which contains many handwriting samples from Parkinson’s disease (PD) patients. [8], the author Describes The outcomes exhibit that engine and autonomic side effects are the most enlightening for assessing the personal satisfaction of Parkinson's infection patients. [9], authors applied Linear Discriminant Analysis Support (LDA), Support Vector Machines and k-nearest neighbor (k-nn) upon local field potentials sensed from an implanted deep brain stimulation device. Authors used leave one out technique feature selection and performed analysis. [10], the author describes in principle learning, rules are commonly instigated in two stages, rule refinement and guideline choice. [11], author presented a study with two main contributions: firstly, they used a multiple regression normalization strategy to identify differences in spatial-temporal gait features between PD patients and control (healthy) individuals. [12], author employed evolutionary algorithms to provide clinically relevant and objective measures to identify PD both in humans and animal models.

In this paper [13], author presented a new method for the diagnosis of PD based on continuous phonation samples, which were used as attributes. [14], author presented an Enhanced Probabilistic Neural Networks (EPNN), a machine learning technique that make use of local decision circles surrounding training samples to control the spread of the Gaussian kernel. [15], author demonstrated a new method based on SVMs and Bayesian networks to separate IPS from APS (atypical parkinsonian syndromes) that makes use of the 18FFDG PET dataset that allows assessing the glucose metabolism of the brain. Their methodology achieved an accuracy rate over 78%, a reasonable result between sensitivity and specificity, suggesting the proposed method is suitable to assist the diagnosis of PD. [16], author proposed to employ a combination between smart home and machine learning technologies to observe and quantify the behavioral changes of PD patients.

III. PROPOSED WORK

Parkinson’s diseases are very common and widespread worldwide. Several researches carried on diagnosis and prediction of Parkinson’s diseases in various fields. Parkinson’s disease is a chronic, progressive and neuron-degenerative illness. Detection and diagnosis of such disease need more data analysis process, which can be handled by data mining techniques. The research work is carried out in a systematic manner with the help of data mining techniques. Machine learning algorithms are used in the proposed work to diagnose the Parkinson’s disease. The overall processes that are involved in this work are shown in the figure 3.1. This research work is based on the input from the UCI repository which involves 240 instances and different type of records. Each record has 48 attributes.

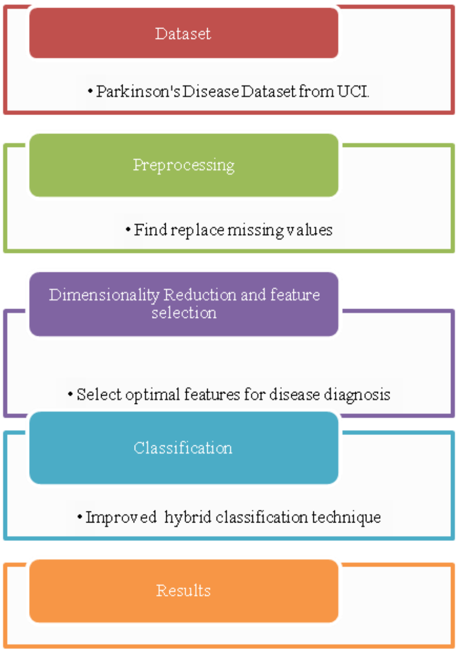


Figure 3.1: Empirical Methods for Parkinson’s disease Classifications

3.1 Process of Parkinson's disease Classification

The collected input data is given as input to the research work to classify the data as risky Parkinson's or low Parkinson's or a normal one by proceeding in the various processes. Process 1 is preprocessing, process 2 is dimensionality reduction and in process 3 classifications and disease detection is done. The research work is based on the input from the UCI repository which involves 240 instances. Each record has 48 attributes. A novel technique is proposed to classify the Parkinson's data set. The classification method is HMEDSVM Classification algorithm.

Process I: Preprocessing

The data taken from UCI repository undergoes preprocessing where missing value and not a number constraint are checked and replaced by value generated by the preprocessing algorithm.

Process II: Dimensionality Reduction

The pre-processed data's are applied for dimensionality reduction process with the help of the hybrid algorithm. The preprocessed data is fed into a hybrid algorithm named as Enhanced multimodal evolution (EME). This algorithm is used for creating subset of the main records. This algorithm is used in the second process.

