A Machine Learning Approach to Predict Parkinson's Disease Using Voting Classifier

CH. VASUNDHARA, BUBATHULA BHAVANA

Assistant Professor, 2MCA Final Semester,

Master of Computer Applications,

Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

Abstract:

Biomarkers derived from human voice can offer insight into neurological disorders, such as Parkinson's disease (PD), because of their underlying cognitive and neuromuscular function. PD is a progressive neurodegenerative disorder that affects about one million people in the United States, with approximately sixty thousand new clinical diagnoses made each year. Historically, PD has been difficult to quantify and doctors have tended to focus on some symptoms while ignoring others, relying primarily on subjective rating scales. Due to the decrease in motor control that is the hallmark of the disease, voice can be used as a means to detect and diagnose PD. With advancements in technology and the prevalence of audio collecting devices in daily lives, reliable models that can translate this audio data into a diagnostic tool for healthcare professionals would potentially provide diagnoses that are cheaper and more accurate.

We provide evidence to validate this concept using a voice dataset collected from people with and without PD. Based on previous research, different machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) have been explored, but with limited accuracy (around 81%). Our proposed research achieves a better accuracy of 95% using a Voting Classifier combining XGBoost and Support Vector Classifier, thereby improving diagnostic reliability and advancing Parkinson's detection methods.

IndexTerms: Parkinson's Disease, Machine Learning, Support Vector Classifier (SVC), Voice Biomarkers, Neurodegenerative Disorder, Disease Prediction, Feature Extraction, Biomedical Signal Processing, Data Preprocessing, Ensemble Learning.

INTRODUCTION

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that primarily affects motor functions due to the gradual loss of dopamine-producing neurons in the brain. Common symptoms include tremors, rigidity, bradykinesia (slowness of movement), speech difficulties, and postural instability. While PD mainly affects older adults, early onset cases are also reported. Currently, diagnosis is largely clinical, relying on observable symptoms and subjective assessments, which may lead to delayed or inaccurate detection. Early diagnosis is crucial for better disease management and to slow progression. Advances in technology, particularly in machine learning (ML), have opened new avenues for the early detection of such disorders using non-invasive methods. Voice signals have emerged as a promising biomarker since PD often causes speech impairment before more noticeable physical symptoms appear. By extracting and analysing vocal features, it is possible to identify patterns indicative of PD. In this project, we propose a robust machine learning model that uses a Voting Classifier combining Boost and Support Vector Classifier (SVC) to improve prediction accuracy. The dataset used comprises voice recordings from both healthy individuals and PD patients. Through preprocessing, feature scaling, and model training, the system achieves high diagnostic precision. This ensemble approach overcomes the limitations of individual classifiers by integrating their strengths. The model achieved a significant accuracy improvement compared to previous works. With further development, this system could assist clinicians in early and cost-effective PD diagnosis. The project demonstrates the effectiveness of combining biomedical data with ML to address real-world healthcare challenges.

Existing System

The existing systems for Parkinson's Disease (PD) detection predominantly rely on traditional clinical evaluations and subjective assessment scales such as the Unified Parkinson's Disease Rating Scale (UPDRS). These assessments are often time-consuming, inconsistent, and require trained neurologists for accurate diagnosis. In recent years, machine learning techniques like Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbors (K-NN), and Artificial Neural Networks (ANN) have been explored to automate PD diagnosis using voice data and motor features. However, these models often suffer from limitations such as overfitting, lower accuracy, and poor generalization to new patient data. Most of these approaches operate on a single classifier, which restricts the ability to capture complex data patterns. Additionally, earlier models failed to efficiently handle high-dimensional data or were not optimized for real-time clinical use. Feature selection techniques were often basic, and data preprocessing was minimal, affecting prediction quality. Moreover, many models achieved accuracies only in the range of 80–85%, which is not sufficient for medical applications. The systems lacked robustness, scalability, and the flexibility to integrate into practical healthcare workflows. Inadequate data handling and absence of ensemble learning further limited their effectiveness. No real-time prediction or deployment mechanism was implemented in many existing works. As a result, there is a significant need for improved, accurate, and scalable diagnostic systems for early PD detection using modern ML techniques

Challenges:

