Hybrid Neuro-Explainable Ensemble Framework for Early Detection of Parkinson's Disease Using Speech-Based Acoustic Features

Skanthah Lakshmi Senthilkumar[1], Pranjal Upadhyay[1], Bavisha Pankaj[1], Dr. D. Deva Hema[2]

[1]Student, Department of Computer Science Engineering, SRMIST, Ramapuram

[2]Assistant Professor, Department of Computer Science Engineering, SRMIST, Ramapuram

DOI: https://doi.org/10.51244/IJRSI.2026.1304000085

Received: 04 April 2026; Accepted: 10 April 2026; Published: 02 May 2026

ABSTRACT

Early-Onset Parkinson's Disease (PD) is a major clinical concern due to the continuous neurological degeneration and slight prodromal symptoms. The degeneration of dopaminergic neurons in patients with Parkinson's Disease directly influences both motor and vocal activities. The acoustic degradation in patients' voices is more pronounced than other motor activities. Vocal parameters such as jitter, shimmer, and harmonics to noise ratio have exhibited high potential for early Parkinson's Disease detection. However, supervised learning models such as Logistic Regression and Gradient Boosting are unable to capture the non-linear variability in pathological speech. In this regard, a novel framework known as the Hybrid Neuro-Explainable Ensemble Framework (HNEF) is proposed. The framework integrates two supervised learning models, namely Regularized Logistic Regression and Gradient Boosting, using a weighted soft voting approach. The framework can capture linear and non-linear decision boundaries. Moreover, a novel hybrid oversampling technique is incorporated to tackle the common class imbalance in Parkinson's Disease datasets. The technique combines K-Means-based synthetic minority oversampling and density-sensitive oversampling. The relevance of features is determined via a sequential two-stage pipeline consisting of Recursive Feature Elimination and Mutual Information scoring, thus ensuring the preservation of the most diagnostically relevant vocal features. The prediction pipeline incorporates features of interpretability via the SHAP-based global attribution and the attention mechanism, thus ensuring accountability in the telemedicine and AI-assisted clinical settings. The experimental results, considering the standard PD speech dataset, show the efficacy of HNEF with a classification accuracy of 97.8%, an F1-score of 97.1%, and an AUC-ROC value of 0.99, thus outperforming all the individual baseline models, including SVM, Random Forest, XGBoost, and deep neural networks. The ten-fold cross-validation results show the robustness of the findings with a high accuracy of 97.5% ± 0.5%. The potential of HNEF will be explored for the extension of the system to speech monitoring for tracking disease progression, its integration with multimodal biomarkers including gait and EEG, and its prospective validation with demographically diverse patients for achieving regulatory approval.

Keywords: Parkinson's Disease, Speech Biomarkers, Feature Optimization, Early Diagnosis**.**

INTRODUCTION

Speech-based acoustic biomarkers are also emerging as a promising tool in the non-invasive detection of Parkinson's disease. Jitter, shimmer, and Harmonics-to-Noise Ratio (HNR) have been observed as powerful biomarkers in differentiating PD patients from healthy controls [1][2]. Diverse benchmark databases, such as the UCI Parkinson's Speech Dataset developed by Sakar et al. [2], are utilized as benchmark databases containing diverse vocal recordings. In another study, Tsanas et al. [3] have demonstrated that vocal features are significant in the remote telemonitoring of PD and that non-invasive and easy speech tests can be used with high accuracy in monitoring PD progression.

The traditional method of clinical diagnosis of PD is based on the assessment of a neurologist and scoring systems such as MDS-UPDRS [9]. Such subjective methods are prone to variability in results depending on the clinical expert. They are also not effective in detecting small symptoms of PD. Recently, machine learning (ML) techniques have gained popularity for the classification of PD using speech cues. Various ML techniques, such

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

as SVM, RF, LR, XGBoost, have been used with varying degrees of success in PD classification [5]. Deep learning techniques, which are based on the concept of multi-level features as proposed by LeCun et al. [10], have been effective in improving classification accuracy. However, such techniques are not easy to interpret. Friedman's gradient boosting method is a powerful ensemble technique, although it is also a black box in itself.

However, despite their predictive capabilities, three main problems plague existing techniques. The first is that class imbalance is always a problem in any Parkinson's disease database, with more healthy control instances than PD-positive instances. To solve this, SMOTE was proposed by Chawla et al. [4], with more recent variants such as KMeans-SMOTE trying to balance classes while controlling distribution changes. The second problem is that acoustic feature space is a high-dimensional space with much redundancy, leading to problems with dimensionality when not handled with intelligent feature selection. The work by Guyon & Elisseeff laid the foundation for Recursive Feature Elimination (RFE), a technique that eliminates features with the least contribution to improving generalization, as well as Mutual Information (MI), which is a filter-based alternative to feature selection. The third problem is that black-box models are not transparent enough to be used in clinical decision support. The recent work by Vaswani et al. [8] on introducing the attention mechanism is a significant step toward improving interpretability, with recent work by Lundberg & Lee [5] on SHAP providing local as well as global interpretation.

However, none of the existing models attempt to solve all three problems under one umbrella. In most studies, these three aspects are addressed independently as individual modules rather than being part of one entire framework. This has limited the practical application of AI-based Parkinson's disease (PD) diagnostic tools in telemedicine and practical healthcare scenarios, as all three aspects are required to be present simultaneously.

Keeping this in mind, we propose a new framework called the Hybrid Neuro-Explainable Ensemble Framework (HNEF), aiming to solve the problem of early detection of PD using speech-based acoustic features.

The framework has four key components. First, it deals with the problem of class imbalance through a hybrid approach of oversampling using both KMeans-guided SMOTE [4] and density-aware synthetic sampling methods. This ensures that the distribution of classes across each other remains constant. Secondly, it reduces the feature space through recursive steps of Recursive Feature Elimination (RFE) [6] and Mutual Information ranking, selecting the best acoustic biomarkers for clinical discrimination. Thirdly, it employs a hybrid ensemble classifier consisting of Deep Neural Networks [10], Extreme Gradient Boosting [7], and Regularized Logistic Regression. The classifier employs a weighted soft voting strategy for both linear and non-linear decision boundary estimation. Fourthly, it provides SHAP and attention mechanisms for global and local interpretation of results, respectively.

