A novel approach using machine learning for Parkinson Disease Detection using Biomarkers

Nangoud H T[1],Mrs.Priyanka Padki[2],Satyam Kumar Jha[3], Satyam Kesharwani[4]

1 . Department of Computer Science and Engineering, BNMIT, Bengaluru

2.Assistant Professor, Department of Computer Science and Engineering, BNMIT Bengaluru

3. Department Of Computer Science and Engineering, BNMIT, Bengaluru

4.Department of Computer Science and Engineering, BNMIT,Bengaluru

Abstract: *Parkinson's disease is a neurological illness that affects the nervous system and gets worse with time. Parkinson's disease, on the other hand, has no particular diagnosis and can only be identified by a number of motor symptoms. Patients with Parkinson's disease reported vocal deterioration in almost 90% of cases. In this study, we present a voice and speech signal data-based model for PD identification. The PD Despite having fewer data points, the voice data set employed in this experiment has a high degree of dimension. Our proposed model contained methods that are useful in PD predictionby applying combination of various machine learning techniques . The proposed model very efficiently predicts the parkinsons disease in a patient based on speech biomarkers with high accuracy .*

[Keywords: Parkinson's Disease detection ·Machine Learning ·SVKN Algorithm]

INTRODUCTION

More than 10 million people worldwide suffer with Parkinson's disease. It is the second most common neurological ailment after Alzheimer's[1]. A deterioration in motor and cognitive function is the fundamental characteristic that makes Parkinson's disease distinct. A single test cant be utilized to establish a diagnosis. Instead, doctors must do a thorough clinical study of the patient's medical history. Unfortunately, this diagnostic approach is totally incorrect. The accuracy of early diagnosis (symptoms for less than 5 years) is just 53%, according to the National Institute of Neurological Disorders [2]. Even while an early diagnosis is crucial for effective therapy, it is not significantly more accurate than conjecture. These challenges [3] are the driving force for our work on a machine learning strategy to correctly diagnose Parkinson's, which makes use of a dataset of different speech characteristics (a non-invasive yet distinctive tool) from the University of Oxford. Why are there speech features? Since almost every Parkinson's patient experiences significant vocal degeneration and speech is very indicative and characteristic of the condition, it makes sense to use voice to diagnose the illness (inability to generate prolonged phonations, tremor, hoarseness). Additionally advantageous are voice analysis' low cost, ease of clinical use, and non-invasive nature[4].

Literature Review

The use of speech and voice data in the identification of Parkinson's disease has been the subject of several articles in recent years. An author proposed a model for identifying Parkinson's disease, and a comparison of neural networks, decision trees, regression, and DM neural was done [6]. On MFCC voice recording samples, a study used an SVM classifier and the (LOSO) validation approach to distinguish between patients with Parkinson's disease and healthy people. A two-stage attribute selection and classification model is discussed in another paper[8] and is used to identify Parkinson's illness. Ali proposed a two-dimensional simultaneous sample and feature selection technique for the early detection of PD[1].

Using ensemble learning strategies, a separate author combined numerous classifiers and attained an accuracy of 86 percent [9]. Additionally, models with incredibly high PD detection accuracy have been suggested in other articles. For instance, one study was able to achieve an accuracy of 99.5 percent[12] by combining weighted clustering with a complex valued artificial neural network in their recommended model. The outcomes of these experiments are biased despite their excellent precision. The investigations made advantage of sparse data points and several voice recording samples for each subject [10][11]. PD detection has been the subject of several studies, but more study is required to develop models that are more accurate, resilient, and effective.

Dataset

The public was given access to the data set used to assess the suggested model as PD To train and evaluate the Parkinson's Data, the University of Oxford contributed a speech data collection of 195 occurrences (147 patients with Parkinson's disease and 48 controls). 22 characteristics (components that may be indicative of Parkinson's, such as frequency, pitch, and amplitude/period of the sound wave) were employed to follow the changes in those aspects in line with the PD. Additionally, the MJ Star Fox Foundation Dataset was utilised to train the data for a number of additional speech attributes and to add variation to the dataset.