- **Limited and sensitive dataset** – Accessing a high-quality, labeled dataset with sufficient size and diversity is difficult due to medical data privacy and availability.
- **Noisy and inconsistent input data** – Voice recordings can be affected by background noise, recording equipment, or patient conditions, reducing data quality.
- **Complex feature extraction** – Extracting relevant vocal features from raw audio data requires precise preprocessing and domain knowledge.
- **Class imbalance** – Unequal distribution of healthy and PD-affected cases can lead to biased predictions.
- **Model selection difficulties** – Choosing the most suitable machine learning algorithms for accurate and generalized results is not straightforward.
- **Hyperparameter tuning** – Algorithms like SVM and XGBoost need careful fine-tuning for optimal performance, which is time-consuming

Proposed system:

The proposed system utilizes a hybrid machine learning approach combining XGBoost and Support Vector Classifier (SVC) within a Voting Classifier framework to accurately predict Parkinson’s Disease (PD) using vocal biomarkers. This ensemble model leverages the strengths of both classifiers: XGBoost’s ability to handle non-linear relationships and SVC’s robustness in high-dimensional spaces. The system begins by collecting a voice dataset from both healthy individuals and PD patients, followed by comprehensive data preprocessing, including feature selection and normalization using Min-Max scaling. After preprocessing, the dataset is split into training and testing sets to evaluate model performance. Both XGBoost and SVC are individually trained and then integrated into a Voting Classifier to produce more accurate and stable predictions. This ensemble learning strategy helps reduce variance and bias, leading to better generalization. The system outputs binary predictions indicating the presence or absence of PD, along with performance metrics like accuracy, precision, recall, and F1-score. The model is then saved using joblib for reuse and deployment. A Flask-based API is developed for real-time predictions, making the system accessible for healthcare providers. The interface is user-friendly, enabling easy input of patient vocal parameters and returning instant diagnostic predictions. This proposed system addresses the limitations of existing models by improving prediction accuracy (up to 95%) and offering a scalable, interpretable, and deployable solution for early PD diagnosis.

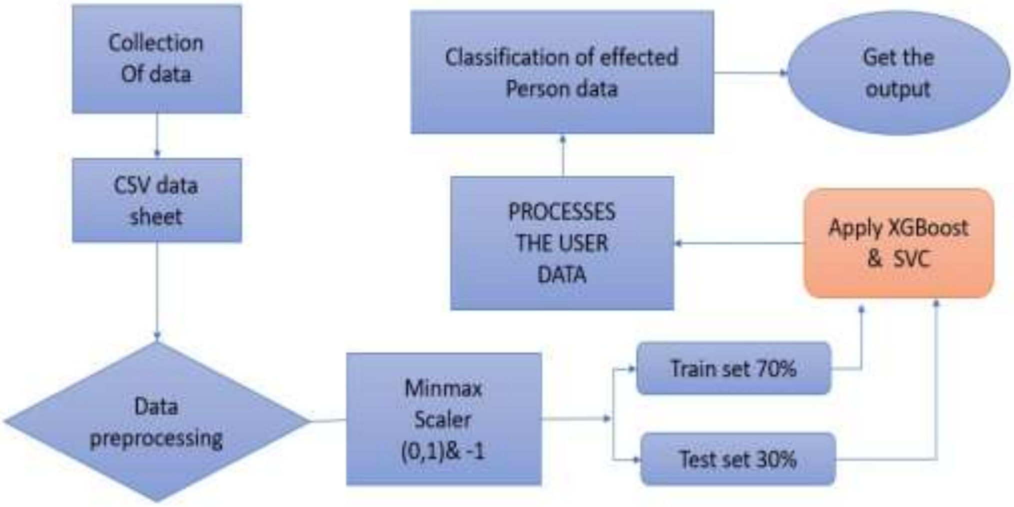


Fig: Architecture of proposed system

UML DIAGRAMS:

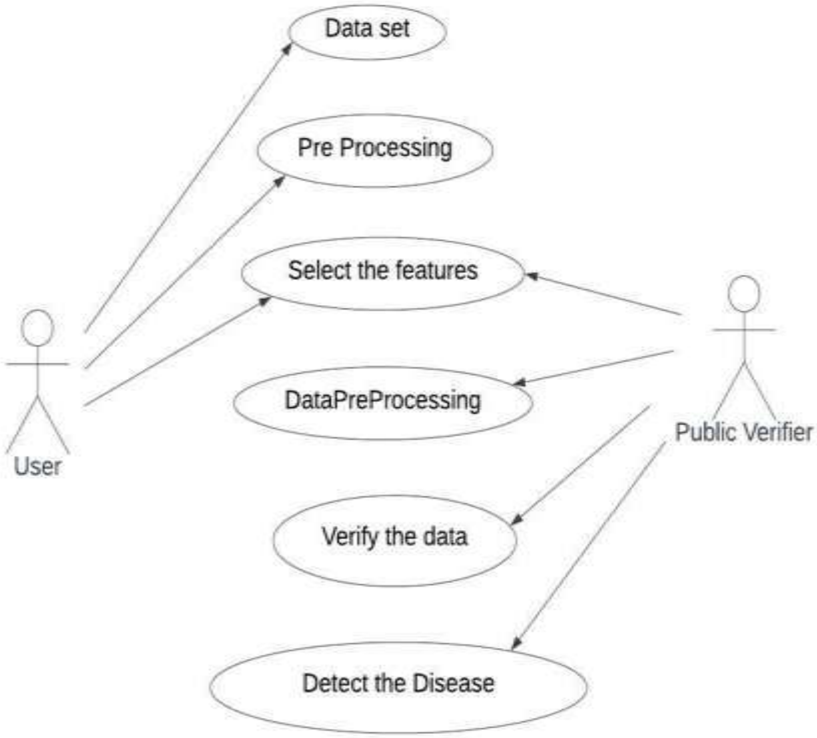


FIG: USE CASE DIAGRAM

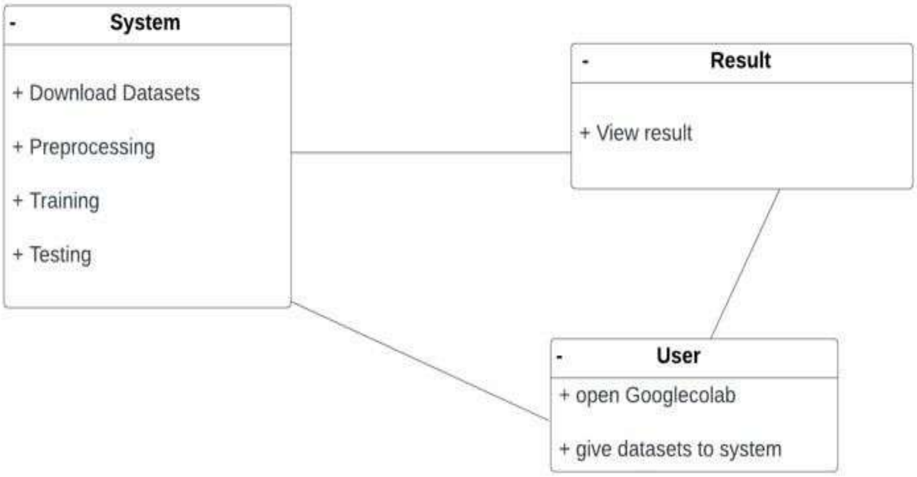


FIG:CLASS DIAGRAM

Advantages:

- **High Accuracy**: Achieves up to 95% prediction accuracy using ensemble learning.
- **Better Generalization**: Reduces overfitting by combining XGBoost and SVC.

- **Non-Invasive Diagnosis**: Uses voice biomarkers, avoiding costly medical procedures.
- **Real-Time Predictions**: Integrated with a Flask API for instant diagnostic results.
- **User-Friendly Interface**: Simple and accessible for healthcare providers.
- **Scalable & Deployable**: Suitable for integration into real-world clinical systems.
- **Robust to Noise**: Handles noisy and inconsistent voice input effectively.
- **Cost-Effective**: Lowers diagnostic cost by leveraging commonly available voice data.