Process III: Classification

The results from the process 2 are applied to the classification process. The proposed work utilizes a new hybrid classification algorithm named as Hybrid Multimodal Evolution Dense Support Vector Machine (HMEDSVM) classification algorithm to check the ability. The fitness is measured in this case. The results of the proposed work are classified into three stages given below.

a. Low Parkinson's
b. Risky Parkinson's
c. Normal

All the proposed methods differ in two aspects. They are the usage of the hybrid algorithm as well as the classifier algorithm. Initially, the hybrid algorithm used is Hybrid Multimodal Evolution and Dense Support Vector Machine is used for classification.

3.1.1 Data Set Collection

The fist process is the collection of data for disease classification. For Parkinson's disease, there are several types of datasets available such as speech dataset, handwriting dataset etc., in this proposal, data mining techniques are used, so selection of appropriate dataset is much important. The database directory from the UCI learning repository contains a Parkinson's database, corresponding test set, and corresponding documentation. The chosen Parkinson's data set contains attributes which could be used to classify the Parkinson's disease records supplied by the UCI repository. Experiments are performed on the Parkinson's datasets that are available at the UCI website for researchers to study and carry on such experiments. The following table 3.1 shows the characteristics of Data set collected from UCI repository.

Table 3.1: Characteristics of Dataset

Data Set Type	Multivariate, Domain-Theory
Attributes Type	Categorical, Real
Associated Tasks	Classification
Number of Instances	240
Number of Attributes	48

To ensure that the patterns derived are as accurate as possible, it is essential to improve the quality of the datasets in the pre-processing stage. Most real-life data sets contain a certain amount of redundant and duplicate data, which does not significantly contribute to the formation of important relationships. This redundancy not only increases the dimensionality of the data set and slows down the data mining process, but also affects subsequent grading performance. So, attribute selection is more important process. The attribute selection is the process of eliminating redundant attributes that are considered irrelevant for the task of mining. However, the presence of attributes those are not useful for classification may interfere with the attributes relevant to degrading classification performance. This is due to the noise provided by these additional attributes and increases the level of difficulty. The objective of attribute selection is therefore to search for a worthy set of attributes that produce comparable classification results to the case when all the attributes are used. In addition, a smaller set of attributes also creates less complicated, easily understandable and even visualized humans. It has to be noted that for a data set with n attributes, there are 2n–1 possible subsets. Therefore, an exhaustive search for an optimal set of attributes would be time-consuming and computationally expensive if n is large.

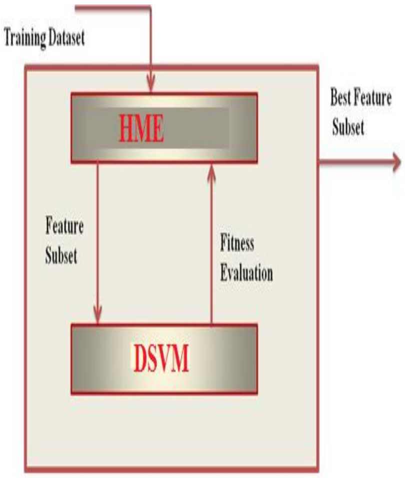
3.1.2 Process I: Preprocessing

This part is important because in the data mining method, the quality of data set is pruned to obtain the significant result. In this part all the missing values are processed, and the outlier values are detected. Preparing the input data is often the most complicated part. Part of the complexity is the challenge of choosing the right data and the right examples for the training process. The pre-processing step is necessary to resolve several types of problems including noisy data, redundant data, missing data values, etc. The high quality data will lead to high quality results and reduced costs for data mining. Missing data should be pre-processed so as to allow the whole data set to be processed by a required algorithm. Generally, most of the benchmarking algorithms work well on the discrete attributes. If the features are continuous, the algorithms can be integrated to create discrete attributes

3.1.3 Process II: Dimensionality Reduction

Data mining is the computerized analyses of enormous volumes of data with a focus to observe and identify the associations and information that are already resume being an implication of the data. The discovery of essential and targeted information from the relatively hefty data through Data mining process may facilitate the use of parallel and distributed computation ambience to improvise both presentation and excellence of the data chosen. The unique characteristics of medical databases that pose challenges for data mining are the privacy-sensitive, heterogeneous, and voluminous data. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the curse of dimensionality. Dimension reduction is a compilation of statistical strategies that minimizes the measurement of the data simultaneously, by sustaining the sequence of relevant information. Generally, government agencies, industries and academia are known for executing on high-dimensional data. At this point, it is noteworthy that the enormous dimension of hefty volume act with adequate potentiality to fetch two issues among the assortment of data. Primarily, to prevail over the predicament of dimensionality, this is obviously allowed to exhibit how the spaces of high dimensionality are intrinsically meager and scanty even amidst the entire voluminous observations made. Secondly, this determines the prudent and frugal distribution of information within the procured data.

