Grenze International Journal of Engineering and Technology, January Issue

Explainable Artificial Intelligence and Multilingual Voice Analysis for Early Parkinson's Disease Detection: A Comprehensive Review and Proposed Framework

Surekha M[1], Mohit Shukla[2], Markanday Sahu[3] and Navaneet Chaturvedi[4]
[1-4]Department of Computer Science, JSS Academy of Technical Education, Noida, India
Email: surekha@jssaten.ac.in, mohitshukla2004.ms, markanday.sahu20, nc05ac152007}@gmail.com

***Abstract*— Parkinson's disease (PD) is a progressive neurodegenerative disorder with ever-increasing worldwide prevalence imposing important clinical and socioeconomic burdens. Where motor symptoms, the target of conventional diagnosis, occur relatively late in disease, speech and voice changes have become an earlier feature. These patterns of reduced pitch variation, monotonous prosody, jitter, shimmer, breathiness and imprecise articulation indicate laryngeal control dysfunction. Machine learning (ML) and deep learning (DL) enable automatic PD detection using acoustic features and neural architectures. However, challenges persist: lack of interpretability, linguistic bias from English-dominant datasets, and vulnerability to recording variability. This paper reviews advances in voice-based PD detection, emphasizing Explainable AI (XAI) and multilingual modeling. We also present a framework integrating classical ML, deep models, self-supervised encoders, and interpretability tools such as SHAP, LIME, integrated gradients, and attention maps.**

***Index Terms*— Parkinson's Disease, Explainable AI, Speech Processing, Voice Biomarkers, Deep Learning, Multilingual Modeling.**

I. INTRODUCTION

Parkinson's disease is a chronic and progressive neurodegenerative condition characterized by the loss of dopaminergic neurons in the substantia nigra, causing disturbances to motor control. Core symptoms involve tremor, bradykinesia, rigidity, and postural instability, but potential non-motor features include olfactory deficits, sleep disturbances, cognitive decline, and speech impairments evident in early stages that can seriously impact quality of life. The prevalence of PD is increasing with global aging of populations, thereby driving up public health and socioeconomic burdens. Early diagnosis importantly offers opportunities for timely therapy, better symptom management, and improved long-term planning; however, clinical diagnosis often happens well after clear motor symptoms have appeared.

One of the earliest detectable non-motor markers of PD involves speech and voice changes. Hypokinetic dysarthria, typically manifested by reduced loudness, monotonic pitch, imprecise articulation, breathiness, and flattened prosody, reflects the neuromuscular deficits that affect respiration, phonation, and articulation. These abnormalities may be captured with acoustic metrics including jitter, shimmer, harmonic-to-noise ratio, formant variations, and Mel-frequency cepstral coefficient patterns.

Grenze ID: 01.GIJET.12.1.582

With increased audio processing and machine learning, voice recordings of computational models are promising for the detection of PD by using common devices. The classical approaches are based on engineered features while the classi- fiers are either SVMs or Random Forests, whereas the deep learning architecture learns representations that are richer in nature-CNNs, LSTMs/GRUs, and transformers-aided by self-supervised pretraining. However, the major challenges in- clude limited model interpretability and dataset bias towards English speakers. Multilingual representation learning and state-of-the-art XAI techniques have to be integrated as a solution to the problem at hand.

II. Literature Review

Research on voice-based Parkinson's Disease detection has spanned several decades and converged across signal process- ing, machine learning, and clinical research. This review synthesizes contributions under four lenses-acoustic biomarker characterization, classical machine learning pipelines, deep learning and representation learning, and explainability efforts applied to speech models. Emphasis is given to multilingual research and the limitations that restrain real-world deploy- ment.

