World Journal of Otorhinolaryngology - Head and Neck Surgery

WJO HNS World Journal of Otorhinolaryngology-Head and Neck Surgery

REVIEW ARTICLE OPEN ACCESS

Applications of Artificial Intelligence in Neurological Voice Disorders

Dongren Yao[1] | Aki Koivu[1] | Kristina Simonyan[1,2]

[1]Department of Otolaryngology-Head and Neck Surgery, Massachusetts Eye and Ear and Harvard Medical School, Boston, Massachusetts, USA | [2]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA

Correspondence: Kristina Simonyan (Kristina_Simonyan@meei.harvard.edu)

Received: 13 March 2025 | **Accepted:** 19 March 2025

Funding: This study was funded by the National Institute of Neurological Disorders and Stroke (R01NS088160), National Institute on Deafness and Other Communication Disorders (P50DC01990, R01DC011805).

Keywords: artificial intelligence | deep learning | neurological voice disorders | speech analysis

ABSTRACT

Neurological voice disorders, such as Parkinson's disease, laryngeal dystonia, and stroke-induced dysarthria, significantly impact speech production and communication. Traditional diagnostic methods rely on subjective assessment, whereas artificial intelligence (AI) offers objective, noninvasive, and scalable solutions for voice analysis. This review examines the applications, advancements, challenges, and future prospects of AI-driven methods in diagnosing, monitoring, and treating neurological voice disorders. We analyze recent advances in AI-based voice analysis, including machine learning, deep learning and signal processing techniques, and evaluate their effectiveness based on existing literature. AI models have demonstrated high accuracy in detecting subtle voice impairments, enabling early diagnosis of voice disorders, and predicting treatment response. Deep learning methods, particularly convolutional and transformer-based networks, have been effective in extracting meaningful biomarkers from acoustic or other modality data. Despite these promising advances, challenges remain, including limited high-quality data sets on some rare neurological voice disorders, ethical concerns regarding patient privacy, and the need for broad clinical validation. Further research should focus on developing standardized data sets, improving the ability of the AI model to learn representations, and enhancing its generalizability. With further development, AI-driven data analysis has the potential to transform the early detection and management of neurological voice disorders.

1 | Introduction

1.1 | Brief Overview of Neurological Voice Disorders

Neurological voice disorders arise from disruptions of the neural control of voice, setting them apart from structural and functional voice pathologies, such as vocal fold nodules or muscle tension dysphonia. They have a considerable impact on a significant proportion of the population, particularly among aging demographics where conditions such as Parkinson's disease (PD), dystonia, and voice tremor (VT) are more prevalent. For instance, laryngeal dystonia (LD) is estimated to affect approximately 1 per 100,000 individuals, with a higher prevalence among women and an average onset age of 40–50 years [1]. Dysarthria in PD affects up to 90% of patients, often worsening as the disease progresses [2]. VT, frequently associated with essential tremor, impacts approximately 20%–30% of patients, with prevalence increasing after the age of 60 [3]. These disorders significantly impair speech production, making it

Dongren Yao and Aki Koivu contributed equally to this study.

difficult for affected individuals to communicate in personal and professional contexts. The impact of these conditions on patients' lives is significant, often resulting in social withdrawal, diminished self-esteem, challenges in finding employment, and reduced quality of life. From an economic perspective, the recurrent need for clinical evaluations, ongoing treatments with botulinum toxin injections or speech therapy, and productivity losses due to vocal limitations impose a substantial economic burden on patients, caregivers, and healthcare systems.

As noted above, neurological voice disorders reflect disruptions in neural regulation of voice and speech control, including regions of the sensorimotor cortex, basal ganglia, and cerebellum, which coordinate the fine-tuned motor commands required for normal vocal fold vibration and tension. Central alterations in the neural network controlling voice and speech production may result in a wide range of clinical symptomatology. For instance, LD is characterized by intermittent vocal fold spasms, accompanied by a strained, strangled, or breathy voice quality. Rhythmic oscillations in pitch or intensity that compromise vocal stability are typical in VT, while voice and speech symptoms in PD are characterized by hypokinetic speech marked by reduced vocal intensity, monotone pitch, and imprecise articulation. Stroke-related dysarthria may lead to deficits in pronunciation, loudness, rate, or intonation during speaking.

The identification and quantification of the voice and speech characteristics in each of these disorders are critical for their diagnosis and targeted therapies. However, the overlapping symptoms that can appear across multiple conditions may require a laborious diagnostic process and often lead to a misdiagnosis due to the lack of objective markers. For the majority of neurological voice disorders, the current standard of the diagnostic work-up includes detailed multidisciplinary assessments involving laryngologists, neurologists, and speech-language pathologists [4]. Moreover, the diagnostic consensus between clinicians may depend on their specialized expertise and training. Even when an accurate diagnosis is achieved timely, current therapies frequently offer only temporary relief or show variable efficacy from one patient to another, further highlighting the lack of universally effective long-term solutions. These trends in the current clinical care of patients with neurological voice disorders underscore the pressing need for objective, scalable diagnostic methods and more durable, individualized treatment approaches for patients affected by neurological voice disorders.

1.2 | Rationale for AI in the Management of Neurological Voice Disorders

The present diagnostic paradigm for neurological voice disorders, which places considerable reliance on expert perceptual evaluations, is inherently subjective and subject to inter-rater variability [5]. This variability can result in inconsistencies and inaccuracies in diagnosis. Subtle acoustic characteristics, such as mild tremor, minor pitch fluctuations, or slight breathiness, are often overlooked, especially in the early stages of a disorder. Beyond diagnosis, treatment personalization remains a significant challenge. The selection and adjustment of individualized treatment plans, such as botulinum toxin dosing or tailored speech therapy, are largely based on the trial-and-error approach. These limitations emphasize the critical need for objective, data-driven methodologies to enhance both diagnostic precision and therapeutic personalization.

Artificial intelligence (AI)-driven solutions offer a variety of significant advantages. By leveraging machine learning (ML) algorithms to process large volumes of acoustic data, they can deliver standardized assessments that reduce subjective variability and capture subtle features, such as minor pitch fluctuations or subtle rhythmic tremor, often missed during acoustic perceptual evaluations. Predictive models built on these data may be powerful in forecasting individual patient trajectories and guiding therapy selection, transforming what is currently a trial-and-error process into one informed by evidence-based insights. Furthermore, AI systems facilitate continuous monitoring through tools like wearable sensors or smartphone applications, enabling clinicians to track voice changes in real time and adjust interventions promptly. This shift towards precision medicine, with its emphasis on patient-centeredness, is ultimately aimed at enhancing efficacy, improving clinical outcomes, and improving patients' quality of life.

AI-driven algorithms demonstrate proficiency in the processing and analysis of substantial data sets, encompassing acoustic signals, patient records, and longitudinal monitoring data [6, 7]. These systems have the capability to identify patterns and subtle anomalies, such as microvariations in pitch or spectral features, that are difficult or impossible to detect perceptually. By leveraging them, evaluations become more standardized and reproducible, thereby reducing the subjectivity and variability that is often present in traditional assessments. Additionally, these algorithms may be well-suited for integration into remote telemedicine platforms, allowing patients to record and submit voice samples from home or outside of the clinic. This not only expands access to care, especially for individuals in remote or underserved areas but also facilitates continuous monitoring and early intervention without the need for frequent in-person visits. These capabilities highlight the transformative potential of AI in creating a more efficient, equitable, and data-driven approach to the long-term management of neurological voice disorders.

ML models have the unique capability to integrate diverse data sources, including acoustic features, imaging data (e.g., nasolaryngoscopy, laryngeal videostroboscopy, or MRI of brain and vocal tract), patient demographics, and medical history [8, 9]. By synthesizing this multidimensional information, ML models may uncover relationships and trends that inform evidence-based recommendations. For example, they can identify correlations between specific acoustic markers and disease severity or progression, enhancing diagnostic accuracy and treatment planning [10]. The potential of these models is further augmented by predictive analytics, which employs historical data patterns to forecast disease progression or treatment responses. This capacity enables healthcare providers to anticipate changes in symptoms, implement proactive interventions, and adapt therapy to the evolving needs of the individual patient. Such personalized insights ensure that patients receive optimized care, improving outcomes while reducing unnecessary treatments and minimizing trial-and-error

approaches. These capabilities make ML an invaluable tool in advancing precision medicine for neurological voice disorders.

Furthermore, the integration of wearable devices, smartphone applications, and onsite recording systems augmented with AI analysis offers a transformative approach to the continuous monitoring of voice quality [11]. These technologies may enable patients to record and monitor their voices regularly and employ advanced AI algorithms that systematically analyze the data to detect early indications of deterioration, such as minor tremor, pitch irregularities, or vocal instability. By providing clinicians with early alerts, these technologies may facilitate timely intervention and improve outcomes. This, in turn, would facilitate earlier intervention, improve treatment outcomes, and allow for proactive adjustments to therapeutic strategies. Moreover, these AI tools may promote patient engagement by offering timely feedback, personalized progress reports, and tailored voice exercises that patients can perform independently. By empowering individuals to take an active role in their voice care, these technologies would ultimately enhance adherence to treatment plans and improve overall health outcomes. Thus, the integration of continuous tracking, early detection, and patient-centered feedback systems positions AI as a transformative tool in advancing both clinical practice and patient autonomy in the management of neurological voice disorders.

In this review, we explore advancements in AI for diagnosing and treating neurological voice disorders. We begin by outlining the fundamental aspects of voice and speech analysis. Next, we summarize common AI techniques that can be applied to neurological voice disorder research. After this, we mention data set collection and quality control. We then present our PRISMA-based review findings, highlighting general trends and providing an in-depth analysis of the individual methods identified. Finally, we discuss the validation, regulatory, and ethical considerations of deploying these complex models in clinical practice, as well as the challenges that must be addressed for future advancements.

2 | Fundamentals of Speech and Voice Analysis

Speech production is a complex process involving the coordinated interaction of the respiratory, vocal fold, and articulatory muscles. Sound is produced by airflow from the lungs, which passes through the vocal folds, causing them to vibrate and thereby forming the fundamental frequency (F_0), or pitch, of the voice. Subsequent to this initial production of sound, the acoustic properties are refined through resonances in the vocal tract, which encompasses the pharynx, oral cavity, and nasal cavity. These resonances contribute to the creation of the distinctive spectral qualities that characterize spoken language. The articulation process, involving the tongue, lips, and jaw, refines the sound into distinct phonemes and words. The underlying neurological mechanisms involve the coordination of muscle movement by the brain and peripheral nerves, with parameters such as pitch, intensity, and timing being adjusted [12]. In the context of neurological voice disorders, disruptions in these neural pathways can result in altered or unstable phonation, highlighting the critical role that precise neuromuscular coordination plays in healthy speech production.

Speech waveform is defined as time-varying pressure signals, generated by the movement of air through the vocal tract during phonation. This signal carries rich acoustic information that reflects the physiological and neurological processes underlying voice production. Among the fundamental parameters, F_0 is a key measure representing the rate of vocal fold vibration, perceived as the pitch of the voice. Variations in F_0 can indicate normal intonation patterns or disruptions, such as monotone speech in neurological conditions like Parkinsonian dysarthria [13]. Intensity, defined as the amplitude of the waveform, corresponds to the loudness of the voice and reflects vocal effort and respiratory support. A reduction in intensity may be indicative of a weakened respiratory drive or incomplete vocal fold closure, which are common features of various voice disorders. Another critical parameter is duration, which captures the temporal aspects of speech sounds, including their length and timing. Variations in duration can highlight prosodic abnormalities or impaired speech rhythm. Because of these reasons, sustained vowel tasks are frequently used for data collection.

2.1 | Common Acoustic Features in Voice Analysis

Time-domain features are essential for analyzing phonatory behavior and vocal stability due to their ability to capture temporal variations in the speech signal. *Jitter*, a measure of cycle-to-cycle variations in the F_0, reflects irregularities in vocal fold vibration. Elevated jitter has been associated with neurological conditions such as PD, indicating impaired motor control of the vocal tract muscles [14]. Similarly, *shimmer* quantifies the cycle-to-cycle amplitude variability, revealing instability in vocal intensity [15]. Increased shimmer may indicate vocal fold lesions or neuromuscular dysfunction affecting the closure and tension of the vocal folds. Voice onset time (VOT), defined as the interval between the release of a stop consonant (e.g., /p/, /t/) and the onset of vocal fold vibration, provides insights into the timing and coordination of phonatory and articulatory systems. Abnormal VOT patterns, such as prolonged or inconsistent intervals, can be indicative of neuromuscular control disruptions as seen in stroke-related dysarthrias [16].

Frequency-domain features analyze the spectral characteristics of the speech signal, offering insights into the resonances and vibratory patterns of the vocal tract. Formants are defined as resonant frequencies that are produced by the shape and configuration of the vocal tract during speech. Each formant is associated with a particular articulatory position. The initial two formants, designated as F_1 and F_2, play a pivotal role in the characterization of vowels, while higher formants contribute to the clarity and richness of speech. Deviations in formant frequencies or bandwidths may indicate structural changes or neurological deficits that alter vocal tract resonance. Harmonics, generated by the periodic vibration of the vocal folds, represent the integer multiples of the F_0 and contribute to the tonal quality of the voice. The harmonic-to-noise ratio (HNR) has been identified as a significant metric for evaluating voice quality, as it quantifies the proportion of harmonic components relative to noise in the signal. A reduced HNR, reflecting increased noise in the signal, often correlates with pathologies

such as incomplete vocal fold closure, vocal tremor, or neurological conditions affecting phonation. Cepstral peak prominence (CPP) is also a feasible metric for voice quality, where the frequency-domain information is transformed to the cepstral domain with a log transformation and an inverse Fourier transform (IFT). The result is a time-like domain where the periodicity of the signal becomes easier to detect. The cepstral peak is found, representing the F_0, and its height is compared to the average noise background surrounding it. CPP is another metric designed to measure vocal stability, and it has been linked to impaired motor control of speech muscles [17].

