e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

SVM-BASED APPROACH FOR PARKINSONS DETECTION USING VOCAL DATA

Vaishnavi[*1]

[*1]Student, Department of Master of Computer Application, N.M.A.M Institute of Technology, Nitte, Udupi, Karnataka, India

DOI : https://www.doi.org/10.56726/IRJMETS37926

ABSTRACT

Parkinson's disease (PD) is usually identified using clinical observations and data from clinical studies, including the identification of many motor symptoms. Contrary to popular belief, traditional diagnostic approaches may be subject to prejudice because they depend on the interpretation of movements that can be feeble to see with human eyes and hence difficult to characterize, potentially leading to misinterpretation. Meanwhile, Parkinson's disease's first non-motor symptoms may be mild and caused by a range of other disorders. As a consequence, these symptoms are commonly overlooked, making early Parkinson's disease identification challenging. To overcome these challenges and improve PD diagnosis and assessment procedures, a machine learning algorithm for PD categorization has been applied to the vocal dataset consisting of 24 features. The report findings will shed light on the detection and classification of Parkinson's disease using the application of machine learning techniques and will aid in the creation of a highly accurate model as well as an effective tool for the disease's early detection and management.

Keywords: Parkinson’s disease, Vocal dataset, Machine Learning, Support vector machine, Detection

I. INTRODUCTION

The development of Parkinson's illness is brought on by the increasing neuronal deteriorations that produce dopamine in the brain causing both motor and non-motor signs such as tremors, bradykinesia, stiffness, and speech difficulty. Parkinson's disease frequently manifests as speech difficulty that can impact several elements of a person's speech, including vocal quality, pronunciation, fluency, and prosody. As a result, speech analysis is now recognized as an intriguing marker for Parkinson's disease identification. Machine learning methods, notably support vector machine learning (SVM), have shown considerable promise in properly identifying Parkinson's disease using voice data in recent years. SVM is a strong machine-learning technique that uses characteristics to categorize data into distinct classes. Models based on SVM have been used to categorize Parkinson's disease patients as well as healthy individuals using speech signal parameters such as basic frequency range, energy, and spectral characteristics. This method has demonstrated exceptional diagnostic accuracy, with excellent specificity, sensitivity, and precision. The selection of relevant features and the selection of ideal SVM hyperparameters, however, are essential for the effectiveness of SVM models. In this context, this paper aims to explore the efficacy of SVM algorithms for PD identification using voice data. The paper will concentrate on choosing relevant features and optimizing the Support Vector Machine hyperparameters for PD identification using voice data.

II. LITERATURE REVIEW

In [1] a supervised ML technique was proposed that utilized PCA for obtaining features using SVM as a method of classification to detect those with Parkinson's illness The primary purpose of this approach was to discover people who would be diagnosed with Parkinson's disease. The trials were carried out using clinical and demographic data from a number of patients. When compared to previous relevant research, the findings demonstrated that the suggested technique was highly effective in identifying Parkinson's disease patients. Mostafa et al. [2] recently attempted to improve PD diagnosis by employing various feature evaluation and categorization techniques. They used a multi-agent system to evaluate various features using DT, NB, NN, RF, and SVM as five classification methods. They conducted various trials with native and filtered datasets to evaluate the suggested technique. The findings demonstrate that by identifying the most useful set of features, this strategy increased the effectiveness of ML methods. Furthermore, the authors of [3] presented a Parkinson's disease expert system based on gathered attributes from recordings of patients' speech. To complement the replication-based experimental design they developed a Bayesian classifier technique to address reliance. 80 patients' voice recordings were used in the experiments, and 50 percent of patients had Parkinson's disease. Finding out which subjects had the disorder and which didn't was the main objective. Ipsita Bhattacharya et al [5], for instance, used the Weka data mining technique to distinguish between healthy and PD individuals. They used SVM, a method of supervised machine learning, for classification. The dataset underwent data pre-processing before classification. Various kernel values were used with LibSVM to gain maximum practical precision. The accuracy of the kernel with RBF and multi-kernel SVM was 60.8696%, while the accuracy of the linear kernel of the SVM was 65.2174%. By identifying dysphonia, Max A. Little et al. [6] suggested a novel method for differentiating between Parkinson's patients and control participants. In their work, the new, reliable dysphonia measure known as pitch period entropy (PPE) was developed. The data, which comprised 195 continuous vowel phonations, was collected from 31 people (23 PD patients as well as 8 healthy controls). Three steps make up their methodology: Calculation of features, pre-processing, selection of features, and classification. Support vector algorithms (SVM) were used with the linear kernel for classification. Their suggested model has an accuracy percentage of 91.4%. Arvind Kumar Tiwari [7], respectively, "Machine Learning Based Techniques for Parkinson's Disease Prediction," To forecast Parkinson's disease, minimal redundancy maximal relevancy algorithms for feature selection were employed in this system to identify the most significant characteristic among all features. This feature selection technique, combined with Random Forests, yielded a success rate of 90.3% and an accuracy of 90.2%. In addition, numerous approaches [8] were applied to anticipate Parkinson's illness. While other studies combined machine learning and deep learning to improve illness prediction, these methods used several kinds of machine learning methods and feature selection techniques to improve the estimation of Parkinson's condition.

