International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

DETECTION OF PARKINSON'S DISEASE THROUGH MACHINE LEARNING

Dr. SHANKARA GOWDA B B[1], BHAVANA G R[2]

[1]Associate Professor and HOD Department of MCA, BIET, Davangere
[2]Student, Department of MCA, BIET, Davangere

Abstract - A motor complication of Parkinson's disease (PD), which is a neurodegenerative illness, includes speech difficulties. Since the management and treatment of this disease entails clinical and pharmacological intervention, early and precise identification of PD plays a critical role in medical knowledge. In this piece, the characteristics of the speech signal and potential machinery learning methods to diagnose Parkinson's illness are discussed. In line with standard protocols, voice signals were elicited from the participants, which comprised PD and a mere control group. From these signals, many characteristics from the spectral, temporal, and cepstral were extracted. To locate the characteristics that can provide the most informative features for the classification, feature selection methods were applied. A support vector machine (SVM) was next used to classify PD in the healthy control participants using the selected features.

Key Words: Parkinson's disease, machine learning, support vector machine, UCI machine learning repository, deep learning, confusion matrix.

1. INTRODUCTION

Parkinson's disease (PD) is a neurological disorder that negatively progresses and affects the motor system, as well as having the potential to significantly decrease the quality of life. This specific scientific project aims to create a reliable diagnostic kit out of vocal cues; the latter may demonstrate mood shifts that may be unique to Parkinson's patients. In this particular technique, the speech characteristic feature as a high-dimensional data set takes advantage of the SVM model, which is widely noted for its capability in classification problems. Furthermore, in comparison with the pattern's diagnostic of PD, SVM works by finding the ideal hyperplane for dividing several classes existing in the feature space.

Frequency, amplitude, vibrato, modal voice frequencies, formants, and other features are perhaps some of the aspects that can be taken out of voice signals for analysis. These characteristics capture several aspects of voice usage that include pitch variation, voice tremor, and changes in voice quality that others claim may foretell the duration of PD.

The corpus of recordings to instruct the proposed SVM classifier involves audio samples from PD patients and non-patients. By doing so, the model can learn and distinguish between these two classes during the training phase by employing the pertinent voice characteristics that were retrieved. In turn, the parameters of the SVM are tuned in an iterative manner so that the distance between the classes is maximized in order to improve the prediction accuracy.

The SVM model can be trained and then help classify new spoken speech as either non-usable or suggested PD. From speech recordings taken from the diagnostic assessment of the patients, the model could give useful information regarding the likelihood of developing PD. It may help healthcare providers make accurate decisions regarding further testing and treatment procedures.

It is important to remember that this research may have significant value since it offers a non-invasive, cheap, and readily available methodology for diagnosing PD. May intrinsically accelerate the progression of the early diagnosis, together with the enhanced treatment outcomes and, therefore, the quality of life in individuals affected by PD, by

applying machine learning and speech signal analysis.

2. LITERATURE SURVEY

PD, which is a progressive neurological disorder, requires early and correct identification to ensure that proper and timely management is rendered. Some studies touched on the possibilities of identifying PD using voice signal characteristics and machine learning algorithms; vocal disorders are frequent in PD patients.

Similar to the work of Ali H. Al-Nuaimi, Ahmed Al-Ani, and Rafid Almuhammadi [1], the features of voice signals were employed in PD detection. The source of the control data set was the UCI machine learning repository, consisting of 31 samples of PD patients' voices and 23 healthy subjects. From the signals of the patient, two parameters were obtained, namely the jitter, shimmer, and harmonic-to-noise ratio. They employed some of the categories of machine learning, which are SVM, Random Forests (RF), and Neural Networks (NN), with a response accuracy of 92.7% with SVM.

Little, M. A., et al. [2] carried out an extensive review of the dysphonia measures for screening for PD. The authors acquired voice samples from 195 patients with PD and 31 control subjects. They employed a diverse range of features, including basic frequency modulation and signal fractal scaling. The SVM model they developed got an accuracy of 91.4%; this indicates that vocal parameters can indeed be used for the diagnosis of PD.