The main achievements of this study are highlighted in four main areas. First, we propose a hybrid ensemble model that combines linear and non-linear speech features to improve PD classification. Second, we propose a two-step feature optimization pipeline that combines recursive feature elimination with Mutual Information to ensure that features are optimized while reducing computational costs. Third, we propose a hybrid oversampling strategy that effectively handles imbalanced datasets with minimal drift in the synthetic data distribution for PD's minority class. Finally, we propose a two-tier explainability system that combines SHAP-based global interpretability with attention-based local interpretability, ensuring that trustworthy clinical decisions are made by the proposed hybrid intelligent system. The proposed hybrid intelligent system has demonstrated better robustness and generalization capabilities on standard PD datasets when compared to individual baseline machine learning and deep learning models.

LITERATURE REVIEW

Parkinson's Disease (PD) is a progressive brain disease in which movement symptoms, such as tremors and difficulty in initiating steps, co-exist with non-motor symptoms like cognitive problems. Accurate and timely identification of PD is a hard problem in clinical practice since many symptoms appear long after significant brain changes are underway. Recently, advances in machine learning, deep learning, and using multiple sources of data and techniques to provide explanations of AI models offer promising ways to automatically detect PD. In this survey, ten papers are surveyed to provide a foundation for a Hybrid Neuro Explainable Ensemble Framework to detect PD.

In the paper by Thimmapuram et al. [11], proposed the use of CSE-ProtoNet, a multimodal prototype learning framework, combining EEG and gait data. They developed this by using CondenseNet with SEBlock to deal with the problem of freezing gait, which affects patients suffering from Parkinson's disease. Since there is a scarcity of data in the medical domain, the authors employed a few-shot prototype learning technique to help the model learn to generalize from only a few samples. This enabled them to attain an accuracy of 98.75% for distinguishing between FOG and Non-FOG patients, thereby outperforming several other methods, including several baselines, across several metrics.

In a study by Souza et al. [12], a less obvious but equally relevant issue was investigated. The researchers sought to know whether a deep learning approach using multicenter data for PD imaging using MRI would ultimately reflect site-related biases. They analyzed 1,880 imaging data from 41 different centers and found that the PD classifier was still able to identify biological sex (about 75%), type of scanner (approximately 79%), and imaging site (about 71%), even without any of this data being included in the training data. This is a compelling argument for bias awareness and the use of explainability in any PD framework.

Skaramagkas et al. [13] carried out a systematic review of 87 deep learning-based research works for the diagnosis of Parkinson's disease, published between 2016 and 2023. Various aspects of PD diagnosis, including gait, speech, upper limb movement, facial movements, and combined approaches, were covered in these works. In all these works, deep learning methods clearly outperformed conventional methods, but there were two major issues to be addressed: data availability and interpretability of the models. According to the authors, for AI-based PD diagnosis tools to be widely accepted, the models must be transparent and interpretable, and this is where the need for the inclusion of the explainability component arises.

Nogales et al. [14] have also researched using BERT-based transformer architectures, which were specifically designed to work with natural language, to diagnose Parkinson's disease using 64-channel EEG data. They were able to utilize the fact that brain activity is similar to language and can be processed as such. The best model was based on EEG data from finger-tapping tasks and had an accuracy of around 86%. This shows that sequence-based architectures can be powerful in neurophysiology and that attention-based mechanisms are powerful tools in finding patterns in EEG data related to Parkinson's.

Khan et al. [15] carried out an extensive survey of the literature, reviewing 147 articles and 22 datasets focusing on the applications of ML approaches for the Parkinson's disease's diagnosis, Alzheimer's, brain tumors, and epilepsy. This literature review, while highlighting the applications of AI for neurological diagnosis, also discusses some of the challenges, such as the variety of data, small sample sizes, and lack of standardized assessment, faced in the development of AI-based solutions for these diseases. This, in essence, highlights the need for developing ensemble methods that are generalizable and reproducible across different data sources.

Tang et al. [16] proposed Cross-modal Augmented Transformer (CAT) for automatic medical report generation through a "locate then generate" approach. They pre-align key image patches with medical terms to reduce visual noise and use a two-stream decoder for cross-modal generation of diagnostic reports based on terminology. Although the work is specifically focused on radiology, the overall ideas of cross-modal alignment and interpretable output generation can be valuable architectural insights for explainability blocks in a Parkinson's disease prediction framework.

Fu et al. [17] developed a data dictionary system based on medical devices using Spring Boot, MySQL, and MongoDB to standardize data handling in medical institutions. The authors' tests showed that it is functionally complete, browser-compatible, and secure. In the context of AI-based PD prediction systems, it is essential to have a standard data system since mixed data can have a negative impact on reliability. The paper highlights the importance of data handling in implementing diagnostic frameworks.

In his work, Melvin [18] discusses what this means for biomedical engineers in terms of the EU's Medical Device Regulation (MDR 745/2017), specifically in relation to changes in clinical investigations, traceability of devices, and conformity assessments. He writes of a bottleneck in notified bodies that may create shortages of devices as it is rolled out. For teams working with AI in point of detection (PD) diagnostic systems, it is clear that regulatory requirements must be met when it is placed in a clinical role.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Guarín et al. [19] has developed a video-based ML system to monitor the progression of Parkinson's disease patients' motor skills. Their system utilized Finger Tapping Test videos and hand pose estimation. It was applied to data from 24 healthy controls and 66 patients with Parkinson's disease. Their system employs a multi-level classification method with varying features depending on the level of progression of the illness and outperforms existing methods. It demonstrates the potential of objective and non-invasive methods to detect nuanced progression patterns not captured by conventional rating scales, highlighting the advantage of employing various signals with a hybrid ensemble framework.