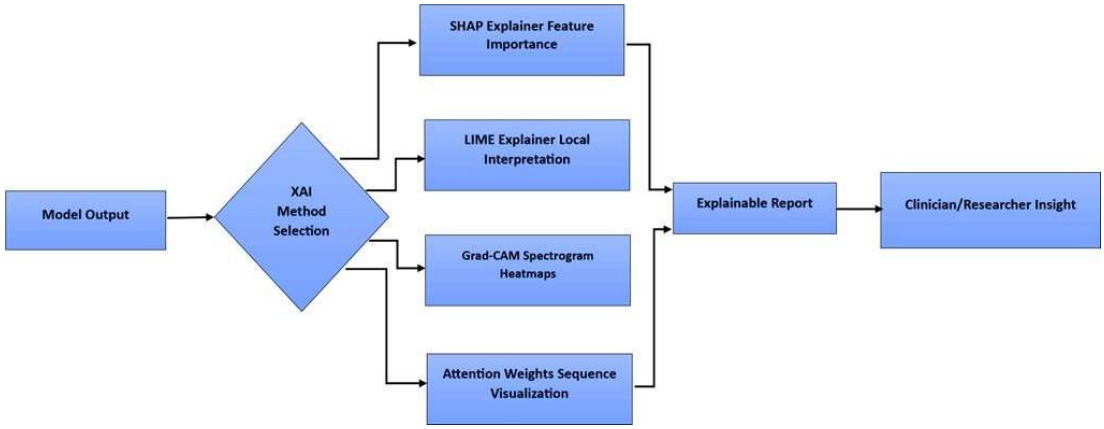

Figure 1: Overview of Explainable AI pipeline

A. Acoustic Biomarkers and Physiological Basis

Understanding the speech impairments of PD is critical for computational detection. PD affects neuromuscular control of the larynx and articulators, generating a wide range of acoustic anomalies. Early work, such as Tsanas et al. (2012), iden- tified nonlinear dysphonia measures that correlated with clinical severity. Increased jitter and shimmer indicate instability in vocal fold vibration, whereas decreased HNR indicates breathiness or incomplete closure of the glottis. Formant shifts and spectral distortions indicate imprecision in articulation. Prosodic disturbances, including reduced variability in pitch, a restricted range, slowed speech rate, and reduced intensity, further typify the speech characteristics of PD. Most ML-based detection systems rely on the following extracted features using Praat, openSMILE, and librosa.

B. Classical Machine Learning Approaches

Early works related to the voice detection of PD were focused on using engineered acoustic features and classical machine learning models. The UCI Parkinson's and Oxford corpora provided access to reproducible experiments using curated feature vectors. Support vector machines, Random Forest, k-NN, and gradient boosting methods were among the usual classifiers. Tsanas et al. showed that nonlinear dysphonia measures coupled with statistical summaries could yield impres- sive accuracy in the estimation of PD severity. SVM was quite effective for small datasets and high-dimensional spaces, while ensemble methods included Random Forest and XGBoost, which captured nonlinear feature interactions rather well. However, such techniques heavily depend on a manual feature engineering process and cannot capture intricate patterns in the temporal-spectral plane of raw audio.

C. Deep Learning and Representation Learning

Deep learning has redefined PD voice detection by allowing models to learn informative representations directly from raw audio or spectrograms instead of relying on handcrafted features alone. Convolutional neural networks

represent extremely effective options for extracting local time–frequency patterns related to dysphonia in Mel-spectrograms, while recurrent architectures, such as LSTMs and GRUs, model longer-term temporal dependencies across speech frames. Hybrid archi- tectures like CNN–LSTM and CNN–BLSTM integrate these strengths by providing both localized feature extraction and more broad contextual modeling. Attention mechanisms further improve localized highlighting of diagnostically relevant speech segments, as does the transformer-based encoder. Self-supervised models like wav2vec 2.0 and HuBERT emanated from ASR but have been adapted to clinical datasets, offering robust pretrained embeddings for improved performance when there is a lack of resources. Studies such as those by Vasquez-Correa, Ren, and Arora demonstrate that deep archi- tectures coupled with multilingual training and transfer learning outperform classical pipelines and better capture subtle temporal-spectral biomarkers of PD.