The strategies of dimension reduction edify to be the best means to tackle these predicaments to several extents by minimizing the set of variables to a less significant set of either existing or unique variables. Here the unique variables evolve into linear combinations or perhaps, even non-linear combinations of the existing ones. Data Visualization is comprehensively credible only when the unique dimension is relatively minute in size as this extends its assistance for data methoding substantially. This is absolutely feasible if the response is specified. If not it may face a futile termination. The figure 3.2 shows the method for dimensionality reduction. Original data set after preprocessing is received as input to generate the subset. The generated subset of data is evaluated to check for the goodness of the subset. If the stopping criterion is not reached the next subset is generated until the subset is evaluated to be good and the stopping criteria are reached. Finally the valid data result is achieved.



3.2.4 Process III: Classification

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. The term could cover any context in which some decision or forecast is made on the basis of presently available information. Classification method is recognized method for repeatedly making such decisions in new situations. If the problem is a concern with the construction of a method that will be applied to a continuing sequence of cases in which each new case must be assigned to one of a set of pre defined classes on the basis of observed features of data. Creating a classification method from a set of data for which the exact classes are known in advance is called method recognition or supervised learning. The contexts in which a classification task is crucial include, three main historical strands of research can be identified:

- Statistical
- Machine Learning
- Neural Network

All the above groups have some objectives in common. They have all attempted to develop methods that would be able to handle a wide variety of problems and to be extremely general used in practical settings with proven success.

3.2. SVM Classification

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for classification and regression challenges. However, it is mainly used in classification problems. In this algorithm, each data element is traced as a point in the n-dimensional space (where n is the number of available functions) with the value of each feature which is the value of a coordinate date. Then, classification is performed by finding the risky-plane that differentiate the two classes very well which is shown in the following figure 3.3.

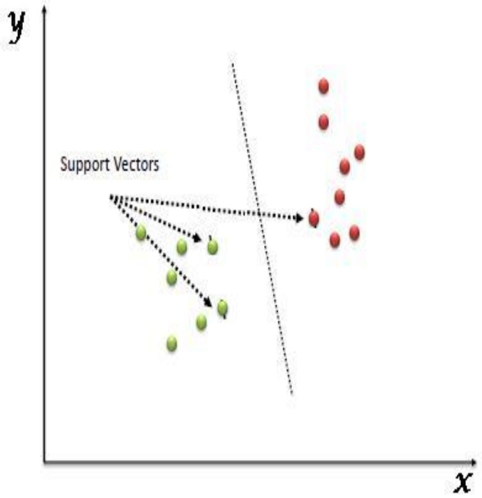


Figure 3.2: Support vectors with two classes separated by risky-plane

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (risky-plane/ line).

HMEWITH DSVM Classification

In today's data mining applications, support vector machines (SVM) are considered a must try—it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace. In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating riskyplane $f(x)$ that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance xn can be classified by simply testing the sign of the function $f(xn)$; xn belongs to the positive class if $f(xn) > 0$. Because there are many such linear riskyplanes, what SVM additionally guarantee is that the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the riskyplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the riskyplane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an infinite number of riskyplanes, only a few qualify as the solution to SVM. The reason why SVM insists on finding the maximum margin riskyplanes is that it offers the best generalizationability. It allows not only the best classification performance (e.g., accuracy) on the training data, but also leaves much room for the correct classification of the future data. To ensure that the maximum margin riskyplanes are actually found, an SVM classifier attempts to maximize the following function with respect to _**w** and *b*:

$$L_P = \frac{1}{2}\|\vec{\mathbf{w}}\| - \sum_{i=1}^{t}\alpha_i y_i(\vec{\mathbf{w}}\cdot\vec{\mathbf{x}}_\mathbf{i} + b) + \sum_{i=1}^{t}\alpha_i$$

where t is the number of training examples, and αi , i = 1, . . . , t, are non-negative numbers such that the derivatives of $L\ P$ with respect to αi are zero. αi are the Lagrange multipliers and $L\ P$ is called the Lagrangian. In this equation, the vectors _**w** and constant *b* define the riskyplane. In SVM Jinyan and Huiqing, 2002; Mahmoud and Awaida, 2009 finds the riskyplane with maximum margin in between two classes. The Support Vector Machine (SVM) is actually based on learning with kernels some of which form the support vectors. A great advantage of this technique is that it can use large input data and feature sets. Thus, it is easy to test the influence of the number of features on classification accuracy.