Compared to time or frequency domain features, Mel-frequency cepstral coefficients (MFCCs) extend the scope of basic spectral analysis by incorporating perceptual aspects of human hearing. MFCCs represent the most widely used features in speech processing, derived from the speech signal's spectral characteristics. MFCCs transform the speech spectrum into a compact, low-dimensional representation that captures the most critical aspects of vocal tract resonances. The process begins by dividing the speech signal into short time frames, followed by the application of a Fourier Transform to obtain the spectrum. The spectrum is then mapped onto a Mel scale, which models the human ear's nonlinear perception of pitch. The subsequent application of a logarithmic transformation and a discrete cosine transform (DCT) to the Mel-scaled spectrum results in the computation of MFCCs as coefficients that efficiently summarize the spectral envelope. MFCCs have been shown to be particularly effective for voice analysis due to their ability to provide a concise yet highly descriptive representation of vocal tract dynamics, focusing on the frequencies most relevant to human perception. This makes them robust to variations in recording conditions and speaker characteristics. Their compactness reduces computational complexity, while their capacity to capture vocal tract shape information makes them standard inputs to many AI-based voice processing models. These models utilize MFCCs to perform tasks such as voice disorder detection, speaker recognition, and phoneme classification with high accuracy, making MFCCs indispensable in modern speech analysis workflows. The three common types of features in voice data are presented and compared in Table 1.

2.2 | Signal Processing and Enhancement Techniques

Raw voice recordings often contain various distortions and artifacts that can obscure the true characteristics of a patient's voice. Common issues include background noise, such as ambient environmental sounds or electrical hums, which can mask subtle voice features critical for diagnosis. Reverberations, caused by sound reflecting off surfaces in the recording environment, can blur the clarity of the voice signal, distorting the temporal and spectral characteristics necessary for accurate analysis. Channel distortions may arise from differences in recording equipment or microphone quality, which can introduce additional noise or alter the frequency response of the recorded signal.

Given these challenges, robust preprocessing is essential to ensure that the extracted acoustic features genuinely reflect the patient's vocal characteristics, rather than being influenced by these external factors. The elimination of noise, the normalization of volume levels, and the filtration of undesirable frequencies are pivotal steps in preprocessing, as they enhance the clarity and reliability of the voice signal, thereby facilitating more precise analysis. This step is crucial for training AI models, as it minimizes the risk of erroneous conclusions drawn from distorted data and ensures that any observed variations in voice quality are truly indicative of the patient's condition. A foundational technique employed in this process is *normalization*, which involves adjusting the amplitude of audio signals to a consistent range. This technique ensures uniformity across recordings, thereby minimizing the impact of variations caused by different recording equipment, environments, or speaker loudness. By standardizing the signal's amplitude, normalization preserves the underlying acoustic features while reducing variability that could affect model performance. Another essential step is *silence removal*, which aims to eliminate non-speech segments from the audio. Silence or background noise has the potential to diminish the significance of extracted features and compromise the efficacy of AI models. Voice activity detection algorithms are commonly used for this purpose, identifying voiced regions based on energy levels or spectral characteristics [18]. By retaining only meaningful speech segments, silence removal ensures that the analysis is focused on the most informative parts of the signal. *Filtering* further enhances signal quality by reducing noise and eliminating unwanted frequencies that might obscure important acoustic features. *Spectral subtraction*, for instance, involves estimating and subtracting the noise's spectral profile, effectively suppressing unwanted background sounds. Other techniques, such as low-pass and high-pass filters, remove frequencies outside the typical speech range (below 50 Hz and above 8 kHz), while notch filters target specific noise sources, such as electrical hum at 50 or 60 Hz. These filtering techniques are critical for ensuring that the signal retains its core speech characteristics without interference from extraneous noise.

TABLE 1 | Differences between three common types of features in voice data.

Feature type	Derived from	Dimensionality	Focus	Applications
Time domain	Raw waveform	Low	Vocal fold stability	Voice pathology detection
Frequency domain	Spectral analysis	Medium-high	Resonance and tonal structure	Articulation and resonance
MFCCs	Perceptual spectrum	Low	Spectral envelope and perception	AI-based speech processing

Abbreviations: AI, artificial intelligence; MFCCs, Mel-frequency cepstral coefficients.

In addition to traditional filtering methods, an increasing number of people are resorting to AI algorithms for the purpose of noise reduction in the context of speech processing. Specifically, *denoizing autoencoders* learn to map noisy speech to clean counterparts by training neural networks on paired data sets, effectively handling complex noise patterns [19]. Similarly, generative adversarial networks (GANs) use a generator-discriminator framework to generate high-quality, natural-sounding speech, excelling in nonstationary noise environments [20]. Convolutional neural networks (CNNs) have been employed for their ability to identify spatial patterns [21], while recurrent neural networks (RNNs) have been utilized for their capacity to discern temporal dependencies [22]. These advanced deep-learning approaches have been widely adopted for speech enhancement tasks. Advanced transformer models are also being explored for their ability to capture both short- and long-term dependencies in noisy speech [23]. Additionally, end-to-end models like *Wave-U-Net* process raw waveforms directly, bypassing traditional time-frequency representations for high-fidelity restoration [24]. These sophisticated methods can complement traditional techniques to refine noise reduction further, ensuring that extracted acoustic features reflect the true voice characteristics.

Evaluating the quality of denoised speech is essential to ensure that noise reduction techniques effectively enhance the signal while preserving its intelligibility and naturalness. Several metrics are commonly used for this purpose, each addressing different aspects of speech quality. The Perceptual Evaluation of Speech Quality (PESQ) is a widely used objective metric that simulates human auditory perception to compare denoized speech with a reference clean signal. PESQ considers time–frequency distortions and alignment, providing a score ranging from −0.5 to 4.5, where higher values indicate better quality. This metric is particularly useful in telecommunication and audio enhancement research for assessing overall speech quality. Another crucial metric is the Short-Time Objective Intelligibility (STOI), which focuses on measuring how easily denoised speech can be understood. STOI evaluates the similarity between the short-time spectral envelopes of the clean and denoised signals, providing a score between 0 and 1, with higher values reflecting better intelligibility. This metric is widely employed in speech enhancement applications, such as hearing aids, where intelligibility is a primary concern. SNR can be used in this context as a crucial metric that quantifies the ratio of the desired speech signal power to the residual noise power.

Apart from noise reduction, data augmentation is a prevalent preprocessing step in training voice models. It assists in expanding data sets and enhancing the generalization of AI models, especially in scenarios with limited number of training observations. By artificially modifying existing data, augmentation introduces acceptable variability that simulates real-world conditions, enhancing model robustness. *Pitch shifting*, for instance, involves altering the F_0 of speech while preserving its temporal structure. Raising or lowering the pitch mimics variations associated with factors such as age, gender, or emotional states, enabling models to adapt to diverse speaker characteristics. *Time-stretching*, another augmentation technique, involves altering the duration of speech without changing its pitch. This modality simulates different speaking rates or tempo variations. This technique helps models handle both fast and slow speech patterns, which are often observed in pathological or emotional speech. Adding simulated noise is another powerful approach, replicating real-world conditions such as background chatter, wind, or electronic hum. By mixing clean speech with controlled noise levels, models are able to discern between the voice signal and irrelevant sounds, improving performance in noisy environments. Noise can be tailored to specific applications, such as clinical environments or outdoor settings, to further enhance robustness.

In natural settings, audio volume is rarely consistent. This is why warping the frequencies of the audio can also be used as an augmentation method. Vocal tract length perturbation (VTLP) is a method that can be used for this purpose. It simulates variations in the vocal tract length across different speakers, and it is based on warping the frequency domain with a piecewise linear warping function, where the warping parameters are chosen at random [25]. Combined with those mentioned preprocessing and these augmentation strategies, they contribute significantly to building reliable and adaptable AI systems for disordered voice analysis.

2.3 | Time–Frequency Representations

Time–frequency representations play a pivotal role in speech analysis, as they facilitate the decomposition of audio signals into their temporal and spectral components, thereby providing a detailed perspective on the evolution of frequencies over time. These representations are particularly crucial for ML models, which depend on rich and structured inputs to extract meaningful features. A prominent technique employed in this field is the short-time Fourier transform (STFT), which divides the speech signal into overlapping time frames and calculates the Fourier transform for each segment. This process yields a matrix of spectral data, which offers insight into the temporal variation of the signal's frequency content. While STFT provides high-resolution information, it exhibits a fixed trade-off between time and frequency resolution based on the window size, which can limit its flexibility. This is why both resolutions should be parameterized with respect to the learning task.

A spectrogram is a visual representation of the STFT output, in which the intensity of the frequencies is displayed as a color or grayscale image. Spectrograms serve as powerful inputs for the following model, transforming the raw waveform into a structured format that emphasizes temporal and spectral patterns. These patterns are often indicative of specific speech characteristics, making spectrograms an essential tool for voice disorder analysis, speech recognition, and other tasks. Mel-spectrograms are a subsequent enhancement of spectrograms, mapping the frequency axis onto a Mel scale to emulate the human auditory system's nonlinear perception of pitch. This transformation emphasizes frequencies most relevant to human hearing while compressing higher frequencies, providing a more perceptually relevant representation. The compactness and alignment with the auditory perception of Mel-spectrograms make them a preferred choice for AI models, as they focus on features that are both meaningful and computationally efficient to process.

The employment of time–frequency representations, including STFT, spectrograms, and Mel-spectrograms, enables AI models to effectively capture the temporal dynamics and spectral characteristics of speech. These representations provide a comprehensive input space, thereby enhancing the capacity of neural networks to learn intricate patterns. Consequently, they are considered essential components in contemporary speech and voice analysis workflows.

Since time–frequency representations encode spectral information over multiple time frames and frequency bins, they would suffer from high-dimensional issues, particularly when used as inputs for ML models. To address this challenge, dimensionality reduction techniques have been employed, including feature selection, which preserves the most significant variables, and feature projection, which generates new variables by combining the original ones. For feature selection algorithms, they can be categorized into embedding methods, which integrate feature selection during model training phase, like *Lasso regularization* and *Random Forests*. Statistical-based measurement filters, such as *chi-squared tests* and *Pearson's correlation*, assess the relationship between each feature and target variable independently of the ML model. Wrappers, for instance, Recursive Feature Elimination (RFE), calculate different feature subsets to find the most effective combination. For feature projection techniques, including manifold learning (e.g., t-SNE, UMAP), linear discriminant analysis (LDA), and principal component analysis (PCA), transform the data into low-dimensional spaces while preserving essential structures within certain constraints. These techniques are designed to streamline the feature set, enhancing computational efficiency, and mitigating the risk of overfitting. This ensures that the retained features provide a robust representation of the voice signal. This optimization is critical for building effective models in speech and voice disorder analysis.

3 | Overview of AI Techniques With AI-Based Voice Studies

3.1 | ML and Statistical Modeling in Voice Studies

ML and statistical modeling form the backbone of many voice analysis systems, providing mechanisms to identify, categorize, and interpret vocal data effectively. Classical ML algorithms, such as support vector machine (SVM), Random Forests (RF), and Gaussian mixture model (GMM), have been widely adopted due to their interpretability and relatively low computational demands. SVM has demonstrated efficacy in discriminative tasks, such as differentiating between healthy and pathological voices, by creating decision boundaries that maximize class separability [26]. RF and Decision Tree (DT), on the other hand, are ensemble methods that offer enhanced robustness and the ability to rank the importance of features, rendering them ideal for feature-based voice studies [27]. GMM has proven effective in tasks like speaker verification, modeling probability distributions of vocal characteristics to distinguish individuals with high accuracy [28].

Statistical approaches complement ML methods by providing tools for predictive modeling and uncertainty quantification. Regression models are commonly used to predict voice quality metrics or assess the progression of vocal conditions [29, 30]. Bayesian frameworks have been shown to be particularly useful in clinical settings to quantify uncertainty and provide probabilistic interpretations of predictions [31]. This capability is critical for making reliable decisions in high-stakes applications, such as determining treatment plans for neurological voice disorders. While these classical methods have been instrumental in advancing voice analysis, they are often limited by their reliance on predefined features, which may overlook subtle patterns and interactions in the data.

3.2 | Deep Learning Architecture for Voice Analysis

Deep learning (DL) has brought significant advancements to voice analysis, offering the ability to learn complex patterns directly from raw data. In contrast to conventional methods that depend on handcrafted features, DL models automatically extract hierarchical representations, enabling a more nuanced understanding of voice signals. Among these architectures, CNNs are particularly effective for analyzing spectrograms and Mel-spectrograms, which represent speech in the time–frequency domain. CNNs demonstrate a high degree of proficiency in capturing spatial patterns, such as formant transitions and spectral peaks, making them ideal for tasks like voice pathology detection, emotion recognition, and prosody analysis. For example, CNNs have been used to identify spectral irregularities in pathological voices, thereby providing clinicians with diagnostic insights [32].

RNNs, including long short-term memory (LSTM) networks, gated recurrent units (GRUs), and variations based on these models, are particularly suited for capturing the temporal dynamics of speech. These models have proven to be highly effective in the context of sequential data, facilitating applications such as continuous monitoring of vocal health and speech synthesis. Specifically, RNNs can model how vocal characteristics change over time, which is crucial for detecting conditions like VT or tracking the progression of neurological disorders such as PD [33]. The ability of RNNs to retain long-term dependencies makes them valuable for understanding speech patterns in both clinical and nonclinical settings.

