Telemedicine-Driven Early Detection of Parkinson's Disease: The Machine Learning Approach *

*

1st Kanakala Sri Harika
Dept of Computer Science and Engineering
sriharikakanakala1@gmail.com

***Abstract*—In recent years the growth of Parkinson's Disease has increased day by day for elderly people aged 45-85. It is a brain disorder there is no cure for this disease. It occurs due to the loss of neurons in the brain. People with this disease have difficulty walking, speaking, and swallowing. There are many speech signals processing for the detection of PD in early stages. In this paper, we will deal with people with PD and healthy individuals. The data used in this paper is audio data. The paper highlights the use of ML techniques in telemedicine for early detection. The models used here are K-nearest neighbor(KNN), Support Vector Machine(SVM), Random Forest(RF), and Decision Tree(DT). The most accurate results are obtained from three models that are KNN, DF, and RF which have an outstanding accuracy of 100%. This accuracy is obtained by using the SMOTE over-sampling technique which gives balanced data from the unbalanced dataset. This technique helped to improve the accuracy of the models. Using RELIEF feature selection algorithm SVM gained accuracy up to 92.31%. This feature selection algorithm extracted 13 relevant features from the dataset. In our research, we aimed to use ML in telemedicine, for the early detection of PD without any physical visits.**

***Index Terms*—Parkinson's Disease, MDVP dataset, KNN, DT, RF, SVM.**

I. Introduction

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disorder that is occurred by the loss of dopamine-producing neurons in a specific area of the brain known as the substantia nigra. Dopamine is a neurotransmitter that plays a crucial role in transmitting signals in the brain that control movement. The reduction in dopamine levels can occur due to genetic anomalies, environmental toxins, and aging.
The features of PD are motor symptoms which include tremors, bradykinesia, muscular rigidity, and postural instability in addition to this PD has different non-motor symptoms such as anxiety, depression cognitive impairment, and autonomic dysfunction. PD disease can also develop dysarthria a disability of motor speech that affects articulatory, prosodic functions, and respiratory.

PD mostly occurs in elderly people, with the rates of PD increasing significantly between the ages of 50 and 60. Some symptoms progress slowly are they do not notice them easily. Tremor is the first symptom that people with PD will notice. In the beginning stage, the tremor will show up in n just one arm or leg or on one side of the body. It can impact the movement of the chin, lips, and tongue. This leads to difficulty in swallowing, chewing, and speaking.

Currently, there is no cure for PD because of the loss of dopamine-producing neurons in the brain and once these neurons are damaged they cannot be regenerated. The are some treatments available to manage PD that can help to improve the quality of life for the people who are affected by this disease. There are some common treatments for Parkinson's disease including Dopamine Replacement, COMT Inhibitors, Deep Brain Stimulation (DBS), Physical Therapy, Speech Therapy, and Lesioning Procedures.
In recent study has focused on various factors to enhance the condition and improve its diagnosis, treatment, and management. Some factors that focus on recent research are genetic factors, neuroinflammation, neuroprotection, non-motor symptoms, telemedicine, and advanced imaging techniques.

II. PROBLEM STATEMENT

PD, a neurodegenerative condition that disproportionately affects the elderly, is one of the most common neurological disorders. Parkinson's patients have mobility and speaking difficulties, which makes it difficult to visit them in person for therapy or monitoring. For necessary care, medication modifications, and disease progression monitoring, these difficulties make it difficult for patients to attend healthcare institutions frequently. Given the rising global aging population, early identification of PD is essential for enabling patients to live more normally.

As a result, it's even more crucial to accurately and early diagnose PD, both in person and without a doctor's visit. Solving this problem is not just about healthcare, it's something society needs to do because it affects the health of a big group of people. The fact that there are more older people makes it important to find better ways to detect PD early, even from far away, and with a lot of accuracy.

Identify applicable funding agency here. If none, delete this.

