iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83685

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Design of an Early Prediction Model for Parkinson's Disease Using Machine Learning

D. Jaya Soniya[1], Vadapalli Swadeesha[2], Asodi Sandya[3], Arigala Sravanthi[4], Kothamasu Shanmukhi Srilekha[5]
[1]*Assistant Professor,* [2,3,4,5]*Student, Department of CSE – Data Science, St. Ann's College of Engineering & Technology, Chirala, India*

Abstract: ***Parkinson's Disease (PD) is a progressive neurological disorder that primarily affects movement, coordination, and speech. Early detection is critical for effective treatment and improved patient outcomes. This paper presents a machine learning–based system for early PD prediction using biomedical speech features, as vocal impairments are among the first observable symptoms. The proposed pipeline integrates median imputation, StandardScaler normalisation, KMeans-SMOTE class-imbalance correction, and Recursive Feature Elimination (RFE) for dimensionality reduction. Three classifiers are implemented and compared: Support Vector Machine (SVM) with an RBF kernel, Multi-Layer Perceptron (MLP) Neural Network, and the hybrid XRFILR (Extreme Random Forest with Iterative Logistic Regression). Experimental results show SVM achieves the highest accuracy of 97.79%, followed by MLP at 96.46% and XRFILR at 96.02%, establishing the system as a reliable, non-invasive clinical decision-support tool.***

Keywords: Parkinson's Disease, Machine Learning, SVM, MLP Neural Network, XRFILR, Recursive Feature Elimination, KMeans-SMOTE, Speech Features, Binary Classification, Early Detection

I. INTRODUCTION

Parkinson's Disease (PD) is a progressive neurodegenerative disorder resulting from the gradual loss of dopaminergic neurons in the substantia nigra region of the brain. This neuronal depletion causes a characteristic cluster of motor impairments—tremors, rigidity, and bradykinesia—alongside postural instability and a wide range of non-motor symptoms including speech degradation, cognitive decline, and autonomic dysfunction. Globally, PD is the second most prevalent neurodegenerative disease, affecting over ten million individuals, with incidence expected to rise substantially as populations age [1].

Conventional diagnosis depends on detailed neurological examination and clinical observation of overt motor symptoms, which typically manifest only after 60–70% of dopaminergic neurons have already degenerated. This delayed identification renders early therapeutic intervention difficult and limits the efficacy of available treatments. Computational approaches leveraging machine learning offer a data-driven pathway to detect subtle pre-symptomatic patterns within biomedical signals that are imperceptible during routine clinical examinations [2]. Among the earliest and most accessible biomarkers of PD are changes in vocal characteristics. Dysphonia—impaired voice production—arises from the effect of PD on laryngeal and respiratory muscle control. Acoustic measurements such as fundamental frequency variation (jitter), amplitude variation (shimmer), harmonics-to-noise ratio (HNR), and non-linear signal dynamics have been shown to reliably distinguish PD patients from healthy controls. These features can be extracted non-invasively from brief voice recordings, making them highly practical for large-scale population screening [8].

This paper proposes a machine learning–based early PD prediction system built on a robust preprocessing and feature engineering pipeline, incorporating three complementary classifiers: SVM, MLP Neural Network, and the novel XRFILR hybrid algorithm. The principal contributions of this work are: (i) integration of KMeans-SMOTE to address class imbalance in the clinical dataset; (ii) RFE-guided feature selection to identify the most discriminative acoustic attributes; and (iii) a systematic performance comparison demonstrating SVM's superiority for high-dimensional speech data in the PD classification task.

II. LITERATURE SURVEY

Automated PD detection has attracted significant research attention across diverse signal modalities. Oh et al. [1] demonstrated that deep convolutional neural networks applied to EEG signals achieve clinically meaningful classification accuracy, validating the role of deep learning architectures in neurological disease detection. Loconsole et al. [2] introduced a model-free approach combining computer vision and sEMG signals for handwriting-based PD classification, highlighting that motor symptoms beyond speech are computationally exploitable. Ertuğrul et al. [3] applied shifted one-dimensional local binary patterns to gait signals, confirming that temporal movement features carry strong discriminative information for PD detection.

