IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

Machine Learning for Early Detection of Parkinson's Disease

Anushree A[1] , Sowmya M S[2]

M C A Student, Department of Computer Application, Bangalore Institute of Technology, Bangalore, India[1]

Assistant Professor, Department of Computer Application, Bangalore Institute of Technology, Bangalore, India[2]

ABSTRACT: Parkinson's disease is a severe neurological condition that primarily affects elderly persons and impairs speech and movement. For patients, it presents difficulties because it makes daily duties more difficult..Detecting Parkinson's disease early is really important so doctors can start treatment sooner and help patients live better lives. Researchers are using technology called machine learning to help with this. Computers are used in machine learning to learn from data and provide predictions. Researchers used voice recordings from 30 participants with Parkinson's disease and 30 participants without the condition in this investigation.

They taught computers to recognize patterns in these recordings. After testing four different methods of training the computers Random Forest, K-nearest neighbors, Support Vector Machine, and logistic regression models they discovered that Random Forest was the most effective in identifying Parkinson's illness.

It was right almost 92% of the time. This means it correctly identified Parkinson's disease in 95% of the people who actually had it. By demonstrating how effectively this technology functions in telemedicine, the researchers expect to (using technology to treat patients from a distance), moredoctors and patients will use it.

I. INTRODUCTION

Parkinson's disease (PD) is a common brain condition that affects movement. It results in symptoms such as sluggish movements, tremors, and stiff muscles because dopamine levels in the brain are reduced. Because there is no known cure, slowing the disease's course requires early detection and specialized therapy.

One early sign of PD is vocal cord issues, which can be easily measured through voice tests done remotely, like recording a vowel sound or speaking a sentence. Even before other symptoms manifest, these tests can aid in the early diagnosis of Parkinson's disease. Deep brain stimulation is one therapy that doctors can use to help manage Parkinson's disease symptoms by increasing dopamine levels once the disease has been identified early. While there's no cure yet, early detection followed by the right treatment can reduce tremors and balance problems, allowing patients to live more normally. This research focuses on using machine learning (ML) techniques to detect PD from audio recordings of patients' voices. The Random Forest model showed promising accuracy (91.83%) in identifying PD from specific voice characteristics. This method could revolutionize telemedicine by allowing patients to monitor their condition using simple voice recordings on their phones, without needing frequent visits to a clinic. The project intends to improve telemedicine solutions for Parkinson's disease (PD) by comparing several machine learning (ML) models, hence facilitating faster and easier diagnosis for patients who have mobility challenges.

1.1 Literature survey

Scientists have investigated a number of approaches to predict Parkinson's disease (PD), including the use of MRI scans, genetics, walking patterns, and, more recently, the analysis of auditory deficits.

Previous studies focused on genetic data achieved accuracies around 88.9% using SVM models, but newer research described in this paper improves on this with 91.83% accuracy using the same model. This demonstrates how much more successful audio data is at detecting Parkinson's disease (PD) than genetic data.

Other research employed distinct methodologies, such as deep learning models applied to voice data or keystroke analysis.

However, some of these methods were complex and required specialized software like MATLAB.. On the other hand, this study makes advantage of Python's open-source capabilities, which are quicker and more effective.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

Deep learning models are widely used for Parkinson's disease (PD) detection; nevertheless, they frequently demand high computational and memory resources. Principal Component Analysis (PCA) is one of our methods for streamlining the data and enhancing the functionality of more basic machine learning models, such as random forest, logistic regression, KNN, and SVM.

Prior studies also examined many facets of Parkinson's disease, including brain scans and movement patterns. Certain research sought to increase detection by thoroughly analyzing different biomarkers, or to lessen the subjectivity of physician
assessments.
In order to categorize the conditions of PD patients, these machine learning models were applied to their audio data in this study. Our objective is to improve PD detection.