The emergence of Transformer models has further advanced the DL adoption in voice studies. Unlike RNNs, transformers use self-attention mechanisms to capture long-range dependencies more efficiently, making them suitable for processing large and complex data sets. Pre-trained models, such as Wav2Vec 2.0, HuBERT, and Whisper, leverage self-supervised learning to extract meaningful representations from unlabeled speech data [34–36]. These models can be fine-tuned for specific tasks, including voice pathology detection, speaker verification, and emotion recognition. Wav2Vec 2.0, for example, demonstrates a particular aptitude for identifying subtle acoustic anomalies in pathological voices by capturing both temporal and spectral dependencies [37], while HuBERT focuses on harmonic and prosodic features, making it highly effective for emotion and prosody analysis [38].

Graph convolutional networks (GCNs) are emerging as powerful tools in voice analysis, particularly for tasks like speech emotion recognition. GCNs demonstrate a high degree of proficiency in the modeling of non-Euclidean data structures, such as relationships between different acoustic features or temporal frames. For example, studies have proposed GCN-based architectures that integrate temporal and spatial features, leveraging adaptive graph structures to model interrelationships within speech data [39]. Additionally, approaches combining GCNs with attention mechanisms, such as skip graph convolutional networks (SkipGCNs) or graph attention networks (GATs), have achieved state-of-the-art results by dynamically focusing on relevant local and global interactions. These methods demonstrate the potential of GCNs to enhance voice analysis tasks by capturing complex dependencies within speech signals.

Autoencoders are another key architecture in DL for voice analysis, particularly for tasks like noise suppression and anomaly detection. Denoising autoencoders learn to reconstruct clean speech signals from noisy inputs, offering robust solutions for speech enhancement. Variational autoencoders (VAEs) are particularly useful for generating synthetic voice samples and identifying rare anomalies in voice data, making them valuable for augmenting data sets and supporting research in rare voice disorders.

GANs have also emerged as a powerful tool in voice studies. By leveraging a generator-discriminator framework, GANs can produce high-quality synthetic speech and enhance noisy signals. Models like speech enhancement GAN (SEGAN) have demonstrated exceptional performance in improving speech clarity, even in challenging acoustic environments [40]. Additionally, GANs are being explored for voice synthesis applications, enabling the generation of realistic voice samples for training and testing AI models [41].

Recently, the integration of large language models (LLMs), such as Whisper, GPT, and BERT-based speech models, has opened new possibilities in multimodal analysis. These models combine acoustic and textual data, providing a holistic approach to voice studies. For example, Whisper, developed by OpenAI, integrates speech recognition and audio transcription capabilities, enabling detailed analysis of both spoken content and vocal characteristics. When paired with CNNs or other transformers, LLMs offer a comprehensive framework for tasks, like voice disorder detection, emotion analysis, and speaker profiling. Furthermore, hybrid systems that combine traditional DL architectures with LLMs are particularly promising [42]. For instance, features extracted from CNNs, or transformers can be augmented with textual insights from LLMs, enhancing the model's ability to capture both acoustic and semantic nuances. Multimodal approaches, integrating voice data with other modalities, such as laryngeal imaging or electroencephalography (EEG) signals, are further pushing the boundaries of what DL can achieve in voice studies [43, 44].

As DL architectures continue to evolve, their applications in voice analysis are becoming increasingly sophisticated. These technologies are transforming the field, from diagnosing voice disorders to enhancing speech quality and enabling real-time monitoring. Future developments in this field are likely to focus on the integration of DL with edge computing for on-device analysis, enabling real-time applications in wearable devices and telemedicine, and advancing the personalization of AI models to better cater to individual patient needs. These advancements underscore the transformative potential of DL in voice studies and its critical role in the future of vocal health and communication technologies.

4 | Data Collection

4.1 | Importance of High-Quality Data

The emergence of LLMs, such as GPT and Whisper, has revolutionized speech and language processing by leveraging vast data sets and powerful architectures to perform generalized tasks. These models can transcribe, classify, and analyze speech with impressive accuracy, showcasing their potential for a wide range of applications. However, training robust AI-based voice models for specific applications like voice disorder detection still requires accessible, representative, and well-annotated data sets. LLMs trained on extensive collections of general-purpose text or audio may lack the requisite granularity and domain-specific insights needed to recognize subtle acoustic patterns unique to voice disorders. High-quality data sets that capture diverse voice samples, including those with specific pathologies, are essential to fine-tuning or adapting these generalized models for clinical tasks. Furthermore, ensuring the accuracy and representativeness of these data sets is crucial for ensuring the reliability of models across different demographics, recording environments, and levels of disorder severity. This, in turn, reduces potential biases and enhances the fairness of AI-based systems.

The challenges of collecting voice disorder samples remain significant. Rare conditions, such as LD or ALS-related dysarthria, inherently limit the availability of samples, making it difficult to amass a large and balanced data set. The lack of data collection standardization complicates the accumulation of good-quality data. Ethical and privacy constraints further complicate the situation, as speech data frequently contains personally identifiable information, necessitating strict compliance with data protection regulations. Additionally, the annotation of these data sets requires clinical expertise to ensure the precision of labels denoting disorder type, severity, and accompanying symptoms. This process is resource-intensive, requiring collaboration with experienced clinicians and access to specialized diagnostic tools.

These challenges highlight the persistent necessity for targeted data collection efforts, even in the era of LLMs. Initiatives, such as multi-center data sharing, federated learning frameworks, and synthetic data generation, have the potential to overcome these barriers, facilitating the creation of high-quality data sets tailored for voice disorder analysis and advancing the clinical applicability of AI-based voice models.

4.2 | Speech Corpora and Voice Analysis

The development of robust AI-based voice models heavily relies on high-quality data sets, and several well-known databases have been established to support pathological speech analysis. For

instance, the Massachusetts Eye and Ear Infirmary (MEEI) Voice Disorders Database is one of the earliest and most frequently cited collections, featuring a variety of pathological voice samples with detailed acoustic features and corresponding diagnoses. Another prominent resource is the Saarbrüecken Voice Database (SVD), which includes recordings from individuals with various voice disorders, annotated with demographic information, for example, age and gender, as well as clinical diagnoses. The TORGO Data set, which contains speech and articulatory data on seven subjects with either cerebral palsy or ALS, is another notable resource. Similarly, the Parkinson's Voice Initiative Data set offers voice samples from PD patients, emphasizing early-stage detection and progression monitoring. Recently, Bridge2AI Voice Consortium has released their V1.0 data set of voice as a biomarker of health, which contains a neurological cohort of 31 cases of PD and 25 cases of Alzheimer's, dementia, or mild cognitive impairment [45].

When selecting a speech corpus, several criteria must be considered to ensure the data set is representative and suitable for training AI models. Diversity in demographics such as age, gender, and severity levels of disorders is crucial to avoid biases and improve model generalization. The recording conditions, including the type of equipment used and the environment in which the samples were collected, also play a significant role in ensuring data consistency. Furthermore, the incorporation of high-quality ground truth labels, such as confirmed diagnoses, severity ratings, and acoustic measurements, is imperative. These labels serve as the foundation for supervised learning and model evaluation.

The importance of expert annotation in establishing gold-standard references cannot be overstated. Clinical experts use standardized protocols and diagnostic tools to ensure that the labels accurately reflect the pathology being studied. Partnerships with physicians, voice and speech clinics, and research institutes frequently result in the generation of proprietary data sets comprising meticulously labeled samples. These collaborations not only expand the availability of diverse data sets but also enhance the quality of annotations through access to specialized expertise and equipment. Annotation consistency is a critical factor in building reliable corpora. Rating scales such as the Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) scale or the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) are commonly used for subjective evaluations of voice quality. However, inter-rater reliability can be a challenge, as different clinicians may interpret, and rate voice samples differently [5]. Addressing this challenge through training sessions, calibration exercises, or the involvement of multiple raters for each sample helps mitigate variability and ensures consistent labels, which are essential for accurate model training and evaluation. By adhering to these principles in data set selection, annotation, and standardization, the field can progress toward creating robust and fair AI systems for voice analysis.

5 | AI-Driven Diagnostic and Screening Tools

5.1 | Screening Method

In this section, we present the findings from our systematic review of studies that employed ML or DL methods for analyzing neurological voice disorders. The studies included in this review were carefully screened based on predefined criteria, as outlined in the PRISMA framework. The screening method is illustrated in Figure 1A and is designated as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). A comprehensive search was conducted in PubMed and Web of Science, encompassing publications from January 2010 to January 2025. The search terms included “Machine Learning,” “Deep Learning,” “Neurological Voice Disorder,” “Voice Disorder,” “Laryngeal Dystonia,” “Spasmodic Dysphonia,” “Parkinson's Disease,” “PD,” “Essential Tremor,” “Voice Tremor,” “Tremor,” “Vocal Fold Paresis,” “Vocal Fold Paralysis,” “Vocal Paresis,” “Voice,” “Dysphagia,” “Acoustic,” and “Speech”, both in isolation and in combination. The search field was “All Fields” for both Web of Science and PubMed. This yielded a total of 2258 articles. After removing duplicates, 1346 articles remained. However, 1007 articles were still excluded during the title and abstract screening, and 20 more were excluded during the full-text screening. Ultimately, 319 studies were selected for meta-analysis. Figure 1B illustrates the most common ML/DL pipeline for neurological voice disorders diagnosis and related tasks. It should be noted that different modality data require distinct preprocessing steps. Subsequent to this, the pipelines undergo modifications, yet they can be organized as follows: feature reduction or representation, classifier or model training, and performance evaluation. As mentioned above, feature reduction is necessary for dealing with high-dimensional data. Representation learning with deep neural networks can also be treated as similar functions. In a supervised learning situation, the training of a model aims to find a decision boundary, that is, a hyperplane that distinguishes between different categories. Alternatively, the model can estimate the joint distribution over data and labels. The performance of new testing data can be depicted by various metrics, including accuracy, precision, sensitivity, specificity, and the area under the receiver operating characteristic curve (ROC AUC). It is important to note that the efficacy of a model is measured by various metrics, each providing a unique perspective on the model's performance.

As illustrated in Figure 2, the survey contains several key aspects. It reveals a trend of researchers increasingly adopting ML and DL methodologies for the analysis of neurological voice disorders. From 2010 to 2015, the fields of ML and DL exhibited minimal activity, with ML demonstrating slightly higher levels of activity. Between 2016 and 2019, both fields began to show growth, with ML experiencing a faster initial rise. However, from 2020 onward, DL experienced a marked increase, surpassing ML by 2023 and reaching its height with 43 studies, in contrast to ML's 41. ML had an earlier peak in 2021 with 32 studies but subsequently began to decline. It is evident that DL methods, with their advanced representation capabilities, have emerged as the preferred approach when employing AI algorithms for classification or prediction tasks. Furthermore, ML has 217 studies, making up a larger proportion of the chart. DL accounts for 154 uses, forming a smaller share. Among the specific ML techniques, SVM dominated with 134 studies, followed by FS with 58, RF with 57, and k-NN with 50. Other techniques include Boosting (31), Logistic Regression (LR, 26), Naïve Bayes (NB, 25), and Decision Trees (DT, 20), which have relatively smaller shares. Additionally, the “Others” category accounts for 43 uses, capturing less common methods. CNNs were the most frequently used model, with 79 studies,

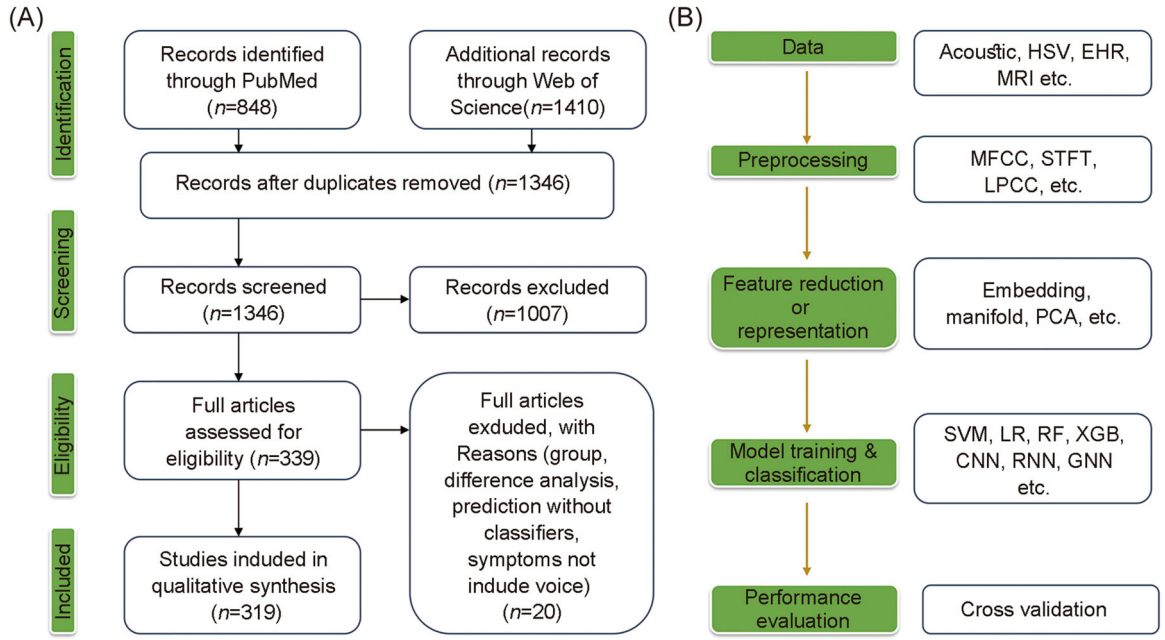

FIGURE 1 | (A) Literature search results for each screening step. (B) Summary of standard machine learning or deep learning steps for neurological voice disorders classification. CNN, convolutional neural network; EHR, electronic health records; GNN, graph neural network; HSV, high-speech videoglottography; LPCC, linear prediction cepstrum coefficients; LR, logistic regression; MFCC, Mel-frequency cepstral coefficients; MRI, magnetic resonance imaging; PCA, principal component analysis; RF, random forests; RNN, recurrent neural networks; STFT, short-time Fourier transform; SVM, support vector machine; XGB, extreme gradient boosting.

significantly surpassing other models. ANNs followed with 41 uses, indicating their continued importance. RNNs accounted for 25 uses, while Transformers were used 15 times, reflecting the growing adoption of attention-based models. The utilization of AE and GAN, with five studies each, indicates their specialized application in specific domains. GNN demonstrates a lower frequency, with only two occurrences, suggesting their limited integration into current models. The majority of studies utilized acoustic data, encompassing vowel phonation or voice sentence recordings. The overview of data sets used in the reviewed articles is depicted in Table 2. The employment of an endoscopic high-speed (HS) camera represented a cutting-edge approach, enabling the capture of rapidly moving vocal folds during phonation. Researchers were endeavoring to provide reliable vocal fold diagnosis through the integration of HS video and AI models. Some studies incorporated multimodal data to diagnose neurological voice disorders.