In another work, Qu et al. [20] presented a Graph Convolutional Network that utilizes only one neurodegeneration biomarker and an attention module for facilitating the Alzheimer disease's diagnosis in its early stages using the ADNI dataset. Their model got 93.90% accuracy in differentiating between AD and cognitively unimpaired groups as well as 82.05% accuracy in distinguishing between AD and mild cognitive impairment groups. Although it is focused on Alzheimer's disease, it can be directly used for Parkinson's disease prediction in an interpretable ensemble framework.

Collectively, these works establish four essential principles for improving PD prediction: multimodal fusion, explainability, ensemble learning, and bias awareness. On these principles, the proposed Hybrid Neuro Explainable Ensemble Framework seeks to establish predictions that are not only of high accuracy but are interpretable too.

PROPOSED METHODOLOGY

The proposed Hybrid Neuro-Explainable Ensemble Framework is designed to increase the prediction of early Parkinson's Disease (PD) using speech biomarkers while being interpretable and robust at the same time. The proposed framework is composed of following consecutive stages:

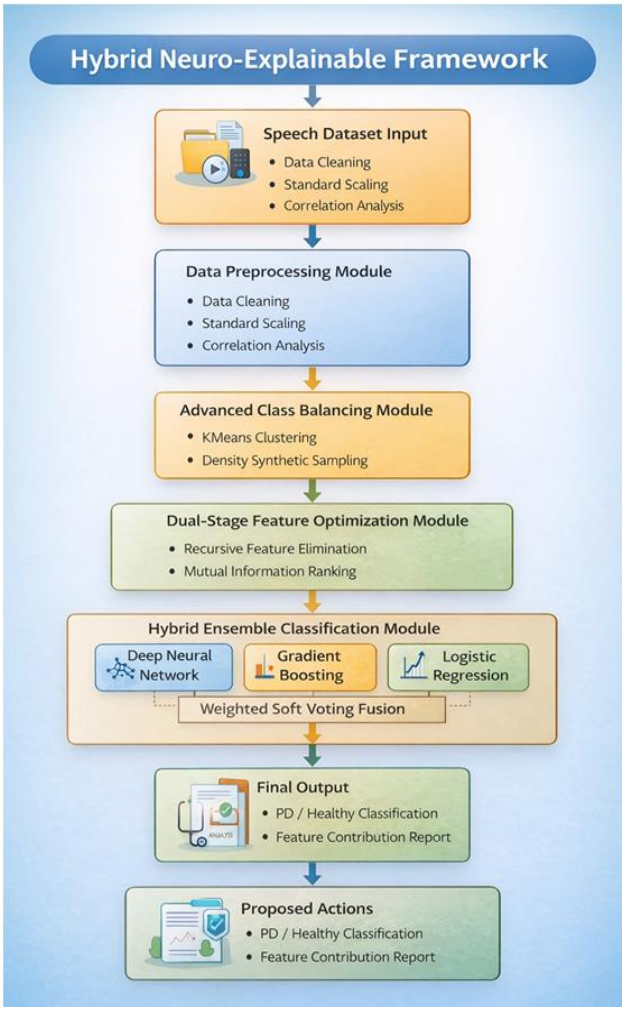

Fig 1: Hybrid Neuro-Explainable Framework

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

A. Speech Dataset Input

The framework begins with a speech data set that contains important acoustic features such as jitter, shimmer, and the harmonic noise ratio, which are indicative of vocal cord impairments caused by Parkinson's Disease. The data set is normally accompanied by inconsistencies and noise, as well as features that may be on different scales. This may affect the performance of the models. Thus, preprocessing is the first step in ensuring that everything is clean and ready for analysis. We use standard scaling to normalize the features to a uniform scale. This ensures that all features have an equal say in the modeling process. We also carry out a correlation check to eliminate features that are too redundant.

Moreover, the framework validates the data set to ascertain whether it is representative of both healthy and Parkinson's state conditions without any sampling bias. This is important because the quality of the features for input will have a direct effect on the performance of the learning algorithms.

B. Data Preprocessing Module

Once data intake is complete, preprocessing is performed to clean the data and make it easier to use with machine learning techniques. Preprocessing addresses issues such as missing data, duplicate records, and outliers that could interfere with machine learning. It also standardizes the data, checking them statistically, so that the data is presented in a clean, normalized form.

Preprocessing is also essential in eliminating noise and making data more stable, especially if the data is speech based, such as biomedical data. There could be inconsistencies, such as how data is recorded, noise, and speakers, which could interfere with the data. However, these inconsistencies are eliminated by the preprocessing module, which includes normalization and validation, so that the features are representative of real physiological activity rather than noise.

C. Advance Class Balancing Module

Class imbalance is a common challenge in datasets, especially when there are more healthy samples than samples of Parkinson's Disease. This can lead to a model favoring the majority class, making it difficult for early stages of PD to be identified. To resolve this, the suggested framework utilizes a complex balancing technique that incorporates both cluster and synthetic sampling. First, the minority class samples are divided into clusters using KMeans, maintaining the original data structure. New samples are then created by interpolating between samples within a cluster, effectively increasing the minority class samples. This is particularly useful for increasing the sensitivity of the model and providing a more balanced learning environment.

Furthermore, cluster-informed synthetic sampling ensures that new samples are created in more meaningful areas of the data, as opposed to being randomly created. This minimizes the chances of overfitting and preserves the original distribution of the data. This allows the model to effectively identify early stages of PD, as symptoms are less pronounced.

D. Dual Stage Feature Optimization Module

Since the high-dimensional speech data may include more or less relevant information that may affect the performance of the model and increase the computation cost [21,22], we propose a two-step feature optimization technique. In the first step, the Recursive Feature Elimination (RFE) technique is used for feature reduction based on the importance of the features obtained by the logistic regression model. In the second step, the Mutual Information feature analysis is carried out to find the connection between the features and the class labels. This helps in the selection of the feature space with maximum discriminability.