Telemedicine can identify Parkinson's disease (PD), facilitating patients' remote health monitoring. According to their first findings, the K nearest neighbor model has the highest accuracy (91.83%) and highest sensitivity, suggesting that it has the ability to accurately diagnose Parkinson's disease (PD) using speech recordings.

II. PROPOSED METHODOLOGY

This study collects voice data from Parkinson's patients using PPMI and UCI datasets, focusing on features like jitter, shimmer, and MDVP during vowel phonations. Four models are trained using the cleaned data: SVM, KNN, Random Forest Regressor, and Logistic Regression. The remaining 25% of the data is used for testing.

These models learn to classify audio recordings as either PD or healthy based on voice frequency variations.
Metrics including sensitivity, precision, accuracy, confusion matrix, and ROC-AUC score are used to evaluate the models' performance.

The research aims to identify key voice attributes for PD classification and understand how data imbalance affects accuracy. Three approaches are tested: training on the full dataset, using PCA for attribute selection, and training on a balanced dataset. Overall, the goal is to enhance PD detection through voice analysis with machine learning, potentially improving early diagnosis and treatment outcomes.

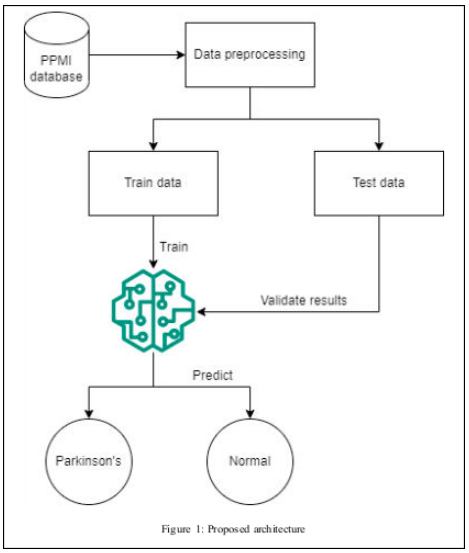


Figure 1: Proposed architecture

Fig 1. illustrates the generic process implemented. It demonstrates the stages of data ingestion from PPMI database, separation of data into testing and training sets, training of four models on data and validation of results using test data.
Algorithm for approach 1: Models are trained on 22 attributes of data
* Collect MDVP audio data from PPPMI and UCI databases
* Perform data analysis to detect skew, imbalance and distribution of variables in data
* Utilize the Standard Scaler to scale the data to a common range.
* Split dataset into testing and training sets, where training data is 75% of total
* Train SVM, logistic regression, random forest and KNN models.

International Journal of Innovative Research in Computer and Communication Engineering

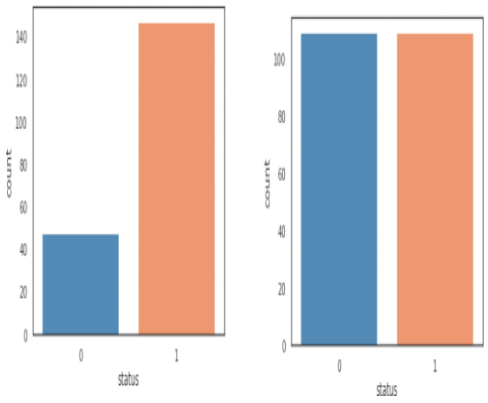
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

Algorithm for approach 2: Analysis of Principal Components (PCA)
(PCA) is applied to identify 5 key attributes
* Collect MDVP audio data from PPPMI and UCI databases
* Perform data analysis to detect skew, imbalance and distribution of variables in data
* Resize the information to a common range using Standard Scaler
* Identify variance in every column of informationand analysis of principal componenet(PCA) to identify 5 most relevant features to model training, out of 22 attributes.
* Split dataset into testing and training sets, where training data is 75% of total
* Retrain Random Forest, Logistic Regression, and SVM and KNN models.
* Compare classification results using confusion matrix, ROC-AUC curve and accuracy
Algorithm for approach 3: Imbalance removal in dataset
* Collect MDVP audio data from PPPMI and UCI databases
* Perform data analysis to detect skew, imbalance and dissemination of variables in data
* The dataset is imbalanced, with 109 records of PWP and 40 records of normal people, as illustrated in