For early detection of PD, the use of ML algorithms is common in the medical field. ML enables software to train data and develop outstanding results. For detecting PD, several data are applied to ML models. ML models help to combine trained data of people who are suffering from PD and healthy people with testing data of people who are suffering from PD and healthy people to identify accurate results.

Recent telemedicine studies have focused on some factors that enhance healthcare delivery effectiveness and improve diagnosis, treatment, and accuracy. Some key factors in recent telemedicine research include remote Monitoring and Wearable Devices, Interoperability, patient acceptance and satisfaction, technology integration, telemedicine effectiveness, economic considerations, and several other factors used to advance telemedicine in healthcare.

III. LITERATURE SURVEY

This research mainly focuses on audio data for early detection. While previous research by Doneti Sowmya [1] has focused on predicting PD using SVM and data mining techniques we can predict symptoms like gait, tremors, and micrographia and had an accuracy of 86%. This paper lacked the use of preprocessing techniques and feature selection that would help to improve the ML model. The research paper I have done describes an improved SVM model with an accuracy of 92.31% using the SMOTE oversampling technique this addresses the issue of class imbalance. Using audio data is more useful for classifying PD than using genetic data.

Ahmed M. Elshewey[8] has focused on hyperparameter tuning through Bayesian Optimization significantly improves the performance of the models compared to their default parameters and the data which used 23 features and 195 instances. After hypertuning with BO-RF achieved 89.7%, and our research paper outperformed RF model accuracy by 100% using the SMOTE oversampling technique. Luca Parisi[7] has focused on the effectiveness of m-ark (a kernel function)coupled with SVM (without Lagrange multipliers) for supervised ML-based classification, which can aid in early PD detection based on speech features.

Athanasios Tsanas*[3] extracted more non-linear dysphonia features and applied four feature selection algorithms(LASSO, mRMR, RELIEF, and LLBFS), to identify relevant features, using the SVM classifier. The paper consists of voice recordings from 43 subjects and extracted 263 samples.
Max A. Little1[4] proposed a new measure called Pitch Period Entropy (PPE), which is used to assess characteristics of vocal pitch or fundamental frequency (F0) in speech. The paper improved results by introducing a novel measure (PPE), carefully selecting relevant features, using non-standard measures, applying a powerful classification algorithm (SVM), and employing bootstrap resampling for validation. The study highlights that SVM with Gaussian RBF kernel combined with the new PPE measure, which is used for improving the classification of PD based on voice analysis.

Mohammad Shahbakhi[5] the results of their PD diagnosis using voice analysis by employing feature selection with Genetic Algorithm (GA), selecting SVM with a Gaussian radial basis kernel, and tuning key parameters. The authors employed a GA to select the most useful and powerful features from a set of 14 features related to speech factors. They found that certain combinations of features, including "Fhi (Hz), Fho (Hz), jitter (RAP), and shimmer (APQ5) features," led to the highest accuracy. Salim Lahmiri[6] implemented SVM and combined it with ROC or Wilcoxon-based ranking techniques and achieved an accuracy of 92.21%, which was obtained when using 14 patterns selected by the Wilcoxon-based pattern ranking technique. This paper focused on patients aged 50-65 years.

Based on our literature review, we have implemented a PD classification model on audio data. Through our review, we will try to improve the detection of PD through telemedicine. Now, our research aims to explore RF, KNN, DT, and SVM models to classify PD patients with audio data.

Here we used the RELIEF feature selection technique to identify and select relevant features in a dataset and eliminate irrelevant or duplicate data and this technique works well for classification problems. We are using The SMOTE Over Sampler technique is an oversampling used to address class imbalance, Oversampling involves increasing the number of samples in the minority class to match the number of samples in the majority class. Using the SMOTE oversampling technique it shows that KNN model and RF are the best-performing models with an accuracy of 100% which outperforms all other models, the decision tree model with a 98.3% accuracy. Using RELIEF feature selection SVM model with an accuracy of 92.31%.