TABLE I: UPDATED SUMMARY OF RELATED WORKS AND THEIR MAPPING TO THE PROPOSED FRAMEWORK

S.No.	Authors	Year	Dataset / Source	Key Contribution	Mapping to Proposed Framework
1	Shen et al.	2025	Clinical Voice Dataset (Sci Rep)	Explainable AI using SHAP for PD voice analysis	XAI + Modeling
2	Dentamaro et al.	2024	PPMI Multimodal Dataset	Multimodal DL + Explain-ability for early PD detec- tion	Multimodal fusion + XAI
3	Ngo et al.	2022	Systematic Review	Comprehensive review of speech biomarkers for PD	Baseline biomarker insights
4	Favaro et al.	2023	Multilingual Speech Corpora	Interpretable multilingual biomarkers for PD	Multilingual modeling
5	Jeong et al.	2024	Multilingual Speech Recordings	ML-based PD classification using linguistic variability	Feature extraction + Multilingual
6	Saravanan et al.	2022	Review Study	AI-based PD diagnosis roadmap	Validates ML/DL di-rections
7	Escobar Grisales et al.	2023	PD Speech/Language Dataset	Deep models for articulatory + linguistic impairments	Modeling + Speech-language insights
8	Jeancolas et al.	2021	PD Voice Corpus	X-vector based quantitative biomarkers	Feature learning
9	Pandey et al.	2024	Voice Signal Dataset	CNN-LSTM model for PD speech classification	Deep modeling
10	van Gelderen et al.	2024	Systematic Review	Survey of DL techniques for PD speech detection	Research validation
11	La Quatra et al.	2025	Bilingual PD Speech Corpus	Dual-head model for bilin-gual PD detection	Multilingual + Modeling
12	Quan et al.	2021	PD Speech Dynamics Dataset	Dynamic acoustic features + DL model	Temporal feature modeling
13	Hires et al.	2022	Augmented Voice Recordings	Voice-specific data augmen-tation for PD detection	Robust preprocessing
14	Dixit et al.	2023	Review Paper	Comprehensive review of AI models for PD	Support for XAI + DL integration
15	Majda- Zdancewicz et al.	2024	Speech Phonation/Prosody Dataset	Analysis of phonation, artic-ulation, prosody in PD	Biomarker analysis
16	Khanom et al.	2024	Multi-feature PD Dataset	Explainable ensemble learning for PD (XEMLPD)	XAI + Ensemble modeling
17	Velu & Jaisankar	2025	ML-based PD Dataset	Early prediction using ML classifiers	Classical ML module
18	Shyamala & Navamani	2024	PD Voice Dataset	Efficient ML-based prediction model	ML baseline improve-ments
19	Veetil et al.	2024	Language independent Dataset	Robust PD detection across languages	Language-independent modeling
20	Meghraoui et al.	2021	Pathological Voice Dataset	New preprocessing tech-nique for PD phonation analysis	Preprocessing + Biomarker extraction

D. Multilingual Datasets and Cross-lingual Generalization

Multilingual robustness is a critical requirement for creating PD voice screening systems to work globally. Although several initiatives have developed multilingual datasets, publicly available corpora remain dominated by English, such as the smartphone-based mPower dataset. Resources like PC-GITA and Neurovoz add Spanish and Spanish–Portuguese samples, respectively. More recent efforts have been directed to Hindi, Mandarin, Arabic, and other languages. Phonetics, prosody, and tonal patterns vary across languages, so monolingually trained models often fail when applied cross-lingually. Performance drops without domain adaptation are

consistently reported. Techniques such as domain adversarial training, language-aware normalization, and multilingual pretraining mitigate these distribution shifts.

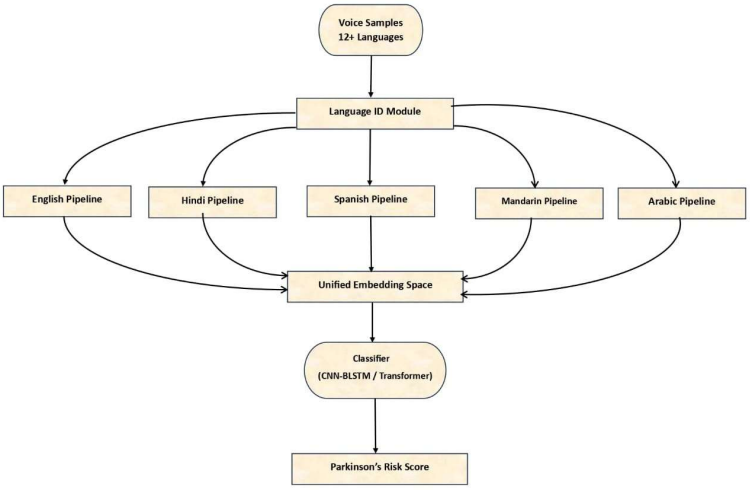

Figure 2: Multilingual PD speech datasets and cross-lingual generalization challenges

