International Journal of Computer Applications (XXX - XXX)
Volume XXX - No.XXX, June 2026

Psychoacoustic Framework for Automated Dysarthria Severity Assessment Via Hybrid Acoustic Ensembles

Reddipalli Shashank[1], Gopi Kumar Jha[1], Nirbhay Singh[2], and Ramesh K. Bhukya[1]
Department of Electronics and Communication Engineering,
[1]Indian Institute of Information Technology Allahabad, Deoghat, Jhalwa, Prayagaraj-211015, (U.P.), India.
[2] Guru Ghasidas Vishwavidyalaya, Koni, Bilaspur-495009, Chhattisgarh, India.

ABSTRACT

Dysarthria is a neurological motor speech disorder that affects speech production and intelligibility through impairments in articulation, phonation, and prosody. Reliable severity assessment is important for clinical diagnosis and long-term monitoring, yet existing automated systems rely heavily on *Mel-Frequency Cepstral Coefficients* (MFCCs), which may not adequately capture subtle pathological speech variations. This work investigates *Bark-Frequency Cepstral Coefficients* (BFCCs) as an alternative acoustic representation for dysarthria severity classification. A hybrid 193-dimensional feature vector is constructed by combining BFCCs with Mel-spectrogram, Chromagram, Spectral Contrast, and Tonnetz features to characterize complementary articulatory, spectral, and prosodic properties of dysarthric speech. Experiments are conducted on the TORGO corpus, with *Voice Activity Detection* (VAD) applied for silence removal and *Adaptive Synthetic Sampling* (ADASYN) used to address class imbalance. The extracted features are evaluated using a diverse set of *machine learning* (ML) models, including *k*-NN, *Decision Tree* (DT), *Support Vector Machines* (SVM), PCA, *Random Forest* (RF), AdaBoost, LogitBoost, CatBoost, LightGBM, XGBoost, and SGD. Results show that BFCC-based hybrid features provide strong discriminative capability across severity levels. Among the evaluated models, RF achieved the highest classification accuracy of 98.33%, followed by LightGBM (98.20%) and XGBoost (97.68%). The results indicate that BFCCs capture pathological speech characteristics more effectively than conventional cepstral representations and, when combined with ensemble learning, enable accurate and robust dysarthria severity classification.

General Terms

Human Factors, Dysarthric Speech Recognition, Speech Assessment, Speech Enhancement, Speaker Adaptation, Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS)

Keywords

Dysarthria Severity Detection, Dysarthric Speech, Speech Disorders, Machine Learning, Acoustic Features, Speech Analysis, Classification.

1. INTRODUCTION

Human speech production requires coordinated, synchronized movement of more than 100 muscles in the respiratory, laryngeal, velopharyngeal, and oral structures, which are constantly active under precise neural control. Damage or a decline in the structure of any part of the central or peripheral pathways involved in this type of communication renders this channel faulty [18]. The main clinical defect in this neuromuscular failure is dysarthria, with no structural defect of the vocal organs, but rather an inability to appropriately actuate them. It is a very heterogeneous group of motor speech disorders which result from an acute neurological lesion (ischemic or hemorrhagic stroke, traumatic brain injury, cranial nerve lesions) or a progressive neurodegenerative condition such as *amyotrophic lateral sclerosis* (ALS), parkinson's disease, Huntington's disease, Friedreich's ataxia, multiple sclerosis, etc. [22]. Since various neurological diseases affect different parts of the motor system, there are several clinical types of dysarthria. Hypertonicity, strained-strangled voice, and slow speech rate are the most commonly observed characteristics of the speech, which is called *spastic dysarthria* (SD), when a bilateral *upper motor neuron* (UMN) lesion is present, and hypernasality has been quantified using modulation spectral features [28]. Damage to *lower motor neurons* (LMN) results in flaccid dysarthria characterized by severe breathiness, rapid air wastage, short phrases, and consistent hypernasality [14]. With cerebellar pathology, the dysarthria is ataxic, with irregular breaks in articulation and a "scanning" speech style, including increased equal stress and variation in pitch and/or intensity. The pathologies of the basal ganglia are divided into hypokinetic dysarthria (Parkinson's disease: quiet, monotone, fast speech with blurring consonants) [3] and hyperkinetic dysarthria (Huntington's disease: involuntary arrest of voice, variation in loudness) [33]. ALS results in a mixed spastic-flaccid presentation because motor neurons are degraded in both the upper and lower motor neurons. Speech analysis can be useful in different ways at different times in a disease. In stroke and severe *traumatic brain injury* (TBI), the dysarthria will have a sudden onset, and early objective testing will help to set functional baselines to inform rehabilitation. In contrast, neurodegenerative diseases have a gradual onset with subclinical deterioration of the bulbs occurring well before diagnosis in a "prodromal phase" that is characterized by a compensatory reduction in rate or increase in respiratory effort that occurs without

awareness to the patient. These behaviors help to maintain intelligibility in the short-term but help to camouflage the progressive neuromuscular deterioration. Once the compensatory mechanisms have been exhausted, communication ability will rapidly diminish. Measurement of small changes in the acoustic signal when this happens can help to provide early intervention, voice banking, proactive therapy, and timely assistive technology deployment [14, 33].

Automated dysarthria severity assessment can be used for a number of important applications. In longitudinal monitoring, remote telehealth can capture patients' speech samples using a short speech token on a smartphone, providing the clinician with ongoing objective data without physically visiting the clinic. Automated acoustic features that capture variability in the vocal tract, fundamental frequency, and changes in energy across the spectrum have much more fine-grained dimensions than the coarse integer scales used in clinical trials such as the *Unified Parkinson's Disease Rating Scale* (UPDRS) or *ALS Functional Rating Scale-Revised* ALSFRS-R for gauging treatment response [26]. Interactive biofeedback platforms can provide objective visual feedback in real time on pitch range, breathiness, and consonant contrast during home practice for the purpose of rehabilitation. Automated speech severity analysis in neurosurgery can inform the setting of Deep Brain Stimulation parameters to avoid unwanted current spread to neighboring corticobulbar tracts, which can affect limb motor function and exacerbate hypokinetic dysarthria. In pathological speech processing, the goal of feature extraction is to separate the distortion of the speech signal from the linguistic information. The typical tools used are *Mel frequency cepstral coefficients* (MFCC); however, the MFCC design removes fine spectral information, which is counterproductive for a diagnostic signal, especially when breathiness, hoarseness, and spirantization are the key characteristics of the disorder [3]. Joshy et al. [19] evaluated MFCCs and *Constant-Q cepstral coefficients* (CQCCs) and deep acoustic features on UA-Speech and TORGO, and obtained marginal improvements in using CQCCs over the optimized MFCC baselines, with their best CNN-MFCC pipeline attaining 96.18% accuracy in severity classification. Bhat et al. [4] claimed that the nature of dysarthric speech changes fundamentally on the level of speech timbre and was able to achieve 93% accuracy with a *Random Forest* (RF) on TORGO with attack time, spectral centroid, and spectral flux descriptors. Falk et al. [14] used modulation spectral features that measure the rate of change of the energy in the sub bands, and achieved 95% three-class spastic dysarthria accuracy on the *Nemours* database. Other psychoacoustic scales (other than Mel) have been less explored in pathological speech. The Bark scale, which is based on critical bandwidths for loudness perception [35] has better discriminative ability in lower frequencies, which are the most affected by incomplete tongue closure in severe dysarthria [30]. Vasquez-Correa et al. [33] studied Bark-band energies for Parkinsonian assessment, where they achieved better results than MFCCs for hypokinetic dysarthria detection, but they did not use the *Bark-Frequency Cepstral Coefficients* (BFCCs) [35] in conjunction with other complementary spectral contrast and tonal descriptors in a single vector.

Due to the limited size of datasets in medical corpora [30], the choice of classifiers for dysarthria studies is constrained. The *support vector machines* (SVMs) and RFs served as good baselines, and Bhat et al. [4] achieved good results when modeling boundaries in lower-dimensional feature spaces, in a non-linear fashion. In pathological speech, the effectiveness of the SVM lies in its ability to build optimal separating hyperplanes in high-dimensional acoustic feature spaces, and the RBF kernel is especially useful because it provides an implicit infinite-dimensional mapping that can represent the highly non-linear, overlapping decision boundaries typical of multi-class dysarthria severity. Othman et al. [1] systematically compared *k-nearest neighbor* (*k*-NN), SVM, and RF on UA-Speech and found that the SVM with an RBF kernel achieved the best macro F1-score of 0.88, and the RF also performed better at generalizing across the imbalanced classes. In addition, the ensemble property of RFs also helps reduce variance through bootstrap aggregation, which is very useful when noisy acoustic measurements are used in clinical recording situations. *k*-NN is very simple to compute but very sensitive to the curse of dimensionality and has been shown to perform poorly when the feature space has more than 50 dimensions without careful tuning of distance metrics. *Logistic Regression* (LR) and *Linear Discriminant Analysis* (LDA) have also been used as interpretable probabilistic baselines, but are not well suited to the overlapping severity distributions found in multi-class dysarthria tasks [21] due to their assumption of linear separability in the feature space. It was found that tackling class imbalance without intervention biases models towards the majority classes, and that oversampling with *synthetic minority over-sampling technique* (SMOTE) increases accuracy by 5-8% [11]. Subsequently, *gradient boosting* (GB)-based methods replaced the same for tabular classification. Regularized sequential boosting was introduced by *eXtreme Gradient Boosting* (XGBoost) [8], and it was applied on English, Korean, and Tamil dysarthric speech and prosodic features by Lee et al. [34], which produced an average F1 score of 67.14%. XGBoost's second-order gradient optimization and L_1/L_2 regularization also help it avoid overfitting on small clinical corpora, and its built-in missing-value handling is useful when acoustic features cannot be reliably calculated from poor-quality speech segments. *Light Gradient Boosting Machine* (LightGBM) and *Categorical Boosting* (CatBoost) improved scalability via histogram-based split search and native handling of categorical features, respectively, dramatically reducing training time while maintaining classification performance on tabular acoustic features [29]. One important feature of CatBoost is ordered boosting, which helps avoid leakage of the target variable during tree construction and is particularly important for speaker-independent dysarthria classification, where the speaker identity can be a confounding variable [13]. LogitBoost implemented probabilistic boundaries, which are better suited to overlapping severity distributions, and thus generated calibrated posterior class probabilities that are more interpretable for clinicians than hard class labels [15]. An earlier ensemble method, *Adaptive Boosting* (AdaBoost), was also tested for related speech pathology tasks but was reported to be more sensitive to mislabeling of severity ratings in clinical datasets, where there is often substantial interrater disagreement [31]. this work extend by evaluating 10 classifiers on a unique Bark-scale hybrid feature vector, providing the most complete comparison of classifiers to date on this feature vector. To ensure a fair and reproducible comparison between the classifiers, all classifiers are evaluated using the same stratified 5-fold cross-validation, with ADASYN oversampling applied only within each training fold to avoid data leakage, and hyperparameters optimized via Bayesian optimization [2]. Dysarthric corpora are inherently imbalanced in that there are many more healthy controls than severely impaired speakers. Typical techniques include SMOTE and *adaptive synthetic sampling* (ADASYN) [17, 11]. SMOTE with *principal component analysis* (PCA) was shown to improve F1-scores for the minority class by up to 12% on UA-Speech. Waveform augmentation (time stretching, pitch shifting, and noise addition) and SpecAugment (masking of the spectrogram) increased robustness by up to 5% in accuracy in the presence of limited training data [32].