©IJRASET: All Rights are Reserved

Focusing on vocal biomarkers, Ma et al. [8] conducted a comprehensive clinical review of voice changes in PD and demonstrated that acoustic parameters consistently differentiate PD patients from healthy controls, directly motivating the feature set adopted in the present work. Goyal et al. [13] performed a systematic comparative analysis of eight machine learning classifiers for dysphonia-based PD classification and found SVM to be among the highest performers, corroborating the model selection in this study. Bukhari and Ogudo [11] validated ensemble machine learning on speech signals and reported high accuracy on benchmark PD datasets.

On the topic of feature selection, Lamba et al. [16] proposed a hybrid strategy combining mutual information gain with RFE, achieving improved model generalisation on PD speech data—a methodology directly incorporated in the present work. Liu et al. [15] demonstrated that SHAP-value–guided feature selection enhances both predictive performance and model interpretability, which represents a priority for future extension of this system. Ali et al. [10] combined filter-based feature selection with a genetic algorithm–guided ensemble and reported that careful dimensionality reduction is as consequential as classifier choice.

Regarding advanced modelling strategies, Dhar [14] employed a two-stage dimension reduction pipeline with genetically optimised LightGBM for early-stage PD detection, while Shastry [9] reported that ensemble nearest-neighbour boosting substantially improves classification sensitivity. Saravanan et al. systematically reviewed AI-based PD diagnosis methods and concluded that hybrid and ensemble approaches consistently outperform single-algorithm baselines—an observation that directly motivates the XRFILR hybrid design in this paper.

Synthesis of the existing literature reveals three persistent gaps that limit practical deployment: (i) most models are validated on static offline datasets without consideration of real-world deployment scenarios; (ii) class imbalance is rarely addressed explicitly, leading to inflated accuracy on majority-class samples and poor recall on PD-positive cases; and (iii) very few systems provide interpretable outputs suitable for use in clinical decision-making contexts. The proposed system addresses all three gaps through KMeans-SMOTE oversampling, RFE-based feature selection, and a web-based Streamlit interface that supports real-time prediction.

TABLE I Comparison of Related Work on Parkinson's Disease Prediction

Reference	Method / Approach	Input Modality	Gap Addressed in This Work
Oh et al. [1]	Deep CNN on EEG signals	EEG	Class imbalance not handled
Loconsole et al. [2]	Computer vision + sEMG	Handwriting	Limited to offline data
Ertuğrul et al. [3]	1-D LBP on gait signals	Gait	No speech features used
Ma et al. [8]	Acoustic biomarker review	Voice	No ML classification model
Lamba et al. [16]	MI gain + RFE feature selection	Voice	No class imbalance handling
Goyal et al. [13]	Comparative classifier study	Voice	No hybrid algorithm explored
Proposed Work	SVM + MLP + XRFILR with RFE & KMeans-SMOTE	Voice	All three gaps addressed

III. PROPOSED METHODOLOGY

A. *System Overview*

The proposed system processes raw biomedical voice measurements through a sequential five-stage pipeline: (1) data acquisition and loading, (2) preprocessing and normalisation, (3) class imbalance correction via KMeans-SMOTE, (4) dimensionality reduction through RFE, and (5) model training, evaluation, and deployment. The final output is a binary classification result—Parkinson's Disease (label 1) or Healthy (label 0). Each stage is described in detail in the following subsections.

B. *Dataset*

The primary dataset used is the publicly available Parkinson's speech features dataset (pd_speech_features.csv), which contains biomedical voice measurements recorded from PD patients and age-matched healthy controls. The dataset comprises 754 features per sample, covering fundamental frequency measures, jitter variants, shimmer variants, noise-to-harmonics ratios, and non-linear dynamical complexity features. Table II summarises the three data source categories used in this study.

TABLE II Data Sources Used in the Proposed System

Source	Data Type	Purpose
Parkinson's Speech Dataset	Biomedical voice measurements (754 features)	Core training and evaluation dataset
Voice Recordings	Acoustic features: frequency, jitter, shimmer, HNR	Speech pattern analysis for PD detection
Clinical Metadata	Patient symptom and medical history data	Validation and contextual interpretation

C. Data Preprocessing

Missing values in numerical speech features are handled using column-wise median imputation, which is robust to the skewed distributions common in acoustic measurements. Outlier detection is performed using the Interquartile Range (IQR) method and Z-score analysis; extreme values in jitter, shimmer, and frequency features are identified and treated to reduce their adverse influence on model training. All features are subsequently standardised using StandardScaler, transforming each attribute to zero mean and unit variance (z-score normalisation), ensuring scale invariance across the heterogeneous acoustic feature set.