E. Explainability and Clinical Trust

Explainability is crucial for clinical adoption of PD voice models, because black-box decisions cannot be trusted in clinical settings. Feature-level XAI tools such as SHAP and LIME assign importance scores to acoustic features including jitter, shimmer, and MFCCs, with the goal of helping clinicians understand which measures influenced the model's prediction. For deep models using spectrograms or raw audio, methods such as Grad-CAM, Integrated Gradients, SmoothGrad, and attention visualization highlight critical time–frequency regions. While some studies suggest alignment between XAI model outputs and known physiological markers, the use of standardized XAI practice is limited. Furthermore, few studies assess explanation reliability or stability, and without this, trust and clinical integration remain limited.

F. Benchmarking and Evaluation Practices

Further, without universally accepted benchmarks or standard evaluation protocols, the comparison across studies is com- plicated. The differences in dataset splits, of speaker-independent versus sample-wise, preprocessing including filtering and silence trimming, feature extraction pipelines, and performance metrics ranging from accuracy to AUC and sensi- tivity/specificity to F1 inflate or render incomparable claims. Speaker-disjoint cross-validation, multi-metric reporting, calibration checks, and validation on held-out corpora remain the best practice. Newer benchmarks, such as UPDSB, in coordination with corpus releases, attempt to present uniform tasks; their adoption is still growing.

G. Ethical, Societal, and Practical Considerations

Ethical concerns include sample bias, privacy, and potential harms from false positives or negatives. Voice is personally identifying; consent, de-identification, and secure handling are essential. Symptom misclassification has the potential to cause distress or lead to unnecessary clinical investigations or missed treatment opportunities. The accessibility of the technology in low-resource contexts depends on the development of multilingual datasets and on ensuring models are robust across a variety of devices and recording environments. Participatory data collection strategies, open dataset governance, and inclusion of diverse demographic cohorts are increasingly recommended by researchers.

H. Synthesis and Research Gaps

In summary, this literature review shows that (1) acoustic biomarkers are physiologically valid indicators for PD; (2) deep learning models, especially with pretraining, are superior to many classic pipelines; (3) multilingual modeling and domain adaptation become essential in global deployment; and (4) explainability mechanisms need to be systematically integrated and validated to build clinician trust. There are clear gaps in the literature regarding large-scale, multi-language benchmark datasets, standardized protocols of evaluation, and rigorous

XAI validation linked to clinically meaningful phenomena. The proposed framework in Section III is focused on addressing these gaps.

III. METHODOLOGY

A. Proposed Framework Overview

Framework In summary, we propose a modular framework that involves multilingual audio capture, language identifica- tion, feature extraction, deep representation learning, explainability modules, and clinician-oriented reporting. The system shall be agnostic to devices-such as smartphones, laptops, and clinic microphones-and shall support two major modes of operation: (1) screening mode for population-level triage and (2) monitoring mode for longitudinal patient tracking.

The pipeline stages are:

- Data acquisition: Recordings are made of short standardized tasks: sustained vowel /a:/, sentence reading, diado- chokinetic syllable repetition, and spontaneous speech. Metadata includes language, age, gender, type of device used, and clinical annotations if available.
- Preprocessing: Noise reduction - spectral subtraction or Wiener filters, trimming silence, resampling to a single rate - for example, 16 kHz, amplitude normalization. Optionally, vocal tract length normalization and per-language normalization statistics are kept.
- Language identification (LID): A lightweight LID model classifies the sample into language/family categories. The LID output informs language-specific normalization and model routing.
- Feature extraction: Both handcrafted and learned features. Handcrafted features include jitter, shimmer, HNR, MFCCs, formant statistics, prosodic descriptors (pitch contour metrics, speech rate). For the learned features, Mel- spectrogram patches, raw waveform frames, and embeddings from pre-trained encoders are extracted (wav2vec 2.0, HuBERT).
- Modeling: A multilingual backbone is trained with several strategies, including joint multilingual training where language labels are added as auxiliary signals, domainadaptive finetuning from self-supervised speech encoders, and ensemble fusion that combines outputs from classical ML and deep learning models. The model architectures involved include CNN–BLSTM hybrids and transformer encoders with attention pooling.
- Explainability: Two parallel streams of explanations complement each other. Feature-level XAI, SHAP/LIME, presents the important engineered features, while input-level XAI-saliency maps, integrated gradients, and attention heatmaps-present time-frequency regions in spectrograms that drive decisions. The explanations are cross-validated for plausibility against known clinical biomarkers.
- Output and clinician interface: The system outputs a score for the probability of PD, associated confidence inter- vals, and visual explanations. A summary report will itemize which biomarkers contributed to that decision, e.g., increased jitter and decreased pitch variability, model calibration metrics, and suggested next steps, e.g., clinical referral, follow-up recording.