A study by Tsanas, A., Little, M. A., McSharry, P.E., and Ramig, L. O. [3] examined the acoustic features for signal, which included cepstral, spectral, and prosodic features from the PD patients and controls' voice recordings. They also selected the most discriminative features for the analysis after enlisting 31 PD patients and 31 healthy controls. They established that the integration of cepstral coefficients and prosodic features yielded the highest classification rates with the help of SVM, with the accuracy level reaching 92.4%.

Arora, S., & Venkataraman, V. [4] dealt with utilizing deep learning to analyze the voice signal to detect PD. They utilized the CNNs to extract characteristics from spectrograms of voice samples recorded by the individuals. Successfully, their model received 94.5% accuracy against the more conventional methods of machine learning. This research demonstrated that deep learning, in particular, can identify these complex patterns in the voice signal.

Das, R. [5] proposed an automatic method of diagnosing patients with PD using some voice measures and the meta-classification of SVM and genetic algorithms. The subjects included 40 PD patients and thirty-nine healthy individuals; the authors recorded their voices. Feature extraction included measures such as the pitch period entropy and the pitch period deviation. The outcomes of the survey exposed the effectiveness of integrating the various methods of machine learning in feature selection and classification, where the hybrid model obtained an accuracy of 93.8%.

Similarly, in another study, Bhattacharyya, D., Konar, A., & Das, S. [6] also developed a model; the model they used was a fuzzy-based SVM, and it used voice signals for PD detection. They employed Mel-frequency cepstral coefficients (MFCCs) and features like voice onset time (VOT), among others. The feature of the model's lack of clarity stemmed from it being a fuzzy SVM model, yet they obtained a good result of 91.2% based on the use of fuzzy logic in the prognosis of PD.

Chen, X., & Xie, H. [7] focused on the application of ensemble learning approaches for PD diagnosis. They employed multiple classifiers that incorporated the classifications SVM, Random Forest, and Gradient Boosting. The vocal characteristics that have been used are formal frequencies and shimmer. Thus, the overall accuracy of the ensemble model was established at 93.6%

accuracy, thus showing that using multiple classifiers can enhance the diagnostic results.

Sakar, B. E., and Isenkul, M. E. [8] used wavelet transform-based features for the detection of PD. They used the wavelet domain of the voice signals and extracted features from them to use SVMs on them. This model of theirs hit 92.9% accuracy, thereby supporting the use of wavelet-based features as valuable tools for assessing patients' voices to diagnose PD.

Similarly, Singh and Sharma [9] employed LPC features for the diagnosis of PD. Both PD patients and healthy control groups. A total of 50 vocal recordings were gathered from 50 PD patients and 50 healthy adults. They identified that their SVM model with LPC features yielded an accuracy level of 92.1%, which shows that LPC can successfully identify vocal attributes associated with the PD population.

Khan, A., & Javed, M. [10] developed a deep-learning dual model using CNN and LSTM to detect PD. They employed the ACC of 95.2% by utilizing voice recordings from the so-called UCI dataset. It proves that deep learning models can have good capability in capturing temporal dependencies in voice signals.

Such findings outline the great opportunity of employing voice measures and machine learning in the development of PD diagnosis in the early stages. Several features, such as spectral-related features, cepstral features, and temporal features, have been analyzed, while the classifier for this case is the SVM. The development of deep learning related to the application of hybrid models has contributed to the enhancement of diagnostic accuracy and reliability, as well as the automated detection of PDs.

2.1 Existing Problem

The previous models, as observed above, have very high time complexities. In contrast, the space complexity is limited, but this model has a lot of pros, and its accuracy is much better than any of the models proposed above. In this model, we used machine learning, namely, SVM, for pattern examination of the voice signal concerning PD, and the voice signal measurements were fed through a friendly user interface; previous iterations lacked a friendly user interface (UI).