figure 2(a). The imbalance is resolved by up sampling [23] the minority class to reach 109 records each, as seen in figure 2 (b).
* Utilize the Standard Scaler to scale the data to a common range.
* Split dataset into testing and training sets, where training data is 75% of total
* Retrain the KNN, SVM, random forest, and logistic regression models.
* Compare classification results using confusion matrix, ROC-AUC curve and accuracy



Voice measurements were taken from 31 people, including 23 with Parkinson's disease. Patients ranged from 46 to 85 years old, while healthy individuals averaged 23 years. Each person's voice was recorded about 6 times across 195 sessions, lasting 1 to 36 seconds per recording.

Information processing
Information cleaning was done to tidy up the dataset and handle missing information. Figure 3 illustrates how the noise-to-harmonic ratio (NHR) increases as Parkinson's disease progresses, indicating poorer voice quality

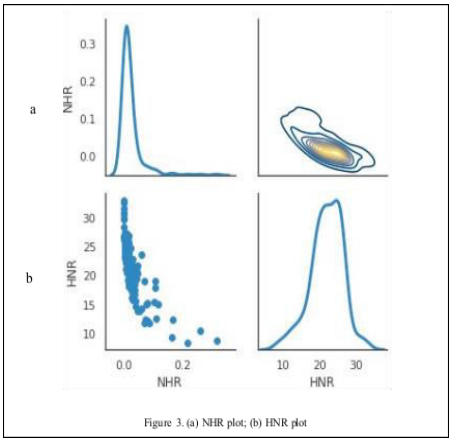


Figure 3. (a) NHR plot; (b) HNR plot

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

Figure 4 shows a box plot of 22 different attributes from the dataset. It uses blue for normal records and orange for records from Parkinson's patients (PWP). In the PWP group, the plot indicates more outliers in the NHR (noise-to-harmonic ratio), suggesting higher speech noise. The HNR (harmonic-to-noise ratio) also has several outliers below the median, showing differences in these voice features compared to normal records.

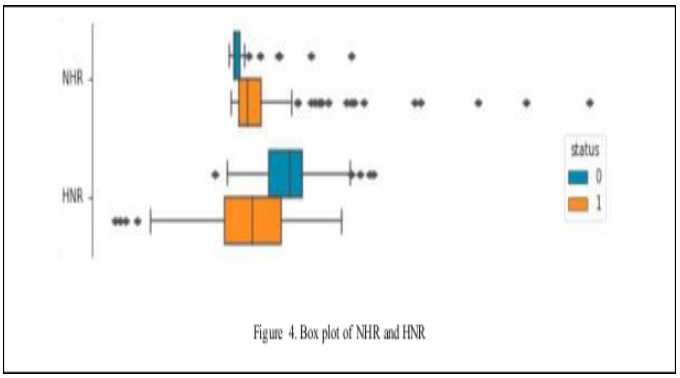


Figure 4. Box plot of NHR and HNR

Figure 5 displays a pair plot of shimmer data, comparing the voice shimmer characteristics between Parkinson's patients (PWP) and healthy individuals. It shows:
- Shimmer: APQ3 and Shimmer: DDA are positively correlated.
- Shimmer: APQ5 and Shimmer: APQ3 have a left-skewed relationship.
These relationships highlight differences in voice shimmer patterns between Parkinson's patients and healthy individuals.