IV. DATASET

The data for this research used MDVP audio data which is collected from PPMI and UCI databases. This data contains both Parkinson's and healthy patient's voice modulations. The data consists of 195 sustained vowel phonations from 31 male and female subjects, of which 23 were diagnosed with PD. The diagnosed patients are of age 45-85. Collected 6 phonations from each subject and that range from 0-36sec. The data consists of 147 vowel phonations from the person who was diagnosed with PD and 48 from healthy individuals.

The data is converted into CSV ASCII format. The 'status' column says that 1 is for PD and 0 is for healthy persons. The goal of this paper is to determine the person who is suffering from PD.

/P:Shimmer	MDVP:Shimmer(dB)	...	Shimmer:DDA	NHR	HNR	status	RPDE	DFA	spread1	spread2	D2	PPE
195.000000	195.000000	...	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000
0.029709	0.282251	...	0.046993	0.024847	21.885974	0.753846	0.498536	0.718099	-5.684397	0.226510	2.381826	0.206552
0.018857	0.194877	...	0.030459	0.040418	4.425764	0.431878	0.103942	0.055336	1.090208	0.083406	0.382799	0.090119
0.009540	0.085000	...	0.013640	0.000650	8.441000	0.000000	0.256570	0.574282	-7.964984	0.006274	1.423287	0.044539
0.016505	0.148500	...	0.024735	0.005925	19.198000	1.000000	0.421306	0.674758	-6.450096	0.174351	2.099125	0.137451
0.022970	0.221000	...	0.038360	0.011660	22.085000	1.000000	0.495954	0.722254	-5.720868	0.218885	2.361532	0.194052
0.037885	0.350000	...	0.060795	0.025640	25.075500	1.000000	0.587562	0.761881	-5.046192	0.279234	2.636456	0.252980
0.119080	1.302000	...	0.169420	0.314820	33.047000	1.000000	0.685151	0.825288	-2.434031	0.450493	3.671155	0.527367

Fig. 1. Collection of Dataset

The data in this dataset is unbalanced, the data is balanced through SMOTE oversampling technique.

V. METHODOLOGY

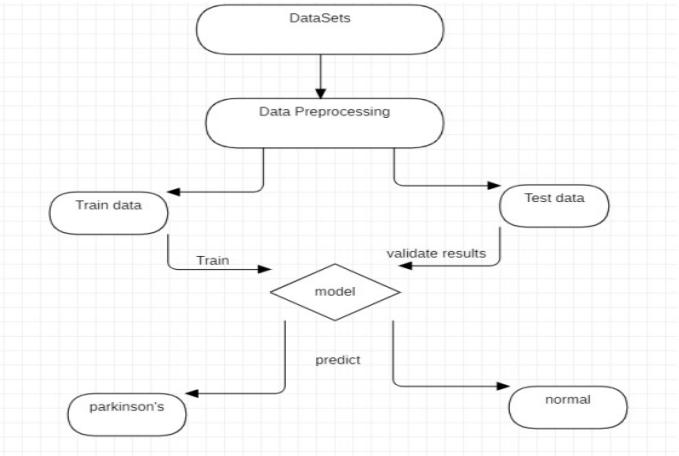

Fig. 2. Proposed Methodology

First, we collected MDVP audio data from the PPMI database. The dataset contains information about jitter, shimmer, and MDVP of vowel phonations. The data is pre-processed and analyzed. Four models are selected SVM, KNN, RF, and DT these models are trained on 75% of data and tested on 25% of data, and they are predicted based on accuracy, Cross Validation Score, precision score, recall score, F1-Score Score. As this is voice data it may contain background noise for increasing accuracy it is better if the dataset doesn't contain any other noises

A. Data Preprocessing

Data Preprocessing is used to clean the data from noisy data. The data in this dataset is Uncovering and Mitigating Bias in PD Class-Imbalanced Dataset. Data consists of ¾ of the instances being Parkinson's patients and ¼ are benign.

a) Feature Engineering: "The attributes 'name' and 'status' are removed from the table, and following the removal, the 'status' label is assigned as the target feature."

b) Feature Extraction: The most relevant features are extracted from the data using the RELIEF feature selection algorithm. The selected features are trained and tested to give accurate results.