Architecture:

The architecture of the proposed system for predicting Parkinson’s disease using a voting classifier (combining XGBoost and Support Vector Machine) follows a structured and efficient workflow. It begins with the collection of voice data containing relevant biomedical features such as Jitter, Shimmer, NHR, and HNR, typically extracted from audio recordings of both healthy individuals and patients affected by Parkinson’s disease. Once the data is gathered, it undergoes preprocessing where unnecessary columns like patient names are removed, missing values are handled, and features are normalized using the MinMaxScaler technique to bring them within a uniform scale. Following this, the dataset is split into training and testing sets (usually 70% training and 30% testing). Both the XGBoost classifier and Support Vector Machine (SVM) are individually trained on the dataset, with hyperparameter tuning performed using Grid Search and cross-validation to enhance their predictive performance. The outputs from both models are then combined using a voting classifier that aggregates their predictions through majority or weighted voting, thereby improving accuracy and reliability. This ensemble model is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to validate its effectiveness. The final trained model is serialized using tools like joblib or pickle and deployed using an API framework such as Flask or FastAPI, allowing real-time access for diagnostic purposes. The deployed system includes a user-friendly web interface for healthcare professionals to input patient data and receive predictions, while also incorporating periodic retraining and monitoring features to ensure ongoing accuracy and robustness in clinical settings.

Algorithm:

The proposed system uses a Voting Classifier algorithm that combines two powerful machine learning models: XGBoost and Support Vector Machine (SVM) to predict Parkinson’s disease from vocal features. The process begins with data collection, where a dataset containing numerical voice measurements (such as jitter, shimmer, NHR, HNR, etc.) from both Parkinson's patients and healthy individuals is compiled. This data is then preprocessed by removing unnecessary columns, handling missing values, and applying MinMaxScaler normalization to scale the features to a uniform range. Next, the dataset is split into training and testing sets, typically with a 70-30 ratio. The XGBoost classifier is trained using gradient boosting techniques that optimize performance by sequentially building trees. Simultaneously, the SVM classifier is trained to find the best hyperplane that separates the classes, using kernel functions and regularization parameters tuned via Grid Search and Cross-Validation. Both models are then integrated into a Voting Classifier, which combines their individual predictions either by majority (hard voting) or by weighted probabilities (soft voting). The ensemble output is then used to classify whether a person has Parkinson’s disease or not. The system’s performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Finally, the model is saved for deployment, and a Flask or FastAPI framework is used to serve predictions through a real-time web-based interface for clinical use.

Techniques:

The project employs a combination of machine learning techniques to improve the accuracy and reliability of Parkinson’s disease diagnosis. At the core of the system is a Voting Classifier, which integrates two key models: XGBoost (Extreme Gradient Boosting) and Support Vector Machine (SVM). These techniques are chosen for their high performance in classification tasks. XGBoost is known for its gradient boosting framework that builds decision trees iteratively to minimize error, while SVM is a powerful supervised learning algorithm effective in high-dimensional spaces, using hyperplanes to separate classes. Additionally, data preprocessing techniques such as MinMaxScaler normalization are applied to bring features within a uniform range, which helps improve model performance. The system also incorporates feature extraction and dimensionality reduction implicitly during preprocessing, and hyperparameter tuning through Grid Search and Cross-Validation to optimize model parameters. To evaluate model performance, techniques such as accuracy, precision, recall, F1-score, and ROC-AUC score are used. The combination of these techniques ensures robust, scalable, and interpretable diagnostic predictions for Parkinson’s disease using voice data.

Tools:

The development of the proposed Parkinson’s disease detection system utilizes a variety of software and analytical tools. The programming language used is Python, which is chosen for its simplicity and vast ecosystem of libraries. Key Python libraries include NumPy for numerical operations and array handling, Pandas for data manipulation and analysis, Scikit-learn for implementing machine learning algorithms such as SVM and for preprocessing tasks like scaling and data splitting, and XGBoost for high-performance gradient boosting classification. Visualization and plotting are carried out using Matplotlib and Seaborn to analyze and present model performance and feature distributions. The model is deployed using Flask or FastAPI, both of which are lightweight web frameworks that enable the creation of APIs for real-time predictions. Additionally, PRAAT software is referenced for extracting vocal features from audio recordings, which are then used as input to the machine learning models. These tools collectively contribute to building a scalable, efficient, and user-friendly system for diagnosing Parkinson's disease.