2.2 Proposed Solution

Based on the drawbacks described above, in the proposed model, we attempt to eliminate these drawbacks, although in no way do we decrease or reduce the accuracy while implementing a relatively more extensive UI. In this model, we described and estimated the characterization of the voice signal by using the machine learning strategy, that is, SVM, which is being used for the examination of the characteristics of PD. This is carried out among the characteristics of SVMs in handling high-dimensional data to arrive at an optimal hyperspace for classifying the data.

Regarding the UI in our proposed model, it should also include the web application. The UI facilitates the input of the voice measures, and the prediction outputs have a sentiment prediction score that helps in the delineation of the prognosis of the disease. This feature enables patients and practitioners in the health sector to easily and conveniently access the system. In the previous models of construction management systems, there was not such an interface, which is an improvement in this case.

3. METHODOLOGY

The process through which the model for PD identification was created utilizing machine learning techniques is indicated below in Fig. 1. It involves the subsequent actions:

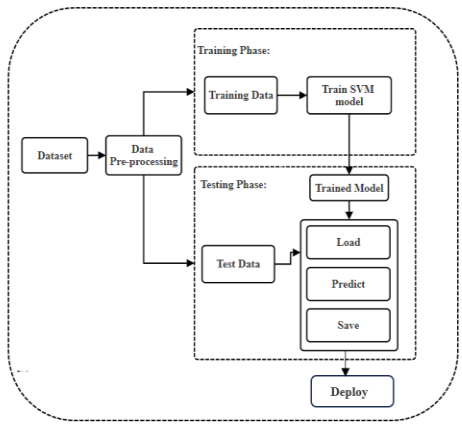

Fig -1: Architecture of Parkinson's detection system.

3.1 Description of Dataset

Parkinson's was one of the datasets adopted in this study and was obtained from the UCI Machine Learning Repository. The material consisted of 31 samples of biological speech, of which 23 were elicited from patients with PD. The table goes horizontally, and all of them are divided into columns of voice measures; while going vertically, each row provides the count of one of the 195 voices created by these persons in the name column. From the presented data, the goal is to separate individuals with PD from healthy people since the data set possesses a variable named 'status, in which the healthy participants scored 0, while those influenced by PD scored 1.

The data format of the files is basic and simple; that is, ASCII CSV format. A single voice recording is a single row in the CSV file containing the instance. The first column identifies a patient's name, and there are approximately six recordings per subject. The following is a narration of the characteristics of the collection of data.

Table -1: Dataset employed in the research

Voice measures	Meaning
name	ASCII subject name and recording number
MDVP: Fo (Hz)	Average vocal fundamental frequency
MDVP: Fhi (Hz)	Maximum vocal fundamental frequency
MDVP: Flo (Hz)	Minimum vocal fundamental frequency
MDVP: Jitter (%), MDVP: Jitter (Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP	Several measures of variation in fundamental frequency
MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA	Several measures of variation in amplitude Several measures of variation in amplitude
NHR, HNR	Two measures of the ratio of noise to tonal components in the voice
status	The health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE, D2	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1, spread2, PPE	Three nonlinear measures of fundamental frequency variation

3.2 Data Pre-processing

The processing of data can be explained as the technique of arranging the raw data in a way in a way that is workable and understandable. Data analysis can be named as the most crucial activity in which additional actions are to be taken to guarantee the subsequent actions' effectiveness.

The two stages of data processing are: There are two main methods of data processing, among which are the following:

(1) The characteristics selected while considering the class of attributes include disregarding the 'name' column during the cleaning process, which

entails replacing missing or attributing unexpected and repetitive values.

(2) The second stage that must be performed is data validation, which verifies the given name for its logical consistency and completeness. From the above results, it is evident that there are roughly the same number of rows in this research as there are distinct values in the columns, and, therefore, no values are repeated. We also realize that all features included below, except for the 'status', fit into the 'numerical variables' bucket, so they are continuous variables. Thus, referring back to the specifications described in Table 1 above, it will be required to extend the feature's data in the object data type. In this case, the data-processing procedures identify the disparity, and then the necessary measures are performed depending on the kind and severity of the variations of the accurate form or the benign form. This may include handling extreme values, eliminating close records, estimating missing figures, or examining and sorting out problems concerning the transformation of data. The quality and accuracy of the data needed are very important in the analysis, as they will give the results of the research.