To rigorously assess the proposed framework, one needs to examine the existing pathological speech corpora available, each of which

tackles a specific pathological or clinical dimension. The famous databases include the UA-Speech corpus, which is generally used for isolated word recognition using multiple microphone arrays [21]; the Nemours database, consisting of short sentences from 11 male speakers with varying degrees of dysarthria [25]; the Spanish spoken by Parkinson's patients by the PC-GITA corpus [27]; and the Korean speech command database for smart homes with different degrees of dysarthria level known as the QoLT/Home-Service corpus [9]. These databases have made a major contribution to the field, but the present study uses the TORGO database of dysarthric articulation [30]. This particular corpus was chosen for two main acoustic and clinical benefits. The recordings are made with head-mounted microphones to reduce room reverberation; the *signal-to-noise ratio* (SNR) ratio is high, precisely as it is needed from a mathematical point of view to ensure the stability of the BFCCs.
Nevertheless, the standard in use remains quite subjective, based on SLP-perceptual assessment of intelligibility, such as the *sentence intelligibility test* (SIT) or the *Frenchay dysarthria assessment* (FDA). These protocols are still time-consuming, susceptible to inter- and intra-rater variability, and susceptible to listener fatigue, personal bias, and room acoustics [12]. This variability makes it difficult for them to be sensitive, in detecting slight disease progression. In recent years, automated systems have evolved from simple perturbation measurements (jitter, shimmer) to the use of complex *machine learning* (ML) pipelines including boosting algorithms XGBoost [8], class imbalance mitigation [11], and systematic comparisons between classifiers [1]. Traditional ASR systems remain inadequate, as they are designed not to detect acoustic degradation but to compensate for it [24]. The acoustic boundaries between neighboring impairment categories are very non-linear and overlapping, making it difficult to apply feature engineering to multi-class severity classification. Although MFCCs have been the predominant representation so far [4, 19], the triangular Mel-scale filter banks they use to maximize linguistic invariance tend to smooth out fine-grained spectral anomalies. Micro-tremors, sudden changes in the configuration of the vocal tract, and slight widening of the formant bandwidth that signal early neuromuscular impairment often are lost in this process [10], making reliable separation of mild from moderate impairment difficult. In this work, BFCCs are investigated, offering finer low-to-mid frequency resolution in the region of the formants F_1 and F_2, which are important psychoacoustic parameters for the clinical manifestations of dysarthric degradation, namely articulatory undershoot and vowel centralization. The system used a 193-dimensional hybrid feature vector that accounts for BFCCs, the energy of the *Mel-spectrogram*, *Chroma* pitch-class profiles, *Spectral Contrast*, and *Tonnetz* harmonic mappings [16], and tested it on ten classifier architectures, which the system hypothesized best suited to capture the non-linear articulatory patterns of pathological speech. ADASYN is used in this work, as, unlike uniform oversampling, it focuses on generating synthetic data for hard-to-classify instances, which, in the multi-class severity setup, helps better preserve the inter-class boundary. Just high accuracy is not enough to adopt clinically; clinicians need understandable decision support. Vasquez-Correa et al. [33] used SHAP values to interpret the predictions made using RF for Parkinson's dysarthria assessment, and found that the most influential features were jitter, shimmer, and HNR. There are built-in feature importance metrics in LightGBM and XGBoost that serve similar purposes. Physiologically motivated features (BFCCs, Spectral Contrast, Chroma) allow contributions from the different features to be interpreted as concrete acoustic dimensions of speech pathology.
The remainder of this paper is organized as follows: Section 2 provides a description of the database and its statistics. Section 3 provides the feature extraction techniques used for analysis. Section 4 provides a brief description of each ML model used for the investigation. Section 5 presents the experimental results and analysis. Finally, Section 6 concludes the investigation and summarizes the work.

2. DATABASE

TORGO Corpus (English): The primary database of dysarthric articulation utilized for this study [30]. It contains recordings by 15 speakers: 8 speakers with dysarthria due to neurologic illness *Cerebral Palsy* (CP) and *Amyotrophic Lateral Sclerosis* (ALS), and 7 age and gender matched healthy controls. This corpus was chosen because it consists of head-mounted microphones (necessary for achieving a high Signal-to-Noise Ratio (SNR), which is important for the stability of BFCCs) and because it covers both isolated words and continuous speech, providing a comprehensive assessment of prosodic deficits. This dataset contains approximately 23 hours of English speech collected from speakers with dysarthria (caused by CP or ALS) and matched healthy controls (HC). This dataset is categorized into four severity classes: controlled, medium, low, and very low, and their class distribution is shown in Fig. 1. The detailed distribution of the TORGO dataset is shown in Table 1. Speech recordings are of the types: Isolated Words, Non-Words, Sentences, Paragraphs, and Picture Descriptions.

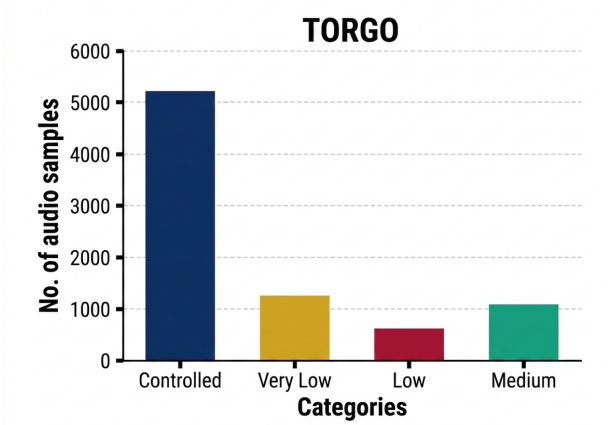

Fig. 1. TORGO database class distribution across severity levels.

3. FEATURE EXTRACTION

For feature extraction, the speech features came up with a 193-dimensional hybrid feature vector to capture various aspects of speech pathology. This vector contains a model of the source-filter mechanism of speech production. The feature extraction is used the spafe library [23] to extract 40 BFCCs. The Bark scale is based on the psychoacoustic theory of loudness perception and critical bandwidths suggested by Zwicker [35]. It simulates the frequency selectivity of the ear by dividing the audible spectrum into 24 critical bands, each corresponding to a different cochlear processing channel. The feature extraction pipeline proceeds as follows:
Voice Activity Detection (VAD): Pathological speech datasets may include long periods of silence or respiration between words [20, 7, 5]. Unless corrected, classifiers may learn to model silence

Table 1. TORGO dysarthric speaker utterance distributions.

Speaker ID	Group	Gender	Severity	Utterances	Speech Duration	Sessions	Recording Types
F01	Dysarthric	Female	Severe	228	Included in ∼14 h dysarthric speech	Multiple	Head Mic, Array Mic
F03	Dysarthric	Female	Moderate	1097		Multiple	Head Mic, Array Mic
F04	Dysarthric	Female	Mild	675		Multiple	Head Mic, Array Mic
M01	Dysarthric	Male	Severe	739		Multiple	Head Mic, Array Mic
M02	Dysarthric	Male	Severe	772		Multiple	Head Mic, Array Mic
M03	Dysarthric	Male	Mild	806		Multiple	Head Mic, Array Mic
M04	Dysarthric	Male	Severe	659		Multiple	Head Mic, Array Mic
M05	Dysarthric	Male	Moderate-Severe	610		Multiple	Head Mic, Array Mic
FC01	Control	Female	N/A	N/R	Included in ∼9 h control speech	Multiple	Head Mic, Array Mic
FC02	Control	Female	N/A	N/R		Multiple	Head Mic, Array Mic
FC03	Control	Female	N/A	N/R		Multiple	Head Mic, Array Mic
MC01	Control	Male	N/A	N/R		Multiple	Head Mic, Array Mic
MC02	Control	Male	N/A	N/R		Multiple	Head Mic, Array Mic
MC03	Control	Male	N/A	N/R		Multiple	Head Mic, Array Mic
MC04	Control	Male	N/A	N/R		Multiple	Head Mic, Array Mic
Total				**5586**	**∼23 h**	**15 Speakers**	**Head/Array Mic**

N/R denotes "Not Reported" in the official TORGO documentation. The TORGO corpus contains 15 speakers (8 dysarthric and 7 control speakers), with approximately 23 hours of recorded English speech collected with both head-mounted and array microphones. Reported utterance counts correspond to those commonly used for dysarthric speakers.

profiles rather than pathology. An energy-based VAD was implemented using the librosa framework. *Root Mean Square* (RMS) energy is calculated for short-term frames, mathematically defined as:

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x^2(n)} \tag{1}$$

Where $x(n)$ is the discrete amplitude of the speech signal and $N = 2048$ is the frame length. Segments with an E_{RMS} lower than a dynamic threshold—computed relative to the global maximum energy of the specific utterance—are considered non-speech background noise and are discarded systematically.