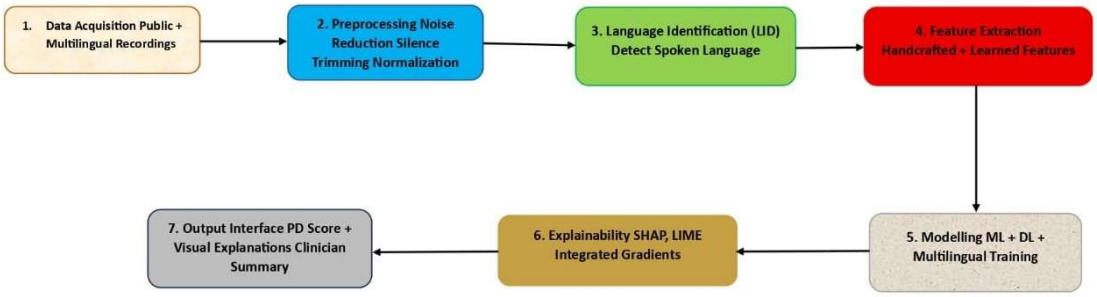

Figure 3: Overall workflow of the proposed PD detection model

B. Language-Independent Biomarkers and Normalization

A core principle is to focus on language-independent biomarkers-features whose physiological relation to PD is consis- tent across languages. Examples include jitter, shimmer, HNR, and some MFCC dynamics. Normalization is nevertheless required: language-specific mean-variance statistics and calibration with languagespecific healthy controls remove con- founding due to phonetic variation.

C. Model Training and Evaluation Protocol

We adopt speaker-disjoint splits and nested cross-validation to ensure fair evaluation. Performance metrics include ac- curacy, precision, recall, F1-score, area under the ROC curve, and calibration metrics such as Brier score and expected calibration error. We emphasize external validation on held-out languages that have not been seen during training to probe true generalization. Ablation studies determine the contribution of (1) handcrafted features, (2) pre-trained embeddings, and (3) explanation regularization-training encouraging explanations to align with clinical markers.

IV. Discussion

Voice-based PD detection demonstrates great potential for early screening and continuous monitoring, but some technical, linguistic, and clinical issues must be overcome before real-world deployment. A prominent technical issue is inconsis- tent recording conditions, since clinical microphones differ significantly from smartphones. Robust preprocessing, data augmentation, domain adaptation, and in-app recording quality checks may, nevertheless, reduce such domain shifts. Lin- guistic variability creates another barrier, as phonetic and prosodic differences between languages may be misattributed to pathology. Language identification, per-language normalization, or multilingual and transfer learning are thus crucial, and this needs diversified data collection efforts.

First, there is a need for interpretable models that clinicians can trust. Meaningful explanations, such as feature-level contributions or spectrotemporal saliency maps, need to be checked for reliability by perturbation tests and clinician-in- the-loop review. Better practices for benchmarking are also in order: speaker-disjoint splits, multi-metric reporting, and calibration assessment. Ethical considerations include privacy-preserving data handling, transparent consent, and careful communication of results. In conclusion, strong multilingual data, explainable modeling, and stringent evaluation are essential for clinical adoption.

V. Conclusion

This review explored voice-based Parkinson's Disease detection with a focus on explainability and multilingual general- ization. Acoustic biomarkers like jitter, shimmer, harmonic-to-noise ratio, patterns in MFCC, and prosodic features are physiologically grounded indicators of vocal dysfunction resulting from PD and provide a solid basis for computational analysis. While these engineered features are still useful for classical machine learning models because of their inter- pretability when data is scarce, deep learning approaches-and in particular those using self-supervised speech encoders-are more effective at capturing the complex temporal-spectral patterns characterizing pathological speech. There are two main challenges to clinical translation: first, large, diverse multilingual datasets are limited, and second, the integration of re- liable explainability methods is insufficient. Monolingual bias risks unfair performance across language groups. The use of black-box deep models reduces clinician trust. A unified framework will leverage multilingual data collection, hybrid feature-learning pipelines, domain-adaptive transfer learning, and layered XAI using feature attribution and saliency-based visualization. Future efforts should commit to multilingual benchmarking, robust evaluation, explanation validation, and longitudinal studies to enable the equitable and clinically dependable screening of PD.