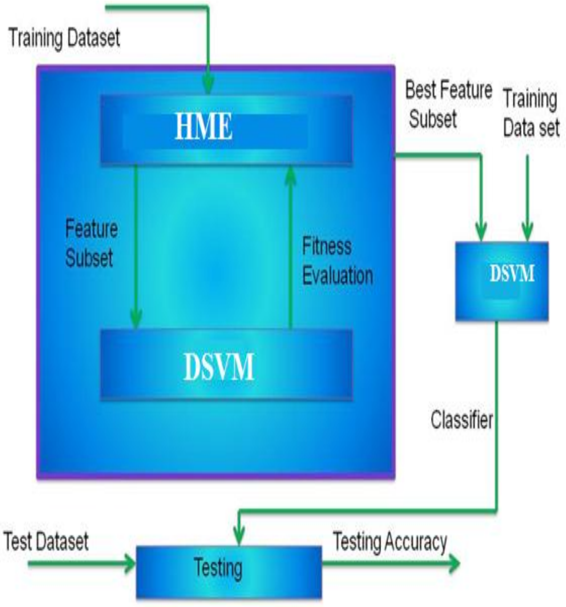


Figure 3.12: Framework of HMEDSVM with Dense feature

The framework of HMEDSVM with dense classification is shown in figure 3.12. The Parkinson's dataset is given as input as a training data to the Hybrid Multimodal Evolution algorithm. A feature subset is extracted by using the DSVM method. Based on the fitness evaluation function, the output is to provide the best feature subset. This output along with the other training data set, the DSVM classifier is used to perform the classification of the Parkinson's data as "Risky Parkinson's" and Low Parkinson's". The classification accuracy is verified in the testing process by comparing the test dataset with that of the classified Parkinson's data to produce the accuracy of the testing results. Figure 3.13 shows the implementation on Classification.

Various steps involved in the implementation are

- Load data from the database
- Preprocessing stage
- Feature selection stage where the errors are stabilized
- Classification stage where the data is classified as Risky, low and normal one.
- Show performance stage where the multiple performance metrics, confusion matrix, hit/miss values are shown

On proper implementation of HMEDSVM with dense feature, the classification accuracy is evaluated to be 67.97%.

IV. RESULTS AND DISCUSSSION

4.1 Introduction

A HMEDSVM classifier produces output with two class values or labels, such as Yes/No and 1/0, for given input data. The class of interest is usually denoted as "positive" and the other as "negative".

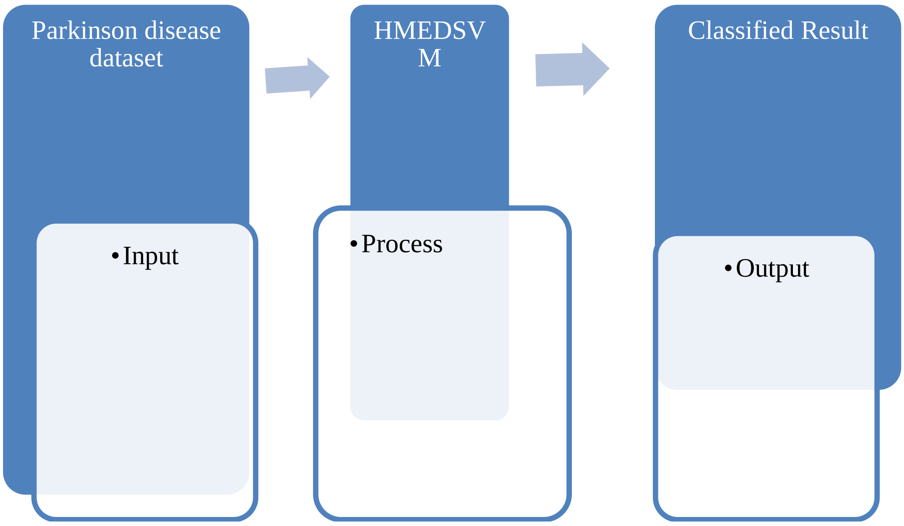


Figure 4.1: HMEDSVM Classifier

A HMEDSVM classifier produces output with two classes for given input data.