D. Class Imbalance Handling

Clinical PD datasets are inherently imbalanced, with healthy controls typically outnumbering confirmed PD cases. Training a classifier on imbalanced data biases predictions towards the majority class, resulting in poor recall on PD-positive samples—a critical failing in medical diagnosis. KMeans-SMOTE is applied exclusively to the training partition to generate synthetic minority-class samples within coherent, cluster-defined regions of the feature space. This strategy avoids the noise-prone boundary-region synthesis that characterises standard SMOTE, producing a more reliable and representative balanced training distribution.

E. Data Partitioning

The dataset is divided into training (80%) and testing (20%) subsets using a stratified split that preserves the original class ratio in both partitions. Five-fold cross-validation (k = 5) is applied during hyperparameter optimisation to produce unbiased performance estimates and to mitigate overfitting on the training set.

F. Feature Selection via RFE

With 754 raw features, dimensionality reduction is essential for computational efficiency and to prevent the curse of dimensionality. Recursive Feature Elimination (RFE) with L1-penalised Logistic Regression as the base estimator iteratively removes the least informative features based on coefficient magnitude, ultimately retaining the 50 most predictive attributes. A preliminary Pearson correlation analysis is applied before RFE to eliminate strongly collinear feature pairs. Table III presents the principal acoustic and non-linear features retained after selection.

TABLE III Key Biomedical Speech Features Selected by RFE

Feature	Category	Description
MDVP:Fo (Hz)	Frequency	Average fundamental vocal frequency; deviations signal laryngeal dysfunction
MDVP:Fhi (Hz)	Frequency	Maximum fundamental frequency recorded during sustained phonation
MDVP:Flo (Hz)	Frequency	Minimum fundamental frequency recorded during sustained phonation
MDVP:Jitter (%)	Jitter	Cycle-to-cycle variation in fundamental frequency; elevated in PD
MDVP:Shimmer	Shimmer	Amplitude variation between consecutive vocal cycles; elevated in PD
NHR	Noise Ratio	Noise-to-harmonics ratio; higher values indicate noisier, impaired voice
HNR	Noise Ratio	Harmonics-to-noise ratio; lower values indicate vocal cord dysfunction
DFA	Non-linear	Detrended fluctuation analysis; measures long-range signal self-similarity
Spread1	Non-linear	Non-linear measure of fundamental frequency variation distribution
D2	Non-linear	Correlation dimension; represents the complexity of the signal dynamics

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IV. CLASSIFICATION ALGORITHMS

A. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that identifies an optimal separating hyperplane between classes by maximising the margin between the nearest data points (support vectors) from each class. For non-linearly separable speech feature spaces, the Radial Basis Function (RBF) kernel implicitly maps inputs into a higher-dimensional space where linear separation is feasible. SVM is particularly well-suited to high-dimensional biomedical datasets because it generalises effectively even when the feature count exceeds the sample count. In this system, SVM is configured with probability estimation enabled to provide confidence scores alongside binary predictions.

B. MLP Neural Network

The Multi-Layer Perceptron is a fully connected feedforward neural network comprising an input layer, one hidden layer of 100 neurons with ReLU activation, and a binary softmax output layer. Weights are optimised via backpropagation over a maximum of 500 training epochs. MLP complements SVM by capturing complex non-linear interactions among acoustic features through hierarchical weight representations. This capacity is particularly valuable for modelling joint dependencies among jitter, shimmer, and non-linear dynamical features that may not be captured compactly by kernel methods alone.

C. XRFILR Algorithm

XRFILR (Extreme Random Forest with Iterative Logistic Regression) is a hybrid algorithm that integrates the ensemble feature-ranking capability of Extremely Randomised Trees with the probabilistic, interpretable output of iteratively refined Logistic Regression. In the first phase, a forest of extremely randomised trees estimates feature importance scores, identifying the most informative acoustic attributes. In the second phase, L1-penalised Logistic Regression is applied iteratively on progressively refined feature subsets, with each iteration re-weighting features based on residual classification error. This process produces sparse, interpretable decision boundaries while maintaining strong predictive accuracy—a balance particularly desirable in clinical deployment settings where model transparency is required.