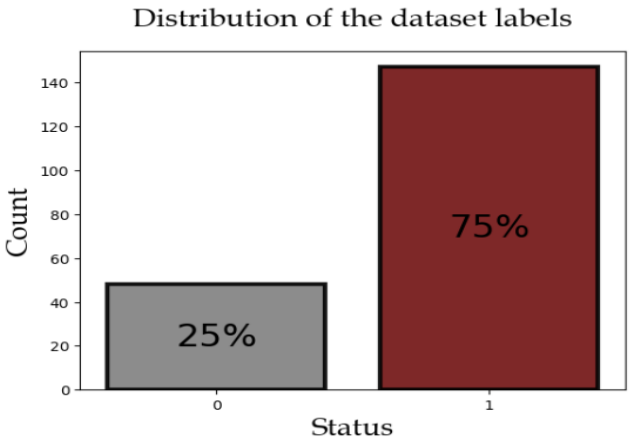

Fig. 3. Imbalanced collection of Dataset

The status of this dataset is 147 individuals who have PD and 48 individuals are benign.

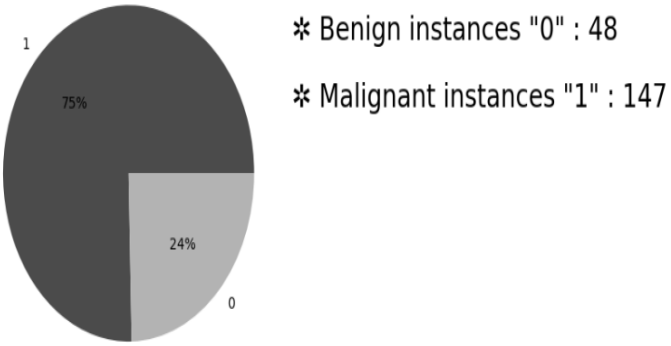

Fig. 4. Imbalanced Dataset

As shown in Fig 4. The dataset appears to be unbalanced, with a significantly higher number of samples in the Malignant category compared to the others. To address the issue of class imbalance, an over-sampling data balancing technique will be used.

As shown in Fig 5. The SMOTE over-sampler technique is used to address the class imbalance in the dataset and this technique is used to balance the data in the dataset. After using this technique we achieved balanced data. Before pre-processing the data the over-sampling technique is used and it will balance the data and give accurate results.