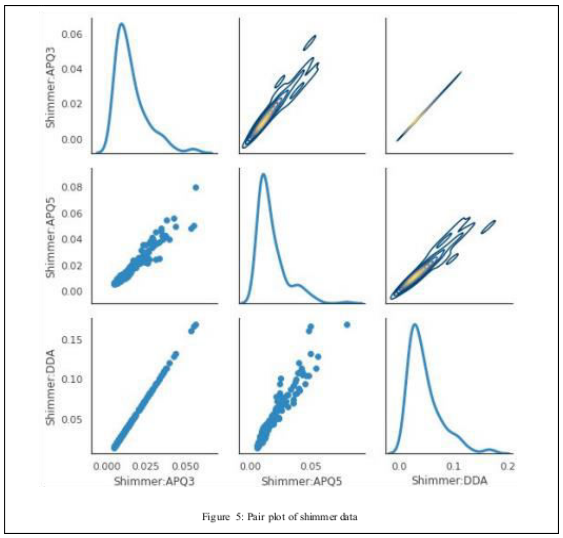


Figure 5: Pair plot of shimmer data

III. Model training

This research paper examines Random Forest, Support Vector, and Logistic Regression classifiers. and K nearest neighbours' models in 3 approaches:
* Complete dataset of 195 records and 22 attributes
* Dataset with 195 records and 5 attributes after Principal Component Analysis (PCA)
* Balanced dataset with 109 records and 22 attributes

3.1 Logistic regression for classification

Logistic regression is a technique for machine learning that forecasts categorical outcomes based on input variables. It uses a logistic curve to compute probabilities between 0 and 1, which is ideal for scenarios where relationships aren't linear but follow A more intricate pattern, like in audio data used to classify Parkinson's disease. The activation function of logistic classification has been illustrated in figure 6.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

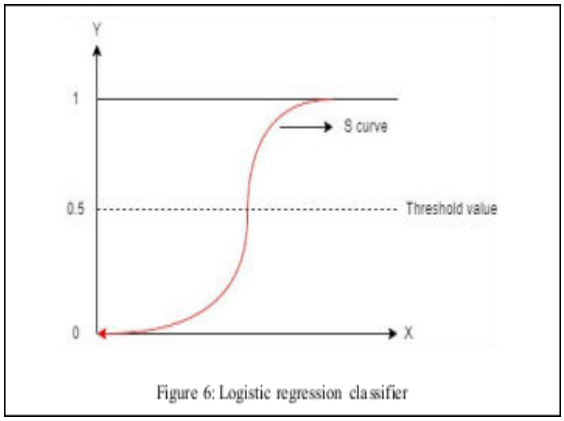


Figure 6: Logistic regression classifier

3.2 Random Forest classifier

Random Forest classification algorithm

One kind of machine learning algorithm for handling regression and classification issues is called Random Forest. This study uses a random forest classifier, which generates several decision trees depending on various dataset subsets. Every decision tree makes an individual forecast, and the ultimate prediction is arrived at by a vote of all the predictions. Random forests employ the average forecast from all trees to increase accuracy and lower the chance of overfitting, in contrast to traditional decision trees where one tree's prediction may predominate. The model grows more resilient as additional trees are added to the forest because it takes into account a wider variety of viewpoints from the various trees.

Figure 7 likely illustrates how this ensemble of decision trees works together in the random forest classifier, showing how predictions are aggregated to produce a more reliable final output.