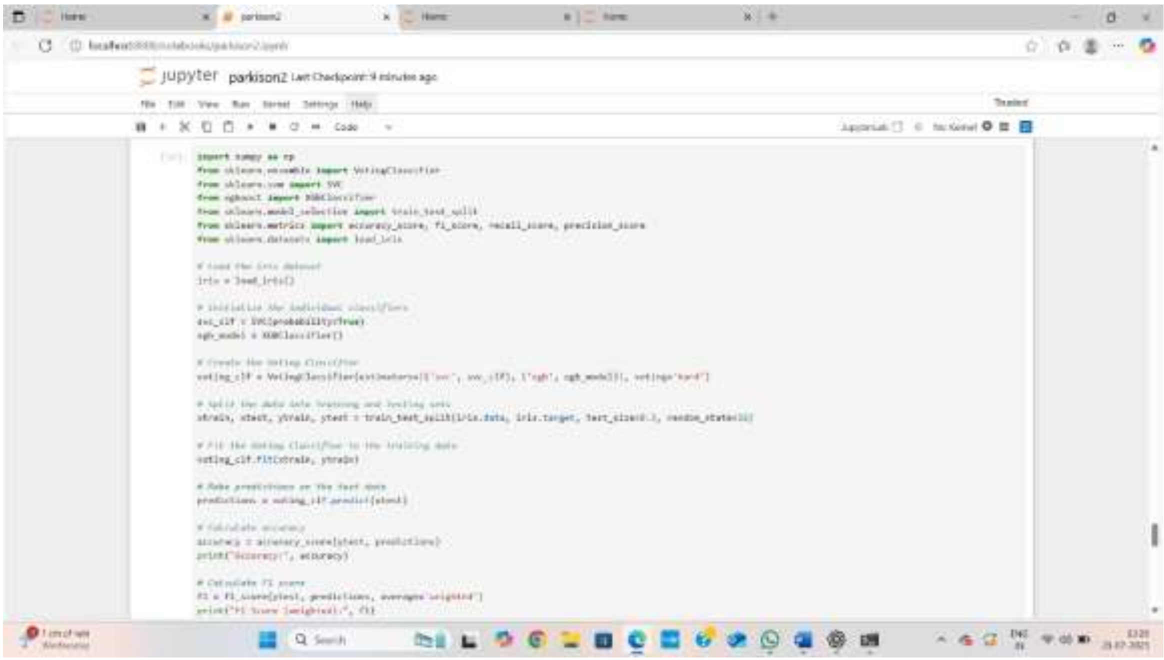
Methods:

The proposed system adopts a range of methods to accurately detect Parkinson's disease from vocal features. It begins with the data collection method, where vocal measurements from both healthy individuals and Parkinson's patients are gathered and structured into a dataset. This is followed by data preprocessing methods, which include handling missing values, removing irrelevant columns (like patient names), and applying MinMaxScaler normalization to standardize feature ranges. For model building, the system employs supervised machine learning methods, specifically XGBoost and Support Vector Machine (SVM), which are then combined using a Voting Classifier method to improve classification performance through ensemble learning. Hyperparameter tuning is conducted using Grid Search along with Cross-Validation to optimize model parameters. The models are then evaluated using performance measurement methods such as accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, the system incorporates deployment methods using Flask or FastAPI to provide real-time access to the model through an API. These methods together contribute to a reliable, efficient, and scalable solution for early-stage Parkinson's disease detection.

METHODOLOGY

Input:

The machine learning model in the provided code consists of the Iris dataset, which contains features representing the measurements of iris flowers, such as petal and sepal lengths and widths, along with corresponding class labels indicating the species. The dataset is split into training and testing subsets, where 70% of the data is used for training the model, and 30% is reserved for testing. Two classifiers are initialized for the prediction task: a Support Vector Classifier (SVC) and an XGBoost Classifier. These classifiers are combined into a Voting Classifier, which aggregates the predictions from both models using a 'hard' voting strategy. This means the final prediction is based on the majority vote of the individual classifiers' outputs. The input data is preprocessed, and then the models are trained and tested to calculate performance metrics like accuracy and F1 score.