III. SVM ALGORITHM

The SVM (Support Vector Machine) technique is a supervised approach to learning employed in both categorization and regression techniques. SVM classification seeks to identify the hyperplane that divides data points from various classes in a space with high dimensions. The closest data points to the hyperplane are called support vectors and play an important role in determining the hyperplane. The SVM algorithm maximizes the margin between the support vector and the hyperplane, resulting in improved generalization efficiency on new data. If the data is unable to be separated in a linear manner the technique of SVM can transfer it to a higher-dimensional space in which it can be segregated linearly by using a kernels function. The input data are transformed by the kernel function into a novel feature space where a hyperplane can be located. Once the hyperplane has been discovered, the SVM algorithm may categorize new points of data side along the hyperplane they fall. If the data point lies on the hyperplane's positive side, it is assigned to one class; if it lies on the negative side of the hyperplane, it is assigned to the opposite class. SVM classification is a sophisticated machine-learning technique that works well with modest numbers of training samples and can handle high-dimensional data. SVM has been applied successfully in a variety of fields, including bioinformatics, the classification of images, and the processing of natural languages.

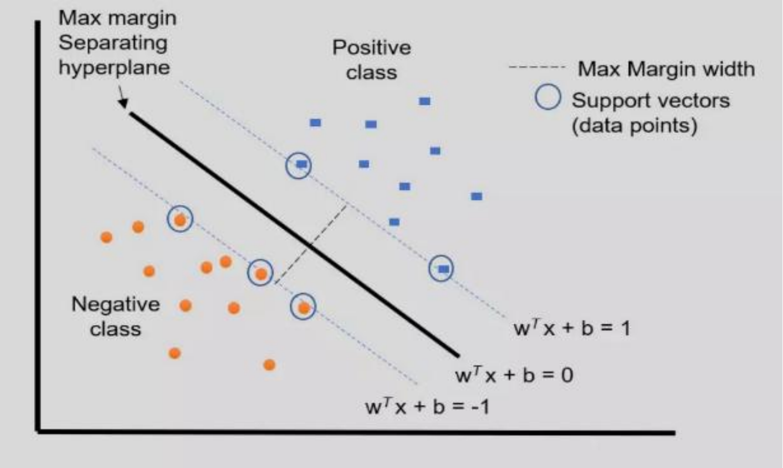

Fig 1: Support vector machine

IV. METHODOLOGY

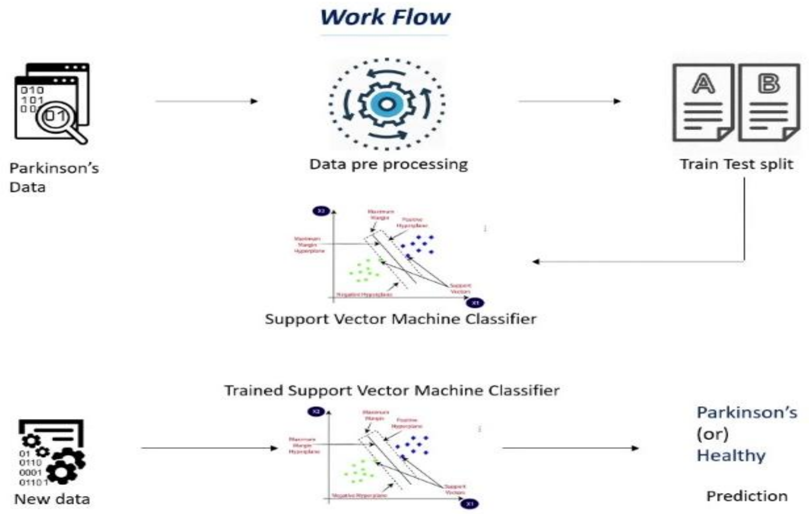

Fig 2: Workflow diagram

1) Data Collection: Obtaining dataset information from both Parkinson's disease patients and healthy individuals. A variety of speech patterns may be included in the material.

Data set information

Source: Max Little of the University of Oxford

This dataset includes biological voice measurements from 195 individuals, Parkinson's illness (PD) is present in 48 of them. In the table, each column indicates a distinct vocal measurement in order and each row represents one of the 195 voice recordings of these folks. The main goal of the data is to distinguish between Parkinson's disease patients and those in good health using a status section with a numerical value of 0 for good health and 1 for PD.