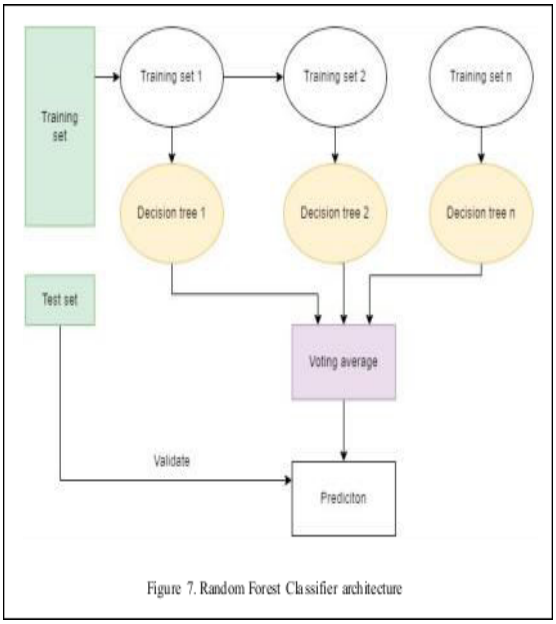


Figure 7. Random Forest Classifier architecture

3.3 Support vector machine (SVM)

Support Vector Machine (SVM a machine learning algorithm applied to tasks involving classification. It creates a hyperplane in a high-dimensional space to separate different categories of data based on their features. SVM works well for Parkinson's disease voice data because it can handle non-linear relationships using a technique calledthe kernel trick, which creates a higher-dimensional space out of the data. Because SVM defines the It is memory-efficient and uses a decision boundary with a subset of training data points called supportvectors.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

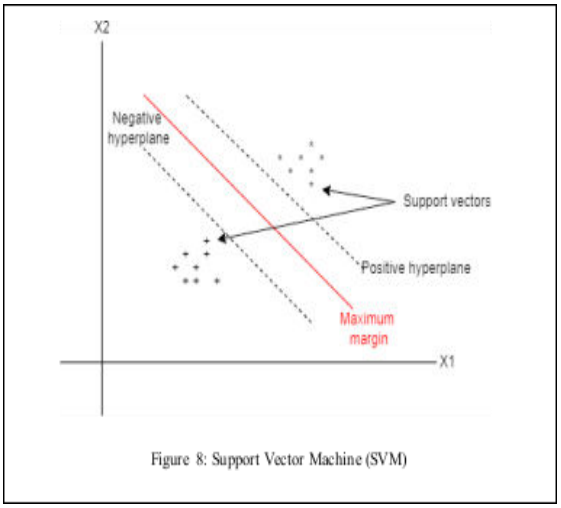


Figure 8: Support Vector Machine (SVM)

3.4 K nearest neighbors (KNN)

K Nearest Neighbors (KNN) is a method of machine learning that clusters similar data points together based on their features. It doesn't assume any specific patterns in the information and works well with small datasets. For instance, in a dataset of 109 audio recordings, KNN efficiently creates two clusters: one for Parkinson's Disease (PWP) and another for healthy individuals. This method is effective because it learns directly from the data without needing to construct a complex model beforehand. It's especially helpful in situations where the dataset is balanced, meaning it has roughly equal numbers of examples for each category, allowing KNN to accurately categorize fresh data points based on their similarity to existing examples.

IV. MODEL EVALUATION

To Identify the top-performing model. among three approaches with nine trained models, several essentialmetrics are used for evaluation. These metrics include the ROC-AUC curve, confusion matrix, recall, accuracy, precision, and F1 score.

The ROC-AUC curve is a graphical representation that assesses a model's ability to distinguish between classes by plotting the ratio of false positives to true positives. A greater AUC value denotes superior performance in classification tasks. The confusion matrix provides a detailed breakdown of a model's predictions, displaying false positives, false negatives, real positives, and true negatives. Accuracy measures the overall correctness of the model's predictions, while precision quantifies the percentage of all anticipated positives that were accurately forecasted as positives. Recall, which is a synonym for sensitivity, quantifies the model's ability to distinguish positive examples from all real positives.

The F1 score combines precision and recall into a single metric, offering a balanced assessment of a model's accuracy. Together, these metrics provide comprehensive insights into the effectiveness of every model, assisting in the choice of the best strategy for the particular categorization task.

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)

Accuracy=(2*precision*recall)/(precision+recall)

Table 1 presents the performance results of four different models after applying Approach 1, where each model was trained using 22 attributes from the MDVP dataset.