Pre-emphasis and Framing: The raw signal is pre-emphasized to give higher weight to the high-frequency formants ($\alpha = 0.97$) and then split into 25 ms frames with 10 ms overlap.

Bark Integration: The power spectrum is warped to the Bark scale (z) using the empirical approximation:

$$z = 13\arctan(0.00076f) + 3.5\arctan\left(\left(\frac{f}{7500}\right)^2\right) \tag{2}$$

Filter Bank Application: The power spectrum is filtered using a bank of 64 triangular filters. The Bark scale focuses the filter density in the low-frequency range (200–1000 Hz).

Decorrelation: The Discrete Cosine Transform (DCT) is used to transform the log-energies of the filter bank outputs to get the final cepstral coefficients. The n-th coefficient is calculated as:

$$c_n = \sum_{k=1}^{M} \log(E_k)\cos\left[\frac{\pi n(k-0.5)}{M}\right] \tag{3}$$

where E_k represents the energy of the k-th Bark filter, and $n \in [1, 40]$.

Clinical significance: Clinical deficits in the neuromotor system typically cause "vowel centralization," which is a collapse of the Vowel Space Area (VSA) due to the inability to move the tongue to the extreme positions of vowels like /i/, /u/, and /a/. Since F_1 and F_2 reside primarily in the low-to-mid frequency range, the Bark scale's finer resolution allows the classifier to detect subtle articulatory undershoot that standard MFCCs tend to smooth over.

Mel-Spectrograms: Computed the mean of 128 Mel-spectrogram bands in order to capture the high-frequency frication noise that is caused by imprecision in articulating consonants. The bands are averaged over time to get a static vector of the global energy distribution, which is mapped to the Mel scale using:

$$f_{Mel} = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

In dysarthric speech, this frequently reveals a spectral tilt—reduced energy at high frequencies due to poor glottal closure and weak articulatory contact.

Chroma Features: Extracted 12 Chroma features (Pitch Class Profiles) to quantify prosodic rigidity. The 12 different pitch classes p are obtained from the frequency f with the following relation:

$$p(f) = \left\lfloor 12\log_2\left(\frac{f}{f_{ref}}\right)\right\rfloor \pmod{12} \tag{5}$$

where f_{ref} is the reference tuning frequency (e.g., $C_0 \approx 16.35$ Hz). When a person speaks in a healthy fashion, the fundamental frequency (F_0) changes dynamically to express stress, which distributes energy into multiple Chroma bins. In dysarthric speech, hypertonic laryngeal muscles lock the voice to a narrow pitch range, producing static high-energy peaks in the Chroma vector—capturing prosodic rigidity that purely spectral features miss.

Spectral Contrast: Employed 7 bands of Spectral Contrast to measure voice quality. The feature is based on dividing the spectrogram into octave-scale sub-bands k and calculating:

$$C_k = \log(\max(S_k)) - \log(\min(S_k)) \tag{6}$$

where S_k is the RMS (root mean square) energy in band k. Low values for contrast mean that the noisy spectrum is flat, which correlates with high levels of breathiness due to incomplete closure of the glottis.

Tonnetz Features: The 12-dimensional Chroma vector $c \in \mathbb{R}^{12}$ is mapped to a 6-dimensional geometric space of intervals: the circle of fifths, minor thirds, and major thirds. The projection for the i-th

Table 2. Feature composition of the hybrid acoustic vector used for dysarthric speech analysis.

Feature Type	Dimensions
BFCC	40
Mel-Spectrogram Bands	128
Chroma Features	12
Spectral Contrast Bands	7
Tonnetz Features	6
Total	**193**

Tonnetz dimension is given by:

$$t_i = \sum_{p=0}^{11} c_p \cdot w_{i,p} \tag{7}$$

where $w_{i,p}$ are the transformation weights which map the pitch classes to the 2D spatial coordinates of the Tonnetz geometry. This mathematical projection proves to be very useful in the detection of pitch instability and micro-tremors in dysarthric speech.
The complete contents of the feature vector are listed in Table 2. The hybrid acoustic feature vector comprises spectral, perceptual, harmonic, and tonal descriptors [6]. BFCC captures perceptually relevant frequency information, Mel-spectrogram features represent time-frequency energy distributions, Chroma features encode pitch-class information, Spectral Contrast measures spectral peak-valley variations, and Tonnetz features characterize harmonic relationships. The concatenation of these descriptors results in a 193-dimensional feature vector.

4. METHODOLOGY

k-**NN:** It is a non-parametric, instance-based learner that stores the complete training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ and classifies an unseen sample $\mathbf{x}^*$ by identifying its k nearest neighbors. This lazy learning strategy is practical in small-data medical settings, where parametric models risk overfitting during training. In this work, the initial value set $k = 3$ after cross-validation over $k \in \{1, 3, 5, 7, 9\}$, finding that $k = 3$ offered the best bias-variance balance for four-class problem. Smaller values (especially $k = 1$) overfit on ADASYN-generated minority samples; larger values blurred boundaries between adjacent severity levels. Pairwise similarity is measured by *Euclidean distance*:

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}, \quad n = 193 \tag{8}$$

The predicted class is:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{j=1}^{k} \mathbb{I}\big(y^{(j)} = c\big) \tag{9}$$

where $\mathcal{C} = \{0, 1, 2, 3\}$, $y^{(j)}$ is the label of the j-th nearest neighbor, and $\mathbb{I}(\cdot)$ is the indicator function. A known limitation of *k*-NN in high-dimensional spaces is distance concentration: as dimensionality $n \to \infty$, the ratio $d_{\max}/d_{\min} \to 1$, meaning all samples become approximately equidistant from any query point. In 193-dimensional space, this manifests as confusion between the *healthy control* and *very low* intelligibility classes.
Decision Tree (DT): DTs partition the feature space recursively using information-theoretic criteria. Unlike distance-based or margin-based methods, they build an explicit rule-set through binary decisions on feature thresholds. The following steps were followed for the implementation: (1) *Tree construction*: At each internal node, the algorithm selects the feature A and threshold t that maximize *Information Gain* (IG). Node impurity is measured using the *Gini Index* (GI):

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2 \tag{10}$$

where p_i is the proportion of samples in class i at node S, and $c = 4$ is the number of severity classes. The GI equals zero when a node is perfectly pure and approaches its maximum when classes are uniformly distributed. The splitting criterion selects the feature-threshold pair (A, t) maximizing *Gini Gain* (GG):

$$GG(S, A, t) = GI(S) - \frac{|S_{left}|}{|S|} GI(S_{left}) - \frac{|S_{right}|}{|S|} GI(S_{right}) \tag{11}$$

where S_{left} and S_{right} are the subsets split by $A \leq t$ and $A > t$, respectively.
(2) *Stopping criteria:* The tree grows until; all samples in a node share the same class ($Gini = 0$); maximum depth is reached (`max_depth=None`); or minimum sample thresholds are met (`min_samples_split=2`, `min_samples_leaf=1`).
(3) *Prediction:* For a test sample $\mathbf{x}$, the tree crosses from root to leaf evaluating the learned rules at each node. The prediction is the majority class in the leaf:

$$\hat{y} = \arg\max_{i} \sum_{j \in L(\mathbf{x})} \mathbb{I}(y_j = i) \tag{12}$$

where $L(\mathbf{x})$ is the set of training samples that reach the same leaf as $\mathbf{x}$.
(4) *Clinical interpretability:* DTs produce explicit diagnostic rules, such as "IF BFCC_5 ≤ -2.3, and Spectral_Contrast_2 > 1.8, then severity = Severe." This transparency can help clinicians validate model decisions-useful for clinical acceptance.
(5) *Feature space partitioning:* DTs partition the space into axis-aligned hyperrectangles. This means rules like "BFCC_12 > 0.5" rather than complex linear combinations. The axis-aligned constraint limits performance when true decision boundaries are diagonal, but it identifies which individual acoustic features are most discriminative—a useful insight for speech pathology research.
(6) *Overfitting:* A single tree can grow arbitrarily complex to fit training data, causing high variance. The model achieved 93.25% accuracy—respectable, but short of ensemble methods.
SVM: SVMs are maximum-margin classifiers extended to the multi-class problem via *one-versus-one* (OvO) decomposition, training $\binom{4}{2} = 6$ binary classifiers and resolving conflicts through voting. The core SVM optimization builds a decision boundary in a kernel-induced *Reproducing Kernel Hilbert Space* (RKHS) that maximizes the geometric margin between the closest class clusters. The primal soft-margin optimization with slack variables $\xi_i \geq 0$ is:

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \tag{13}$$

subject to $y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where $C = 0.1$ is the regularization coefficient. A small C value enforces a wider margin at the cost of more boundary misclassifications, reducing overfitting to the acoustic noise inherent in dysarthric recordings. Because severity boundaries are non-linear in the original 193-dimensional space, and used the *radial basis function* (RBF) kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp\big(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\big), \quad \gamma = 0.001 \tag{14}$$

This computes inner products in $\mathcal{H}$ without explicitly computing $\phi(\mathbf{x})$. The small γ (broad kernel) allows each support vector to influence a wide region of feature space, producing smooth boundaries suitable for modeling gradual acoustic degradation across severity levels. The dual decision function is:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i\in\mathcal{S}} \alpha_i y_i \, K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{15}$$

where $\mathcal{S}$ is the set of support vectors, $\alpha_i \geq 0$ are Lagrange multipliers from the *Karush-Kuhn-Tucker* (KKT) conditions, and b is the learned bias term.