2) Feature extraction: Applicable attributes from the dataset are extracted. The mean, average, standard deviations, and variance-like statistical measures can be included. Other characteristics can be retrieved depending on the kind of data collected.

3) Data Splitting: Dividing the data into datasets for testing and training. The training set will be the one utilized for training the Support Vector Machine model, while the test data set would be used to assess the model's performance.

```
X_training_data, X_testing_data, Y_training_data,
Y_testing_data=train_test_split(X,Y,test_size=0.2,random_state=7)
```

4) Data Pre-processing: Noise or oddities from the data are removed. This could include normalization, scaling, or filtration.

```
scale = StandardScaler ()
X_training_data=scale.transform(X_training_data)
X_testing_data=scale.transform(X_testing_data)
```

5) Model Training: Use the training dataset to train the SVM model. This stage entails picking the right kernel function and tweaking the model's hyperparameters.

```
model = svm.SVC(kernel = 'linear')
model.fit(X_training_data,Y_training_data)
```

6) Model Evaluation: Examine the produced SVM model's effectiveness on a test dataset. Accuracy, recall, F1 score, accuracy, and the area beneath the curve (AUC) are examples of performance measurements.

Classification report

Model Accuracy: 92.3%

Table 1. Classification report

	precision	recall	f1-score	support
0	0.89	0.80	0.84	10
1	0.93	0.97	0.95	29
accuracy			0.92	39
macro avg	0.91	0.88	0.90	39
weighted avg	0.92	0.92	0.92	39

Based on the metrics provided, the model had a total accuracy of 0.9237, or 92%. This means that the model accurately classified 36 examples out of 39 predictions.

Precision represents the proportion of genuine positives among all predicted positives.

For each class, recall reflects the fraction of genuine positives among actual positives.

The score known as F1 is the harmonic mean of recall and accuracy, which provides a balance between these two measurements.

The total amount samples present in each class is indicated in the support column.

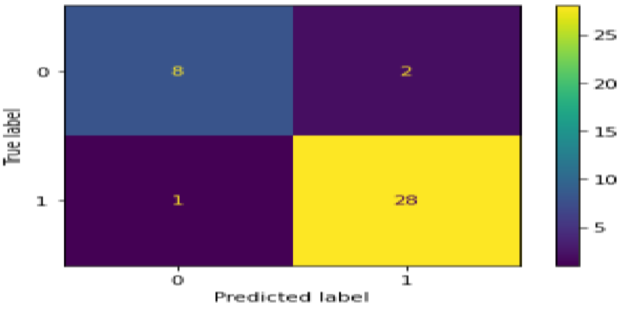

Fig 3: Confusion matrix

Model Deployment: Using the trained model generated by SVM to detect Parkinson's disease in a real-world situation. Stream lit allows the development of a web-based user interface for quick access and assessment of Parkinson's illness.