Table1.ResultsofApproach1:22attributestraining

Metric	LogisticRegression	RandomForest	SVM	KNN
Accuracy	83.67%	91.83%	85.71%	85.71%

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

Precision	1.0	0.95	1.0	0.95
Recall	0.83	0.86	0.84	0.86
ROCAUCcurve	0.636	0.701	0.682	0.701

Random forest combines forecasts derived from several decision trees to improve accuracy. It evaluated an average of 100 decision trees in this particular case. The confusion matrix shown in Figure 9 further details its performance, showing it correctly identified 38 people with PD (true positives) and 7 healthy individuals (true negatives), with 4 instances of misclassification (false negatives).

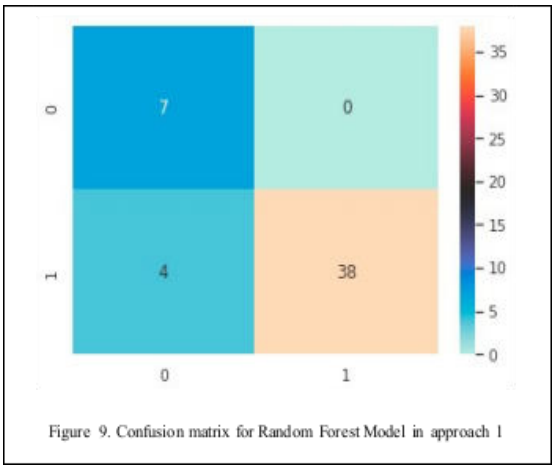


Figure 9. Confusion matrix for Random Forest Model in approach 1

Table 2 presents the performance outcomes of four models after applying Approach 2, which involved using Principal Component Analysis (PCA) to reduce the dataset to five major attributes: MDVP, Shimmer, Jitter, PPE, and RPDE.

Table2.Resultsof Approach2:5attributesafterPCA

Metric	LogisticRegression	RandomForest	SVM	KNN
Accuracy	83.67%	83.67%	91.75%	83.67%
Precision	1.0	1.0	1.0	0.92
Recall	0.83	0.90	0.86	0.90
ROCAUCcurve	0.636	0.818	0.727	0.779

Help Vector Machine used an L1-support with a linear kernel, which is particularly suitable for PCA-derived data. It efficiently finds the ideal hyperplane (decision boundary) with high accuracy. The confusion matrix in Figure 10 details

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

its performance, showing that SVM correctly classified 38 people with PD (true positives) and 5 healthy individuals (true negatives), with 6 instances of misclassification (false negatives).

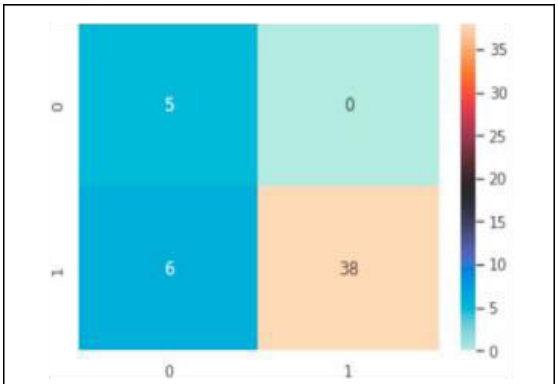


Figure 10. Confusion matrix for Support Vector Model in approach 2

Table 3 presents the results of Approach 3, where models were trained on a dataset balanced with a same quantity of records for both normal individuals and patients with Parkinson's Disease (PWP).

Table3.Resultsof Approach3:Balanced dataset

Metric	LogisticRegression	RandomForest	SVM	KNN
Accuracy	85.71%	85.71%	81.63%	91.83%
Precision	0.89	0.89	0.82	0.95
Recall	0.92	0.92	0.94	0.95
ROCAUCcurve	0.811	0.811	0.817	0.883

K Nearest Neighbors model performed best in the balanced dataset scenario due to its high precision and recall of 0.95 and 0.92, respectively. With an equivalent quantity of records for both categories, KNN efficiently identifies similarities and differences, making its classification of Parkinson's Disease faster and more accurate. The confusion matrix in Figure 11 illustrates its performance, showing that KNN correctly classified 36 people with Parkinson's (true positives) and 9 healthy individuals (true negatives), with 2 false negatives and 2 false positives.