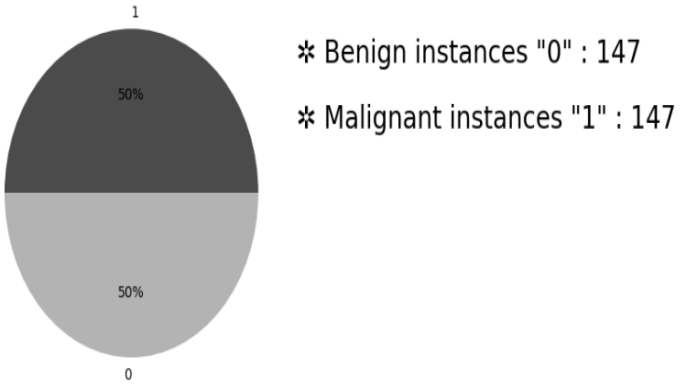

Fig. 5. Over-Sampled Dataset

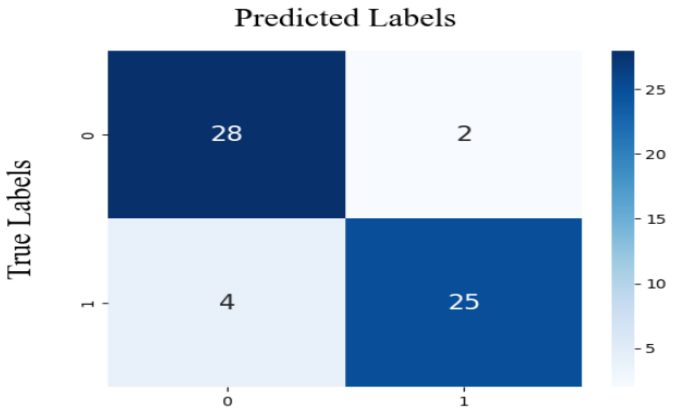

Fig. 6. Confusion matrix for Support Vector Model in approach 2

VI. MODEL TRAINING

In this research, we are focusing on 4 models SVM, KNN, DNN, and RF. These models are evaluated based on 3 approaches:

- The models will be trained and evaluated on the unbalanced dataset.
- The models are trained and evaluated on the over-sampled data.
- The models are trained and evaluated using the features selected by the Relief algorithm.

a) Support vector Machine(SVM): : SVM is a supervised machine learning model used for classification analysis. It is used to handle high-dimensional data.
In this research, Since PD voice data is not linearly separable we use an SVM linear kernel with a value of 10 for the 'C' parameter. SVM constructs a hyperplane in a high-dimensional space to separate data points into different classes. It is used for linear and non-linear classification as it handles higher-dimensional spaces using kernel functions. The 'C' parameter controls the balance between margin width and the penalty for misclassification. A smaller 'C' allows a wider margin and reduces overfitting.

Here In this model, Fig 6. the first approach is model trained and evaluated on unbalanced data the data in this dataset has noisy data which leads to overfitting of the model.

In the next approach, the model is trained and evaluated on over-sampled data. Here SVM applied to SMOTE Over Sampler technique is used for detecting class imbalance by over-sampling.
IB Fig 7. The final approach employs an SVM model trained and evaluated using features specifically selected by the Relief algorithm.

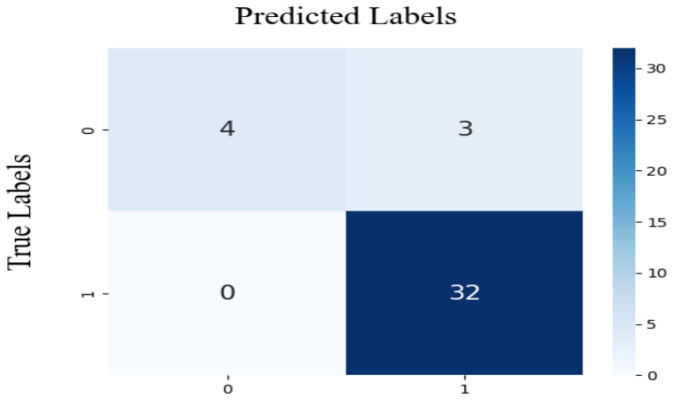

Fig. 7. Confusion matrix for Support Vector Model in approach 3

b) K-nearestNeighbor(KNN): KNN is a non-parametric machine-learning algorithm used for classification. It works best for balanced audio data of 109 records due to the small dataset size. It efficiently forms two distinct clusters for Parkinson's Disease with PWP (People with Parkinson's) and healthy data.
To address overfitting caused by imbalance in the data, we employed the SMOTE over-sampling technique to create a more balanced dataset. In Fig 8. KNN with balanced data gained maximum accuracy.
In Fig 9. The final approach employs a KNN model trained and evaluated using features specifically selected by the Relief algorithm. After employing the KNN model to the selected features there is more overfitting in the model. In the future, we need to use optimization or regularization techniques to reduce overfitting.