V. LIBRARY UTILIZATION

NumPy: NumPy serves as a Python module for numerical computing that supports arrays, matrix structures, and mathematical functions. It is one of the most extensively used libraries for scientific computation, analyzing data, and machine learning. Key features of the NumPy library are NumPy arrays, mathematical functions, algebraic operations, and broadcasting.

Pandas: Pandas serve as a freely available data analysis and manipulation package written in Python. It makes working with organized data like tabular, time series, & observational information simple and efficient.

Pandas has a variety of manipulation and analysis of data features, including filtration, classification, merging, combining, pivoting, and shaping data. It also offers sophisticated functions for dealing with data that is missing, historical data, and textual information.

Sklearn: A well-liked Python package for machine learning is sci-kit-learn, sometimes referred to as Sklearn. It offers straightforward and effective tools for analysis and mining. It works nicely with other Python ecosystem libraries and is developed on top of the SciPy, NumPy, and Matplotlib packages.

Streamlit: The Open-source Python module Streamlit can be utilized to easily build web-based applications for machine learning projects and data science. It is made to make it simple for you to convert existing scripts written in Python into interactive web applications that can be distributed to others. It can build interactive dashboards, visualizations, and online interfaces for your machine learning and data analysis models using Streamlit. It offers a straightforward API for building customized widgets and components, enables quick iteration, and lets you see your results in real-time.

Pickle*:* An option for serializing and deserializing objects in Python into a continuous sequence of bytes is provided by the pickle library module in Python. An object is transformed into a byte stream which may be stored or sent across a network during the serialization process. It takes a continuous stream of bytes and turns them back into an object using a process called deserialization. When storing and retrieving objects or data structures to or from disc files, the pickle modules are frequently utilized. It is often used to save a program's current state to ensure it can be continued at a later time.

VI. ADVANTAGES

Early detection: SVMs are able to be trained to detect minor changes in the voice that may signal Parkinson's disease in its early stages, enabling earlier intervention and therapy.

Consistency: SVMs are objective and precise in their processing of speech data because they rely on mathematical methods. This lowers the possibility of human error or prejudice in diagnosis.

Cost-effective: Because SVMs are capable of being trained on enormous amounts of voice data, they provide a low-cost method of screening for Parkinson's illness on a wide scale.

Non-invasive: Because voice data can be gathered non-invasively, it provides an easy and non-threatening method of diagnosing Parkinson's disease. This is particularly useful for patients who are afraid or unable to undertake more intrusive procedures.

VII. RESULTS AND DISCUSSION

With this model, early detection and intervention of Parkinson's' illness is facilitated with an accuracy rate of 92.3% on a dataset of 195 voice samples with 147 samples from healthy people and 48 samples from people having Parkinson's disease. Support Vector Machine (SVM) classifier had been trained on 80 percent of the data set and evaluated on 20%. Overall, these findings indicate that SVM can detect Parkinson's illness using voice data. It is crucial to note, however, that the classifier's performance can be affected by a number of variables, such as the dataset's quality and quantity, the features chosen, and the SVM algorithm's selection of appropriate hyperparameters.

VIII. FUTURE ENHANCEMENTS

Future research may lead to the development of numerous improvements in Parkinson's disease identification utilizing voice data. A significant quantity of high-quality data is required for machine learning algorithms to learn from for improving accuracy. The ability to train more accurate predictions for Parkinson's detection may be attainable if additional data, particularly data from a wider spectrum of patients, become accessible. For those with Parkinson's disease, real-time evaluation of voice data may be used to deliver prompt feedback. They could keep an eye on their symptoms and modify their treatment as needed.