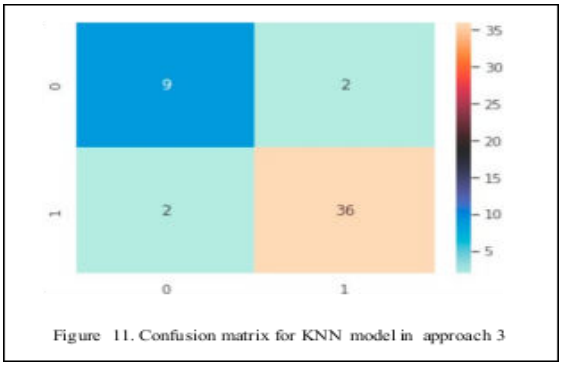


Figure 11. Confusion matrix for KNN model in approach 3

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207043 |

V. FINDINGS AND CONVERSATION

It talks about how several models of machine learning perform in classifying Parkinson's Disease using vowel phonation data.

Random Forest Classifier achieves 91.835% accuracy and correctly identifies 95% of Parkinson's cases (sensitivity). It's considered excellent because it gives equal importance to all 22 attributes in the dataset, making it robust against varying data characteristics and ensuring reliable predictions without false positives.

After applying PCA to simplify the dataset, SVM reaches an accuracy of 91.836% and a sensitivity of 94%. SVM excels in handling outliers and performs well in separating data into different categories.

K Nearest Neighborsis known for its ability to classify information without making any assumptions aboutpatterns, KNN also performs strongly in datasets where Parkinson's and non-Parkinson's cases are equally represented.forward, the study suggests enhancing classification by combining audio data with REM sleep data, recognizing that audio alone may not suffice as a reliable biomarker for Parkinson's Disease.

REFERENCES

1.Chen, R., Herskovits, E. H. (2007). Machine-learning techniques for building a diagnostic model for very mild dementia. Neuro Image, 27(4), 1078-1087.
2.Rana, B., Barua, S., Das, D., Yadav, D., & Pal, S. (2020). Parkinson’s disease diagnosis using a novel hybrid feature selection method based on ML techniques. International Journal of Neuroscience, 130(1), 89-99.
3.Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37(2), 1568-1572.
4.Almasi, S. A., Rahmani, A. M., & Hosseinzadeh, M. (2019). Parkinson's disease detection using deep learning algorithm based on Autoencoder model and fine-tuning. Informatics in Medicine Unlocked, 16, 100196.
5.Prashanth, R., Roy, S. D., Mandal, P. K., & Ghosh, S. (2014). High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. International Journal of Medical Informatics, 84(6), 444-457.
6.Mostafa, S. A., Mustapha, A., Mahmuddin, M., Sulaiman, S. N., Hamid, S. H. A., & Deris, S. (2018). Development of an automated diagnostic tool for Parkinson's disease based on hybrid fuzzy expert system and artificial neural network. Neural Computing and Applications, 30(4), 1285-1302.
7.Ozcift, A., & Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Computer Methods and Programs in Biomedicine, 104(3), 443-451.
8.Solana-Lavalle, G., Mena-Maldonado, E., & Cervantes, J. (2020). Parkinson’s disease detection based on video analysis: A systematic review. Applied Sciences, 10(19), 6935.
9.Chen, S. G., Lin, C. H., Wu, M. J., & Chen, C. H. (2020). Using deep learning with a neuro-fuzzy system to construct a global feature extractor for the diagnosis of Parkinson’s disease. Sensors, 20(15), 4296.
10.Vaishali, K., & Pachaiyappan, S. (2021). Predicting Parkinson’s disease using machine learning techniques: A survey. Materials Today: Proceedings, 37, 3528-3531.
11.Li, Y., Zhang, H., & Wu, T. (2017). Data-driven gait analysis using a machine learning approach. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(8), 1533-1542.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details