PCA + k-NN: To address KNN's vulnerability to distance concentration in 193 dimensions, prepend PCA as a dimensionality reduction stage. PCA finds a linear orthogonal transformation that projects data onto a subspace maximizing retained variance, via eigendecomposition of the sample covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{193\times193}$:

$$\mathbf{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \tag{16}$$

The eigenvectors $\{\mathbf{u}_j\}$ define the principal axes of the data distribution, and corresponding eigenvalues $\{\lambda_j\}$ quantify variance along each axis. Retain the top-d eigenvectors $\mathbf{U}_d$ satisfying:

$$\frac{\sum_{j=1}^{d}\lambda_j}{\sum_{j=1}^{193}\lambda_j} \geq 0.95 \tag{17}$$

This threshold preserves dominant articulatory variance while discarding components that mainly encode inter-frame and inter-speaker noise. Each input is projected as:

$$\tilde{\mathbf{x}} = \mathbf{U}_d^\top(\mathbf{x} - \boldsymbol{\mu}) \in \mathbb{R}^d \tag{18}$$

This projection eliminates correlated inter-band redundancy from adjacent Mel-spectrogram bins and overlapping BFCC coefficients, and concentrates clinically relevant variance into a compact subspace. k-NN with $k = 3$ is then applied to $\tilde{\mathbf{x}}$, where reduced dimensionality restores meaningful Euclidean proximity.

PCA's limitation is that its linear projection is blind to class labels—the principal components maximizing total variance do not necessarily separate the four severity classes. Non-linear interactions between BFCC bands and Spectral Contrast may be discarded if they lie along low-variance directions. This explains the accuracy gap between PCA + k-NN (94.24%) and ensemble methods, and motivates supervised non-linear dimensionality reduction (e.g., *linear discriminant analysis* or RF-based feature importance) in future work.

RF: RF is a bagging ensemble method that trains $T = 201$ DTs, exploiting the bias-variance decomposition: while individual trees are high-variance, averaging their predictions reduces variance without increasing bias. The odd number ($T = 201$) eliminates the majority-vote tie in the four-class problem.

Each tree h_t trains on a bootstrap resample $\mathcal{D}_t$ and evaluates only $m = \lfloor\sqrt{193}\rfloor = 13$ randomly selected features at each candidate split. This feature subsampling actively decorrelates tree predictions –by forcing each tree to specialize on different feature subsets, the method amplifies the variance reduction from averaging. Node splits maximize Information Gain under Shannon entropy:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{19}$$

$$\Delta H(S, A) = H(S) - \sum_{v} \frac{|S_v|}{|S|} H(S_v) \tag{20}$$

where p_i is the class proportion, A is the candidate splitting feature, and S_v is the subset that satisfies the split condition v. Entropy is preferred over Gini impurity for this multi-class problem as it is more sensitive to low-probability class membership, which helps detect the rare *very low* intelligibility class after ADASYN oversampling.The ensemble vote is:

$$\hat{y} = \arg\max_{c} \sum_{t=1}^{T} \mathbb{I}(h_t(\mathbf{x}) = c) \tag{21}$$

Ensemble variance satisfies:

$$\text{Var}\big[\bar{h}(\mathbf{x})\big] = \rho\,\sigma^2 + \frac{1-\rho}{T}\,\sigma^2 \tag{22}$$

where ρ is the average pairwise Pearson correlation between individual tree predictions and σ^2 is the single-tree variance. As $T \to \infty$, the second term vanishes and the irreducible ensemble variance is governed solely by $\rho\sigma^2$. Random feature selection directly reduces ρ by preventing dominant BFCC features from appearing in every tree.

Post-hoc feature importance scores, computed as mean decrease in impurity, reveal that the first 10 BFCC coefficients and 3 Spectral Contrast bands carry the highest importance weights — empirical validation of the clinical hypotheses motivating our feature design.

AdaBoost: AdaBoost constructs a strong classifier through the sequential composition of $T = 100$ weak learners. Unlike bagging, boosting creates a directed chain of error correction, making it suited to refining boundaries between acoustically adjacent severity classes.

The implementation used the decision stumps (depth-one trees) as base learners. At iteration t, each training sample carries weight $w_i^{(t)}$ initialized uniformly at $1/N$. After fitting stump h_t, misclassified samples receive amplified weights:

$$w_i^{(t+1)} = \frac{w_i^{(t)}\exp(-\alpha_t\, y_i\, h_t(\mathbf{x}_i))}{Z_t} \tag{23}$$

where Z_t is a normalization constant. The confidence weight assigned to learner h_t is:

$$\epsilon_t = \sum_{i:\, h_t(\mathbf{x}_i)\neq y_i} w_i^{(t)}, \qquad \alpha_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t} \tag{24}$$

Note that $\alpha_t > 0$ only when $\epsilon_t < 0.5$ (stump performs better than chance). The final classifier is:

$$F(\mathbf{x}) = \text{sign}\left(\sum_{t=1}^{T}\alpha_t\, h_t(\mathbf{x})\right) \tag{25}$$

For the four-class problem ($K = 4$), system use *stagewise additive modeling using a multi-class exponential loss* (SAMME) extension:

$$\alpha_t^{\text{SAMME}} = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t} + \ln(K-1) \tag{26}$$

This correction ensures that a K-class random guesser receives zero weight. Early stumps typically split on features separating Healthy Controls from dysarthric classes; later stumps refine the boundary between low and medium intelligibility that the initial ensemble misclassifies.

Stochastic Gradient Descent **(SGD):** The SGD classifier is a general-purpose linear model that minimizes regularized empirical risk via first-order stochastic updates. It establishes the extent to which the 193-dimensional space is linearly separable across the four severity classes—a useful diagnostic for understanding feature space structure.

With logistic (log-loss) objective and L_2 regularization λ, the model parameters $\boldsymbol{\theta} \in \mathbb{R}^{K\times d}$ (one weight vector per class in decomposition of one-vs. rest) are updated per randomly selected training sample $(\mathbf{x}_i, y_i)$:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t \Big[\nabla_{\boldsymbol{\theta}}\, \ell\Big(y_i, f(\mathbf{x}_i; \boldsymbol{\theta}^{(t)})\Big) + \lambda\, \boldsymbol{\theta}^{(t)}\Big] \tag{27}$$

For the multi-class case, system used the minimize softmax cross-entropy loss:

$$\ell(y_i, \mathbf{f}) = -\log\left(\frac{\exp(f_{y_i})}{\sum_{c=1}^{K}\exp(f_c)}\right) \tag{28}$$

The learning rate follows an inverse scaling schedule:

$$\eta_t = \frac{\eta_0}{(1+\lambda\eta_0 t)^{\text{pow}}} \tag{29}$$

where η_0 is the initial learning rate and $\texttt{pow} = 0.5$, satisfying the Robbins-Monro convergence conditions. Training runs for a maximum of 1000 epochs with early stopping if loss improvement falls below 10^{-3} over five consecutive epochs. The stochastic gradient updates introduce beneficial noise into the optimization trajectory, preventing the model from settling in sharp minima that do not generalize to the test set. The strong SGD accuracy (94.63%) confirms that a substantial portion of inter-class separability in the hybrid feature space is captured by a linear projection.

LogitBoost: LogitBoost extends AdaBoost by replacing the exponential loss with a binomial log-likelihood objective. This matters for dysarthric severity classification because the class membership probabilities of adjacent severity levels are inherently overlapping and continuous rather than crisp. The binomial log-likelihood is more robust to outliers and provides probabilistic class membership estimates — a clinically useful property for communicating prediction confidence.

LogitBoost performs functional gradient descent in the space of additive logistic models. At each boosting stage m, working responses z_i, and sample weights w_i are derived from the current ensemble prediction $F^{(m)}(\mathbf{x}_i)$ via the logistic sigmoid:

$$p_i^{(m)} = \frac{1}{1+\exp(-2F^{(m)}(\mathbf{x}_i))} \tag{30}$$

$$z_i = \frac{y_i - p_i^{(m)}}{p_i^{(m)}(1-p_i^{(m)})}, \qquad w_i = p_i^{(m)}(1-p_i^{(m)}) \tag{31}$$

The working response z_i captures the normalized discrepancy between the predicted probability and the true label. Samples near $p_i = 0.5$ (uncertain boundary cases) receive large working responses, directing the next weak learner to focus on ambiguous severity boundaries. The weight w_i is maximal when the model is most uncertain and approaches zero for confident predictions—a natural self-regularization mechanism. A regression tree f_m is fitted to the weighted least-squares objective:

$$f_m = \arg\min_{f} \sum_{i=1}^{N} w_i (z_i - f(\mathbf{x}_i))^2 \tag{32}$$

and the ensemble is updated as $F^{(m+1)} = F^{(m)} + \nu f_m$ with shrinkage $\nu = 1.0$. The binomial log-likelihood minimized is:

$$\mathcal{L} = -\sum_{i=1}^{N}\Big[y_i \log p_i^{(m)} + (1-y_i)\log(1-p_i^{(m)})\Big] \tag{33}$$

Configured $T = 100$ boosting rounds in a multi-class one-vs.-rest scheme. The probabilistic base learner directly models posterior class membership, making LogitBoost suited for overlapping acoustic distributions between adjacent severity levels. Its competitive performance (97.52%) confirms that severity boundaries are probabilistic rather than crisp.