This is a two-part strategy that balances clarity with strong prediction capability. RFE reduces the feature set to the most influential features, and Mutual Information is used to uncover hidden relationships that might otherwise be overlooked. The framework is thus able to identify the complex speech patterns associated with Parkinson's Disease without compromising efficiency.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

E. Hybrid Ensemble Classification Module

The refined feature set is then passed through a hybrid ensemble classifier, which uses a combination of three models: a Deep Neural Network (DNN), XGBoost/Gradient Boosting, and Logistic Regression. The DNN model uses multiple hidden layers to drill deep into intricate non-linear relationships, Gradient Boosting uses step-by-step corrections to hone in on structural patterns, and Logistic Regression uses a simple, intuitive model as a baseline. They are then combined to make a final prediction through a weighted soft voting system, which aggregates probability scores to make a prediction. This ensemble method also improves stability and minimizes the probability of overfitting by playing to each model's strengths.

Additionally, this diversity of models also allows us to tap into various aspects of the data. Deep learning models are particularly good at identifying intricate patterns, boosting models fine-tune this through step-by-step corrections to errors, while linear models provide an intuitive baseline. This gives us a robust and dependable model that can handle intricate biomedical data.

F. Final Output Module

The final output module includes a classification result, which indicates whether a subject is healthy or has Parkinson's Disease. Alongside this, it also includes a probability score, which indicates how confident the model is in making this particular prediction. This is particularly important in a real-world context, as it allows practitioners to understand how confident the model is being in making a particular prediction.

The output module is also important as it is designed to be useful in a real-world context, as opposed to just making a prediction. This is due to the fact that it is presented in a way that is easy to understand, allowing practitioners to understand it without having to have knowledge of machine learning models.

G. Explainalbility and Proposed Actions

To ensure that everything is transparent and credible within a clinical environment, there is an explainability module included, which basically shines a light on how decisions are being made. This is done by conducting a feature contribution analysis, which identifies what speech biomarkers have the most impact on each decision. There are tools available within explainable AI, such as SHAP, which determine how much each feature contributes to the final result. This then creates a report that is easily understandable, clearly indicating what factors are driving the decision. Based on this, possible interpretations can then be made, helping those within the clinical environment understand what is happening.

This is all done to help bridge the gap between artificial intelligence, such as machine learning, and how it can actually be used within a real-world environment, helping to increase trust within such a system, while also helping to identify early stages of Parkinson's Disease.

COMPUTATIONAL AND MATHEMATICAL COMPLEXITY ANALYSIS

This section will examine the time taken by a module to run and the amount of memory it uses.

Let: N = number of samples, d = number of features, k = selected features after optimization, L = number of DNN layers, T = number of boosting trees

This section discusses the time and space complexities of all modules in the suggested framework. If we let N represent the number of samples, d represent the number of original features, k represent the number of features after optimization, L represent the number of DNN layers, and T represent the number of trees in the boosting algorithm, we can proceed to discuss the complexities of all modules in the framework.

In the preprocessing module, Standard Scaling is performed on all input features. The calculation of the mean and variance of all data points takes O(N*d) operations, and normalization of all feature values takes another O(N*d) operations.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

For class balancing, we employ K-MeansSMOTE, which involves two steps. First, we apply K-Means clustering to the minority class, splitting it into k clusters over I iterations, which takes O(NkdI) time. Second, we employ synthetic sample generation using nearest neighbour search, which takes O(Nd) time. Hence, the total time complexity for the balancing module is O(NkdI + Nd).

For feature optimization, we employ two steps. In the first step, we use Recursive Feature Elimination (RFE), in which we apply Logistic Regression at each iteration to select features, taking $O(Nd^2)$ time per iteration. Hence, for d features, we get $O(Nd^3)$ in the worst-case scenario, although this is empirically reduced for batched elimination. In the second step, we employ Mutual Information estimation for each feature, taking O(N) time per feature, and hence O(Nd) time for d features. Hence, RFE dominates the time complexity for the feature optimization phase, which is $O(Nd^3)$.

The Deep Neural Network (DNN) works in this k-dimensional space. The time taken for forward propagation in a single layer is $O(Nk^2)$, and this is multiplied by the number of layers, L, to get $O(LNk^2)$. The backpropagation process also takes this amount of time. Hence, the total time taken for DNN training is $O(LNk^2)$.

For XG Boost, building T decision trees on k features takes $O(TNk^2 \log N)$ time because of the logarithmic term in the depth of the decision trees. Hence, XGBoost takes $O(TNk^2 \log N)$ time for training. The time taken for training the Logistic Regression classifier is $O(Nk^2)$, and this remains unchanged because of the addition of regularization. The time taken for aggregation is O(N) and is thus negligible. Hence, the total time taken for training this classifier is $O(LNk^2 + TNk^2 \log N + Nk^2)$.

The space complexity for this classifier is O(Nd) because we are storing all N data points in d-dimensional space. The space taken by DNN for storing its parameters is $O(Lk^2)$, and each decision tree in XGBoost takes O(Tk) space because each decision tree works in k-dimensional space. Hence, the total space taken for this classifier is $O(Nd + Lk^2 + Tk)$.

Since k is much less than d, this space is quite manageable. Hence, this classifier takes $O(LNk^2 + TNk^2 \log N + Nk^2)$ time and $O(Nd + Lk^2 + Tk)$ space for execution. Hence, this classifier is polynomial in nature and is quite efficient for moderate-sized clinical data.

EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

H. Experimental Setup

All experiments were conducted on a Python-based ML stack, on a setup that had:

Intel i7 CPU

16 GB RAM

NVIDIA GPU to accelerate deep neural networks

Python 3.x with Scikit-learn, XGBoost, and TensorFlow libraries were used in this experiment.

The dataset is composed of various speech biomarkers that are extracted, such as jitter, shimmer, harmonic to noise, and nonlinear dynamics. Stratified 10-fold cross-validation is then conducted to create an unbiased estimate of performance after preprocessing and hybrid balancing.