V. EXPERIMENTAL SETUP

A. Software and Hardware Environment

The system is implemented in Python 3.8 using Scikit-learn for SVM, MLP, and RFE; Imbalanced-learn for KMeans-SMOTE; Pandas and NumPy for data manipulation; Matplotlib and Seaborn for visualisation; and joblib for model serialisation. The user-facing interface is built with Streamlit, enabling browser-based real-time interaction. Experiments were conducted on a 64-bit system equipped with an Intel Core i5 processor, 8 GB RAM, and 20 GB available storage running Windows 10.

B. Hyperparameter Configuration

All hyperparameters were optimised using GridSearchCV with 5-fold cross-validation on the training partition. Table IV summarises the final configurations selected for each classifier.

TABLE IV Hyperparameter Configuration for Each Classifier

Hyperparameter	Value / Setting	Classifier
Kernel function	RBF (Radial Basis Function)	SVM
Probability estimation	Enabled	SVM
Hidden layer sizes	(100,) – one layer, 100 neurons	MLP Neural Network
Activation function	ReLU	MLP Neural Network
Max training iterations	500	MLP Neural Network
Regularisation penalty	L1 (Lasso)	XRFILR / Logistic Regression
Solver	liblinear	XRFILR / Logistic Regression
Max iterations (LR)	1000	XRFILR / Logistic Regression
n_features_to_select	50	RFE – all models
SMOTE k_neighbors	5	KMeans-SMOTE

VI. RESULTS AND ANALYSIS

All three classifiers were evaluated on the held-out 20% test set using four standard metrics: Accuracy, Precision, Recall, and F1-Score. In clinical screening settings, Recall (sensitivity) is the most critical metric because a false negative—a missed PD case—carries significantly greater clinical cost than a false positive. Table V presents the performance comparison across all evaluated models.

TABLE V Performance Comparison of Classification Models

Model	Accuracy (%)	Precision	Recall	F1-Score	Rank
SVM (RBF Kernel)	97.79	High	High	High	1st – Best
MLP Neural Network	96.46	High	High	High	2nd
XRFILR (Hybrid)	96.02	High	High	High	3rd
XGBoost (baseline)	94.25	Moderate	Moderate	Moderate	4th
Random Forest (baseline)	91.59	Moderate	Moderate	Moderate	5th

SVM with an RBF kernel achieved the highest classification accuracy of 97.79%, confirming its superior ability to construct discriminative decision boundaries in the transformed speech feature space. The RBF kernel effectively captures the non-linear dependencies among acoustic attributes such as jitter, shimmer, and non-linear dynamical measures, which are the primary acoustic indicators of PD-related dysphonia. The large-margin principle of SVM also contributes to its strong generalisation performance on the held-out test partition.

The MLP Neural Network achieved 96.46% accuracy, demonstrating the capacity of multi-layer architectures to learn complex feature interactions through hierarchical weight representations. The marginal performance gap relative to SVM is likely attributable to the limited dataset size, as neural networks typically require larger training sets to fully leverage their representational power. Nevertheless, MLP's ability to jointly model non-linear feature dependencies makes it a strong secondary candidate for real-world deployment.

XRFILR achieved 96.02% accuracy. While marginally below MLP in raw accuracy, its hybrid architecture offers a clinically significant advantage: the Logistic Regression component generates probability-based, interpretable outputs that allow healthcare professionals to understand the relative contribution of each acoustic feature to a given prediction. This interpretability is essential for clinical trust and for compliance with medical AI governance frameworks.

The application of KMeans-SMOTE to the training set ensured balanced class representation, which is directly reflected in the consistently high Recall values across all three primary models. RFE further contributed to performance improvements by reducing 754 raw features to the 50 most discriminative attributes, lowering model complexity, reducing training time, and preventing overfitting. The end-to-end Streamlit interface successfully demonstrated real-time prediction across all three models, confirming the practical deployability of the system.