CatBoost: CatBoost introduces two algorithmic advances: *Ordered Boosting* (OB) and *symmetric DTs* (SDTs). These address the prediction shift bias in standard GB, where residuals are computed on the same data used to fit the corresponding tree—a form of data leakage that inflates training performance and reduces generalization. On the small TORGO dataset, this bias is acute.

OB: For each training sample $\mathbf{x}_i$, the residual target is computed from a model trained only on samples preceding it in a random permutation σ:

$$r_{\sigma,i} = y_i - F_{\sigma,i-1}(\mathbf{x}_i) \tag{34}$$

Because $\mathbf{x}_i$ never appears in the residual-computing model, the gradient estimates are unbiased. In practice, CatBoost maintains multiple permutations to obtain stable estimates.

SDTs: CatBoost grows oblivious (symmetric) trees where all nodes at a given depth share the same splitting criterion, producing a perfectly balanced binary tree with exactly 2^D leaves. Leaf values use a second-order Newton step:

$$v_\ell = -\frac{\sum_{i\in\ell} g_i}{\sum_{i\in\ell} h_i + \lambda} \tag{35}$$

where $g_i = \partial\ell/\partial\hat{y}_i$ and $h_i = \partial^2\ell/\partial\hat{y}_i^2$ are the first- and second-order gradients of the multi-class cross-entropy loss, and λ is the L_2 leaf regularization coefficient. The L_2 term prevents degenerate leaf values when a leaf contains very few samples—important for the rare *very low* intelligibility class even after ADASYN.

Configured CatBoost with $T = 500$ iterations, learning rate $\eta = 0.1$, and tree depth $D = 6$, giving $2^6 = 64$ leaves per tree. The MultiClass loss function uses a softmax output layer:

$$\hat{p}(c|\mathbf{x}) = \frac{\exp(F_c(\mathbf{x}))}{\sum_{c'=1}^{K}\exp(F_{c'}(\mathbf{x}))} \tag{36}$$

The SDTs constraint offers consistent regularization but may introduce a mild accuracy penalty: oblivious trees must evaluate the same feature at all nodes of a given depth, preventing the flexible asymmetric branching needed to simultaneously resolve formant centralization in BFCC bands and prosodic rigidity in Chroma features.

LightGBM: LightGBM, developed by *Microsoft Research*, reduces training cost through two complementary optimizations: *Gradient-based One-Side Sampling* (GOSS) and *Exclusive Feature Bundling* (EFB). These address the twin bottlenecks of evaluating all split candidates and managing high feature dimensionality.

GOSS: Standard GB weights all training samples equally. GOSS exploits the observation that samples with large gradient magnitudes (currently misclassified instances) carry more information for finding the optimal split than low-gradient samples (already well-fitted). GOSS retains all $|A| = an$ high-gradient instances and randomly samples a fraction b of the low-gradient instances

Table 3. Hyperparameters of all classification models used for the investigation

Model	Hyperparameters	Values
k-NN	Number of neighbors (k); Distance metric	$k = 3$; Euclidean
DT	Split criterion; Maximum depth; Minimum samples split	Gini; None; 2
SVM (RBF)	Kernel; Regularization parameter (C); Kernel coefficient (γ)	RBF; $C = 0.1$; $\gamma = 0.001$
PCA + k-NN	PCA variance retained; Number of neighbors (k)	95%; $k = 3$
RF	Number of trees; Split criterion	201; Entropy
AdaBoost	Number of estimators; Learning rate; Base estimator	100; 1.0; Decision Stump
SGD	Loss function; Regularization; Maximum iterations	Log-loss; L2; 1000
LogitBoost	Number of estimators; Learning rate; Loss function	100; 1.0; Logistic
CatBoost	Iterations; Learning rate; Tree depth; Loss function	500; 0.1; 6; MultiClass
LightGBM	Number of leaves; Learning rate; Number of estimators; Objective	31; 0.05; 500; Multiclass
XGBoost	Maximum depth; L_1 regularization (α); Objective function	4; 10; Softmax

($|B| = bn$), correcting for sampling bias via an amplification factor:

$$\tilde{g}_j(A, B) = \frac{1}{n}\left(\sum_{x_i \in A} g_i + \frac{1-a}{b}\sum_{x_i \in B} g_i\right) \tag{37}$$

The factor $(1-a)/b$ rescales the down-sampled subset to its original statistical weight, thereby keeping split-gain estimates unbiased. In the dysarthric severity classification, GOSS naturally allocates computational resources to boundary cases between adjacent severity levels.

EFB: The 193-dimensional vector contains mutual exclusivity structures. Chroma bins at opposing pitch classes rarely co-activate, and BFCC bands at very different frequency ranges rarely fire together. EFB identifies sets of mutually exclusive or near-exclusive features and bundles them into single compound features, reducing effective feature dimensionality without information loss.

LightGBM uses leaf-wise (best-first) tree growth, always splitting the leaf with the maximum gain regardless of depth. The gain at candidate split (d, v) is:

$$\text{Gain}(d, v) = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{38}$$

where G_L, G_R are summed first-order gradients, H_L, H_R are summed second-order Hessians, λ is L_2 leaf regularization, and to determine the minimum gain threshold for accepting a split. Leaf-wise growth achieves lower training loss per tree relative to level-wise approaches of the same leaf count. LightGBM configured with $T = 500$ trees, 31 leaves per tree, and learning rate $\eta = 0.05$. The small learning rate slows convergence but substantially improves generalization.

XGBoost: XGBoost formalizes the regularized learning objective using a second-order Taylor expansion of the loss function, enabling more accurate gradient estimates than first-order approaches. At boosting step t, XGBoost minimizes:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N}\left[g_i f_t(\mathbf{x}_i) + \tfrac{1}{2}h_i f_t^2(\mathbf{x}_i)\right] + \Omega(f_t) \tag{39}$$

where g_i and h_i are the per-sample first-order gradient and second-order Hessian of the Softmax cross-entropy loss. The Hessian as a sample-level weight improves the functional gradient step precision—samples with high curvature receive more conservative updates, preventing overfitting of boundary cases. The regularization term penalizes tree complexity:

$$\Omega(f_t) = \gamma T + \tfrac{1}{2}\lambda\|w\|^2 + \alpha\|w\|_1 \tag{40}$$

where T is the number of leaves, w is the leaf weight vector, λ is the L_2 ridge penalty, $\alpha = 10$ is the L_1 Lasso penalty, and γ is the minimum gain threshold for leaf splits. Minimizing over w_j gives the closed-form optimal leaf weight and per-leaf gain:

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \qquad \text{Gain}_j = -\frac{G_j^2}{2(H_j + \lambda)} \tag{41}$$

where $G_j = \sum_{i \in \text{leaf}_j} g_i$ and $H_j = \sum_{i \in \text{leaf}_j} h_i$.

Set `max_depth=4` and $\alpha = 10$. The large L_1 penalty drives weights of acoustically uninformative Mel-spectrogram bands and Chroma features—whose prosodic information is already captured by BFCC and Spectral Contrast—toward zero, enforcing automatic feature sparsity. This implicit feature selection is beneficial in medical datasets with correlated or redundant features, and it provides an implicit explanation of which acoustic dimensions drive severity predictions.

5. EXPERIMENTAL RESULTS AND ANALYSIS

k-NN recorded the lowest accuracy at 87.46%. The healthy control class achieves an R of 0.980, indicating that most normal speech samples are correctly identified. Performance drops substantially for the very low dysarthric class, which achieves a P of only 0.772, and R of 0.683, for an F of 0.725. Medium class reaches an F of 0.720, reflecting confusion between mild impairment and borderline healthy speech in the high-dimensional BFCC space. The macro-averaged F of 0.811 confirms that while the feature space encodes discriminative structure, an instance-based learner without dimensionality reduction is insufficient for reliable severity stratification. The performance of the traditional ML classifiers is shown in Table 4.

The DT outperformed *k*-NN, achieving 92.47% accuracy and a weighted F of 0.924. The healthy control achieves P of 0.956 and R of 0.965, and the model recovers substantially better on the dysarthric classes than *k*-NN. Very low reaches an F of 0.840 (from 0.725), and low reaches 0.875 (from 0.720). The improvement reflects the tree's ability to learn explicit non-linear thresholds on individual acoustic features and rules for low-frequency BFCC coefficients that separate severe dysarthric speakers from healthy controls, without needing to measure distances. Per-class, P and R are reasonably balanced, suggesting that the Gini criterion did not favor the majority class; ADASYN oversampling produced a well-balanced training set. The remaining performance gap relative to ensemble methods reflects the well-known high variance of a single unconstrained tree: it fits training partitions closely but generalizes less reliably to test samples from speakers with atypical pathological profiles. The SVM improved upon *k*-NN, achieving 92.23%

Table 4. Performance comparison of k-NN, DT, SVM, PCA + k-NN, and RF classifiers for dysarthria severity detection in terms of Precision (P), Recall (R), F1-score (F), Accuracy, Macro Average, and Weighted Average.