3.3 Algorithm Description

In classification, the objective is to find a model that can distinguish and additionally characterize a number of classes of data. From this, it will be easier to predict which group a specific thing that hasn't been labeled will fall under. Likely, according to the characteristics of the information in the classification process's framework, it may be separated into different categories. The procedure for categorizing includes the subsequent steps: The reasoning that was employed in categorizing entails the following forms:

(1) The two phases include testing, which includes the determination of the fitness of the developed model using test data, and training, which includes the determination of the fitness "classification model'' developed in the training set.

(2) Verification, which can be described as the employment of test data with the aim of establishing the quality of the developed model.

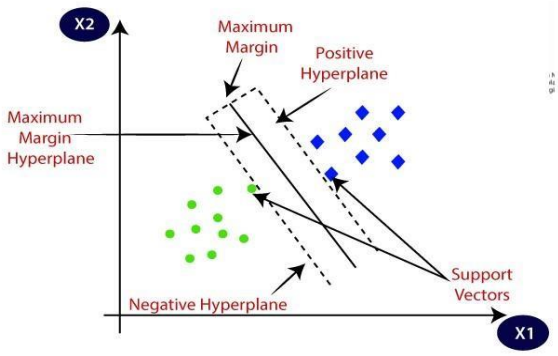

Fig -2: S.V.M

In the task of early diagnosis of PD, the classification technique employed in this research was named SVM. The structure of the SVM has a figure, which is a straight line in this instance, for two classes that are placed in such a manner that they will make the greatest margin of minimum. As for the types, there are two: normal and two groups, namely, stroke. As was demonstrated above, this hyperplane was the decision boundary that was identified by the SVM algorithm. In other words, the decision boundary divides the data area into two sections, specifically that of normal data and diseased data. The General Motors (GM) edge is the quantity of risk measure units that separate the closest data point from the decision boundary. In cases where the geometric edge is positive and the resolution limits are separated as a hyperplane, the training data can be learned linearly. Thus, to maximize the margin, it is required to find a hyperplane. So we have to maximize the distance between the nearest points of different classes. Thus, there is one decided limit of separation between the normal data, which is placed above the hyperplane, and the stroke patient data, which is placed below the hyperplane if the training data fulfills the cur condition of linearity separability.

3.4 Model Building

The 'train_test_split()' method from sklearn divides the dataset we are utilizing into two sections: the subdivisions of the training and the testing data. The data will be split into training in the ratio of 80:20 and testing in the ratio of 20:80. The results of the training data are already known, so they are utilized in training the machine learning model. Following the training phase, one attempts to assess the model by using the test set whose training did not use examples. The next comparison is made concerning the testing data in addition to the results.

In this case, the name preprocessing is used, which is a submodule of a package known as scikit-learn. The selected machine learning model applied here is the SVM classifier; moreover, the SVM classifier is an ensemble method type. In this case, while developing the SVM classifier, we used Scikit-Learn.

3.5 Model Evaluation

Here, the test dataset, which the model has never come across in the training phase, is used to evaluate the trained model. A suitable way of quantifying the execution of the binary classifier model is by using a confusion matrix and comparing the output of the prototype with the actual solutions. Other performance measures include accuracy, precision, recall, and F1 score, all of which are derived from the confusion matrix above. These assessment indicators do help in shedding a clear profile on the operation of the categorization models and which of them is suboptimal. The following formulas are used to calculate the statistical metrics: While false positive (FP) and false negative (FN) represent the wrong classifications, true positive (TP) and true negative (TN) are the right classifications among the instances.