Before the DL era, ML has been essential for understanding relationships between variables, making predictions and drawing inferences. LDA, SVM, k-NN, and similar algorithms help estimate future outcomes based on historical data and identify key relationships between variables. Feature selection or reduction as necessary steps mentioned above are widely used for high-dimensional small data set situations. Many works used PCA and LDA as the most common feature extraction step for dimensionality reduction after voice preprocessing. Ji and Li [46] employed an energy-based feature ranking algorithm with ensemble strategy followed by linear SVM classifier for PD detection. Tsanas et al. [47] compared four different feature selection methods (LASSO, mRMR, RELIEF, and LLBFS) with SVM and RF classifier. They found some dysphonia measures complement existing algorithms in maximizing the ability of the classifiers to discriminate healthy controls from PD patients. Saeedi and Almasganj [48] proposed a wavelet feature extraction approach with adaptive wavelets to sort different types of vocal disorders. Verde et al. [49] showed that opportune feature selection methods can help no matter SVM, Decision Tree, or other classifiers to gain good performance on voice pathology detection. Yang et al. [50] proposed a hierarchical boosting dual-stage feature reduction ensemble model for PD speech recognition on three data sets. This dual-stage mechanism integrates the advantages of traditional feature extraction and feature selection algorithms. Özbay et al. [51] employed Grey Wolf Optimizer, which focused on feature selection by using local search and evolutionary operator, to minimize feature space for automated voice pathology detection. Lamba et al. [52] developed a hybrid feature selection approach for PD detection with mutual information gain and recursive feature elimination. By efficiently applying the feature selection method, the accuracy of the model could be increased by using fewer features.

SVM, as one of the most widely used algorithms in ML, excels in cases with smaller data sets, clear margins, and high-dimensional spaces. The kernel-based mapping trick allows SVM to compute the dot product in high-dimensional space without performing the transformation implicitly. A variation of SVM used for regression tasks is Support Vector Regression (SVR). Unlike classification SVM, which separates classes, SVR aims to find a function that predicts continuous values within a range. Voigt et al. trained preprocessing high-speed video data by SVM with Radial Basis Function (RBF) [67]. These phonovibrogram features on the testing set achieved approximately 81% accuracy for the binary classification tasks. Huang et al. also used glottal images and SVM to identify various vocal fold

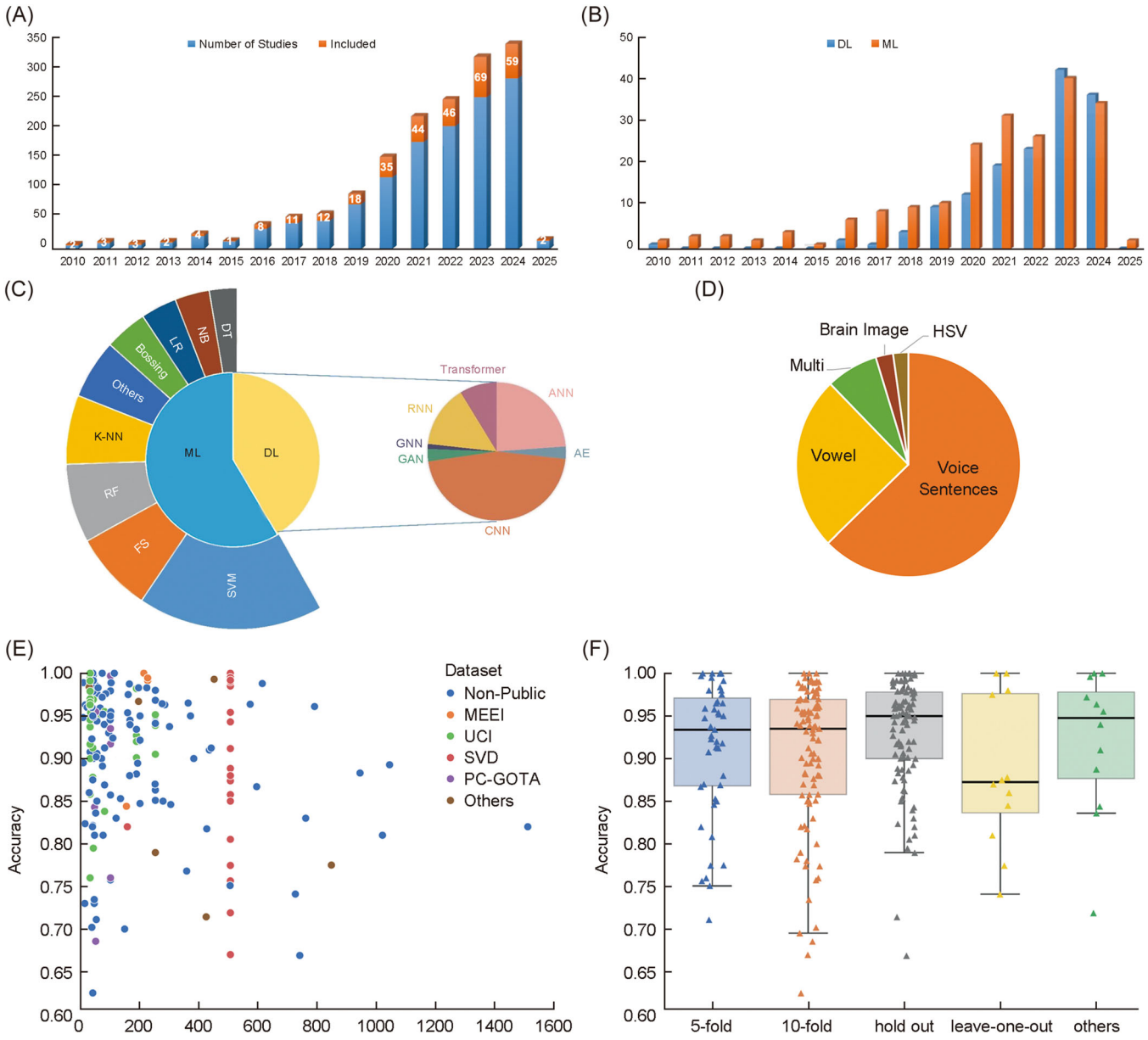

FIGURE 2 | Summary of the selected studies. (A) A total number of papers (blue) and PRISMA-selected papers (orange). The blue blocks represent papers identified through keyword searches, while the orange blocks signify the final included articles following screening and eligibility assessments. (B) Number of deep learning (blue) and machine learning (orange) publications per year. (C) Proportion of deep learning (DL) and machine learning (ML) methods used. (D) Proportion of modality data used. (E) Scatter plot of overall reported accuracy versus the total sample size across different databases. (F) Boxplot of overall reported accuracy based on different cross-validation methods. AE, autoencoder; ANN, artificial neural network; CNN, convolutional neural network; DT, decision tree; FS, feature selection; GAN, generative adversary network; GNN, graph neural network; HSV, high-speech videoglottography; k-NN, k-nearest neighborhood; LR, logistic regression; NB, naïve Bayes; RF, random forest; RNN, recursive neural network; SVM, support vector machine.

disorders with 98.75% accuracy [53]. Saeedi et al. [54] used a genetic algorithm and RBF SVM to classify normal and pathological voices on two data sets. Montaña et al. [55] designed an expert system consisting of feature extraction, feature selection, and SVM to discriminate healthy people from PD in the early stage by using the diadochokinesis test. Lahmiri and Shmuel [56] used SVM with recursive feature elimination to reduce the effect of correlation bias in patterns and the Bayesian optimization technique is also leveraged for optimizing the parameters of RBF kernel. Harimoorthy and Thangavelu [57] proposed an adaptive linear kernel SVM to help detect PD patients. Compared to the existing kernel function as a polynomial, radial basis, and sigmoidal-based function, the new one shows improvements in prediction accuracy, F1-score, and other metrics.

K-Nearest Neighbor (k-NN) as a simple yet powerful non-parametric algorithm can also be used for both classification and regression tasks. There is no need for training with k-NN, the results are very easy to explain and unlike linear models, k-NN makes no assumptions about the underlying data distribution, making it useful for nonlinear problems. Sharma et al. [58] used Rao algorithm for solving optimization during feature

TABLE 2 | Overview of data sets reported in the reviewed articles.

	UCI	SVD	PC-GITA	MEEI	mPower	VOICED	Other	Not public
Number of articles	87	41	17	12	7	5	39	146
Number of subjects	(31, 252)	(100, 506)	(50, 100)	(100, 226)	(424, 4051)	208	(14, 4121)	(76, 805)
Voicing task	Vowel, Speech	Vowel	Vowel	Vowel	Speech	Vowel	Vowel, Speech	Vowel, Speech
Population	Various	German	Spanish	American	American	Italian	Various	Various

Note: The number of subjects is presented as (min, max) for data sets with multiple versions or as a combined count for groups of different data sets.
Abbreviations: MEEI, Massachusetts Eye and Ear Infirmary; PC-GITA, Parkinson's Group – Gimenez, Ituttalde, Torres and Arias; SVD, Saarbrücken Voice Data set; UCI, University of California Irvine; VOICED, Voice ICar fEDertco II.

selection and subsequent choice of parameter k in the k-nearest neighbor classifier. Oguz et al. [59] proposed a feature selection and classification technique for PD based on speech signals using a meta-heuristic algorithm. After generating the feature space by immune plasma algorithm, a k-NN classifier is used to do the final prediction.

Random Forest (RF) as a powerful ensemble learning method can improve predictive accuracy and robustness compared to individual decision trees. The ability to rank feature importance, handle unbalanced data, and withstand noise or missing values make this algorithm widely used for classification and regression tasks. Vaičiukynas et al. used RF as a supervised algorithm to detect PD and also to fuse information in the form of soft decisions, obtained using various audio feature sets from different modalities [27]. Pramanik et al. [60] proposed a new method which undertakes minimum number of decision trees to achieve maximum detection accuracy. This method is tested on two acoustic data sets of PD and HCs with about 95% accuracy. Vaish [61] developed the RF model to analyze existing multi-modal clinical data on PD patients, prodromal PD patients, and healthy controls. Voice analysis is the most robust test compared to other data based on feature ranking by RF. Sheikhi et al. [62] proposed a forest-based ensemble approach that combines rotation forest and random forest algorithms to analyze each patient's voice data and to determine the patient's disease severity.

Boosting as another category ensemble learning technique can improve predictive performance by sequentially training weak learners and combining them into a strong model. These algorithms like Adaptive Boosting, Gradient Boosting Machine, Extreme Gradient Boosting, and so on reduce bias, reweight misclassified samples for imbalanced data sets. Schlegel et al. [63] used Adaboost to find the subset of parameters that can best separate perceptually hoarse, perceptually not hoarse sounding voice, and healthy controls. Tunc et al. applied a two-stage estimation model to discriminate PD from HCs, where the first step is to use a wrapper-based feature selection algorithm to find the most informative speech features, and then to feed the selected set of features into the extreme gradient boosting. Skibińska and Hosek [64] used two types of data to diagnose PD. Acoustic signals were parameterized in the fields of phonation, articulation, and prosody. Video recordings of a face were analyzed in terms of the movement of facial landmarks movement. Both modalities were consequently modeled by the XGBoost algorithm.

Other ML algorithms such as naïve bayes, Extreme Learning Machine, Hidden Markov Model (HMM) are also be used for these tasks. Za'im et al. [65] used online sequential extreme learning machine (OSELM) to detect voice pathology. OSELM is able to learn for the training data through a chunk-by-chunk mechanism with constant and varying lengths. Chandrakala et al. developed a hybrid model with HMM and deep neural network (DNN) for impaired speech recognition. Temporal dynamics of the speech utterances are modeled by HMM and the posterior probabilities are given by DNN. Pramanik et al. [66] used collaborated feature bank and traditional Naïve Bayes, which developed in the Weka machine learning repository, for detection of PD patients.

5.2 | DL and End-to-End Models

DL enables neural networks to automatically learn hierarchical representations from raw data, eliminating the need for handcrafted feature extraction. It has surpassed traditional ML in popularity due to the development of specialized hardware; not heavily rely on manual feature engineering; breakthroughs in model architectures; low cost/threshold of pretrained large models, and the ability to generalize across diverse tasks. Unlike traditional ML methods, end-to-end models learn both feature extraction and task-specific patterns within a single framework. This approach has significantly advanced tasks such as speech recognition, image classification, and natural language processing, where complex structures must be inferred from high-dimensional data.