Test dataset for evaluation

A dataset used for performance evaluation is called a test dataset. It should contain the correct labels (observed labels) for all data instances. These observed labels are used to compare with the predicted labels for performance evaluation after classification.

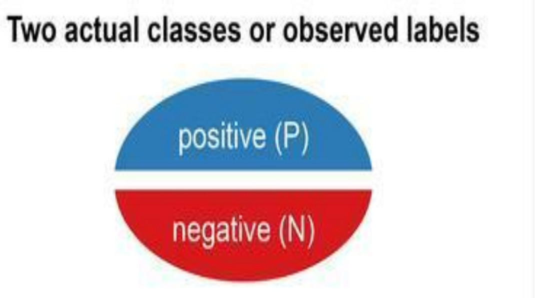


Figure 4.2: Classes of Proposed Classifier

In proposed classification, a test dataset has two labels; positive and negative.

Predictions on test datasets

The predicted labels will be exactly the same if the performance of a HMEDSVM classifier is perfect, but it is uncommon to be able to develop a perfect HMEDSVM classifier that is practical for various conditions.

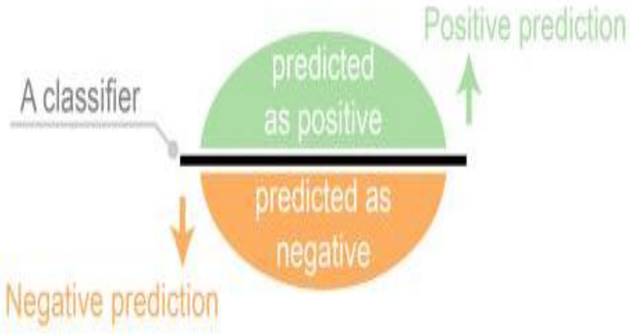


Figure 4.3: Perfect classifier prediction

The performance of a HMEDSVM classifier is perfect when it can predict the exactly same labels in a test dataset. Hence, the predicted labels usually match with part of the observed labels.

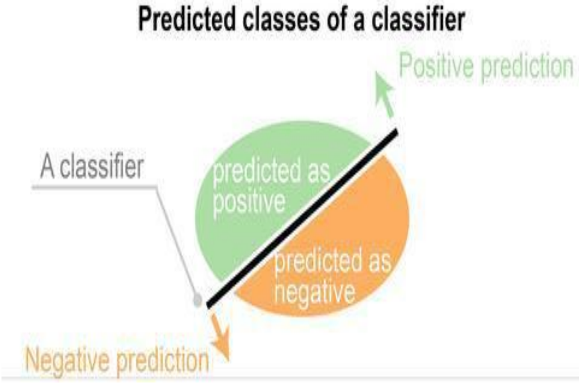


Figure 4.4: Predicted classes of a classifier

The predicted labels of a classifier match with part of the observed labels.

V. RESULTS

The results show that the proposed hybrid classifier models achieves remarkable dimensionality reduction on the medical data and after stabilization, the data are classified as Parkinson's disease and normal classes. The results are evaluated based on ten evaluation metrics and the accuracy of classification using existing SVM model is 67.97%, the accuracy of Naïve bayes model is 97.97% and the accuracy of proposed work is achieved to be 99.89%.

VI. CONCLUSION AND FUTURE WORK

The objective of this research work is aimed to show the classes of Parkinson's disease from the available raw medical dataset for accurate disease diagnosis. The proposed Parkinson's disease classification system will increase the accuracy and detection ratio with the help of data mining techniques. This system will reduce the diagnosis time and false detections. This system is useful for medical domain and also for clinical laboratory analysts which gives accurate and precision results. The results show that the proposed hybrid classifier model achieves remarkable dimensionality reduction on the medical data and after stabilization; the data are classified as normal and disease classes. The results are evaluated based on ten evaluation metrics and the accuracy of classification using SVM technique is 67.97%, the accuracy of naïve bayes model is 97.97% and the accuracy of proposed Hybrid classifier HMEDSVM is achieved to be 99.89%. In future the work can be extended for different classifiers and with various evaluation algorithms. In the preprocessing phase discretitaztion methods can be implemented to improve the various constraints imposed on large dimensional medical data set. In the dimensionality reduction phase values assigned for the control parameters in differential evolutionary algorithm can be modified to improve global optimization. These models could be test in big data scenarios. The futuristic perspective of this approach may rise up high due to its reliable efficacy as it forms the fulcrum of this research.