Proposed Framework

In this paper, we propose a useful approach using PD Speech data. It categorizes Parkinson's Disease. Pre-processing of the PD Speech data collection includes 753 characteristics from a total of 252 participants. Because of this, a very small number of data points are represented by a huge feature space. Data pre-processing is therefore essential to creating classifier models that function well. The following is a description of every technique that was applied:

Rational Regression (RR) : The sigmoid logistic equation is used to describe the likelihood of each class in binary classification. It has weights (coefficient values) and biases (constants). They are represented by a single number (1 for one class and 0 for the other).The model will discover the ideal weights and biases during trainingIt is feasible to calculate the likelihood that data points in a particular class

truly exist using logistic regression. Assume p = P(Y = 1) for a dataset with n feature sets (x1,..., xn) and a binary target class Y. As a result, the following is how the linear relationship between log-odds and characteristics may be expressed: = log b p 1 p, where b stands for the logarithm base and I stands for the technique parameters. One can compute odds using the log-odds exponent. The likelihood that an observation is correct is determined as follows: P equals 1 plus b(0+1x1+...+nxn)

We propose a machine learning model based on the SVKN algorithm. This technique utilizes the anomaly detection and regression capabilities of SVM and classification. There is a common practice in classification to partition samples into training and testing groups based on decision making matrix. Only the true class of each training sample is utilised to train the classifier during training, but the class of each test sample is predicted during testing. It is important to remember that KNN employs the class labels from the training data, making it a "supervised" classification technique. On the other hand, "clustering" approaches, which use unsupervised classification, do not use the training data's class labels. With the 1-nearest neighbour rule, the test sample's predicted class is identical to the real class of its closest neighbour, where mi is x's nearest neighbour if the distance between them is less than one.

The most common true class among the k closest training samples is chosen as the projected class of test sample x for k-nearest neighbours. The D:x decision rule is comprised of this. Testing sample class predictions are tabulated using the confusion matrix, often known as C, which has dimensions. The confusion matrix's diagonal element c is raised by one if test sample x's projected class is correct (i.e., equals). The off-diagonal element c is raised by 1 if the projected class is erroneous, though (i.e., if). The classification accuracy is determined after all test samples have been classified by dividing the percentage of properly recognised samples by the total number of classified samples. The value n-total, where c is a diagonal element of C , indicates the total number of samples that were categorized. Think about research that used machine learning and the X and Y characteristics to categorise 19 samples. The pairwise Euclidean distance between the 19 samples is shown in Table 1The other samples are test samples, whereas sample x11 serves as a training sample. Examples x10 (blue class label), x12 (red class label), x13 (red class label), and x14 (red class label) are the four samples that, when k=4, are most similar to sample x11 (red class label).

When the feature values are modified before classification analysis, a classifier's performance may occasionally improve. The feature transformations of standardization and fuzzification are two that are often utilised. Cross-validated performance calculation. A fundamental rule of classification analysis states that class predictions are not given for data samples used for training or learning. The accuracy will be unreasonably biased in favour of learning if class predictions are produced for training or learning samples. Instead, class predictions are generated for samples that were excluded from the training procedure. Cross-validation, which involves computing classification accuracy for multiple splits of the input samples used during training, is a common method for evaluating the performance of most classifiers. For instance, a set of input samples is divided into 5 divisions during 5-fold (=5) cross-validation training with, to the best degree possible, comparable sample sizes (D1, D2,..., D5). The concept of ensuring uniform class representation across all partitions is known as stratified cross-validation, and it is preferable. In

the beginning, samples from partitions D2,D3,...,D5 are utilised for training whereas samples from partition D1 are used for testing in 5-fold cross-validation. Next, samples from partition D2 are utilized for testing, while samples from groups D1, D3,..., and D5 are used for training. Repeat this until each division has been tested separately. It is also typical to re-partition all of the input samples, let's say 10 times, in order to get a more precise estimate of accuracy. First, a connection between SVM and 1NN is discovered by examination of the SVM classification procedure. This link serves as the theoretical underpinning of SVM-KNN, and Theorem 1 will elaborate on it.SVM classifiers are equivalent to 1NN classifiers, which select one representative point for each class's support vectors.