Early artificial neural networks (ANNs) were not fully end-to-end, as they depended on manually engineered features before being processed by a multi-layer perceptron (MLP) for classification or regression. This limitation hindered their ability to generalize directly from raw data. However, the emergence of DL architectures enabled networks to jointly learn both feature extraction and prediction together. CNNs automated spatial feature learning for images, while RNNs captured dependencies in sequential data without predefined temporal features. More recently, Transformers revolutionized sequence modeling by replacing recurrence with self-attention, leading to state-of-the-art performance across multiple domains. By eliminating the bottlenecks of feature engineering, modern end-to-end models offer greater efficiency, adaptability, and scalability.

5.3 | ANNs

ANN is a model architecture that is inspired by human neurology and is the basis of various contemporary DL methods. In its most elementary formulation, an ANN is a feedforward network containing one layer of nodes with weights and an activation function. Information flows in one direction, is weighted by learnable parameters, summed, and passed through an activation function that introduces nonlinearity. The introduction of nonlinearity is the key aspect of how an ANN can learn complex patterns. The activation function loosely approximates aspects of the action potential mechanism in the human brain. However, unlike real neurons, traditional ANNs do not exhibit temporal dynamics, refractory periods, or spike-based communication. To address these limitations, RNNs have been proposed to model temporal dependencies, while spiking neural networks (SNN) introduce event-driven, spike-based processing. This evolution in neural network design reflects a growing trend toward biologically inspired computation, bringing artificial models closer to the principles of human neural processing.

MLP is a specific type of ANN, that contains fully connected layers. It can be formulated as a shallow learning network with a single weight layer, commonly referred to as the hidden layer, or as a DL network with two or more hidden layers (DNN). The learning process is typically carried out using a loss function, backpropagation, and an optimization algorithm. Due to its capability, MLPs (and other ANNs) are often considered universal function approximators, meaning they can approximate any underlying function given appropriate training data and conditions. However, this ability also presents the challenge of overfitting, where the learned function fits the training data too closely, and fails to generalize to unseen data sets. This is why clinical applications of trained ANNs must always be validated using independent data sets to ensure their practical utility.

In scientific literature, there have been reported results that would indicate that MLP is not the most optimal solution for neurological voice disorder prediction. The first instance of using an MLP to model neurological voice disorders, our review found was proposed in 2010 by Voigt et al. [67]. For the diagnosis of vocal fold paresis by using phonovibrogram features, the authors reported an accuracy around 80% for MLP, and 93% for SVM. However, in 2019, Almeida et al. [68] showcased how a KNN model was selected over an MLP as the classification model for PD detection from sustained vowels and short sentences data. In the case of Parkinsonian speech impairment classification from voice, Zhang et al. [69] demonstrated how an SVM outperformed a single and double hidden layer MLPs. This was also the result of Zehra Karapinar Senturk [70] in 2020, as SVM with recursive feature elimination outperformed a shallow MLP.

There has also been substantial supporting evidence that MLPs, more precisely DNNs, are a good modeling strategy for neurological voice disorders. In 2018, Wan et al. [71] introduced a DNN for PD detection from sustained vowels, and in the leave-one-out cross-validation experiment both 5 layer and 10 layer deep versions outperformed the traditional ML models. Fang et al. [72] then reported in their article how a DNN model was able to outperform an SVM and GMM models in predicting voice disorders such as LD from sustained vowels. In 2021, Pahuja and Nagabhushan [73] demonstrated how a shallow MLP outperformed SVM and KNN models in PD detection from speech. Vital et al. [74] also reported a similar result, where a shallow MLP produced a better 100% accuracy when compared to KNN, SVM, RF, NB, and AdaBoost. After that year in 2022, Jyotiyana et al. [75] also proposed a DNN model for PD classification from acoustic features that had 94.87% accuracy, and it outperformed KNN, DT, RF, LR, and NB.

The studies of shallow and deep MLPs for neurological voice disorders prediction show a split in effectiveness. Early works suggest traditional models like SVM and KNN often outperform MLPs for tasks such as PD detection. However, more recent research highlights the strength of MLPs and DNNs especially, with several studies reporting better performance compared to the more traditional ML models.

5.4 | CNN

DNN models rely on a successful feature extraction step before modeling to perform adequately. This means that they lack translation invariance; small spatial shifts in input data can significantly affect predictions. Also related to this limitation, they struggle to learn spatial hierarchies from data. CNNs were proposed to solve these challenges: they employ filters and

pooling operations to local regions of the input data, effectively reducing spatial dimensions. This gives them improved translation invariance and effectively reduces the number of parameters needed when compared to an MLP. Muhammad et al. [76] were one of the first authors to propose a CNN solution for Neurological voice disorders in 2018, by applying three different regression filters to a sustained vowel octave-spectrogram to construct a multi-channel input for the model. The proposed system achieved a 98.5% accuracy with the SVD database. In 2019, four CNN application articles to neurological voice disorders were proposed. Wodzinski et al. [77] reported more evidence on how a CNN using spectrograms as input performed as good as state-of-the-art models for PD. Hakan Gunduz [78] showcased in his article how parallel convolutional features outperformed a more conventional CNN architecture. The other two studies explored multimodal approaches, which will be discussed in detail in the multimodality section.

In 2020 and 2021 the number of CNN articles related to neurological voice disorders were five and six, respectively. Valeriani and Simonyan [79] proposed a 3D CNN as a microstructural neural network biomarker for dystonia diagnosis in 2020. During that year, Zahid et al. [80], Lee and Choi [81], and Mohammed et al. [82] demonstrated how CNN models outperform previously reported models for PC-GITA and SVD databases. However, Mou et al. [83] also reported that for post-stroke dysarthria tone prediction, an ANN was better when compared to a CNN. Mohammed et al. [84], Hu et al. [85], Majda-Zdancewicz et al. [86], and Ina Kodrasi [87] in 2021 also showcased the applicability of a hybrid CNN model, CNN model for multi-outcome prediction such as spasmodic dysphonia, multi-channel CNN model and multi-input CNN model to neurological voice disorders prediction, respectively. Gupta et al. [88] also demonstrated how ResNet architecture is an improvement over a traditional CNN, and Sonawane and Sharma [89] proposed a CNN for dysarthria severity.

CNN applications continued to increase in number in 2022. Verde et al. [90] showcased a light CNN application to neurological voice disorders mobile health. T. S. Mian [91] showed how the 1D CNN model was better than traditional ML models for screening Parkinsonian voice. Fujimura et al. [92] also proposed a 1D CNN that utilized raw audio instead of spectrograms as input data. Hidaka et al. [93], Reid et al. [94], and Fu et al. [95] showed the applicability of applying established CNN architectures such as EfficientNet and VGG19 for neurological voice disorder prediction. Hireš et al. [96], Mary and Suganthi [97], Quan et al. [98], and Hung et al. [99] also proposed CNN variations for neurological voice disorders prediction: ensemble of CNNs, PCA-CNN model, 2D-1D CNN model and SincNet, a CNN model with learnable sinc functions.

In 2023 even more articles related to CNNs and neurological voice disorders were proposed. The utilization of established CNN architectures continues, as Iyer et al. [100] proposed a PD prediction model based on Inception V3 and Aziz and David [101] investigated ResNet50, DenseNet, MobileNet, ConvNeXt, and EfficientNet architectures for the same task. Skaramagkas et al. [102], Kumari and Ramachandran [103], Liu et al. [104], and Escobar-Grisales et al. [105] proposed novel CNN variations with recurrence plots data as multi-channel input to a CNN, line spectral frequency (LSF) trajectory data input to a CNN, 1D CNN for the multiclass prediction of healthy, hyperfunctional dysphonia and laryngitis and multimodal representation with multiple fusion strategies, respectively. Hireš et al. [106] published a result where XGBoost nor CNN were able to generalize well beyond their training data sets. Xie et al. [107], Chen et al. [108], and Costantini et al. [109] also proposed new variations of already published methods, such as CNN for acoustic features, CNN for sustained vowel data, and CNN as a feature extractor. Ibarra et al. [110] also investigated 2D-CNN, CNN-LSTM, and 1D-CNN architectures for the multitask of PD classification and data set detection.

The usage of established CNN architectures such as ResNets as backbones continued in 2024 for neurological voice disorders prediction [111–115]. Novel solutions were also proposed, as quantum computing-inspired hybrid quantum–classical CNN was proposed for PD prediction by Sha and Rahamathulla [116]. A CNN-based audio processing and PD diagnosis method running on a Field Programmable Gate Arrays hardware was introduced by Majidinia et al. [117]. Akila and Nayahi [118] proposed a novel Parkinson classification neural network (PCNN) that had convolutional layers, a squeeze and excitation (SE) module, an inception block, and a global pooling layer. The authors state that these additions make the model more discriminative, and report PCNN overperforming traditional ML models and a conventional CNN.

CNNs have become one of the main architectures for neurological voice disorders prediction from audio, evolving from spectrogram-based models to advanced architectures like ResNet, EfficientNet, and hybrid CNNs. Researchers have explored multimodal representations, 1D CNNs, and hardware-optimized implementations, with some studies highlighting generalization challenges. Recent innovations include quantum-inspired CNNs and enhanced architectures like PCNN, improving feature extraction and classification.

5.5 | RNN

DNNs and CNNs implement a certain type of memory in the form of trainable parameters, which are adjusted during training. However, this does not consider maintaining any type of temporal memory, where associations between time steps in a sequence are considered. RNN architecture implements such a temporal memory in a form of a recurrent unit that maintains a hidden state. RNNs were highly successful in sequence learning, where the next time step is predicted. After the introduction of the attention mechanism, Transformers are now used for most sequence learning problems. RNNs such as LSTMs are still, however, routinely implemented as medical applications, as they provide a less complex and more established way of modeling sequences when compared to Transformers.

In 2019, one of the first RNN implementations for Neurological voice disorders was introduced by Guedes et al. in their article [119]. The authors used an LSTM network with audio spectrograms to model the temporal variability of dysphonia and other outcomes against healthy volunteers. However, in this study it was the 1D CNN model that performed the best by

producing a marginally better result over the LSTM. This result was partially replicated by Lauraitis et al. [120], where an SVM model performed better when compared to an LSTM in the task of speech impairment detection. Four 2021 articles using RNNs were found by our review, where the target outcome was PD. Two of them used sustained vowels, while the remaining two used speech samples. These articles showcase how a bi-LSTM can be used as an effective feature extractor for classification models [121], an LSTM model can achieve better prediction performance when compared to an ensemble of various ML methods [122], how a ResNet-LSTM hybrid performed favorably when compared to an SVM and RF [123], and lastly how a simple RNN with sech gates reached the same performance as an LSTM and GRU model, however with significantly less model parameters [124].

In 2022 there were a total of five articles focused in RNN models for neurological voice disorders prediction [125–129]. Two of the articles used sustained vowel data, while the rest used speech. Three of the articles reported improved LSTM-based prediction accuracies over traditional RNN, MLP, SVM, DNN, and RF methods. Two of the articles showed a GRU -based solution that performed better when compared to traditional ML and DNN methods. Also, two of the articles proposed a CNN–RNN hybrid method, showcasing the advantage of combining recurrent units with convolutional operations.

2023 had six scientific articles related to neurological voice disorders classification. Once again only one article used sustained vowels for input, while the rest used speech. Javanmardi et al. [130] demonstrated how a CNN could generalize better than an LSTM model, when the specAugment method was used for the training data. Han et al. [131] proposed the first attention-based RNN model for neurological voice disorders prediction, in the form of a bi-LSTM. Three of the articles proposed a hybrid model of either CNN–RNN or RNN–RNN [132–134]. Tayebi et al. [135] also showed how an LSTM can be used for the task of speaker verification, when dysphonic speakers were processed with the model. In 2024, two scientific articles were proposed for neurological voice disorders that utilized RNNs. Both used sustained vowels as input data. Pham et al. [136] showed how an LSTM can perform better when compared to traditional ML methods. Romany F. Mansour [137] also in his article reported how his CNN-ALSTM-based prediction system performed favorably when compared to traditional ML methods.

Based on all the scientific articles that our review found, the three following outcomes can be summarized. First, from all the RNN architectures, LSTMs are the most implemented. Second, hybrid RNN methods that leverage other RNN models or CNNs are reported in literature the most frequently to perform the best. Lastly, there are less and less RNN-based articles published over time. This is in line with the fact that Transformers have emerged as the new standard for sequence learning tasks.

5.6 | Autoencoders, Generative Adversarial Networks and Graph Neural Networks

An autoencoder is an ANN-based architecture used to learn feature encodings from unlabeled data. It consists of two parts: the encoder and the decoder network. The encoder compresses the input data into a smaller feature space, effectively acting as a data bottleneck. The decoder then reconstructs the original input data from this encoding. These networks are trained together, optimizing the reconstruction output to resemble the input as closely as possible.

For Neurological voice disorders, Y. N. Zhang [138] in 2017 proposed a method based on stacked autoencoders (SAE) for the classification of PD from sustained vowel recordings. The SAE was used with time frequency features to extract relevant ones for classification, which was done with a KNN model. This method performed better than classifiers without SAE, by achieving a mean accuracy of over 98%. In 2021, Masud et al. [139] also introduced an autoencoder for PD classification from sustained vowel data. This stacked sparse autoencoder (SSAE) was also used as feature extractor, and from various classification algorithms linear discriminant analysis (LDA) had the best performance, where the SSAE increased the accuracy by 0.96%. The next year, Almasoud et al. [140] described in their article a PD speech classification system, where an RNN-enhanced Graph Long Short-Term Memory Network (GLSTM) was used with LDA and SAE for dimensionality reduction and feature extraction. This system scored 95.4% accuracy, 93.4% F1 score, and 0.865 MCC. In the same year, Khaskhoussy and Ayed also used an autoencoder for feature extraction, and a Gaussian Mixture Model-Universal Background Model (GMM-UBM) for classification [141]. By utilizing sustained vowels, spoken numbers and words, the method achieved 100% accuracy with a test data set of 28 participants. Lastly in 2023, García-Ordás et al. [142] proposed a multi-task MLP network that predicts the UPDRS scale as a regression output, and PD severity as a classification task. The authors state that by learning both of these adjacent tasks, the model is able to perform better in both of them. The severity was predicted with an accuracy of 99.15%, outperforming the state-of-the-art, while the UPDRS prediction had 0.15 MSE.