c) Random Forest: RF is an ensemble learning method in ML that constructs multiple decision trees during training and it is used for classification. It combines the results of multiple individual trees to improve accuracy and reduce the risk of overfitting compared to the single

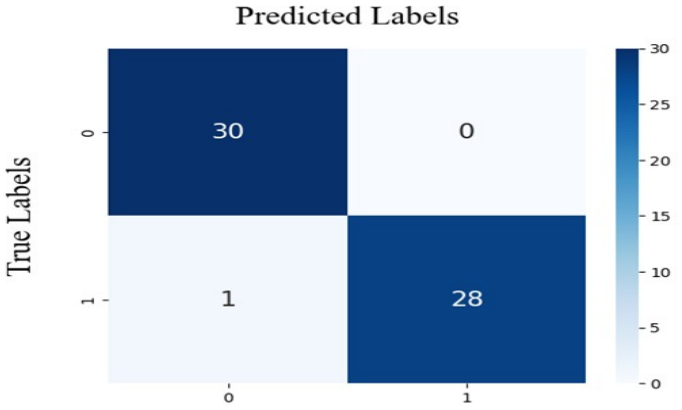

Fig. 8. Confusion matrix for KNN in approach 2"

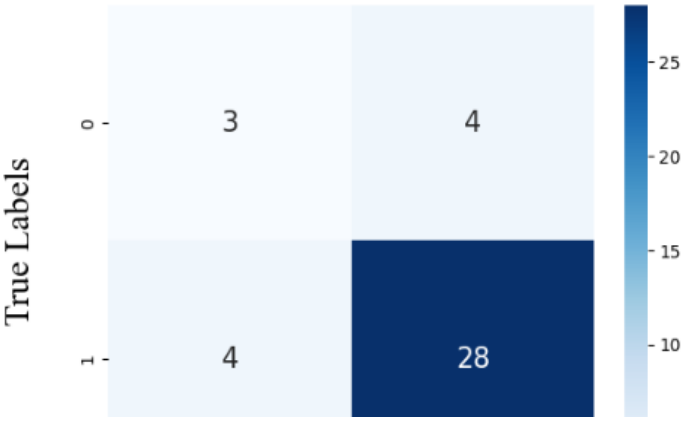

Fig. 9. Confusion matrix for KNN in approach 3"

decision tree. RF is effective for a wide range of datasets and is less sensitive to noise or outliers. "There is a slight overfitting issue in the data, which can be resolved by employing the SMOTE oversampling technique. In Fig 8. RF with balanced data gained 100% accuracy.

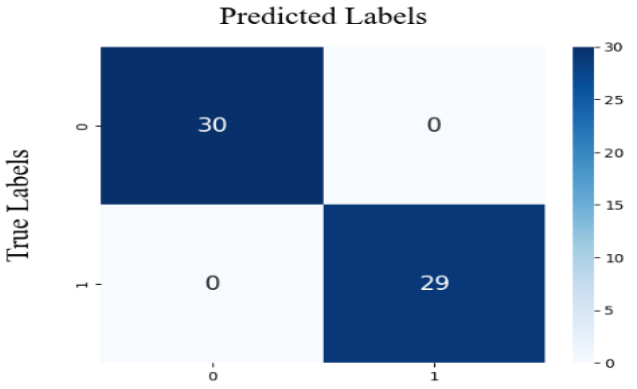

Fig. 10. Confusion matrix for Random Forest in approach 2"

The final approach employs an RF model trained and evaluated using features specifically selected by the Relief algorithm.