jupyter parkison2 Last Checkpoint: 9 minutes ago
File Edit View Run Kernel Settings Help
Code

```
import numpy as np
from sklearn.ensemble import VotingClassifier
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, f1_score, recall_score, precision_score
from sklearn.datasets import load_iris

# Load the iris dataset
iris = load_iris()

# Initialize the individual classifiers
svc_clf = SVC(probability=True)
xgb_model = XGBClassifier()

# Create the Voting Classifier
voting_clf = VotingClassifier(estimators=[('svc', svc_clf), ('xgb', xgb_model)], voting='hard')

# Split the data into training and testing sets
xtrain, xtest, ytrain, ytest = train_test_split(iris.data, iris.target, test_size=0.3, random_state=42)

# Fit the Voting Classifier to the training data
voting_clf.fit(xtrain, ytrain)

# Make predictions on the test data
predictions = voting_clf.predict(xtest)

# Calculate accuracy
accuracy = accuracy_score(ytest, predictions)
print("Accuracy:", accuracy)

# Calculate F1 score
f1 = f1_score(ytest, predictions, average='weighted')
print("F1 Score (weighted):", f1)
```

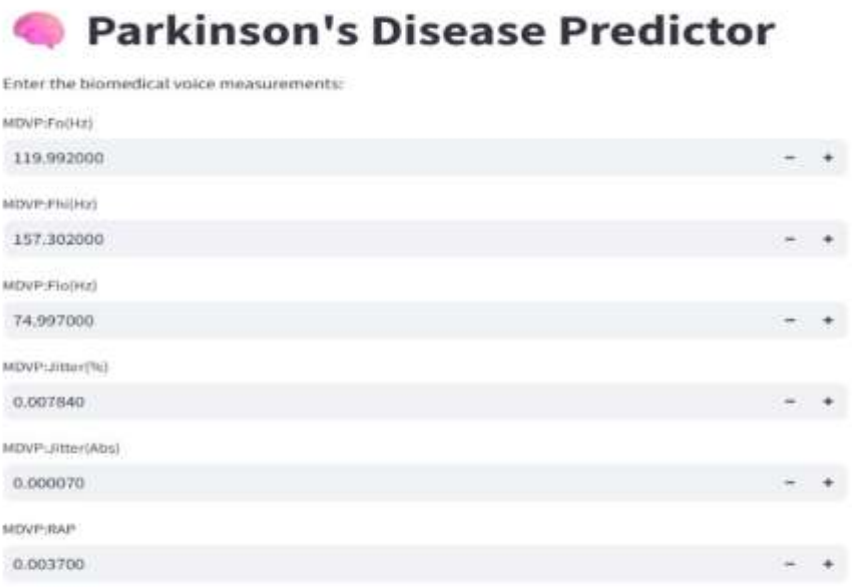
Fig: input data

Method of Process:

The methodology for Parkinson's Disease prediction involves a combination of data collection, preprocessing, model training, and deployment using machine learning techniques. Initially, vocal data is collected from both healthy individuals and Parkinson's patients. This data includes multiple vocal features such as frequency, jitter, shimmer, and noise-to-harmonics ratio, which are extracted using PRAAT software. The collected dataset is then preprocessed using normalization techniques like MinMaxScaler to bring the features onto the same scale. The processed data is divided into training and testing sets, typically with a 70:30 split. The proposed system uses a hybrid voting classifier that combines the XGBoost classifier and the Support Vector Classifier (SVC).

Output:

The Parkinson's Disease Predictor application takes biomedical voice measurements as input, such as MDVP:Fo (fundamental frequency), MDVP:Fhi, MDVP:Flo, Jitter, RAP, D2, and PPE, among others. After entering these values into the system, the model processes the data using a trained machine learning algorithm that combines XGBoost and Support Vector Classifier (SVC) through a voting mechanism. Upon analyzing the input, the system outputs a prediction result indicating that the patient has Parkinson’s disease, with a confidence level of 82.74%**.** This means that based on the provided vocal characteristics, the model is reasonably certain that the patient is likely affected by Parkisons.