Another ANN-based architecture that utilizes two specialized networks in unison is GAN. This model consists of a discriminator and a generator. The generator takes a random input from a noise distribution and produces a synthetic data sample. Both generated sample and a real data observation are fed into the discriminator, which predicts whether an observation is real or fake. The networks are trained together in a zero-sum game: the generator learns to produce increasingly realistic samples, while the discriminator improves its ability to distinguish real from fake. Typically, after training, the discriminator is discarded, and the generator is used to create realistic synthetic data based on the real data distribution. This approach can help augment training data sets and indirectly improve classification performance.

One of the first implementations of GANs in the neurological voice disorders domain was in 2020, when Chui et al. [143] proposed a Conditional Generative Adversarial Network (CGAN) for balancing the training data set by generating observations of underrepresented classes, such as hypokinetic dysphonia. These data were then used to train an Improved Fuzzy c-Means Clustering (IFCM) model for classification. Implementing CGAN as part of the modeling workflow improved the TNR by 10%–12.6% and TPR by 5.8%–16.2% and

contributed towards an increase in performance by 4%–6% compared to the more traditional SMOTE and cost-sensitive learning methods. In the same year, Xu et al. [144] introduced a method called Spectrogram Deep Convolutional Generative Adversarial Network (S-DCGAN) for PD voice data generation. By using ResNet-50 as the classifier, the authors demonstrate that utilizing a spectral normalization (SN) with DCGAN architecture resulted in improved prediction accuracies over DCGAN [145]. The S-DCGAN-ResNet50 hybrid model achieved an accuracy of 91.25% for discriminating PD for healthy volunteers.

Unlike traditional neural networks that operate on grid-like structures (e.g., images or sequences), GNNs leverage message passing to aggregate and update node representations based on their neighbors. This allows GNNs to effectively model structured data. By incorporating graph structures, these models can capture dependencies that may be lost in sequential architectures, making them particularly useful for complex biomedical applications.

Almasoud et al. in 2022 proposed a hybrid model called RNN-Graph-LSTM to classify PD from speech [140]. Each node in the LSTM is connected based on a graph structure rather than a linear sequence. The model obtained a 95.4% accuracy with a test data set. In 2023, Park et al. [44] employed a GNN-based semantics-guided neural network (SGN) model for the task of extracting spatiotemporal features from video for stroke and PD detection, along with a separate time-delay model for the voice modality. This system achieved ROC AUC scores of 0.802 for stroke and 0.780 for PD prediction. In 2024, Zhao et al. [146] collected EEG data from a vocal pitch regulation task, and used graph signal processing-graph convolutional networks (GSP-GCN) method for PD diagnosis. It achieved an averaged classification accuracy of 90.2%, a significant improvement of 9.5% over other DL models such as CNNs, RNNs, and EEGNet.

In summary, autoencoders improve feature extraction by boosting classification accuracies, while GANs address class imbalance and potentially increase performance by 4%–6% when incorporated into a modeling workflow. GNNs also show potential for implementation, outperforming CNNs and RNNs in one study. These advances suggest that hybrid models could further improve diagnostic accuracy and generalizability.

5.7 | Transformers

The transformer architecture is unique because it was designed to solve sequence learning tasks without relying on recurrence or convolution but using multi-head self-attention instead [147]. The utilization of positional encoding and parallelization enables transformers to efficiently capture long-range dependencies and process data faster. The established encoder–decoder architecture from machine translation is still there; however, many models have been proposed where only one of these components is used in the finalized model. In self-attention the elements or tokens in a sequence can attend to all other tokens in the sequence. However, in multi-head self-attention, multiple attention layers are used in parallel with different weight matrices, which allows the model to focus on different aspects of the input sequence. The encoder networks responsibility is to produce rich contextual feature embeddings of the input, and the decoder networks responsibility is to produce outputs from those embeddings.

The first transformer implementations to neurological voice disorders prediction were introduced in 2023. Ribas et al. [148] proposed a neurological voice disorders prediction model that utilized learned representations of transformer encoders of model such as Wav2vec2.0, HuBERT, and WavLM. The article reported an accuracy increase of 4.1% and 15.62% in two experiments over the baseline SVM model. Escobar-Grisales et al. [105] investigated speech and language representations for PD and healthy subject discrimination. The results indicate that speech representations had higher predictive power, which is in line with the fact that PD is mainly a motor disorder, so it affects speech production more than language. Klempíř et al. [149] demonstrated how wav2vec embedding enables the usage of large language corpora when the annotated amount of study data is limited and therefore improving the overall prediction result. Tirronen et al. [150] reported a similar result, where wav2vec embeddings and an SVM model outperformed traditional SVM models. Also, Hemmerling et al. [151] proposed a vision transformer for PD prediction, where the ROC AUC for this task was significantly improved over ResNet18, ResNet50, ResNext50, EfficientNetV1, EfficientNetV2, and the Swin transformer. An exhaustive comparison was also done by Nijhawan et al. [152], who compared a novel transformer model with 10 different ML methods. The results showed how their proposed model outperformed them all, by increasing the average ROC AUC by 1% over the second-best model XGBoost.

The emerging trend of transformers implementations continued in 2024. Zhao et al. [153] showcased how a triplet multimodal network that utilizes transformer blocks can be used for PD screening from voice. The model utilized multiple sustained vowels and free speech data modalities and achieved 99% and 90% accuracies for two test data sets. Klempíř and Krupička [154] in their article used a wav2vec embeddings from speech to predict PD classification and demographic and articulation characteristics as a regression task. The authors showcased how their method overperformed the feature extraction using traditional acoustic features for multi-language data sets. Malekroodi et al. [113] published a result where the Swin transformer was compared against popular CNN architectures in PD prediction from voice, and VGG16 was the best performing one. However, Tougui et al. [155], Irshad et al. [156], and Madusanka and Lee [157] published results where a transformer-based solution overperformed CNN-based methods for PD diagnosis, dysarthric speech recognition, and PD classification, respectively.

In 2023 and 2024, transformer-based models, such as Wav2vec and vision transformers, were successfully applied to neurological voice disorders prediction, outperforming traditional models like SVMs and CNNs. Key studies demonstrated the advantages of using speech representations and multimodal networks for PD classification and prediction, achieving high accuracies and improved results.

6 | AI in Rehabilitation and Therapy

6.1 | Advances in Treatment

Diagnosis lends itself well to classification modeling tasks, as it often involves differentiating between two groups. In contrast, treatment can be defined in numerous ways, adding complexity to ML and AI implementations. Despite these challenges, the application of ML and AI in neurological voice disorders is gradually advancing, with promising developments in areas such as brain MRI analysis, deep brain stimulation (DBS), treatment efficacy estimation and telerehabilitation.

In 2020, Halai et al. [158] published an investigation of brain-to-behavior prediction models for post-stroke aphasia. The authors describe that these models map brain lesion information to specific neuropsychological tests or aphasia types. They tested four different input modalities related to four behavioral factor scores: phonology, semantics, executive-demand and speech quanta, and reported kernel ridge regression, multi-kernel regression for two modalities and relevance vector regression for the last modality as the best ML models for the tasks. The authors claim that by formulating the learning tasks in such a way, the trained models can indicate causal relationships and therefore advance clinical perspective and research. Next year in 2021, Tankus et al. [159] investigated the applicability of ML models to decode deep brain stimulator recordings into spoken vowels for PD patients. Their sparse decomposition-based decoder model named SpaDe was used in conjunction with an implanted stimulator, and the model was able to correctly decode and predict all produced vowels, 96% of all perceived vowels and 88% of all imagined vowels. The authors state that their method presents an important step forward for restoring speech with a brain–machine interface (BMI). In 2023, Suppa et al. [160] also investigated DBS treatment's contribution to worsening dysarthria symptoms in PD patients during sustained vowel tasks. Using an SVM model, the authors state that their results demonstrate a significant worsening of voice in STN-DBS patients, when compared to patients that received L-Dopa-based treatment.

ML and AI can also improve the quantitation of treatment efficacy, when this task is sufficiently complex enough. During 2020, Suppa et al. [161] investigated how cepstral analysis and ML models could differentiate healthy volunteers and adductor LD patients before and after botulinum toxin treatment. The authors discovered that compared to traditional cepstral analysis, ML models had better success in differentiating HV from LD, and also in quantifying the change in symptomatic voice due before and after botulinum toxin. In 2021, Barbera et al. [162] proposed NUVA, a naming utterance verifier model targeted for aphasia treatment monitoring. The authors used a DNN model with bidirectional GRU units to differentiate healthy and aphasic utterances to detect symptomatic voices during a treatment period. The method achieved 89.5% accuracy in a 10-fold cross-validation experiment and performed better than commercially available ASR models. During the same year, Jain et al. [163] introduced in their article a personalized CNN model for differentiating patients with PD, before and after dopaminergic medication. The authors state that this method enables medical practitioners to assess the effect of the most relevant medication for PD in a personalized way. The model achieved 82.35% in a leave-one-out cross-validation experiment. Most recently in 2023, Yao et al. [164] proposed a DL method named DystoniaBoTXNet that can estimate botulinum toxin treatment efficacy in isolated focal dystonia patients, using brain MRI data. The model is based on 3D CNN architecture, and by utilizing clinically relevant brain regions was able to achieve an overall 96.3% accuracy with independent test data sets. This is why the method has significant translational potential in aiding treatment decision-making.

In summary, the integration of ML and AI into the treatment and rehabilitation of neurological voice disorders has demonstrated remarkable potential to enhance clinical outcomes and treatment efficacy research. The advancements discussed from brain-to-behavior prediction models to personalized treatment efficacy quantification illustrate how ML and AI can provide clinicians with powerful tools to improve patient care. As these technologies continue to evolve, they hold the promise of transforming the landscape of neurological voice disorders management, bridging the gap between research and patient-centered applications.

6.2 | Remote and Telepractice Solutions

AI models can also improve the treatment of neurological voice disorders by promoting telerehabilitation. In 2021, Raza et al. [165] proposed an Internet of Things (IoT) framework for PD remote monitoring and disease progression tracking. The authors had a focus of providing low-latency IoT communication for up to 70 simultaneously connected devices, designed to measure disease progression. The authors also report state-of-the-art MAE performance for the associated UPDRS prediction task. During 2022, Mulfari et al. [166] described in their article a CNN model for isolated word recognition, implemented as a mobile phone app. Dysarthria patients would phonate 13 keywords in Italian with the app, and personalized and global versions of the model would predict what those keywords would be. In terms of mild, moderate, and severe cases of dysarthria, both models would exhibit a prediction performance of over 91% accuracy; however, the personalized models would perform better overall, as the best model of that type would have over 98% accuracy with all cases of dysarthria. The authors state that their model could be used as a telerehabilitation method and part of speech therapy related to dysarthria. In 2023, Rahman et al. [167] introduced in their article a teleneurology platform for PD. Users would use a website to complete eight tasks in three domains using their webcams: speech, facial, and motor. A screening assessment model would automatically provide a PD severity score for each domain, and a global score. Also, for additional questions about their scores, users could interact with a GPT-3 -based LLM model that would assist in management and referral tasks related to PD.

Telehealth prediction applications offer a cost-effective, scalable, and accessible solution to neurological voice disorders diagnosis. Point-of-care voice data collection and testing enable more personalized patient care, which then enables continuous monitoring options that enhance treatment. Among the initial papers identified on this topic was a research paper from 2013

by Mandal and Sairam [168], where the authors present a telehealth-based inference system for early-stage PD detection. They used a portable at-home testing device, where the measurements would be shared to a medical center over the internet. These data would be used as the input to the inference system that uses SVM-based feature selection and an ensemble of ML models with Bayes voting to produce prediction results. In 2016, Haydar Ozkan [169] published an article that depicted a telediagnosis method for PD. For this study, the data were collected in a speech laboratory with a computer and a microphone. While seven ML models were tested, KNN had the best 10-fold cross-validation accuracy of 99.1%, while using PCA-derived features. During 2017, Sakar et al. [170] also proposed a telediagnosis method for PD. Patient data were gathered remotely at their homes by monitoring them for 6 months with weekly intervals. By using all this voice data, the best differentiation of PD patients and healthy volunteers was achieved with an SVM model. In the year 2018, Alhussein and Muhammad [171] introduced in their article a voice pathology detection system that was computed using the SVD and MEEI databases. The authors investigated two CNN architectures, the VGG16 Net and CaffeNet, from which the latter achieved better accuracies in multiple comparisons. Also in 2018, Cesari et al. [172] reported a clinical validation of Vox4Health, a mobile health system released as a smartphone app. In the article, the authors evaluate the rule-based algorithm of Vox4Health and compare it against ML models to improve the performance of the app. Logistic model tree (LMT) and SVM both had better accuracy when compared to the more traditional rule-based algorithm.

In 2020, Chén et al. [173] introduced a telehealth framework utilizing smartphone-based data collection to monitor PD progression, employing an elastic net linear regressor. Similarly, Sajal et al. [174] proposed a smartphone application designed for remote assessment of PD, which used voice data and tremor sensor readings to differentiate between PD patients and healthy controls with KNN and SVM models. Tougui et al. [175] also published an article in 2020 that depicts a smartphone app method for PD classification. From the four ML methods they tested, XGBoost had the best accuracy of 95.78% with unseen data.