We looked at the SVM distributions of incorrect samples and discovered that they are typically quite close to the separating hyperplane. This reminds us that in order to increase the classification accuracy, we should use the hyperplane area's information as much as possible. We are aware that samples located close to the region of the separating hyperplane resemble support vectors. In this case, we use the KNN strategy, which considers each support vector as a representative point, as opposed to the SVM technique, which selects just one representative point for each support vector in each class and cannot properly represent the whole class. This implies the possibility of using more relevant data. Particularly, the SVM classifying approach is accessible for samples far from the separating hyperplane whereas the KNN classifying technique is appropriate for data close to it.

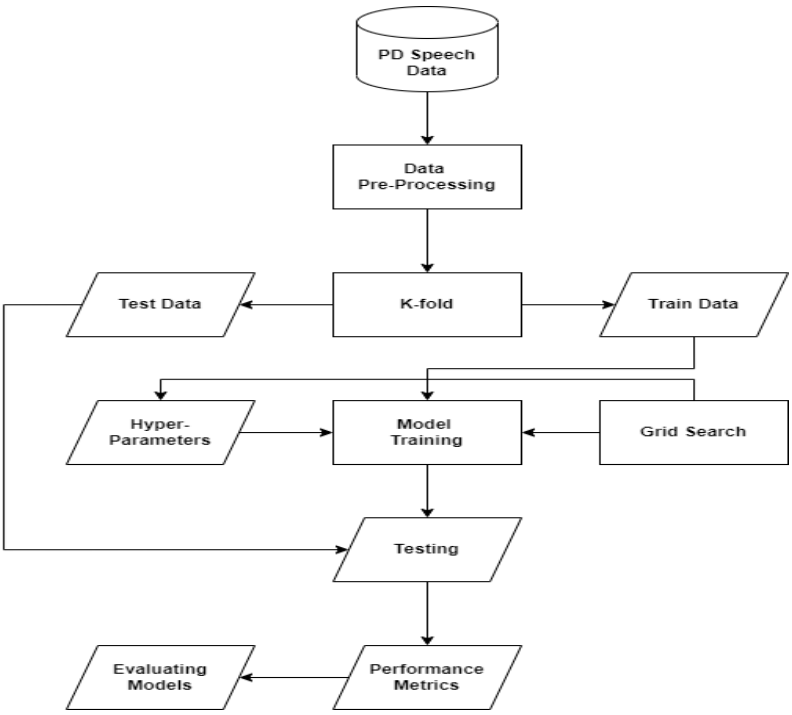

Figure.1. Block Diagram of Proposed Parkinson's Disease Detection Method

In this paper, as shown in the figure(1) a model is proposed that efficiently classifies Parkinson's Disease positive person based on PD Speech data-set.The Data Pre-processing is performed on the PD Speech data-set that contains 753 attributes from 252 subjects in total. This results in a large feature space for a relatively small

number of data points. Thus, data pre-processing is central to high-performing classifier models.

Then we train the data and finally pass them through SVKN algorithm to predict the Parkinson's Disease.

METHODOLOGY

In this model, an algorithm called SVKN is employed. It is created by combining the advantages of the KNN algorithm and SVM. Additionally, data pre-processing is done initially, followed by prediction of Parkinson's using SVKN algorithm.