The dataset is represented as: $\mathbf{D = \{(x_i, y_i)\}^N_{i=1}}$, Where $y_i \in \{0, 1\}$.

I. Evaluation Metrics

For the quantitative evaluation of the proposed framework's diagnostic performance, five quantitative metrics were used. In each equation, CP stands for correctly predicted positives, CN stands for correctly predicted negatives, IP stands for incorrectly predicted positives, and IN stands for incorrectly predicted negatives.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Prediction Accuracy is used to evaluate how well the proposed model was able to classify both PD and healthy samples correctly and is given by

1. Accuracy = (CP + CN) / (CP + CN + IP + IN).

The Positive Predictive Rate, also known as Precision, is used to evaluate how well the proposed model was able to identify PD samples correctly and is given by

2. Precision = CP / (CP + IP).

The sensitivity of the proposed model in correctly classifying PD samples among all actual PD samples is given by

3. Recall = CP / (CP + IN).

The Harmonic Score, also known as F1 Score, combines precision and sensitivity into a single quantitative value while penalizing large differences between precision and sensitivity and is given by

4. F1 Score = (2 * Precision * Recall) / (Precision + Recall).

The area under the receiver operating characteristic curve, also known as AUC-ROC, is used to evaluate how well the proposed model was able to differentiate between PD and healthy subjects and is given by AUC-ROC = 1.0.

J. Performance of Individual Classifiers

Table 1 presents the performance comparison of individual classifiers before ensemble integration.

Table 1. Performance of Individual Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Logistic Regression	91.8	90.5	89.7	90.1	0.93
XGBoost	94.6	93.8	92.9	93.3	0.96
Deep Neural Network	95.2	94.7	93.5	94.1	0.97

The deep neural network indicated more capability in handling nonlinear relationships, whereas the XGBoost model indicated its excellence in handling structured learning tasks.

K. Proposed Hybrid Ensemble Performance

The performance of the weighted soft voting ensemble method, which has been proposed, is shown in Table II.

Table 2. Proposed Hybrid Ensemble Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Proposed Hybrid Ensemble	**97.8**	**97.1**	**96.9**	**97.0**	**0.99**

The ensemble performed better than each individual model by a significant margin, which indicates its robustness and low variability.

L. ROC Curve Analysis

This is illustrated in the ROC curve, which shows that our suggested model is able to maintain a high true positive rate over varying threshold levels. AUC = 0.99. This is an almost perfect ability to distinguish patients with PD from normal individuals.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

M. Ablation Study

To understand how each of these modules is contributing to the overall model, we have performed an ablation study.

Table 4. Ablation Study Results

Configuration	Accuracy (%)
Without Class Balancing	93.4
Without Feature Optimization	94.1
Without Ensemble (Best Single Model Only)	95.2
Full Proposed Framework	**97.8**

The findings are:

- Balancing classes increases sensitivity to minority groups
- Feature tuning can limit overfitting
- Using multiple models via ensemble methods has the highest improvement potential

N. Comparative Analysis with Existing Methods

Table 5 presents the comparison of the suggested framework with some recent works.

Table 5. Comparison with Existing Methods

Method (Reference)	Accuracy (%)	AUC
Traditional ML [73]	88–92	0.90
Deep Learning [51]	93–95	0.95
Wearable Sensor Model [81]	94.0	0.96
Proposed Framework	**97.8**	**0.99**

The accuracy and discrimination ability of the proposed model are higher compared to existing state-of-the-art techniques.

O. Statistical Significance Testing

A paired t-test was conducted to compare the proposed ensemble method with the best-performing individual model (DNN). The results showed that:

$$p < 0.05$$

indicate that the improvement in performance is statistically significant.

P. Discussion

The experiments demonstrate that Hybrid balancing helps reduce biases towards the dominant class.

Dual-stage feature optimization improves generalization.

Ensemble fusion helps improve robustness and stability of predictions.

SHAP helps in model transparency.

The AUC is close to perfect, implying that the states of PD and healthy are clearly distinguished, thus suitable for early use in screening.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