Narendra and Alku [176] proposed a telephone voice data-based assessment of speaker intelligibility that have dysarthria in 2021. The method used acoustic and glottal PCA features derived from voice and a DNN model. The accuracies varied from 40% to 70% with this configuration. The same year, Pah et al. [177] introduced a phoneme-based telehealth solution for PD detection, where acoustic features were collected during sustained phoneme recordings. With an SVM classifier, the authors were able to achieve an accuracy of 84.3% with the best configuration. Also, Carrón et al. [178] proposed a mobile-assisted voice condition analysis system for aiding in the detection of PD. In addition to the smartphone app and its backend server, the authors report that the best accuracy of 92.05% was reached using a Passive Aggressive (PA) classifier. Arora et al. [179] also published an article in 2021 that introduced a smartphone-based symptoms assessment that would differentiate rapid eye movement (REM) sleep behavior disorder (RBD) from PD and healthy volunteers, in addition to predicting self- or researcher-administered clinical scores related to motor function. The authors used the RF model with a leave-one-subject-out cross-validation scheme. The method was able to distinguish RBD from controls with a sensitivity of 60.7% and a specificity of 69.6%, and RBD from PD participants with a sensitivity of 74.9% and a specificity of 73.2%. In the year 2022, Motin et al. [180] proposed a smartphone-based model for PD detection from sustained phonemes collected in a clinical setting. The SVM model had 100% accuracy in the experiments the authors reported. In 2023, Worasawate et al. [181] introduced in their article a CNN model that predicts PD classification using voice samples collected with a smartphone. The CNN architectures tested were LeNet-5, VGGNet-16, and ResNet-50. VGGNet-16 had the best performance, with an F1 score of 98.9%.

Most recently in 2024, Mishra et al. [182] introduced in their article a mobile cloud-based predictive model for estimating the severity of PD symptoms from voice, called PD-DETECTOR. With an accuracy of 96.2%, their method based on a DNN model could be used from a smartphone app with 13 s of latency. Also, He et al. [183] proposed using smartphone-based voice data with various ML models for PD diagnosis. While testing on two different databases, the achieved average diagnosis accuracy was greater than 90% for multiple models, which the authors state demonstrates that voice data-derived biomarkers can be used in a clinical context.

In summary, the evolution of telerehabilitation and telehealth-based approaches for PD detection and monitoring has demonstrated significant advancements over the past decade. Early studies focused on portable testing devices and basic ML models, while more recent research emphasizes smartphone-based solutions, DL techniques, and real-time analysis tools. These advancements have progressively improved diagnostic accuracy, usability, and clinical applicability, showcasing the growing potential of voice and smartphone-derived biomarkers in supporting remote healthcare solutions for PD.

7 | Advanced Applications

7.1 | Voice Restoration and Augmentation

Assistive speech technology can be used to restore a person's ability to communicate. in the case of Neurological voice disorders, AI can propose novel options for reconstructing voice signals. Instead of classification and regression tasks, methods trained for generative tasks can be used to translate symptomatic speech into asymptomatic speech. This can be considered as a speech-to-speech or signal-to-signal modeling task, and it starts with the detection and segmentation of impaired speech. Iliya and Neri [184] proposed in 2016 two methods for the segmentation of silence, unvoiced and steady states. In their work, an SVM model was able to perform the best at segmenting silence and steady states, while an ANN performed the best with unvoiced states. In 2021, Chandrakala et al. [185] proposed a Histogram of States (HoS) and SVM-based method that uses DNN-HMM to learn word lattice triphone embeddings. This approach would then predict words from impaired speech, actively restoring the conveyed language information in a speech-to-text manner. It was able to achieve a 92.61% word accuracy with 15 acoustically similar words data set, 95.68%

word accuracy with 100 words UA-SPEECH data set and 87.11% word accuracy with 50 words TORGO data set. While the method performed exceptionally with dysarthric speech utterances of "very low" and "low" intelligibility levels, the vocabulary used in the experiments was limited when compared to something that would be needed for a casual conversation.

Research efforts in AI-based voice technology grew significantly at this point in time. In 2023, Chu et al. [186] proposed E-DGAN, a model that would generate personalized asymptomatic speech from symptomatic voice data. The model was trained in an adversarial manner while considering speech style, F_0 consistency and silence intervals among other acoustic measures. One of the more desirable features of such models is that it would retain the person's voice characteristics, while improving intelligibility. For Neurological voice disorders such as spasmodic dysphonia, the authors report a word error rate (WER) of 21.7% for E-DGAN, and 78.7% for a variational autoencoder. In the same year, Pan et al. [187] proposed PVGAN, also a GAN-based model for speech data generation. Instead of conversion from symptomatic to asymptomatic, the authors aimed for mimicking pathological voices with PVGAN, so that it could potentially be used as a data augmentation method for other studies. In their comparative experiment results, the authors were able to produce mean 0.801% jitter, mean 4.941% shimmer, mean 19.988 HNR, mean 4.08 MOS, and mean 3.20 PESQ for synthetic functional dystonia speech samples.

7.2 | Multimodal Data Approaches

One of the emerging trends of applying complex modeling methods to the medical field is to consider multimodal modeling, where several data modalities are combined to produce a prediction. The objective is to observe the patient with multiple different measurements in a way that complements the model training process, so that the added utility of new measurements outweighs the added noise or variability of said measurements. New measures can also have a regulating effect on others during modeling, by actively making the model more robust to noise.

One of the earliest implementations of this to the neurological voice disorders space was done in 2017 by Oung et al. [188], where the authors combined motion data of wearable sensors with voice to produce a classification result of healthy control or PD severity with ML classifiers. The data were collected in a clinical setting, and their multimodal approach achieved over 95% accuracy in their modeling task. In 2019, Lo et al. [189] proposed an ML method that utilizes smartphone sensors and voice recordings to predict future disease onset events related to PD and RBD. The authors report a ROC AUCs of over 0.9 in 10-fold cross-validation experiments. In the same year, Vásquez-Correa et al. [190] used motion, voice and handwriting modalities with a CNN model to differentiate between healthy controls and PD patients of three severities. The method performed the best when all modalities were used, achieving 97.6% accuracy with the test data set. During 2020, Chén et al. [173] proposed an ML framework to assess PD progression over 17 days with smartphone sensor and voice data, using an elastic net linear regressor. During that year, Sajal et al. [174] also proposed a smartphone solution, where tremor sensor data were used with voice to differentiate PDs and healthy controls using a KNN and SVM models. The ensemble method of all both data modalities and all model techniques achieved 99.8% accuracy, performing better than individual modalities.

In 2022, Yousif et al. [191] introduced a diagnosis model for PD, based on handwriting and voice modalities and a CNN architecture called VGG19. This method reportedly achieved 99.75% accuracy with the first test data set, and 100% accuracy on the second. Lim et al. [192] also introduced the notion of incorporating facial expression video data to a PD diagnosis procedure with voice in 2022. The authors report a ROC AUC of 0.9 with their validation cohort data set. It was also in year 2022 when Goñi et al. [193] proposed ML-based diagnosis models using four different data modalities. In the paper, a smartphone was used to collected gait, balance, tapping, and voice data. The multimodal method achieved the highest performance, but its 77% accuracy was still unsatisfactory. The same modalities were used by Dotov et al. [194] in 2023; however, data were collected in a clinical setting without smartphones and the modeling technique was linear mixed-effects model. This method improved over the method proposed by Goñi et al. by achieving an 82% accuracy. Gait, handwriting, and voice were also used by Indu et al. in 2023 for PD diagnosis, where the authors proposed a novel modification to the KNN algorithm [195]. Their method had 99.6%, 97.8%, and 94.5% accuracy for gait, handwriting, and voice parameters, respectively. In 2024, Anisha Vaish authored an article where RF models were used to compare PD patients with RBD symptoms and healthy controls, by inspecting independently data modalities such as voice, tapping, handwriting, and gait individually [61]. From them, voice data had the best predictive performance. In the same year, Sivakumar et al. [196] used voice and handwriting modalities with Improved Glowworm Swarm Optimization (IGSO) algorithm to find optimal set of features, then used Radial Basis Functions Networks (RBFN) for classification of PD. This model had 95.78% accuracy with a test data set. Lastly, Kumar et al. [197] introduced EGG as a data modality with voice for voice pathology prediction. Two parallel Invariance Scattering Networks (ISN) were used as feature extractors for the modalities, respectively, and an SVM classifier was used for the final prediction. The method produced ROC AUC values of 0.845, 0.797, and 0.890 for EGG, speech, and multimodal data, respectively.

In summary, recent research highlights the growing adoption of multimodal data in neurological voice disorders diagnosis, reflecting its potential to capture complex diagnostic information. Most studies also reported overall better prediction performances with multimodal data, when compared to single modality solutions. Furthermore, the reported prediction performances in recent studies indicate a level of accuracy that is approaching readiness for clinical implementation.

8 | Validation, Regulatory, and Ethical Considerations

8.1 | Clinical Validation and Standardization

The fundamental way of displaying real-world utility of an AI model is external validation. Models can exhibit excellent

prediction performances within their training data and still fail to generalize to unseen patient populations. Because of this fact, great generalization performance can be considered as the outcome of successful model training and the use of a training data set that adequately represents the target population. In a clinical setting, generalization can be validated by incorporating more test data in the form of other clinical trials. Prediction results can be calculated from these external data sets, which can then be compared against their respective ground truths with various metrics. These metrics are task-specific, and for rehabilitation and therapy tasks they can be highly specialized depending on how the tasks are formulated into prediction tasks. Diagnosis and screening are commonly formulated as binary classification tasks, where well-established confusion matrix metrics such as true positive rate (sensitivity), false positive rate, true negative rate (specificity), and false negative rate are used. Metrics such as accuracy, balanced accuracy or *F*-score, and likelihood rations are also commonly used. Receiver operating characteristic curve (ROC) and the area under this curve (ROC AUC) are also widely used analysis methods as they inspect the full true positive rate and false positive rate range at each classification threshold. However, they should be used if and only if the prediction probabilities are available for analysis, as they are independent from any class probability thresholds used.

The standardization of AI performance metrics in medicine has been discussed for over 30 years [198]. It is as important of a topic as the standardization of other aspects in the clinical AI model development pipeline, data acquisition, preprocessing, de-identification, and label annotation. Coherent performance metric reporting across all clinical trials has the advantage of improved comparability, transparency, and regulatory facilitation during an approval process. While there are active efforts for the standardization of performance metrics in general [199] and in specific clinical fields [200], this study is currently lacking specifically for neurological voice disorders.

8.2 | Privacy and Data Governance

For medical applications of AI, utilizing patient data during model training or during inference in a production environment imposes various considerations for data privacy, security, and compliance to regulations. In addition to the more obvious access, use, and control privacy issues that are directly addressed by regulations such as HIPAA and GDPR, the reidentification of an individual by emerging computational methods, for example, showcases the need for proper, standardized anonymization efforts [201]. This is especially true for voice modality data, as it is highly identifiable and can be collected in a non-private environment with smartphones in telemedicine situations. Methods specifically for voice anonymization have been proposed in the past [202, 203]; however, a gold standard method has yet to emerge. For this to happen, a data representation that preserves essential information in an application-agnostic manner while ensuring sufficient anonymization must be established as a foundational step before widespread adoption can occur.

The systems that implement the model training environment and the production environment need to be secure and compliant with the corresponding data-related regulations. For an AI software product filed to the FDA, these requirements are commonly not unique to AI as any Software as a Medical Device (SaMD) that process patient data needs to be compliant. However, progress is being made in this space for AI as governing bodies become more aware of the nuances of AI software, and propose new guidelines and legislation [204, 205].

8.3 | Interpretability and Trustworthiness of AI Models

In addition to prediction performance, one of the most important aspects of an AI model is interpretability. It is the model's ability to explain its prediction output from the input data in a way that is meaningful for a person to understand. Most AI applications are based on ANNs which compute high dimensional and complex operations during inference, making them difficult to interpret. This is why several interpretability methods have been proposed in the past for them, from ad hoc to highly integrated [206]. These methods are used to investigate the reasoning of the trained model to detect situations where the output prediction is correct, but the irrelevant input data were used for this conclusion. This is called the “Clever Hans” effect and it has been demonstrated to be found in published medical AI applications as well [207, 208].

Clinical decision-making which is based on an AI model's prediction requires the model to be transparent about its reasoning. To build trust and confidence of the model, clinicians need to have a proper understanding of how the prediction was made. This then translates to better patient safety, by ensuring that their AI-based outcome aligns with the established medical knowledge [209]. In the context of neurological voice disorders, SHAP [210], CAM visualization [146], LIME [211], and a combination of methods have been proposed as applicable interpretability methods [212]. However, Mancini et al. [213] reported that despite these efforts the results of an interpretability analysis are still too convoluted for medical domain experts to digest. This highlights the need for methods that balance technical complexity with practical usability in clinical settings. As the AI interpretability field advances towards more generalizable and easier to understand methods, their clinical applications are not far behind.

9 | Challenges to Overcome

9.1 | Data Scarcity and Generalizability

To generalize effectively beyond the training data distribution, ML and AI models need a substantial amount of diverse training data that accurately represents the problem being modeled. Diversity in data is highly important, whether it pertains to geographical, cultural, demographic, biological, or other factors. Unfortunately, as AI continues to advance and is applied to increasingly complex and specialized tasks, collecting enough high-quality data becomes progressively more challenging. This is particularly evident in the medical field, where AI is increasingly being applied to highly specialized areas. We believe that moving forward with AI

applications in the neurological voice disorders field will require greater efforts to address these challenges.