REFERENCES

[1]. B. Senthil Kumar, Dr. R. Gunavathi, “Ranked Cluster Based Adaptive Sampling with Gradient Boosting Classifier for Medical Data”, IOSR Journal of Computer Engineering, 2018.
[2]. A., Miljkovic, D., Konitsiotis, S., Gatsios, D., Lavrač, N., & Robnik-Šikonja, M. (2018). Symptoms and medications change patterns for Parkinson’s disease patient’s stratification. Artificial Intelligence in Medicine (accepted).
[3]. A., Miljkovic, D., Lavrač, N., & Robnik-Šikonja, M. (2018). Analysis of medications changes in Parkinson’s disease progression data. Journal of Intelligent Information Systems. doi:10.1007/s10844-018-0502-y
[4]. Valmarska, A., Miljkovic, D., Konitsiotis, S., Gatsios, D., Lavrač, N., & Robnik-Šikonja, M. (2017). Combining multitask learning and short time series analysis in Parkinson’s disease patients stratification. In Proceedings of the Conference on Artificial Intelligence in Medicine in Europe (pp. 116–125). Springer.
[5]. A., Lavrač, N., Fürnkranz, J., & Robnik-Šikonja, M. (2017). Refinement and selection heuristics in subgroup discovery and classification rule learning. Expert Systems with Applications, 81, 147–162. doi:10.1016/j.eswa.2017.03.041
[6]. J., Papapetrou, P., Asker, L., & Boström, H. (2017). Learning from heterogeneous temporal data in electronic health records. Journal of Biomedical Informatics, 65, 105–119.
[7]. Drotar P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M. Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. Artif Intell Med 2016:39–46.
[8]. A., Miljkovic, D., Robnik-Šikonja, M., & Lavrač, N. (2016). Multi-view approach to Parkinson’s disease quality of life data analysis. In Proceedings of the International Workshop on New Frontiers in Mining Complex Patterns (pp. 163–178). Springer.
[9]. Connolly AT, Kaemmerer WF, Dani S, Stanslaski SR, Panken E, Johnson MD, Denison T. Guiding deep brain stimulation contact selection using local field potentials sensed by a chronically implanted device in Parkinson's disease patients. 7th international conference on neural engineering 2015:840–3.
[10]. A., Robnik-Šikonja, M., & Lavrač, N. (2015). Inverted heuristics in subgroup discovery. In Proceedings of the 18th International Multiconference Information Society (Vol. 178, pp. 41–44).
[11]. Wahid F, Begg RK, Hass CJ, Halgamuge S, Ackland DC. Classification of Parkinson's disease gait using spatial-temporal gait features. IEEE J Biomed Health Inform 2015; 19(6):1794–802.
[12]. Smith SL, Lones MA, Bedder M, Alty JE, Cosgrove J, Maguire RJ, Pownall ME, Ivanoiu D, Lyle C, Cording A, Elliott JHC. Computational approaches for understanding the diagnosis and treatment of Parkinson's disease. Inst Eng Technol Syst Biol 2015;9(6):226–33. (7).
[13]. Hirschauer TJ, Adeli H, Buford JA. Computer-aided diagnosis of Parkinson's disease using enhanced probabilistic neural network. J Med Syst 2015; 39(11):1–1
[14]. Ahmadlou M, Adeli H. Enhanced probabilistic neural network with local decision circles: a robust classifier. Integer Computer-Aided Eng 2010; 17(3):197–210.
[15]. Segovia F, Illán IA, Górriz JM, Ramírez J, Rominger A, Levin J. Distinguishing Parkinson's disease from atypical parkinsonian syndromes using pet data and a computer system based on support vector machines and bayesian networks. Front Comput Neurosci 2015; 9(137):1–8.
[16]. Cook DJ, Schmitter-Edgecombe M, Dawadi P. Analyzing activity behavior and movement in a naturalistic environment using smart home techniques. IEEE J Biomed Health Inform 2015; 19(6):1882–92.

IOSR Journal of Engineering (IOSRJEN) is UGC approved Journal with Sl. No. 3240, Journal no. 48995.

Sherina T.M." Hybrid Multimodal Evolution Dense Support Vector Machine based classification for Parkinson’s disease diagnosis." IOSR Journal of Engineering (IOSRJEN), vol. 09, no. 11, 2019, pp. 15-23.