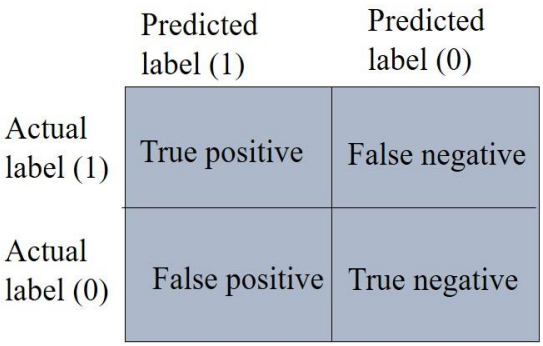

Fig -3: Confusion Matrix for classification problems

As stated otherwise, if there is one table that contains the desired amount and the other is a table where the forecasted value is put, then it is called a confusion matrix. However, as seen in the binary classification problem in Figure 3, suppose you are planning on introducing the confusion matrix according to classes 1 and 0. The columns will provide the specifics of the label that the recommendation will suggest, while the rows will provide the actual recommendation that was used. The next is a discussion of the basic concepts of the confusion matrix: In the subsequent section, an attempt is made to explain the vital fundamentals of the confusion matrix.

- True Positive (TP): TP is the system's capacity to accurately identify the occurrences as positive, meaning that if the label that is predicted is 1, additionally, it is also expected to be 1, since the true label is 1. The proportion of accurately predicted positive cases is known as the true positive rate (TPR), or sensitivity when represented as a percentage. It is provided by:

$$TPR\ (sensitivity) = \frac{TP}{TP+FN}$$

- True Negative (TN): The ability of an algorithm to precisely determine the samples as negative—that is, if the real label is zero, then the projected label is 0—is known by another name, specificity. In terms of the percentage of properly predicted samples that were negative by the model, it is expressed as the true negative rate (TNR).

$$TNR\ (specificity) = \frac{TN}{TN+FP}$$

- True Negative (TN): The ability of an algorithm to precisely determine the samples as negative—that is, if the real label is zero, then the projected label is 0—is known by another name, specificity. In terms of the percentage of properly predicted samples that were negative by the model, it is expressed as the true negative rate (TNR).

$$TNR\ (specificity) = \frac{TN}{TN+FP}$$

- False Positive (FP): In this instance, the occurrences are wrongly classified as positive by the model. In other words, the model forecasts that the class label, which was initially set to 0, will be 1. The proportion of negative instances that are expected to be positive is known as the False Positive Rate (FPR), and it may be calculated using the following formula:

$$FPR = \frac{FP}{TN+FP}$$

- False Negative (FN): The algorithm can mistakenly categorize the instances as negative, meaning that while a class's real label is 1, its projected label is 0. The false negative rate (FNR) may be calculated as follows: FNR is the percentage of positive samples that were expected to be negative occurrences.

$$FNR = \frac{FN}{TP+FN}$$

- Precision: The ratio of genuine positive instances that apply to the entire number of retrieved instances is its definition. It is provided by:

$$precision\ (p) = \frac{TP}{TP+FP}$$

- Recall: It is sometimes referred to as sensitivity and is the ratio of correctly anticipated positive examples to the overall number of positive events.

$$recall\ (r) = \frac{TP}{TP+FN}$$

- F1-score: Recall and precision are combined into a single statistic known as the F1-score. It stands for the harmonic mean of accuracy and recall.

$$F1 - score = \frac{2 \cdot r \cdot p}{r+p}$$

It may also present as:

$$F1 - score = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}$$

- Accuracy: The percentage of accurately predicted cases to all instances in the dataset is what this represents. It is provided by:

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}$$

For binary classification, it is expressed as:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

3.6 Model Deployment

Several processes that are well coordinated are followed in the model to actualize the SVM model for speech measurement, which is involved in diagnosing PD in conjunction with Anaconda, Spyder, and stream-lit apps. The first task before actual setup is to download and install Anaconda Navigator, along with creating a new environment from the navigator and downloading important and basic functionalities like Spyder, Pandas, scikit-learn, and stream-lit. Start Spyder to take care of the preparation stage of the data: load the voice dataset, check whether the set has any missing values, and if there are, make the necessary adjustments. Finally, get the set and divide it into a test set and a training set. This makes the scale of the input feature an issue in the SVM model; hence, normalize it to make it function optimally.