Class	*k*-NN			DT			SVM			PCA+*k*-NN			RF		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Controlled	0.893	0.980	0.935	0.956	0.965	0.961	0.950	0.962	0.956	0.954	0.996	0.974	0.977	0.999	0.988
Very low	0.772	0.683	0.725	0.846	0.835	0.840	0.840	0.827	0.833	0.950	0.866	0.906	0.992	0.951	0.971
Low	0.944	0.799	0.865	0.896	0.862	0.879	0.956	0.910	0.932	0.965	0.878	0.920	0.994	0.931	0.962
Medium	0.838	0.631	0.720	0.878	0.872	0.875	0.861	0.849	0.855	0.971	0.912	0.940	0.997	0.976	0.987
Macro Avg	0.862	0.773	0.811	0.894	0.884	0.889	0.902	0.887	0.894	0.960	0.913	0.935	0.990	0.964	0.977
Weighted Avg	0.871	0.875	0.869	0.924	0.925	0.924	0.922	0.922	0.922	0.956	0.956	0.955	0.984	0.983	0.983
Accuracy		87.46			92.47			92.23			95.60			98.33	

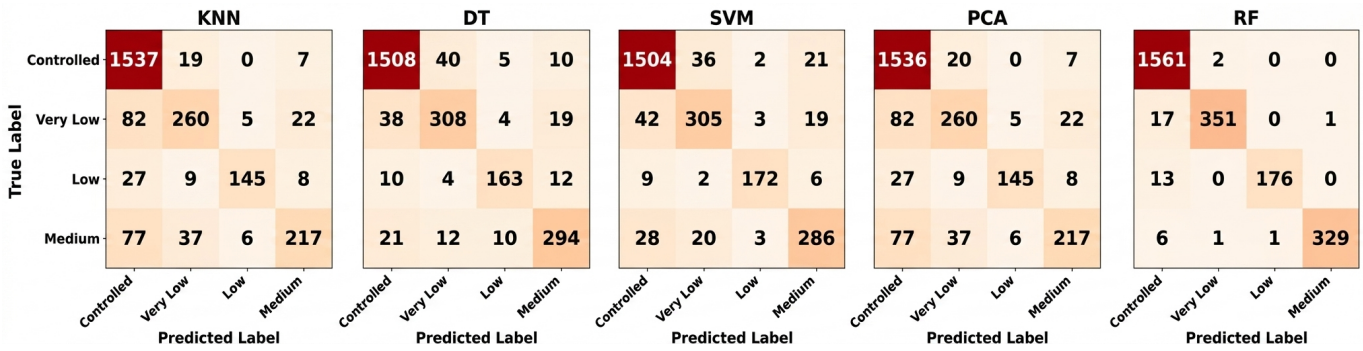

Fig. 2. Confusion matrices for ML-based classifiers using *k*-NN, DT, SVM, PCA + *k*-NN, and RF.

accuracy and a weighted AVG of 0.922, nearly matching the DT but via an entirely different inductive mechanism. The balanced P and R across severity classes, the equalized performance on very low (P (0.840), R(0.827)), and medium (P (0.861), R(0.849)), indicate the RBF kernel learned a stable non-linear decision boundary. The consistent margin-maximization provides measurable improvements over *k*-NN, especially for the acoustically overlapping low- and medium-intelligibility classes. PCA + *k*-NN achieved an accuracy of 95.60%, well above its raw *k*-NN counterpart. Reducing the feature space to its 95% variance-explained principal subspace restored meaningful Euclidean neighborhoods, particularly for very low and medium, whose Fs improved to 0.906 and 0.940, respectively. The healthy control reached an F of 0.974 and a near-perfect R of 0.996, confirming that the dominant variance axes closely correspond to the spectral energy characteristics separating pathological from normal speech. The remaining accuracy gap relative to ensemble methods reflects PCA's inability to preserve non-linear cross-feature interactions between BFCC lower-band components and spectral contrast. RF achieved the highest accuracy among traditional classifiers at 98.33%, with a weighted Avg. of 0.983. Per-class performance was uniformly high: healthy control reached P and R of 0.977 and 0.999; Medium achieved the highest class-level P at 0.997 with R of 0.976. The minority dysarthric classes also showed strong results: very low at $F = 0.971$ and low at $F = 0.962$. The bagging mechanism-201 trees on bootstrap resamples with 13 randomly selected features per split-effectively decorrelated ensemble predictions across the full 193-dimensional acoustic space, yielding stable and robust classification. The entropy split criterion proved particularly suited to the multi-class problem, sensitizing node splits to the rare, very low-intelligibility class.

From Fig. 2, the *k*-NN exhibited the most pronounced misclassification patterns. The confusion matrix shows 82 very low-intelligibility samples misclassified as healthy controls, and 77 medium-intelligibility samples similarly mislabeled. This systematic bias toward the majority class reflects distance concentration in the 193-dimensional BFCC space: the relative distances between the healthy control cluster and dysarthric severity clusters collapse, causing the nearest neighbors of borderline dysarthric samples to include a disproportionate number of healthy control instances. *k*-NN correctly identified 1,537 out of 1,563 healthy control samples but only 260 out of 369 very low intelligibility samples-a class-level R of 0.683, with direct clinical implications as a false negative risk for severely impaired speakers. The DT produced a distinctly cleaner confusion matrix compared to *k*-NN. The Confusion matrix shows that 1,508 healthy control samples are correctly classified, with only 40 misclassified as very low, 5 as low, and 10 as medium in terms of intelligibility. Very low intelligibility shows 308 correct predictions, with 38 misclassified as healthy control and 19 as medium intelligibility. Low intelligibility achieves 163 correct predictions, with 10 misclassified as healthy control, 4 as very low, and 12 as medium. Medium intelligibility shows 294 correct predictions, with 21 misclassified as healthy controls, 12 as very low, and 10 as low. The most striking pattern is the strong diagonal dominance; the tree learned decision rules on individual BFCC thresholds and Spectral Contrast values that sep-

Table 5. The performance comparison of SGD, AdaBoost, CatBoost, LogitBoost, XGBoost, and LightGBM classifiers in terms of P, R, F, and Accuracy. Macro Avg and weighted Avg scores are also reported to evaluate the models under class-imbalanced conditions.

Class	SGD			AdaBoost			CatBoost			LogitBoost			XGBoost			LightGBM		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Controlled	0.966	0.974	0.970	0.964	0.985	0.974	0.971	0.996	0.984	0.972	0.996	0.984	0.980	0.994	0.987	0.982	0.997	0.990
Very low	0.890	0.890	0.890	0.924	0.855	0.888	0.972	0.925	0.948	0.975	0.933	0.953	0.959	0.949	0.954	0.983	0.952	0.967
Low	0.928	0.889	0.908	0.994	0.952	0.973	0.994	0.937	0.965	0.983	0.942	0.962	1.000	0.931	0.964	0.989	0.968	0.979
Medium	0.925	0.909	0.917	0.921	0.930	0.926	0.972	0.939	0.955	0.987	0.945	0.966	0.970	0.955	0.963	0.972	0.951	0.961
Macro Avg	0.927	0.915	0.921	0.951	0.930	0.940	0.977	0.949	0.963	0.979	0.954	0.966	0.977	0.957	0.967	0.982	0.967	0.974
Weighted Avg	0.946	0.946	0.946	0.955	0.955	0.955	0.973	0.973	0.973	0.975	0.975	0.975	0.977	0.977	0.977	0.982	0.982	0.982
Accuracy	94.63			95.52			97.31			97.52			97.68			98.20		

arate classes with reasonable accuracy. The dominant error mode remains cross-class confusion with healthy controls (38 very low and 21 medium samples mislabeled as healthy controls), suggesting that axis-aligned single-feature splits cannot fully resolve the overlapping BFCC low-frequency profiles between severe dysarthric speakers and the healthy control cluster. Notably, cross-dysarthric confusions (very low vs. low, low vs. medium) are relatively low, indicating that within-dysarthric severity boundaries are reasonably well-resolved by single BFCC and spectral contrast thresholds. The high variance of a single unconstrained tree explains why accuracy (92.47%) remains below ensemble approaches; different bootstrap samples would produce different threshold rules, and no single tree generalizes optimally to all test speakers. The SVM substantially reduced majority-class bias by maximizing the kernel-induced margin. Its confusion matrix shows a more balanced error distribution: 42 very low- and 28 medium-risk samples misclassified as healthy controls, a 50% and 64% reduction, respectively, compared to *k*-NN. The RBF kernel's broad influence radius captured a smoother non-linear boundary between spectral energy profiles of healthy and dysarthric speech. The remaining misclassifications concentrate along adjacent class boundaries (very low vs. controlled and medium vs. controlled), consistent with acoustic overlap in prosodic rigidity between mild dysarthric speech and certain healthy speaker patterns in the Chroma and Tonnetz features. PCA + *k*-NN closely mirrors the *k*-NN confusion matrix structurally, 82 very low and 77 medium samples misclassified as controlled, indicating that the linear PCA projection retained the same dominant cluster structure. However, the overall accuracy improvement to 95.60% comes from substantially fewer within-dysarthric confusions: PCA + *k*-NN improved medium R (0.912 vs. 0.631 for *k*-NN), suggesting that the compact PCA subspace reduces intra-dysarthric overlap even though it does not address the controlled confusion for borderline severe dysarthric speakers. RF showed a dramatic reduction in misclassifications across all four classes. Only 17 very low errors are predicted as controlled (compared to 82 in *k*-NN), a fourfold reduction for the clinically most critical minority class. Medium misclassifications drop to just 6 samples mislabeled as controlled, and low shows 13 false positives against controlled. Cross-dysarthric confusions are near-zero: no very low samples are misclassified as low or medium, and only 1 low sample crosses into very low. This near-diagonal structure reflects the decorrelation of tree predictions under feature subsampling, where different trees specialize on different subsets of the 193-dimensional acoustic space and collectively learn fine-grained within-dysarthric severity boundaries that single learners cannot resolve. The confusion matrix analysis of traditional classifiers shows a clear progression from the distance-based confusion of *k*-NN to the robust ensemble separation of RF. A key clinical observation: misclassifications in lower-performing models are not random but systematically biased toward the healthy controlled class; a pattern with direct clinical implications, since false negatives (failing to detect dysarthric severity) carry greater diagnostic cost than false positives.