A. DATASET

The public was given access to the data set used to assess the suggested model as PD To train and evaluate the Parkinson's Data, the University of Oxford contributed a speech data collection of 195 occurrences (147 patients with Parkinson's disease and 48 controls). To track changes in those PD-related features, 22 characteristics (elements that may be symptomatic of Parkinson's, such as frequency, pitch, and amplitude/period of the sound wave) were employed. In order to provide variety to the dataset and train the data for a number of different speech qualities, the MJ Star Fox Foundation Dataset was also used.

B. SVKN ALGORITHM

A connection between SVM and 1NN is discovered by examination of the SVM classification procedure. This link serves as the theoretical underpinning of SVM-KNN, Theorem a will elaborate on it. SVM classifiers are equivalent to 1NN classifiers, which select one representative point for each class's support vectors. We looked at the SVM distributions of wrong samples and found that they are quite close to the parting hyperplane. This gives us that in to increase the classification accuracy, we should use the hyperplane area's information as much as possible. We are aware that samples located close to the region of the separating hyperplane resemble support vectors. Instead of the SVM technique, which selects only one representative point for each support vector in each class and cannot effectively represent the complete class, we utilize the KNN classifier algorithm in this situation, which treats each support vector as a representative point. This implies that more relevant data may be used. Particularly, the KNN classifying technique is good for data close to the separating hyperplane, whereas the SVM classifying method is suitable for samples distant from the hyperplane. The main steps of the new classification method are as follows:

step1 if test = Φ, get z ∈ test, if test = Φ, stop;

step2 calculate g(z) = iyiαik(zi, z) − b;

step3 if |g(z)| > ε, calculate directly f(z) = sgn(g(z)) as output;

if |g(z)| < ε, put it into KNN algorithm to classify;

step4 T ← T − x, go to step1.

"test" denotes both the test set and the empty set in the steps mentioned above. The distance barrier needs to match the criteria of 0 and 1. It should be noted that this approach calculates distance in a feature space with many dimensions. The distance formula used in this case is based on the kernel function and is as follows: (z, zi)2 = 2kn(z, z), Plus 2kn(z, zi) (zi, zi).

RESULTS AND DISCUSSIONS

The proposed model based on SVKN algorithm has provided better performance in terms of accuracy and F1-score in comparison to other contemporary models like SVM and decision tree classifier. In the given data set the pre-processing is done using Gaussian Naive Bayes Algorithm where max Pooling is performed which includes down sampling for removing noise present in the data set and also for null valued features

The given data features are shown in the below figure 2:

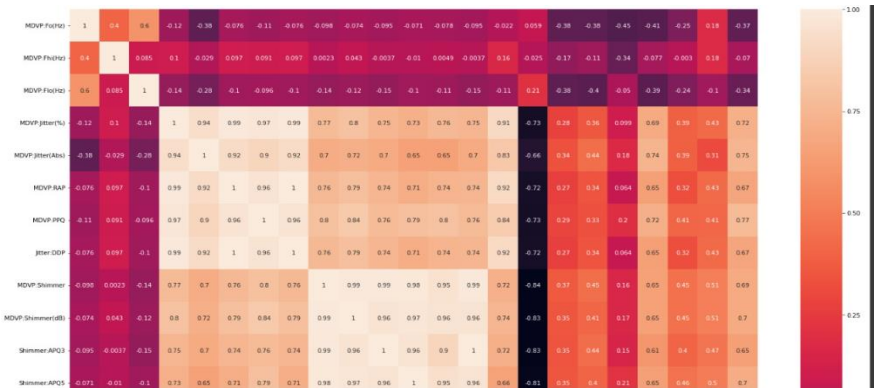

Figure.2. Data set features

Evaluation Standards includes four important value to determine the model's accuracy. True Positive, False Positive, True Negative, and False Negative are the four types.

(i)True Positive:If the suggested framework successfully identifies a sick person, it is considered a true positive and is indicated as TP.

(ii)True Negative: When the suggested framework accurately identifies a healthy person, it is considered a true negative and is shown as TN.

(iii)False Positive:if the proposed framework falsely finds the healthy person as diseased person then it is false positive FP.