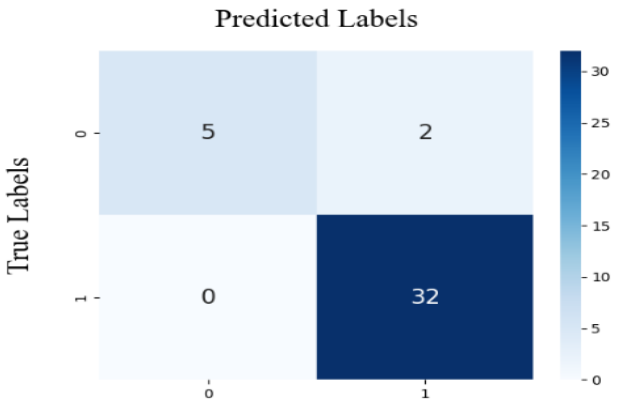

Fig. 11. Confusion matrix for Random Forest in approach 3"

d) Decision Tree: A DT is a supervised ML algorithm used for classification. The algorithm divides the dataset into subsets based on selected attributes, recursively creating a tree structure by selecting the most informative features at each node. The DF model learns to recognize patterns and differences between the two groups based on these(like jitter, shimmer, frequency attributes) features. By traversing the tree structure, the model can make decisions on whether new voice recordings are more closely with those indicating Parkinson's or those from healthy individuals.

There is a slight overfitting issue in the data, which can be resolved by employing the SMOTE oversampling technique. And achieved maximum accuracy. The final

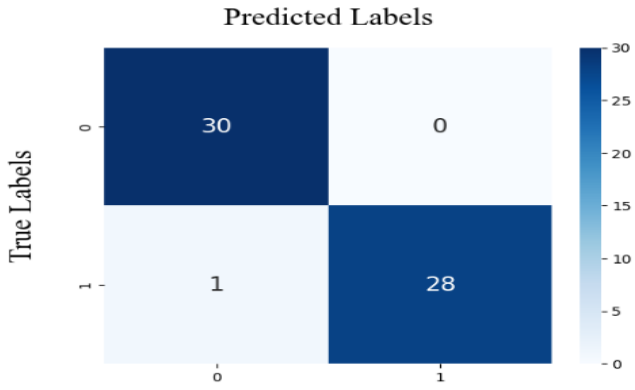

Fig. 12. Confusion matrix for Decision Tree in approach 2"

approach employs a DT model trained and evaluated using features specifically selected by the Relief algorithm. Compared to Approach 3, Approach 2 performed well for this decision tree model.

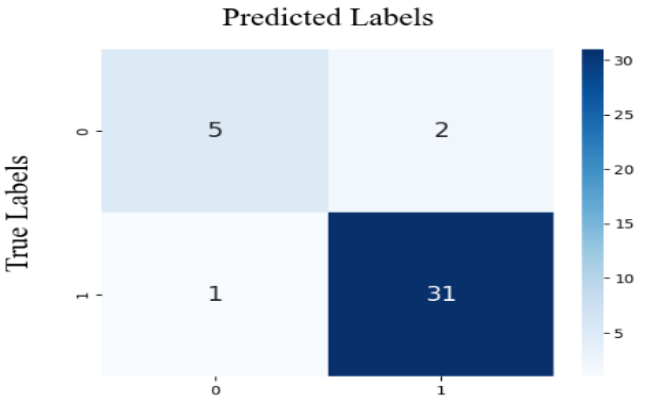

Fig. 13. Confusion matrix for Decision Tree in approach 3"

VII. MODEL EVALUATION

To pick out the first-rate model, we compare the outcomes of three methods and nine fashions educated. In contrast, the metrics chosen are ROC-AUC confusion matrix, accuracy, precision, recollect, and F1 rating. Formulae for those metrics are illustrated in equations 1-three, in which TP stands for True Positives, FP for False positives, TN for True Negatives and FN for False Negatives.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{(TP+TN+FP+FN)}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$\text{F1Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}$$

Table 1 below gives the results of approach 1, which is an unbalanced data

Metric	Random Forest	KNN	SVM	Decision Tree
Accuracy	94.87%	96.79%	88.46%	92.31%
Precision	94.12%	94.12%	88.89%	93.94%
Recall	100%	100%	100.0%	96.88%
F1 Score	96.97%	96.97%	94.12%	95.38%

TABLE I
PERFORMANCE METRICS OF DIFFERENT MODELS IN APPROACH 1

Table 2 below gives the results of approach 2, In this approach data SMOTE over-sampling technique is to create a more balanced dataset and achieved outstanding accuracy.