From the Table 5, the SGD and AdaBoost achieved accuracies of 94.63% and 95.52%, respectively. Both outperform simple distance-based learners but reveal distinct limitations. AdaBoost's re-weighting effectively concentrates later boosting rounds on the boundary between low and medium intelligibility; however, the relatively low R for very low (0.855) suggests that decision stump base learners lack the capacity to fully capture the spectral complexity of severe dysarthric distortions. SGD's competitive accuracy of 94.63% confirms that significant inter-class discrimination is embedded in linear projections of the 193-dimensional vector, while non-linear interactions between BFCC low-band energy and spectral contrast account for the residual gains in ensemble methods. Among the advanced gradient boosting frameworks, LightGBM achieved the highest overall accuracy of 98.20% with a weighted Avg. of 0.982. Very low attained an F of 0.967 (P: 0.983, R: 0.952)-a substantial improvement over both traditional classifiers and weaker boosting variants for this clinically critical minority class. XGBoost followed with 97.68% accuracy; its aggressive L_1 regularization produced a low P of 1.000, meaning every sample predicted as low intelligibility was a true positive, a noteworthy clinical property. LogitBoost attained 97.52% with its highest medium P of 0.987, consistent with its probabilistic working response formulation. CatBoost reached 97.31%, where ordered boosting reduced prediction shift bias on the limited TORGO corpus, though the symmetric tree constraint introduced a mild accuracy penalty. The results clearly demonstrate that ensemble and gradient boosting classifiers outperform traditional models for automated dysarthria severity stratification. The consistent improvement in Fs across all four severity classes as model complexity increases from *k*-NN to LightGBM confirms the non-linear acoustic structure of dysarthric speech and validates the proposed 193-dimensional BFCC-augmented feature representation. The confusion matrices for AdaBoost, SGD, LogitBoost, CatBoost, LightGBM, and XGBoost further characterize the relative diagnostic effectiveness of advanced classifiers as shown in Fig. 3. These models employ iterative learning strategies suited to non-linear acoustic variations in BFCC-based dysarthric features, and their confusion

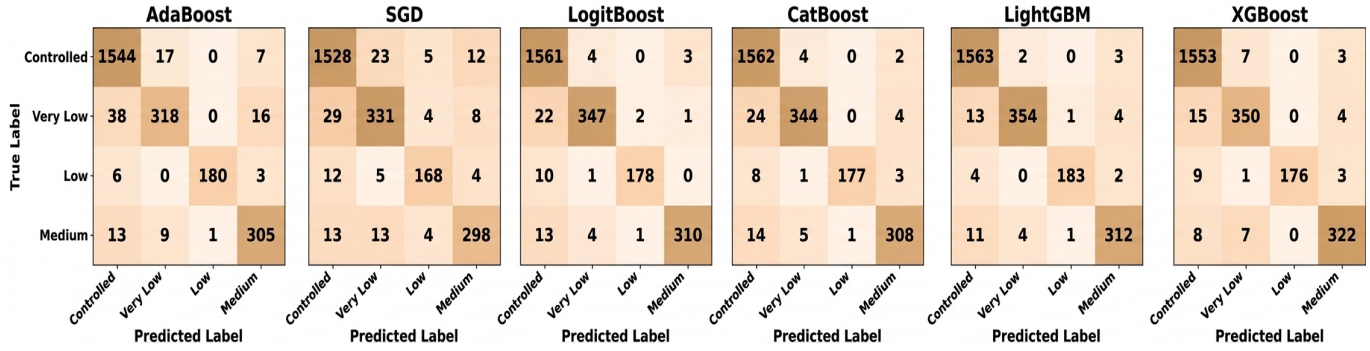

Fig. 3. Confusion matrices for advanced gradient boosting classifiers SGD, AdaBoost, CatBoost, LogitBoost, XGBoost, and LightGBM.

matrices reveal qualitative differences in error distribution. AdaBoost shows 38 very low-intelligibility samples misclassified as controlled and 13 medium-intelligibility samples mislabeled similarly. The very low R of 0.855 reflects the difficulty that decision stumps encounter when resolving subtle formant centralization in severely impaired speech: axis-aligned single-split decisions cannot simultaneously capture the correlated low-frequency BFCC distortions and spectral contrast reductions characteristic of very low intelligibility dysarthric speech. Additionally, 16 very low samples are misclassified as medium, suggesting that sequential re-weighting over-corrects toward the more numerous medium-severity samples. SGD presents the most dispersed confusion pattern, consistent with its role as a linear baseline. The matrix shows 23 controlled, 29 very low intelligibility, 12 low, and 13 medium intelligibility samples incurring cross-class errors across all severity boundaries. A relatively high false positive rate for controlled (12 control samples labeled as medium intelligibility) reflects the linear decision hyperplane's limitation in separating speakers with mild articulatory deviations, where Chroma and Tonnetz features exhibit overlapping distributions that only non-linear modeling can resolve. LogitBoost substantially outperformed AdaBoost, reducing very low misclassifications from 38 to 22 samples predicted as controlled. Its confusion matrix is notably cleaner within-dysarthric boundaries: only 2 very low samples cross into low, and only 1 low sample is misclassified as very low. This reflects LogitBoost's probabilistic logistic base learner, which models posterior class membership directly and provides smooth, calibrated confidence boundaries between overlapping severity levels. CatBoost's confusion matrix shows one of the most balanced error distributions. With only 24 very low and 14 medium misclassifications across all predicted categories, CatBoost demonstrates the effectiveness of ordered boosting in reducing prediction shift bias on the limited TORGO corpus. The near-zero cross-severity confusions confirm that symmetric tree regularization enforces consistent decision boundaries across all severity levels.

XGBoost delivered strong classification performance, with 15 very low and 8 medium misclassifications against the controlled category. The low P of 1.000, from aggressive L_1 regularization ($\alpha = 10$), which means every sample predicted as low intelligibility was a true positive: zero false positives for this category. The controlled medium false negative rate of 8 samples (vs. 11 in LightGBM) reflects XGBoost's second-order Hessian-based gradient estimation, which provides more conservative and stable updates for boundary cases between medium and controlled speech. LightGBM produced the cleanest confusion matrix among all evaluated classifiers. Only 13 very low intelligibility samples are misclassified as controlled, and 11 medium intelligibility samples are similarly mislabeled, the lowest very low and medium false negative counts across all models. Cross-severity confusions are minimal: only 1 very low sample is predicted as low, and only 1 low sample is predicted as medium. This near-diagonal structure reflects LightGBM's GOSS mechanism, which preferentially allocates resources to acoustically ambiguous boundary samples, and its leaf-wise asymmetric growth, which learns deeply specialized decision rules for each severity class.

As model complexity and regularization increase from AdaBoost to LightGBM, the false negative rate for the clinically critical very low intelligibility class decreases monotonically from 38 to 13 misclassifications. This steady reduction validates the proposed BFCC-augmented feature pipeline and confirms that advanced gradient boosting is the most appropriate choice for speech-based dysarthria severity screening when minimizing missed severe impairment cases is the primary clinical objective. The comprehensive evaluation of the ten machine learning models indicates that GB and bagging ensemble techniques significantly outperform traditional algorithms in classifying dysarthria severity. As illustrated in Fig. 4, RF achieved the highest classification accuracy of 98.33%, closely followed by LightGBM (98.20%) and XGBoost (97.68%). The superior performance of these ensemble models can be attributed to their ability to construct non-linear decision boundaries within the complex 193-dimensional hybrid feature space. Furthermore, the application of ADASYN was critical for comprehensive evaluation; by effectively mitigating the inherent class imbalance within the TORGO corpus, ADASYN prevented model bias toward majority classes and ensured high predictive reliability across all severity levels.