(iv)False Negative: A false negative is indicated as FN if the suggested framework depicts a sick individual as the healthy one.

Accuracy, Recall, Precision, and F1-Score are the four metrics that are determined by these four parameters.

Accuracy is crucial metrics for evaluating the proposed framework ,This is given by the formula

$$\text{Accuracy} = \frac{Number\ of\ Diseased\ Persons}{Total\ number\ of\ persons} \quad (1)$$

The classifier that properly classifies propositions as accurate positives defines recall. Given by the formula is this.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

Precision is used to find how positive identification is correct in classification This is given by the formula

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

The classifiers' test accuracy is evaluated using the F1-Score. Given by the formula is this.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The comparison of proposed SVKN -based model and other existing techniques are shown in the below figure.3 and figure.4.the proposed SVKN model achieves a good accuracy of 98% in comparison to existing techniques.

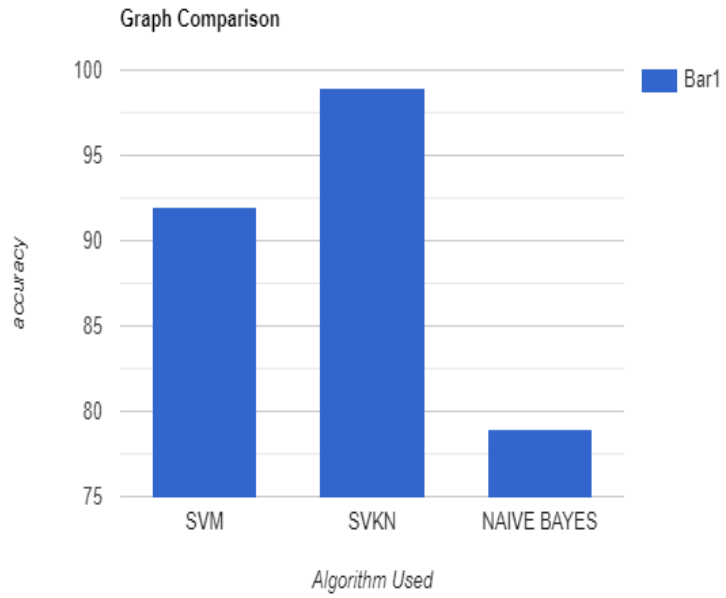

Figure.3.accuracy comparison chart for proposed and other methods

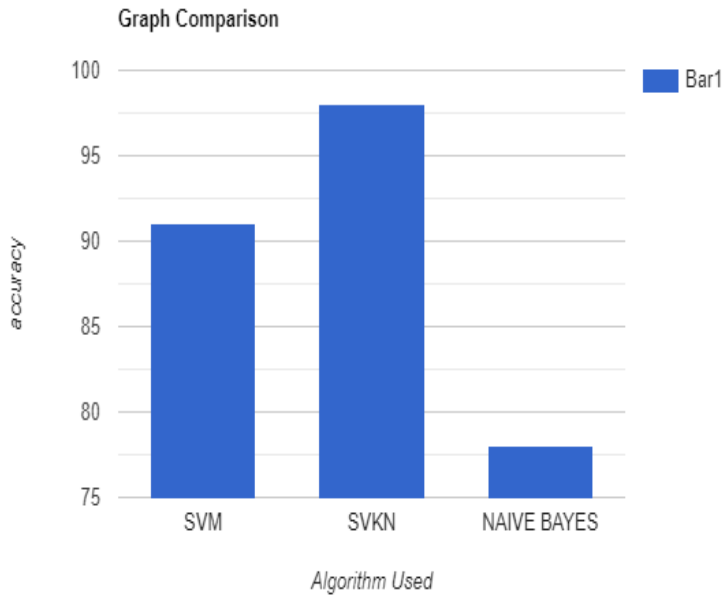

Figure.4. F1-Score comparison table for the suggested and alternative methods

CONCLUSION

In this paper, we proposed a machine learning model based on SVKN algorithm for the purpose of identification of potential patients of Parkinson's disease from the provided data sets. In this proposed method, data pre-processing is done to improve the data quality, reduce noise, and remove the null value attributes for increasing the efficiency of the training data for feature extraction. PD patients and healthy patients are accurately and efficiently identified using the pr0posed model. The proposed framework has been thoroughly tested and compared with other machine learning models in terms of the four key metrics values of Accuracy, Recall, F1-Score, and Precision. The results indicate that the proposed framework outperforms other traditional machine learning techniques in efficient prediction of the Parkinson's disease which would deem very helpful in detecting and also providing early medical care in potential patients.

REFERENCES

[1] Fatih Demir , Abdulkadir Sengur , Ali Ari, Kamran Siddique , (Senior Member, IEEE), Mohammad Alswaitti , (Member, IEEE) (2021). Feature Mapping and Deep Long Short Term Memory Network-Based Efficient Approach for Parkinson's Disease Diagnosis, part-based representation. IEEE Trans. Digital Object Identifier 10.1109/ACCESS.2021.3124765