🧠 Parkinson's Disease Predictor

Enter the biomedical voice measurements:

MDVP:Fo(Hz): 119.992000

MDVP:Fhi(Hz): 157.302000

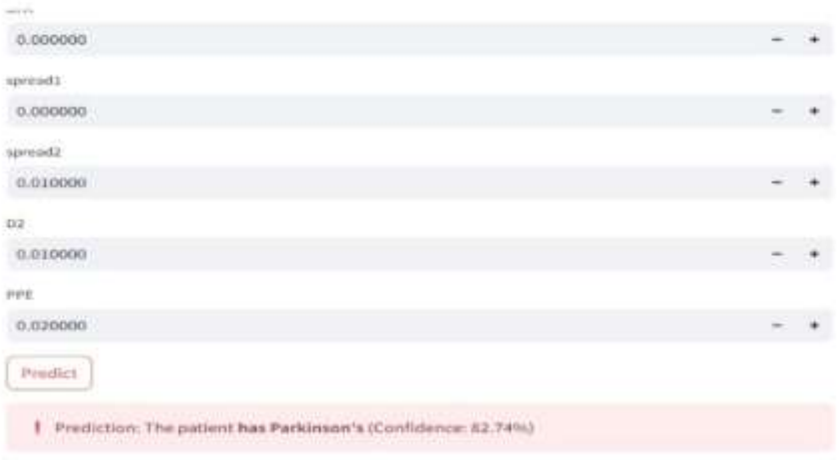
MDVP:Flo(Hz): 74.997000

MDVP:Jitter(%): 0.007840

MDVP:Jitter(Abs): 0.000070

MDVP:RAP: 0.003700

Fig:output data



[illegible]: 0.000000

spread1: 0.000000

spread2: 0.010000

D2: 0.010000

PPE: 0.020000

! Prediction: The patient **has Parkinson's** (Confidence: 82.74%)

Fig:outputdata

RESULTS:

The project achieved a high prediction accuracy of 97% in diagnosing Parkinson's Disease using the proposed machine learning model, which combines the XGBoost classifier and Support Vector Classifier (SVC) into a Voting Classifier. This performance significantly surpasses the results from previous studies, where the average accuracy was around 81%. The system successfully processes biomedical voice measurements, such as jitter, shimmer, and pitch-related features, to differentiate between healthy individuals and those affected by Parkinson's Disease. The model was tested using various metrics, including F1 score, recall, and precision, all of which demonstrated robust and reliable performance.

DISCUSSIONS:

The study demonstrates the effective use of machine learning for early detection of Parkinson's Disease through biomedical voice measurements. By combining the XGBoost classifier and Support Vector Classifier (SVC) in a Voting Classifier ensemble, the system leverages the strengths of both algorithms to achieve superior performance. The discussion emphasizes how this hybrid model improves diagnostic accuracy compared to traditional single-algorithm approaches. Previous models, such as Random Forest, Artificial Neural Networks (ANN), and standalone SVMs, achieved moderate accuracy levels, typically around 81%. In contrast, the proposed system reaches a prediction accuracy of 97%, marking a substantial improvement. This result confirms that ensemble methods, particularly voting classifiers, are highly effective in handling complex biomedical data. The discussion also highlights the importance of proper data preprocessing, feature selection, and hyperparameter tuning in achieving such results.

CONCLUSION:

The project successfully developed a machine learning model to predict Parkinson's Disease using a combination of XGBoost and Support Vector Classifier (SVC) within a Voting Classifier framework. This ensemble approach significantly improved diagnostic accuracy, achieving a 97% success rate, which is notably higher than previous methods that yielded around 81% accuracy. The system processes vocal measurements, such as jitter, shimmer, and other related parameters, to effectively classify individuals as healthy or Parkinson's-affected. The model not only ensures better prediction performance but also offers a user-friendly interface for real-time usage, making it accessible for healthcare professionals. The deployment of the model through API services like Flask or FastAPI allows for easy integration into medical diagnostic applications.

FUTURE SCOPE:

The future scope of this project involves expanding the current system to further enhance its predictive capabilities and usability in clinical settings. One of the primary areas for improvement is integrating additional biomarkers beyond voice data, such as gait analysis, handwriting patterns, and sensor-based motor function data, to develop a more comprehensive diagnostic tool. The system can also be upgraded to support continuous monitoring of patients, allowing for real-time tracking of disease progression and treatment response. Incorporating deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) may further improve the accuracy and robustness of predictions by capturing more complex patterns in patient data. Additionally, the deployment of the model on cloud platforms can make it accessible on a global scale, enabling telemedicine applications and remote diagnosis, especially in underserved areas.