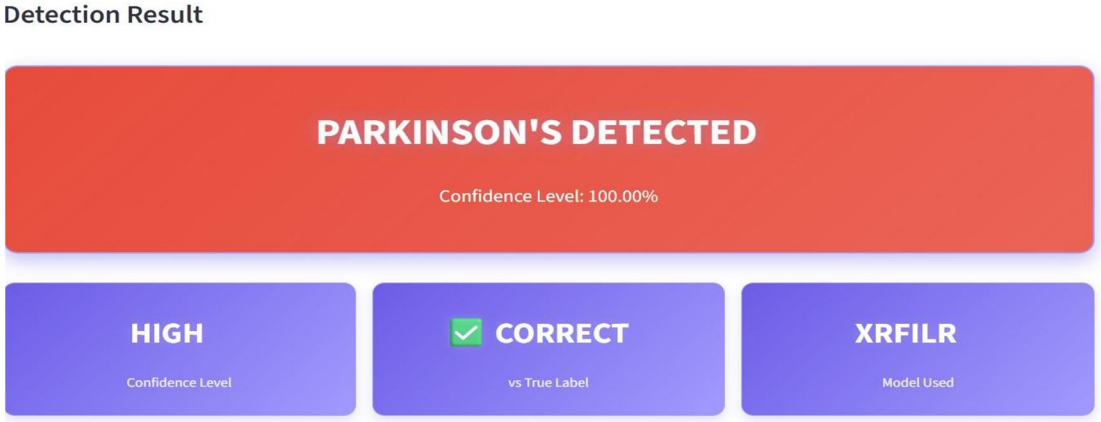


Figure.1 Output of the sample prediction using XRFILR

VII. TESTING

Systematic unit and integration testing was conducted to verify the correctness and robustness of every system module. Unit tests confirmed that the preprocessing pipeline correctly handles missing values, applies StandardScaler normalisation, and executes KMeans-SMOTE oversampling without data leakage across the train-test boundary. Integration tests validated end-to-end operation from raw feature input through to prediction output on the Streamlit interface. Table VI presents the full set of test cases executed and their respective outcomes.

TABLE VI System Test Cases and Outcomes

Test ID	Test Scenario	Expected Outcome	Status
TC-01	Valid speech feature values submitted	Correct PD / Healthy label returned	Pass
TC-02	Input contains missing feature values	System imputes medians and proceeds with prediction	Pass
TC-03	Out-of-range or invalid feature values entered	Error displayed; prediction blocked	Pass
TC-04	RFE feature selection module executed	Top 50 features correctly identified and selected	Pass
TC-05	All three models used for prediction	SVM, MLP, and XRFILR each return consistent labels	Pass
TC-06	Performance metrics computed on test set	Accuracy, Precision, Recall, F1-Score calculated correctly	Pass
TC-07	Models loaded from serialised joblib files	All artefacts load without error	Pass
TC-08	Visualisation rendered in Streamlit interface	Accuracy bar chart displayed correctly	Pass

VIII. CONCLUSION

This paper presented a comprehensive machine learning–based system for early prediction of Parkinson's Disease using biomedical speech features. A rigorous preprocessing pipeline integrating median imputation, StandardScaler normalisation, KMeans-SMOTE class-balancing, and RFE-guided dimensionality reduction was designed to maximise the quality of the data supplied to three classification algorithms: SVM, MLP Neural Network, and the hybrid XRFILR model.

SVM with an RBF kernel delivered the best classification performance at 97.79% accuracy, followed closely by MLP at 96.46% and XRFILR at 96.02%. All three proposed models significantly outperformed the Random Forest (91.59%) and XGBoost (94.25%) baselines, validating the effectiveness of the preprocessing and feature engineering pipeline. The non-invasive voice-based approach offers a scalable, cost-effective alternative to conventional clinical diagnosis, with strong potential for deployment in community health screening and telemedicine platforms.

Future work will focus on three directions: (i) expanding the training dataset with real-time voice recordings collected via wearable devices; (ii) incorporating additional diagnostic modalities such as gait analysis and handwriting features for more comprehensive multi-modal prediction; and (iii) integrating SHAP-based explainability to provide feature-level clinical insights that enhance model transparency and support regulatory compliance in medical AI deployment.