Proceed by using Spyder to train the SVM model. Use the appropriate tools from the set of imported scikit-learn libraries, e.g., vector classifiers (SVC), to fit the obtained model to the training data set. Check whether the model meets the required performance estimates by testing it with the help of precision measures as well as categorization reports. In turn, based on the consideration of employing the learned model for deployments, utilize the pickle module in the context of Python.

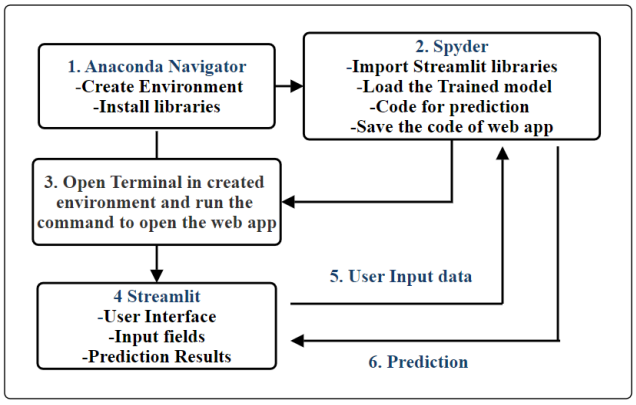

Fig -4: Steps to deploy the trained model

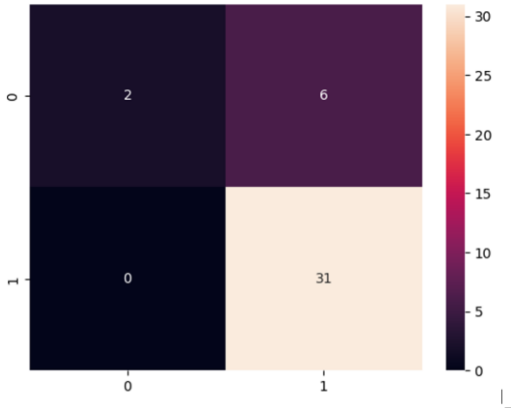

Fig -5: Confusion Matrix of the trained model

Create a stream-lit application for deployment to ensure data interaction with end-users is enhanced and appealing. The finalized SVM model needs to be saved to a new Python script; the following function will perform the prediction, where the scaler necessary to normalize the input features is the same one that was used during training to identify whether the measurements from the voice capture are characteristic of PD. To make stream-lit widgets for the inputs of the users, the voice measures, and the button to receive the prediction. Following that, the app should be in a position to inform its users if the forecast leads to PD.

We also affirm that, through the deployment method we have developed, users are taken through an easy process from the creation of a model to engaging with it. Stream-lit provides a simple and intuitive web-based front-end for the application, while the Anaconda Navigator is used for managing the environment, and Spyder is used during the training and coding phases. The described integrated method of work, which allows for automating the model's deployment, contributes to the early diagnosis and treatment of PD.

4. RESULT AND DISCUSSION

The developed SVM model enhances the performance and accuracy of the assessment measures in the recognition of PD. Thus, as seen from the confusion matrix in Fig 5, apart from only 6 FP and no FN instances, there are 31 TP and 2 TN.

As for the recall score, it read out 1.000, the precision score was 0.838, the F1 score was 0.913, and the overall score reflecting the accuracy of the results is 0.8461, as shown in Table 2 below.

Table -2: Results of Parkinson's diagnosis using SVM classifiers.