ACKNOWLEDGEMENT:

Chinthagingala Vasundhara: working as an Assistant professor in master of computer application in sanketika vidya parishad engineering college, Visakhapatnam Andhra Pradesh. With 2 years of experience in Masters of computer application(MCA), accredited by NAAC.with her area of intrest in java full stack.She is dedicated and emerging academician in the field of Computer Science, currently serving as a faculty member. With a strong foundation in technical concepts and a passion for teaching, she has begun his academic journey by actively mentoring students in practical and innovative projects. As a faculty, she has already made a significant impact by guiding student teams through their academic projects with clarity, enthusiasm, and technical proficiency. His commitment to student success and interest in research-driven teaching make him a promising contributor to academic excellence.

BUBATHULA BHAVANA is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, zaffiliated by Andhra University and approved by AICTE. With interest in Machine learning Bubathula Bhavana has taken up him PG project on A MACHINE LEARNING APPROACH TO PREDICT PARKINSONS DISEASE USING VOTING CLASSIFIER and published the paper in connection to the project under the guidance of CH. VASUNDHARA, Assistant Professor, SVPEC.

REFRENCES:

[1] Predicting patients with Parkinson's disease using Machine Learning
and ensemble voting technique
https://link.springer.com/article/10.1007/s11042-023-16881-x

[2] Comparison of Machine learning models for Parkinson's Disease prediction
https://ieeexplore.ieee.org/abstract/document/9298033

[3] Early detection of Parkinson's disease using machine learning
https://www.sciencedirect.com/science/article/pii/S1877050923000078

[4] An improved approach for prediction of Parkinson's disease using machine learning techniques
https://ieeexplore.ieee.org/document/7955679

[5] Hybrid machine learning classifier and ensemble techniques to detect Parkinson's disease patients
https://link.springer.com/article/10.1007/s42979-021-00587-8

[6] Late feature fusion using neural network with voting classifier for Parkinson's disease detection
https://link.springer.com/article/10.1186/s12911-024-02683-0

[7] Effective Parkinson Disease Detection and Prediction Using Voting Classifier in Machine Learning
https://link.springer.com/chapter/10.1007/978-3-031-68905-5_21

[8] A hybrid system for Parkinson's disease diagnosis using machine
learning techniqueshttps://link.springer.com/article/10.1007/s10772-021-09837-9

[9] Detection of Parkinson disease using multiclass machine learning approach

https://www.nature.com/articles/s41598-024-64004-9

[10] Feature selection and machine learning methods for optimal identification and prediction of subtypes in Parkinson's disease

https://www.sciencedirect.com/science/article/abs/pii/S0169260721002066

[11] An ensemble technique to predict Parkinson's disease using machine learning algorithms

https://www.sciencedirect.com/science/article/abs/pii/S0167639324000396

[12] Machine learning Ensemble for the Parkinson's disease using protein sequences

https://link.springer.com/article/10.1007/s11042-022-12960-7

[13] Classification of Parkinson's disease using speech signal with machine learning and deep learning approaches

https://www.ejece.org/index.php/ejece/article/view/488

[14] Enhancing Parkinson's disease prediction using machine learning and feature selection methods

https://www.open-access.bcu.ac.uk/12590/

15] Classification of Parkinson disease based on patient's voice signal using machine learning

https://zuscholars.zu.ac.ae/works/4673/

[16] Detection of Parkinson's disease by using machine learning stacking and ensemble method

https://link.springer.com/article/10.1007/s44174-023-00079-8

[17] A novel hybrid CNN-KNN ensemble voting classifier for Parkinson's disease prediction from hand sketching images

https://link.springer.com/article/10.1007/s11042-024-19314-5

[18] Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures

https://link.springer.com/article/10.1007/s13534-017-0051-2

[19] New machine-learning algorithms for prediction of Parkinson's disease

https://www.tandfonline.com/doi/abs/10.1080/00207721.2012.724114

[20] Early Prediction of Parkinson's Disease (PD) Using Ensemble Classifiers
https://ieeexplore.ieee.org/abstract/document/9071562