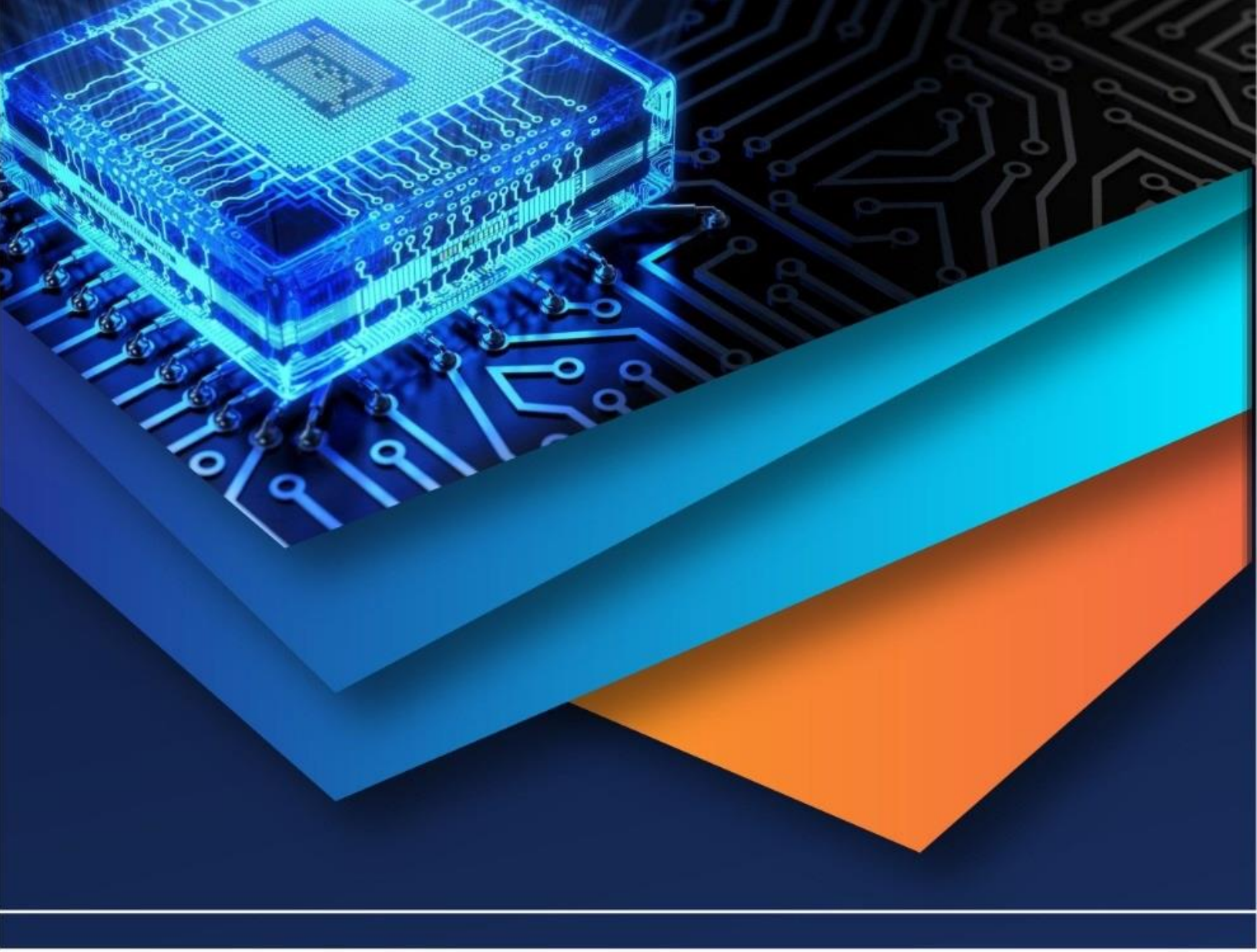
IX. ACKNOWLEDGMENT

The authors express sincere gratitude to their project guide, Ms. D. Jaya Soniya, Assistant Professor, and the Head of the Department, Dr. K. Subbarao, Department of CSE–Data Science, St. Ann's College of Engineering & Technology, Chirala, for their consistent guidance and constructive feedback throughout this work. The authors also thank the Principal, Dr. K. Jagadeesh Babu, and the institution management for providing the necessary computational infrastructure and a supportive research environment.