References

[1] Shen, M., Mortezaagha, P. Rahgozar, A. Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. Sci Rep 15, 11687 (2025).

[2] Dentamaro, V., Impedovo, D., Musti, L. et al. Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database. Sci Rep 14, 20941 (2024).

[3] Ngo, Q.C.; Motin, M.A.; Pah, N.D.; Drota´r, P.; Kempster, P.; Kumar, D. Computerized analysis of speech and voice for Parkinson's disease: A systematic review. Comput. Methods Programs Biomed. 2022

[4] Favaro, Anna, et al. "Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in Parkinson's disease." Frontiers in Neurology 14 (2023): 1142642.

[5] Jeong, Seung-Min, et al. "Machine learning-based classification of Parkinson's disease using acoustic features: insights from multilingual speech tasks." Computers in Biology and Medicine 182 (2024): 109078.

[6] Saravanan, S., et al. "A systematic review of artificial intelligence (AI) based approaches for the diagnosis of Parkinson's disease." Archives of computational methods in engineering 29.6 (2022): 3639-3653.

[7] Escobar-Grisales, Daniel, Cristian David R´ıos-Urrego, and Juan Rafael Orozco-Arroyave. "Deep learning and artificial intelligence applied to model speech and language in Parkinson's disease." Diagnostics 13.13 (2023): 2163.

[8] Jeancolas, Laetitia, et al. ”X-vectors: new quantitative biomarkers for early Parkinson’s disease detection from speech.” Frontiers in Neuroinformatics 15 (2021): 578369.
[9] Pandey, Pandit Vivek Kumar, et al. ”Parkinson disease prediction using CNN-LSTM model from voice signal.” SN Computer Science 5.4 (2024): 381.
[10] van Gelderen, Lisanne, and Cristian Tejedor-Garc´ıa. ”Innovative Speech-Based Deep Learning Approaches for Parkinson’s Disease Classification: A Systematic Review.” arXiv preprint arXiv:2407.17844 (2024).
[11] La Quatra, Moreno, Juan Rafael Orozco-Arroyave, and Marco Sabato Siniscalchi. ”Bilingual dual-head deep model for parkinson’s disease detection from speech.” ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025
[12] Quan, Changqin, Kang Ren, and Zhiwei Luo. ”A deep learning based method for Parkinson’s disease detection using dynamic features of speech.” IEEE access 9 (2021): 10239-10252.
[13] Hireš, Ma´te´, et al. ”Voice-specific augmentations for Parkinson’s disease detection using deep convolutional neural network.” 2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2022.
[14] Dixit, Shriniket, et al. ”A comprehensive review on AI-enabled models for Parkinson’s disease diagnosis.” Electronics 12.4 (2023): 783.
[15] ajda-Zdancewicz, Ewelina, et al. ”Speech signal analysis in patients with Parkinson’s disease, taking into account phonation, articulation, and prosody of speech.” Applied Sciences 14.23 (2024): 11085.
[16] Khanom, Fahmida, et al. ”XEMLPD: an explainable ensemble machine learning approach for Parkinson disease diagnosis with optimized features.” International Journal of Speech Technology 27.4 (2024): 1055-1083.
[17] Velu, K., and N. Jaisankar. ”Design of an Early Prediction Model for Parkinson’s Disease Using Machine Learning.” IEEE Access (2025).
[18] Shyamala, K., and T. M. Navamani. ”Design of an efficient prediction model for early parkinson’s disease diagnosis.” IEEE Access 12 (2024): 137295-137309.
[19] Veetil, Iswarya Kannoth, et al. ”Robust language independent voice data driven Parkinson’s disease detection.” Engineering Applications of Artificial Intelligence 129 (2024): 107494.
[20] Meghraoui, D., et al.: A novel pre-processing technique in pathologic voice detection for Parkinson’s disease. Biomed Signal Process Control 68, 102604 (2021).