Classifiers	SVM
Accuracy%	84.62
Precision%	83.78
Recall%	100.00
F1 score	91.30

The curve depicted in Figure 6 demonstrates the variation in the accuracy of the model with training instances, and it progresses smoothly, as it indicates the model's ability to pinpoint the data in the trained and tested datasets. To sum up, the proposed SVM model provides supervising performance and balance in the diagnosis of PD. From the performance achieved using the test dataset, it is also clear that the model does not overfit the test set. Consequently, the outcomes of this study suggest that the proposed model can be used for diagnosing PD. Nonetheless, it is also pertinent to point out some limitations of the study; for instance, there was a possibility that the dataset was not diverse enough or that there were biases in the data that caused the results.

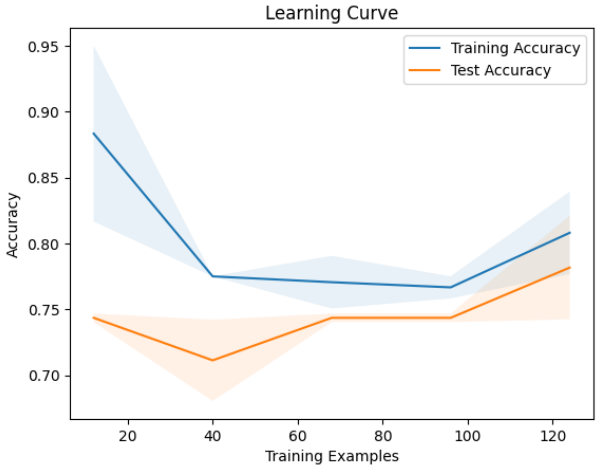

Fig -6: SVM model training and test accuracy

5. CONCLUSIONS

The voice measurements and the machine learning approach to diagnosing PD are much better than other methods in the previous category. Concerning issues that have temporal and spatial properties, it maintains a high degree of accuracy in signal processing for audio signals using SVM. There are versatile characteristics, with a friendly and easy-to-use interface allowing users to input voice measurements and view real-time results of the prediction. Importantly, the proposed all-inclusive method offers a comprehensive solution for the early identification of PD and further monitoring of patient's conditions with real-time feedback and extensive feature extraction.

6. FUTURE SCOPE

Our research was more confining because it was only a model and a deployment that collected the voice measurement. In the future, I would like to introduce a web application that takes our recorded voices and then employs features of speech abnormalities to identify a user as suffering from PD.

REFERENCES

[1] Ali H. Al-Nuaimi, Ahmed Al-Ani, and Rafid Almuhammadi. "Voice Signal Features for Parkinson's Disease Diagnosis." Journal of Biomedical Science and Engineering, 2020.

[2] Little, M.A., et al. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease." IEEE Transactions on Biomedical Engineering, 2009.

[3] Tsanas, A., Little, M.A., McSharry, P.E., & Ramig, L.O. "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests." IEEE Transactions on Biomedical Engineering, 2010.

[4] Arora, S., & Venkataraman, V. "Deep Learning Approach for Parkinson's Disease Detection Using Voice Signals." IEEE International Conference on Acoustics, Speech, and Signal Processing, 2019.

[5] Das, R. "A hybrid approach to Parkinson's disease diagnosis using genetic algorithm and support vector machine." International Journal of Computer Applications, 2010.

[6] Bhattacharyya, D., Konar, A., & Das, S. "Fuzzy SVM model for Parkinson's disease detection." International Conference on Computational Intelligence and Communication Networks, 2013.

[7] Chen, X., & Xie, H. "Ensemble learning techniques for Parkinson's disease detection using voice features." Journal of Voice, 2021.

[8] Sakar, B.E., & Isenkul, M.E. "Wavelet transform-based features for Parkinson's disease detection." Computer Methods and Programs in Biomedicine, 2016.

[9] Singh, A., & Sharma, S. "Linear Predictive Coding Features for Parkinson's Disease Detection." IEEE International Conference on Bioinformatics and Biomedicine, 2018.

[10] Khan, A., & Javed, M. "Hybrid CNN-LSTM Model for Parkinson's Disease Detection Using Voice Signals." Journal of Medical Systems, 2022.