Metric	Random Forest	KNN	SVM	Decision Tree
Accuracy	100%	100%	85.96%	100%
Precision	100%	100%	92.59%	100%
Recall	100%	96.5%	86.21%	96.55%
F1 Score	100%	98.24%	89.29%	98.25%

TABLE II
PERFORMANCE METRICS OF DIFFERENT MODELS IN APPROACH 2

Table 3 below gives the results of approach 2, In this approach the models are trained and evaluated using the features selected by the Relief algorithm

Metric	Random Forest	KNN	SVM	Decision Tree
Accuracy	100%	95.51%	92.31%	100%
Precision	94.12%	87.5%	91.43%	90.62%
Recall	100%	96.5%	100%	90.62%
F1 Score	96.97%	98.24%	95,52%	90.62%

TABLE III
PERFORMANCE METRICS OF DIFFERENT MODELS IN APPROACH 3

VIII. RESULTS AND DISCUSSION

PD classification using vowel phonation data achieved 100% accuracy and 100% precision for the RF, KNN, and DT classifiers. The high accuracy and precision results were attributed to training and evaluating the models using the SMOTE over-sampling technique, ensuring a balanced dataset. From balancing the data we achieved 100% accurate results. This paper also highlights the outcomes of the SVM model that offers an accuracy of 92.31% and a precision of 91.43% after the models are trained and evaluated for the use of the features selected via the RELIEF algorithm. KNN, DT, and RF models are carried out properly for outliers and are robust models. The models are expecting no false positives within the results. Thus, RF, KNN, and DT perform well for MDVP audio data

IX. REFERENCES

1.Doneti Sowmya, Dodla Kavya, J. Rashmitha, V. Satheesh kumar, Preethi Jeevan.Parkinson;s Disease Detection By Machine Learning Using SVM https://www.irjet.net/archives/V10/i1/IRJETV10I1184.pdf

2.Atiqur Rahman, Sanam Shahla Rizvi , Aurangzeb Khan, Aaqif Afzaal Abbasi, Shafqat Ullah Khan, and Tae-Sun Chung.Parkinson's Disease Diagnosis in Cepstral Domain Using MFCC and Dimensionality Reduction With SVM Classifier.https://www.hindawi.com/journals/misy/2021/8822069/

3.Athanasios Tsanas,Max A.Little,Patrick E.McSharry,Jennifer Spielman, Lorraine O.Ramig.Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's

Disease.https://ieeexplore.ieee.org/document/6126094

4.Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig.Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease.https://www.ncbi.nlm.nih.gov/pmc/articles /PMC3051371/

5.Mohammad Shahbakhi, Danial Taheri Far, Ehsan Tahami. Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine.https://cyberleninka.org/article/n/1226851

6.Salim Lahmiri a, Amir Shmuel a.Detection of Parkinson's Disease Based on Voice Patterns Ranking and Support Vector Machine.https://www.sciencedirect.com/science /article/abs/pii/S1746809418302271//

7.Luca Parisi,REenfei MA,Mansour Youseffi. Support Vector Machine for Early Detection of Parkinson's Disease from Speech Signals.https://www.naun.org/main/NAUN/mcs /2021/a142002-007(2021).pdf/

8.Ahmed M. Elshewey 1 Mahmoud Y.Shams Nora El-Rashidy Abdelghafar M. Elhady Samaa M. Shohieb 4, and Zahraa Tarek Bayesian Optimization for Parkinson Disease Classification.https://pubmed.ncbi.nlm.nih.gov/36850682/