6. CONCLUSION

k-NN represents the lower bound at 87.46%, establishing the separability of the raw BFCC-Mel-Chroma feature space without parametric modeling. Its limitations on very low and medium reflect the fundamental inadequacy of Euclidean distance in high dimensions for severity discrimination. The DT (92.47%) and SVM (92.23%) provide comparable intermediate benchmarks: the DT through interpretable threshold rules on individual features, and the SVM through kernel-based margin maximization. Both address different aspects of the high-dimensionality problem but neither achieves the robustness of ensemble diversity. PCA + *k*-NN (95.60%) demonstrates that even a linear reduction of the feature space substantially

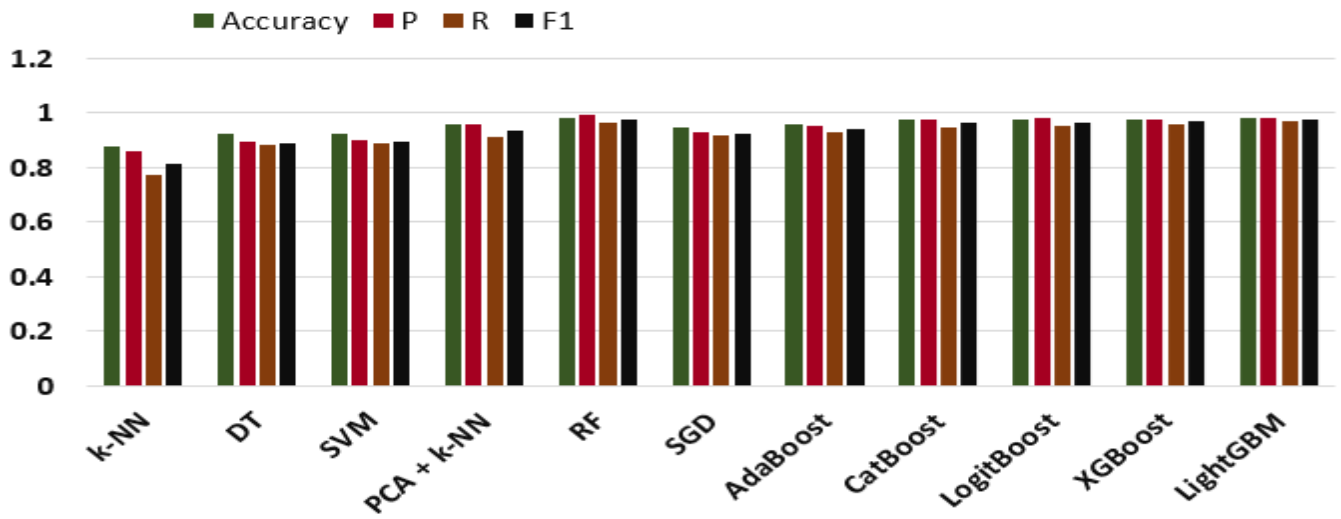

Fig. 4. Comparison of the performance of the evaluated using ML classifiers across four evaluation metrics: accuracy, P, R, and F1.

restores the geometric coherence needed for neighborhood-based classification. The ensemble and boosting classifiers form a second performance tier between 94.63% (SGD) and 98.33% (RF). This tier is characterized by consistent per-class Fs above 0.888 across all severity levels, a threshold ensuring reliable stratification even for the acoustically challenging minority classes. The gradient boosting hierarchy (LightGBM: 98.20% $>$ XGBoost: 97.68% $>$ LogitBoost: 97.52% $>$ CatBoost: 97.31%) reflects the progressive advantage of leaf-wise asymmetric growth and second-order optimization over simpler sequential boosting. RF's 98.33% accuracy is achieved without sequential error correction, and confirms that bagging-based variance reduction is highly competitive on static, time-averaged acoustic profiles when the feature space provides sufficient discriminative structure.

Compared to MFCC-based baselines in prior literature [19], which typically achieve 85–92% accuracy on similar severity classification tasks, our hybrid BFCC-augmented approach yields consistent improvements of 5 to 12 percentage points across all classifier categories. This validates the central hypothesis that the Bark scale's psychoacoustic alignment with critical bandwidths provides superior sensitivity to formant centralization and articulatory undershoot in dysarthric speech. The inclusion of spectral contrast and Tonnetz features further contributed to these gains by independently encoding aspects of breathiness and harmonic stability in dysarthric pathology that BFCCs alone cannot fully capture. From a clinical deployment perspective, the high R scores across all classes, particularly very low R values of 0.855 to 0.952 for ensemble models, also confirm that ADASYN oversampling successfully mitigated class imbalance without introducing systematic biases toward any severity boundary. This P-R balance is essential for a clinical screening tool, where the cost of a missed severe dysarthric case (false negative) substantially outweighs the cost of an over-referral (false positive).

7. REFERENCES

[1] Bassam Ali Al-Qatab and Mumtaz Begum Mustafa. Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features. *IEEE Access*, 9:18183–18194, 2021.

[2] Muhammad Ridha Anshari, Triando Hamonangan Saragih, Muliadi Muliadi, Dwi Kartini, Fatma Indriani, Hasri Akbar Awal Rozaq, and Oktay Yıldız. Performance comparison of adaboost, lightgbm, and catboost for parkinson's disease classification using adasyn balancing. *Jurnal Teknik Informatika (Jutif)*, 6(5):3495–3508, 2025.

[3] Tomas Arias-Vergara, Juan Camilo Vásquez-Correa, and Juan Rafael Orozco-Arroyave. Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. *Cognitive Computation*, 9(6):731–748, 2017.

[4] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kumar Kopparapu. Automatic assessment of dysarthria severity level using audio descriptors. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5070–5074. IEEE, 2017.

[5] Ramesh K Bhukya, SR Mahadeva Prasanna, and Biswajit Dev Sarma. Robust methods for text-dependent speaker verification. *Circuits, Systems, and Signal Processing*, 38(11):5253–5288, 2019.

[6] Ramesh K Bhukya, Aditya Raj, Anshul Kumar, et al. Asvspoof 2021: Detecting spoofed utterances through hybrid features. *APSIPA Transactions on Signal and Information Processing*, 14(1), 2025.

[7] Ramesh K Bhukya, Biswajit Dev Sarma, and SR Mahadeva Prasanna. End point detection using speech-specific knowledge for text-dependent speaker verification. *Circuits, Systems, and Signal Processing*, 37(12):5507–5539, 2018.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[9] Dae-Lim Choi, Bong-Wan Kim, Yong-Ju Lee, Yongnam Um, and Minhwa Chung. Design and creation of dysarthric speech database for development of qolt software technology. In

2011 International Conference on Speech Database and Assessments (Oriental COCOSDA), pages 47–50. IEEE, 2011.

[10] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[11] Camille Dingam, Xueying Zhang, Shufei Duan, Haifeng Li, and Xiaoyu Chen. Imbalanced data classification of pathological speech using pca, smote, and expectation maximization. In *International Conference in Communications, Signal Processing, and Systems*, pages 309–317. Springer, 2021.

[12] Joseph R Duffy et al. *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2012.

[13] Ashit Kumar Dutta and Abdul Rahaman Wahab Sait. A speech disorder detection model using ensemble learning approach. *Journal of Disability Research*, 3(3):20240026, 2024.

[14] Tiago H Falk, Wai-Yip Chan, and Fraser Shein. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, 54(5):622–631, 2012.

[15] Sowmiyalakshmi Ganesh, Thillai Chithambaram, Nadesh Ramu Krishnan, Durai Raj Vincent, Jayakumar Kaliappan, and Kathiravan Srinivasan. Exploring huntington's disease diagnosis via artificial intelligence models: a comprehensive review. *Diagnostics*, 13(23):3592, 2023.

[16] Christopher Harte, Mark Sandler, and Martin Gasser. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 21–26, 2006.

[17] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

[18] Joseph FY Hoh. Laryngeal muscles as highly specialized organs in airway protection, respiration and phonation. In *Handbook of behavioral neuroscience*, volume 19, pages 13–21. Elsevier, 2010.

[19] Amlu Anna Joshy and Rajeev Rajan. Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1147–1157, 2022.

[20] Banriskhem K Khonglah, Ramesh K Bhukya, and SR Mahadeva Prasanna. Processing degraded speech for text dependent speaker verification. *International Journal of Speech Technology*, 20(4):839–850, 2017.

[21] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon R Gunderson, Thomas S Huang, Kenneth L Watkin, Simone Frame, et al. Dysarthric speech database for universal access research. In *Interspeech*, volume 2008, pages 1741–1744, 2008.

[22] Martin Lauritzen, Jens Peter Dreier, Martin Fabricius, Jed A Hartings, Rudolf Graf, and Anthony John Strong. Clinical relevance of cortical spreading depression in neurological disorders: migraine, malignant stroke, subarachnoid and intracranial hemorrhage, and traumatic brain injury. *Journal of Cerebral Blood Flow & Metabolism*, 31(1):17–35, 2011.

[23] Ayoub Malek. Spafe: Simplified python audio features extraction. *Journal of Open Source Software*, 8(81):4739, 2023.

[24] David Martínez, Eduardo Lleida, Phil Green, Heidi Christensen, Alfonso Ortega, and Antonio Miguel. Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing (TACCESS)*, 6(3):1–21, 2015.

[25] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[26] Belinda Ndlovu, Kudakwashe Maguraushe, and Otis Mabikwa. A comparative analysis of machine learning techniques and explainable ai on voice biomarkers for effective parkinson's disease prediction. *Journal of Information Systems and Informatics*, 7(3):2196–2228, 2025.

[27] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, and Elmar Nöth. New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Lrec*, volume 14, pages 342–347, 2014.

[28] Milton Sarria Paja and Tiago H Falk. Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In *Thirteenth annual conference of the international speech communication association*, pages 62–65, 2012.

[29] B Shunmuga Priya, G Chitra, and R Ramalakshmi. Enhancing parkinson's disease diagnosis using smote and lightgbm on imbalanced speech dataset. In *2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)*, pages 1–6. IEEE, 2025.

[30] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46(4):523–541, 2012.

[31] Arman Hossain Shanto, Upoma Deb Suchi, Raida Jafar Chowdhury, Afnan Mohammed, and Fairuz Suhala Reza. *Machine learning-based approach to improving Parkinson's disease diagnosis through voice signal analysis*. PhD thesis, Brac University, 2025.

[32] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data augmentation using healthy speech for dysarthric speech recognition. In *Interspeech*, pages 471–475, 2018.

[33] Juan Camilo Vásquez-Correa, JR Orozco-Arroyave, Tobias Bocklet, and Elmar Noeth. Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of communication disorders*, 76:21–36, 2018.

[34] Eun Jung Yeo, Kwanghee Choi, Sunhee Kim, and Minhwa Chung. Cross-lingual dysarthria severity classification for english, korean, and tamil. In *2022 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)*, pages 566–574. IEEE, 2022.

[35] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013.