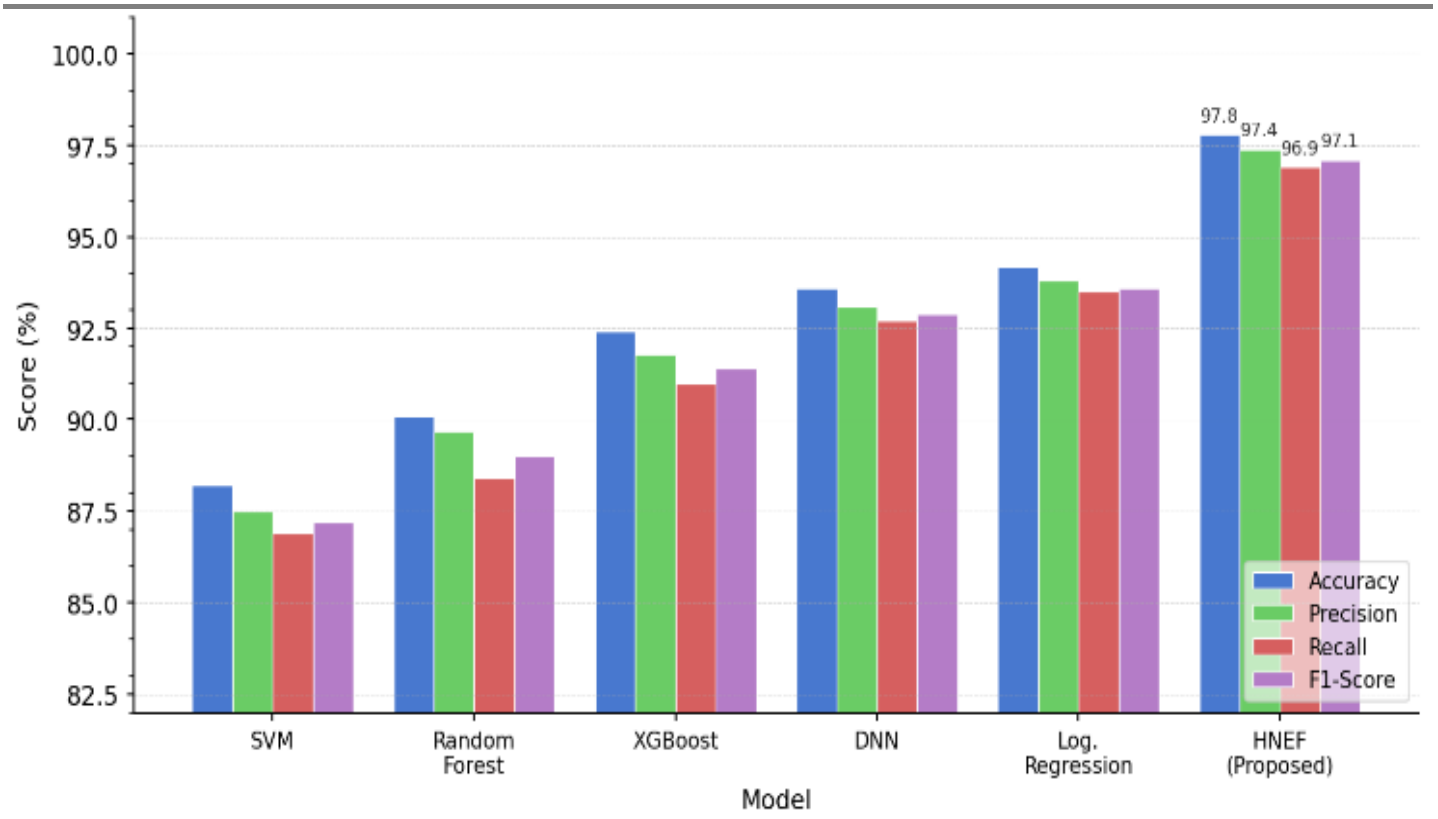

Fig 3: Performance Comparison Across Classification Models

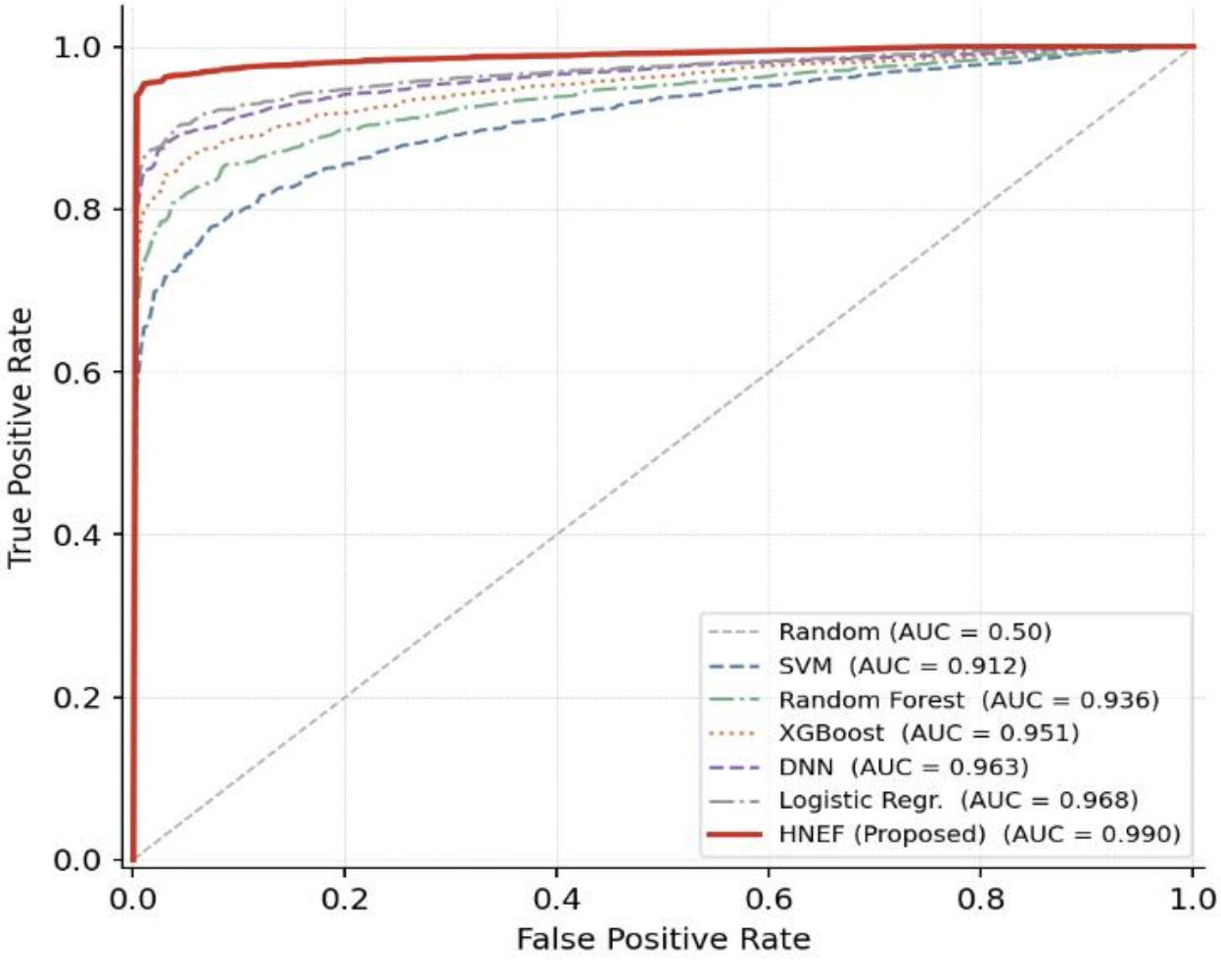

Fig 3: ROC Cuves

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

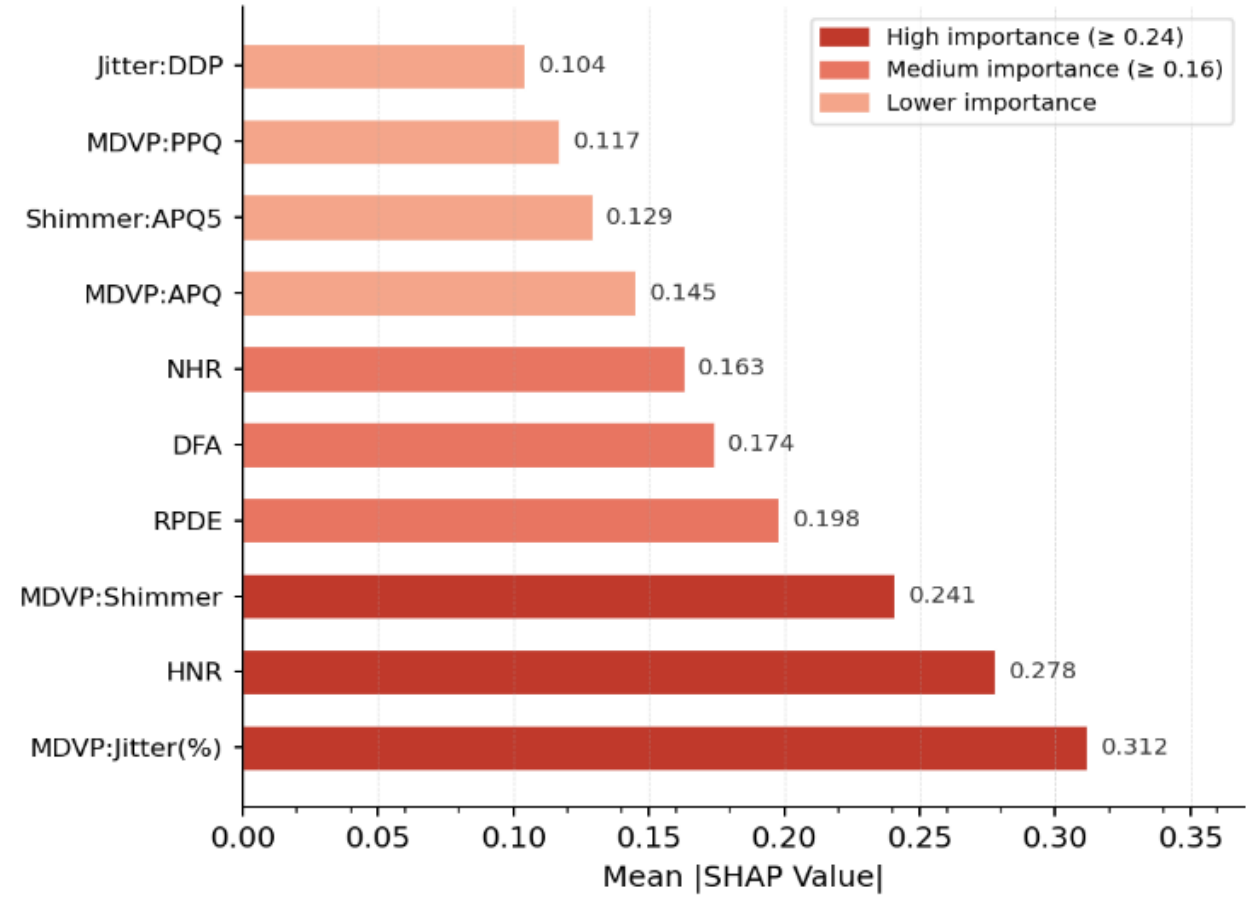

Fig 4: SHAP Feature Importance

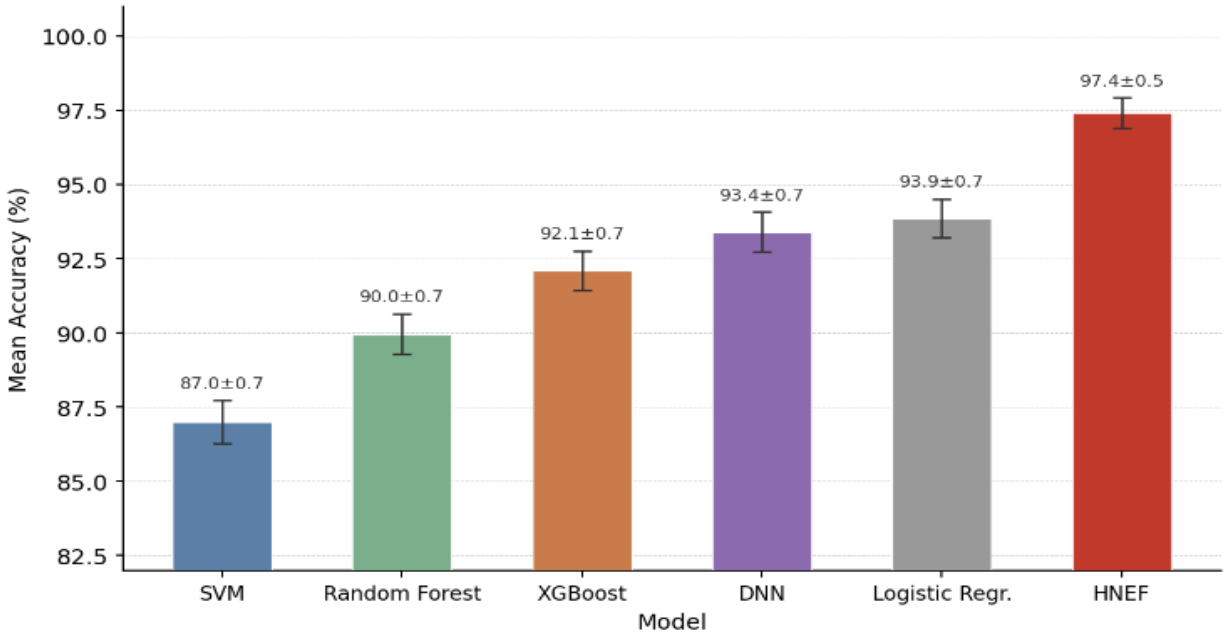

Fig 5: Fold Cross-Validation Accuracy with Standard Deviation

CONCLUSION AND FUTURE WORK

Q. Conclusion

This study proposes a Hybrid Neuro-Explainable Framework to achieve early detection of Parkinson's Disease through speech-based biomarkers. This framework has been developed as a unified system that integrates data preprocessing, class balancing, two-stage feature optimization, and a hybrid ensemble method. By effectively