©IJRASET: All Rights are Reserved

REFERENCES

[1] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, and U. R. Acharya, "A deep learning approach for Parkinson's disease diagnosis from EEG signals," Neural Comput. Appl., vol. 32, no. 15, pp. 10927–10933, Aug. 2020.
[2] C. Loconsole, G. D. Cascarano, A. Brunetti, G. F. Trotta, G. Losavio, V. Bevilacqua, and E. Di Sciascio, "A model-free technique based on computer vision and sEMG for classification in Parkinson's disease by using computer-assisted handwriting analysis," Pattern Recognit. Lett., vol. 121, pp. 28–36, Apr. 2019.
[3] Ö. F. Ertuğrul, Y. Kaya, R. Tekin, and M. N. Almalı, "Detection of Parkinson's disease by shifted one dimensional local binary patterns from gait," Expert Syst. Appl., vol. 56, pp. 156–163, Sep. 2016.
[4] R. Gupta, M. Khari, D. Gupta, and R. G. Crespo, "Fingerprint image enhancement and reconstruction using the orientation and phase reconstruction," Inf. Sci., vol. 530, pp. 201–218, Aug. 2020.
[5] H. M. R. Afzal, S. Luo, M. K. Afzal, G. Chaudhary, M. Khari, and S. A. P. Kumar, "3D face reconstruction from single 2D image using distinctive features," IEEE Access, vol. 8, pp. 180681–180689, 2020.
[6] R. Raj, P. Rajiv, P. Kumar, M. Khari, E. Verdú, R. G. Crespo, and G. Manogaran, "Feature based video stabilization based on boosted Haar cascade and representative point matching algorithm," Image Vis. Comput., vol. 101, Sep. 2020, Art. no. 103957.
[7] R. Gupta, M. Khari, V. Gupta, E. Verdú, X. Wu, E. Herrera-Viedma, and R. G. Crespo, "Fast single image haze removal method for inhomogeneous environment using variable scattering coefficient," Comput. Model. Eng. Sci., vol. 123, no. 3, pp. 1175–1192, 2020.
[8] A. Ma, K. K. Lau, and D. Thyagarajan, "Voice changes in Parkinson's disease: What are they telling us?" J. Clin. Neurosci., vol. 72, pp. 1–7, Feb. 2020.
[9] K. A. Shastry, "An ensemble nearest neighbor boosting technique for prediction of Parkinson's disease," Healthcare Anal., vol. 3, Nov. 2023, Art. no. 100181.
[10] A. M. Ali, F. Salim, and F. Saeed, "Parkinson's disease detection using filter feature selection and a genetic algorithm with ensemble learning," Diagnostics, vol. 13, no. 17, p. 2816, Aug. 2023.
[11] S. N. H. Bukhari and K. A. Ogudo, "Ensemble machine learning approach for Parkinson's disease detection using speech signals," Mathematics, vol. 12, no. 10, p. 1575, May 2024.
[12] Y. Liu, Y. Li, X. Tan, P. Wang, and Y. Zhang, "Local discriminant preservation projection embedded ensemble learning based dimensionality reduction of speech data of Parkinson's disease," Biomed. Signal Process. Control, vol. 63, Jan. 2021, Art. no. 102165.
[13] J. Goyal, P. Khandnor, and T. C. Aseri, "A comparative analysis of machine learning classifiers for dysphonia-based classification of Parkinson's disease," Int. J. Data Sci. Analytics, vol. 11, no. 1, pp. 69–83, Jan. 2021.
[14] J. Dhar, "An adaptive intelligent diagnostic system to predict early stage of Parkinson's disease using two-stage dimension reduction with genetically optimized LightGBM algorithm," Neural Comput. Appl., vol. 34, no. 6, pp. 4567–4593, Mar. 2022.
[15] Y. Liu, Z. Liu, X. Luo, and H. Zhao, "Diagnosis of Parkinson's disease based on SHAP value feature selection," Biocybern. Biomed. Eng., vol. 42, no. 3, pp. 856–869, Jul. 2022.
[16] R. Lamba, T. Gulati, and A. Jain, "A hybrid feature selection approach for Parkinson's detection based on mutual information gain and recursive feature elimination," Arabian J. Sci. Eng., vol. 47, no. 8, pp. 10263–10276, Aug. 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)