[2] Christos Laganas, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Sofia B. Dias, Sevasti Bostantzopoulou, Zoe Katsarou, Lisa Klingelhoefer, Heinz Reichmann, Dhaval Trivedi, K. Ray Chaudhuri,and Leontios J. Hadjileontiadis, Senior Member, IEEE (2021). Parkinson's Disease Detection Based on Running Speech Data From Phone Calls. Citation information: DOI 10.1109/TBME.2021.3116935, IEEE

[3] Changqin Quan , Kang Ren, And Zhiwei Luo (2021). A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech .Digital Object Identifier 10.1109/ACCESS.2021.3051432

[4] ATHANASIOS TSANAS , (Senior Member, IEEE), MAX A. LITTLE ,AND LORRAINE O. RAMIG.Remote Assessment of Parkinson's Disease Symptom Severity Using the Simulated Cellular Mobile Telephone Network. Digital Object Identifier 10.1109/ACCESS.2021.3050524.

[5] WU WANG , JUNHO LEE , FOUZI HARROU , (Member, IEEE), AND YING SUN(2020), Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning, Digital Object Identifier 10.1109/ACCESS.2020.3016062

[6] D. Braga, A. M. Madureira, L. Coelho, and R. Ajith, ``Automatic detection of Parkinson's disease based on acoustic analysis of speech," Eng. Appl.Artif. Intell., vol. 77, pp. 148158,

[7] S. Bilgin, ``The impact of feature extraction for the classication of amyotrophic lateral sclerosis among neurodegenerative diseases and healthy subjects," Biomed. Signal Process. Control, vol. 31, pp. 288294,Jan. 2017.

[8] H Gunduz, ``Deep learning-based Parkinson's disease classication using vocal feature sets," IEEE Access, vol. 7, pp. 115540115551, 2019

[9] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth, ``Deep learning approach to Parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification," in Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., Berlin, Germany, Jul. 2019 pp. 717-720.

[10] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu,``Convolutional neural networks for speech recognition," IEEE/ACM Trans. Audio, Speech Language Process., vol. 22, no. 10, pp. 1533-1545,

[11] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, ``Convolutionalneural network to model articulation impairments in patient with Parkinson's disease," in Proc. INTERSPEECH, Stockholm, Sweden,Aug. 2017, pp. 314

[12] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath,``Deep learning for audio signal processing," IEEE J. Sel. Topics Signal Process., vol. 13, no. 2, pp. 206-219, May 2019.

[13] K.-I. Funahashi and Y. Nakamura, ``Approximation of dynamical systems by continuous time recurrent neural networks," Neural Netw., vol. 6, no. 6, pp. 801-806, Jan. 1993.

[14] S. Hochreiter and J. Schmidhuber, ``Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.