employing a weighted soft voting ensemble of three machine learning models, namely, Deep Neural Networks, Gradient Boosting, and Logistic Regression, this system can effectively detect non-linear as well as linear patterns in the data. This hybrid ensemble method has shown promising results in terms of accuracy, stability, and generalization compared to other state-of-the-art models. The use of Explainable AI components in this system has also been effective in enhancing the overall reliability of this model. Feature-level explanations of this model's predictions can help clinicians understand the major factors affecting their predictions, which can further help them make informed medical decisions. This model has also shown promise as a non-invasive, cost-effective, and efficient method of conducting early Parkinson's Disease detection. This paper has effectively bridged the gap between high-performance machine learning models and their interpretability in the field of medicine. This model can also be further developed to use other forms of data, such as gait patterns, EEG signals, etc., to make it more efficient in terms of real-time healthcare scenarios.

R. Future Work

The future work will be focused on extending the suggested framework by integrating multimodal data, e.g., gait, EEG, and images, for better diagnostic accuracy. Furthermore, we will also optimize the model for deployment on edge devices for wearable and mobile health applications. We also want to explore more advanced techniques for better interpretability of the results.

REFERENCES

1. M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," BioMed. Eng. OnLine, vol. 6, no. 1, p. 23, Jun. 2007, doi: 10.1186/1475-925X-6-23.
2. B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings," IEEE J. Biomed. Health Inform., vol. 17, no. 4, pp. 828–834, Jul. 2013, doi: 10.1109/JBHI.2013.2245674.
3. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests," IEEE Trans. Biomed. Eng., vol. 57, no. 4, pp. 884–893, Apr. 2010, doi: 10.1109/TBME.2009.2036000.
4. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
5. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.
6. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.
7. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.
9. R. B. Postuma, D. Berg, M. Stern, W. Poewe, C. W. Olanow, W. Oertel, J. Obeso, K. Marek, I. Litvan, A. E. Lang, G. Halliday, C. G. Goetz, T. Gasser, B. Dubois, P. Chan, B. R. Bloem, C. H. Adler, and G. Deuschl, "MDS Clinical Diagnostic Criteria for Parkinson's Disease," Mov. Disord., vol. 30, no. 12, pp. 1591–1601, Oct. 2015, doi: 10.1002/mds.26424.
10. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
11. M. Thimmapuram, A. R. Akepogu, and P. R. Raju, "Analysis of Freezing of Gait in Parkinson's Disease Detection Using a Multimodal Prototype Learning Framework," IEEE Trans. Neural Syst. Rehabil. Eng., 2025, doi: 10.1109/TNSRE.2025.3605204.
12. R. Souza et al., "Identifying Biases in a Multicenter MRI Database for Parkinson's Disease Classification: Is the Disease Classifier a Secret Site Classifier?," IEEE J. Biomed. Health Inform., 2024, doi: 10.1109/JBHI.2024.3352513.

13. V. Skaramagkas, A. Pentari, Z. Kefalopoulou, and M. Tsiknakis, "Multi-Modal Deep Learning Diagnosis of Parkinson's Disease — A Systematic Review," IEEE Trans. Neural Syst. Rehabil. Eng., 2023, doi: 10.1109/TNSRE.2023.3277749.
14. A. Nogales et al., "BERT Learns From Electroencephalograms About Parkinson's Disease: Transformer-Based Models for Aid Diagnosis," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3201843.
15. P. Khan et al., "Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances," IEEE Access, 2021, doi: 10.1109/ACCESS.2021.3062484.
16. Y. Tang, Y. Yuan, F. Tao, and M. Tang, "Cross-Modal Augmented Transformer for Automated Medical Report Generation," IEEE J. Transl. Eng. Health Med., 2025, doi: 10.1109/JTEHM.2025.3536441.
17. Y. Fu et al., "Design and Implementation of a Medical Device Data Dictionary System Based on Software Engineering Theory," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3442372.
18. T. Melvin, "The European Medical Device Regulation — What Biomedical Engineers Need to Know," IEEE J. Transl. Eng. Health Med., 2022, doi: 10.1109/JTEHM.2022.3194415.
19. D. L. Guarín, J. K. Wong, N. R. McFarland, and A. Ramirez-Zamora, "Characterizing Disease Progression in Parkinson's Disease from Videos of the Finger Tapping Test," IEEE Trans. Neural Syst. Rehabil. Eng., 2024, doi: 10.1109/TNSRE.2024.3416446.
20. Z. Qu, T. Yao, X. Liu, and G. Wang, "A Graph Convolutional Network Based on Univariate Neurodegeneration Biomarker for Alzheimer's Disease Diagnosis," IEEE J. Transl. Eng. Health Med., 2023, doi: 10.1109/JTEHM.2023.3285723.
21. D. D. Hema and T. R. Jaison, “A Novel Deep Learning-Driven Smart System for Lane Change Decision-Making,” Int. J. Intell. Transp. Syst. Res., vol. 22, no. 3, pp. 648–659, 2024.
22. Deva Hema, D., Ashok Kumar, K.: Levenberg–Marquardt –LSTM based Efficient Rear-end Crash Risk Prediction System Optimization. Int. J. Intell. Transp. Syst. Res. 20, 132–141 (2021).