IX. CONCLUSION

Finally, SVM has demonstrated promising results in diagnosing Parkinson's illness using voice data. SVM classifiers' high accuracy in this experiment suggests that voice data can be a trustworthy source in order to detect Parkinson's disease at an early stage. With this model, machine learning can be used to predict the onset of the illness in the body of a patient, simplifying the procedure for our user. However, it is crucial to remember that SVM performance may be affected by a variety of parameters, including dataset size, reliability, and feature selection. As a result, additional study is required to evaluate SVM's usefulness in detecting Parkinson's illness utilizing speech data on larger sets of data as well as real-world scenarios. SVM may improve the accuracy and efficacy of the diagnosis of Parkinson's illness, thereby allowing for swift diagnosis and therapy.

X. REFERENCES

[1] C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, et al., "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy," Journal of Neuroscience Methods, vol. 222, pp. 230–237, 2014.

[2] S. A. Mostafa, A. Mustapha, M. A. Mohammed, R. I. Hamed, N. Arun Kumar et al., "Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease," Cognitive Systems Research, vol. 54, pp. 90–99, 2019.

[3] L. Naranjo, C. J. Pérez, J. Martín and Y. Campos-Roca, "A Two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," Computer Methods and Programs in Biomedicine, vol. 142, pp. 147–156, 2017.

[4] a supervised ML method was proposed that combined the Principal Components Analysis (PCA) to extract features and SVM as classification method to identify PD patients. The main goal of this method was to determine patients that will be diagnosed with PD or with Progressive Supranuclear Palsy (PSP). The experiments were conducted on data of several patients with clinical and demographic features. The results depicted good accuracy of the proposed method in identifying the PD patients compared to existing related works.

[5] a supervised ML method was proposed that combined the Principal Components Analysis (PCA) to extract features and SVM as classification method to identify PD patients. The main goal of this method was to determine patients that will be diagnosed with PD or with Progressive Supranuclear Palsy (PSP). The experiments were conducted on data of several patients with clinical and demographic features. The results depicted good accuracy of the proposed method in identifying the PD patients compared to existing related works.

[6] a supervised ML method was proposed that combined the Principal Components Analysis (PCA) to extract features and SVM as classification method to identify PD patients. The main goal of this method was to determine patients that will be diagnosed with PD or with Progressive Supranuclear Palsy (PSP). The experiments were conducted on data of several patients with clinical and demographic features. The results depicted good accuracy of the proposed method in identifying the PD patients compared to existing related works

[7] a supervised ML method was proposed that combined the Principal Components Analysis (PCA) to extract features and SVM as classification method to identify PD patients. The main goal of this method was to determine patients that will be diagnosed with PD or with Progressive Supranuclear Palsy (PSP). The experiments were conducted on data of several patients with clinical and demographic features. The results depicted good accuracy of the proposed method in identifying the PD patients compared to existing related works

[8] a supervised ML method was proposed that combined the Principal Components Analysis (PCA) to extract features and SVM as classification method to identify PD patients. The main goal of this method was to determine patients that will be diagnosed with PD or with Progressive Supranuclear Palsy (PSP). The experiments were conducted on data of several patients with clinical and demographic features. The results depicted good accuracy of the proposed method in identifying the PD patients compared to existing related works

[9] M. Abdar and M. Zomorodi-Moghadam, "Impact of Patients' Gender on Parkinson's Disease using Classification Algorithms" Journal of AI and Data Mining, vol. 6, 2018.

[10] Bhattacharya, I., & Bhatia, M. P. S. (2010, September). SVM classification to distinguish Parkinson's disease patients. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (p. 14).

[11] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1010–1022, 2009.

[12] Arvind Kumar Tiwari, "Machine Learning based Approaches for Prediction of Parkinson's Disease," Machine Learning and Applications- An International Journal (MLAU) vol. 3, June 2016.

[13] C. Gao, H. Sun, T. Wang, M. Tang, N. I. Bohnen, et al., "Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease," Scientific Reports, vol. 8, no. 1, pp. 1–21, 2018.