To overcome this, techniques such as transfer learning and domain adaptation have been proposed as effective solutions [214]. Transfer learning involves leveraging pre-trained models, which are typically trained on large and diverse data sets, and fine-tuning them for specific tasks using smaller data sets. This approach greatly reduces the reliance on extensive labeled data in niche domains. For example, in the medical imaging field, models pre-trained on general image data sets can be fine-tuned to identify rare diseases using a relatively small collection of specialized medical images. By reusing features learned from broader data sets, transfer learning enables the development of accurate models even in data-scarce scenarios. Domain adaptation, a specialized form of transfer learning, addresses the challenge of differences between source and target data distributions. Often, data from related domains or broader contexts is more readily available than the specialized data required for a specific task. Domain adaptation techniques bridge this gap by aligning the feature distributions of the source and target domains, allowing models trained on source data to perform effectively in the target domain. This could involve adapting a model trained on publicly available medical data sets to work accurately with data from a particular hospital or region.

9.2 | Technical and Clinical Integration Barriers

One of the concrete challenges of translating a trained model to a real-world medical application is the knowledge gaps related to technology. They can be evident across multiple areas, including the understanding of AI model behavior, its limitations, and how it integrates with clinical workflows. In the past, healthcare professionals may have limited exposure to the technical aspects of AI, making it difficult for them to fully trust or effectively interact with these systems. This lack of familiarity can result in resistance to adopting AI-based tools or hesitation to rely on their output in high-stakes clinical decisions.

Education of the workforce about the AI implementation process is crucial for widespread adoption in the medical domain [215]. This is supported by the growing adoption of AI in society, which is making the technology more accessible to medical professionals than ever before. As a result, this shift should also be reflected in medical education, with health informatics and computer science integrated into the curriculum [215]. AI implementation in healthcare is inherently interdisciplinary, which supports this integration. Promoting collaboration between medical professionals, data scientists, and engineers throughout the design, implementation, and evaluation phases of AI systems can help ensure that the technology is both practical and effective in real-world clinical environments.

10 | Conclusions

AI has emerged as a powerful tool in the diagnosis and monitoring of neurological voice disorders. ML and DL algorithms have demonstrated promising accuracy in detecting neurological vocal disorders and predicting treatment response, offering noninvasive and cost-effective alternatives to traditional diagnostic methods. AI-driven analysis of speech and other modality data has proven particularly valuable in conditions such as PD, LD, and stroke-induced dysarthria, where early detection is crucial for timely intervention.

The overall number of scientific publications on neurological voice disorders outcomes has been increasing annually. Furthermore, research on AI, particularly DL applications in neurological voice disorders, is also expanding, reflecting the growing interest and advancements in this field.

While early studies indicated that traditional models like SVM and KNN outperformed MLPs for neurological voice disorder prediction, more recent research shows that deep neural networks (DNNs) now excel, particularly in tasks like PD detection. CNNs have become the dominant architecture for audio-based neurological voice disorders prediction, and the performance can be enhanced with more advanced backbone model like ResNet and ConvNeXt. LSTMs remain the most implemented RNN architecture, though hybrid methods combining RNNs and CNNs are showing the best results, with a decline in purely RNN-based studies. Nowadays, Transformer-based models like Wav2vec and vision transformers have proven highly effective in neurological voice disorders prediction, surpassing traditional models and achieving improved classification and prediction results.

As of today, there are still challenges for implementing ML and AI in a way that is safe and beneficial from a patient's perspective. Limited training data sets in terms of diversity hinder the model's ability to make confident predictions for marginalized groups. And even with a near-perfect model solution, knowledge gaps related to AI technology for all stakeholders delay widespread adaptation.

To fully realize the potential of AI and ML in neurological voice disorders management, interdisciplinary collaboration is essential. Researchers, clinicians, ML engineers, and policymakers must work together to develop robust, equitable, and clinically validated AI solutions. Future research should focus on improving data set diversity, enhancing model interpretability, and addressing ethical considerations to ensure patient safety and trust. Interdisciplinary collaboration accelerates AI translation into clinical practice, enhancing diagnosis, treatment, and patient outcomes in Neurological voice disorders.

Author Contributions

Study concept and design: Dongren Yao, Aki Koivu, and Kristina Simonyan. Acquisition of data: Aki Koivu and Dongren Yao. Analysis of data: Aki Koivu and Dongren Yao. Drafting the manuscript: Dongren Yao and Aki Koivu. Commenting on the manuscript: Kristina Simonyan. Study supervision: Kristina Simonyan. Obtaining funding: Kristina Simonyan.

Acknowledgments

This study was funded by the grants R01NS088160, R01DC011805, and P50DC01990 from the National Institutes of Health to K.S.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

Kristina Simonyan receives funding from the National Institutes of Health (R01NS088160, R01NS124228, R01DC011805, R01DC012545, R01DC019353, R01DE030464, P50DC01990), the Department of Defense, serves on the Scientific Advisory Board of the Tourette Association of America and the Voice Foundation. The other authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

1. A. Ghanouni, N. Jona, H. A. Jinnah, G. Kilic-Berkmen, S. Shelly, and A. M. Klein, "Demographics and Clinical Characteristics Associated With the Spread of New-Onset Laryngeal Dystonia," *Laryngoscope* 134, no. 5 (2024): 2295–2299, https://doi.org/10.1002/lary.31146.

2. C. Arnold, J. Gehrig, S. Gispert, C. Seifried, and C. A. Kell, "Pathomechanisms and Compensatory Efforts Related to Parkinsonian Speech," *NeuroImage: Clinical* 4 (2014): 82–97, https://doi.org/10.1016/j.nicl.2013.10.016.

3. E. Erickson-DiRenzo, C. K. Sung, A. L. Ho, and C. H. Halpern, "Intraoperative Evaluation of Essential Vocal Tremor in Deep Brain Stimulation Surgery," *American Journal of Speech-Language Pathology* 29, no. 2 (2020): 851–863, https://doi.org/10.1044/2019_AJSLP-19-00079.

4. K. Simonyan, J. Barkmeier-Kraemer, A. Blitzer, et al., "Laryngeal Dystonia: Multidisciplinary Update on Terminology, Pathophysiology, and Research Priorities," *Neurology* 96, no. 21 (2021): 989–1001, https://doi.org/10.1212/WNL.0000000000011922.

5. G. Logroscino, "Agreement Among Neurologists on the Clinical Diagnosis of Dystonia at Different Body Sites," *Journal of Neurology, Neurosurgery & Psychiatry* 74, no. 3 (2003): 348–350, https://doi.org/10.1136/jnnp.74.3.348.

6. L. A. Carrasco-Ribelles, J. Llanes-Jurado, C. Gallego-Moll, et al., "Prediction Models Using Artificial Intelligence and Longitudinal Data From Electronic Health Records: A Systematic Methodological Review," *Journal of the American Medical Informatics Association* 30, no. 12 (2023): 2072–2082, https://doi.org/10.1093/jamia/ocad168.

7. M. J. Bianco, P. Gerstoft, J. Traer, et al., "Machine Learning in Acoustics: Theory and Applications," *Journal of the Acoustical Society of America* 146, no. 5 (2019): 3590–3628, https://doi.org/10.1121/1.5133944.

8. T. Zhang, A. M. Bur, S. Kraft, et al., "Gender, Smoking History, and Age Prediction From Laryngeal Images," *Journal of Imaging* 9, no. 6 (2023): 109, https://doi.org/10.3390/jimaging9060109.

9. M. Eslami, C. Neuschaefer-Rube, and A. Serrurier, "Automatic Vocal Tract Landmark Localization From Midsagittal MRI Data," *Scientific Reports* 10, no. 1 (2020): 1468, https://doi.org/10.1038/s41598-020-58103-6.

10. J. Mei, C. Desrosiers, and J. Frasnelli, "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature," *Frontiers in Aging Neuroscience* 13 (2021): 633752, https://doi.org/10.3389/fnagi.2021.633752.

11. G. Smilarubavathy, S. M. Keerthana, R. Nidhya, T. Priscilla, and D. Pavithra, "Machine Learning in Healthcare: Unlocking Precision Diagnosis and Continuous Monitoring Through Voice Analysis," in *Smart Factories for Industry 50 Transformation*, eds. R. Nidhya, M. Kumar, S. Karthik, R. Anand, and S. Balamurugan (Wiley-Scrivener, 2025), 229–245.

12. K. Simonyan and B. Horwitz, "Laryngeal Motor Cortex and Control of Speech in Humans," *Neuroscientist* 17, no. 2 (2011): 197–208, https://doi.org/10.1177/1073858410386727.

13. S. Skodda, H. Rinsche, and U. Schlegel, "Progression of Dysprosody in Parkinson's Disease Over Time--A Longitudinal Study," *Movement Disorders* 24, no. 5 (2009): 716–722, https://doi.org/10.1002/mds.22430.

14. D. A. Rahn, 3rd, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory Impairment in Parkinson's Disease: Evidence From Nonlinear Dynamic Analysis and Perturbation Analysis," *Journal of Voice* 21, no. 1 (2007): 64–71, https://doi.org/10.1016/j.jvoice.2005.08.011.

15. K. Kitajima and W. J. Gould, "Vocal Shimmer in Sustained Phonation of Normal and Pathologic Voice," *Annals of Otology, Rhinology, & Laryngology* 85, no. 3 pt 1 (1976): 377–381, https://doi.org/10.1177/000348947608500308.

16. P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice Onset Time in Aphasia, Apraxia of Speech and Dysarthria: A Review," *Clinical Linguistics & Phonetics* 14, no. 2 (2000): 131–150, https://doi.org/10.1080/026992000298878.

17. S. Y. Lowell, R. H. Colton, R. T. Kelley, and Y. C. Hahn, "Spectral- and Cepstral-Based Measures During Continuous Speech: Capacity to Distinguish Dysphonia and Consistency Within a Speaker," *Journal of Voice* 25, no. 5 (2011): e223–e232, https://doi.org/10.1016/j.jvoice.2010.06.007.

18. J. Ramírez, J. C. Segura, C. Benítez, Á. Torre, de la, and A. Rubio, "Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information," *Speech Communication* 42, no. 3 (2004): 271–287, https://doi.org/10.1016/j.specom.2003.10.002.

19. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, Speech Enhancement Based on Deep Denoising Autoencoder. 436–440.

20. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems* 27 (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

21. Y. LeCun, B. Boser, J. S. Denker, et al., "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation* 1, no. 4 (1989): 541–551.

22. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-Propagating Errors," *Nature* 323, no. 6088 (1986): 533–536.

23. Y. Wang, A. Mohamed, D. Le, et al., *Transformer-Based Acoustic Modeling for Hybrid Speech Recognition* (IEEE, 2020), 6874–6878.

24. N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition," *Proceedings of ICML Workshop on Deep Learning for Audio, Speech and Language* 117 (2013): 21.

25. D. Stoller, S. Ewert, and S. Dixon, "Wave-u-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proceedings of the 19th ISMIR Conference, Paris, France*, (2018), 23–27, http://www.jiashengli.cn/media/upfile/WAVE-U-NET_20230508025319_310.pdf.

26. M. Markaki and Y. Stylianou, "Voice Pathology Detection and Discrimination Based on Modulation Spectral Features," *IEEE Transactions on Audio, Speech, and Language Processing* 19, no. 7 (2011): 1938–1948.

27. E. Vaiciukynas, A. Verikas, A. Gelzinis, and M. Bacauskiene, "Detecting Parkinson's Disease From Sustained Phonation and Speech Signals," *PLoS One* 12, no. 10 (2017): e0185613, https://doi.org/10.1371/journal.pone.0185613.

28. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing* 10, no. 1–3 (2000): 19–41.

29. T. Pommée, Y. Maryn, C. Finck, and D. Morsomme, "The Acoustic Voice Quality Index, Version 03.01, in French and the Voice Handicap Index," *Journal of Voice* 34, no. 4 (2020): 646.e1–646.e10, https://doi.org/10.1016/j.jvoice.2018.11.017.

30. B. Barsties v. Latoszek, Y. Maryn, E. Gerrits, and M. De Bodt, "The Acoustic Breathiness Index (ABI): A Multivariate Acoustic Model for

Breathiness," *Journal of Voice* 31, no. 4 (2017): 511.e11–511.e27, https://doi.org/10.1016/j.jvoice.2016.11.017.

31. L. Naranjo, C. J. Pérez, Y. Campos-Roca, and J. Martín, "Addressing Voice Recording Replications for Parkinson's Disease Detection," *Expert Systems With Applications* 46 (2016): 286–292.

32. E. Almaloglou, G. S, G. Chrousos, and K. K, "Design and Validation of a New Diagnostic Tool for the Differentiation of Pathological Voices in Parkinsonian Patients," *Advances in Experimental Medicine and Biology* 1339 (2021): 77–83, https://doi.org/10.1007/978-3-030-78787-5_11.

33. Z. K. Senturk, "Layer Recurrent Neural Network-Based Diagnosis of Parkinson's Disease Using Voice Features," *Biomedical Engineering/Biomedizinische Technik* 67, no. 4 (2022): 249–266, https://doi.org/10.1515/bmt-2022-0022.

34. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems* 33 (2020): 12449–12460.

35. W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 3451–3460.

36. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *Proceedings of the 40th International Conference on Machine Learning*, (2023), 28492–28518.

37. F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, *Wav2vec-Based Detection and Severity Level Classification of Dysarthria From Speech* (IEEE, 2023), 1–5.

38. Y. Wang, A. Boumadane, and A. Heba, "A Fine-Tuned wav2vec 2.0/Hubert Benchmark for Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding," preprint, arXiv, arXiv:211102735, 2021.

39. P. Tzirakis, A. Kumar, and J. Donley, *Multi-Channel Speech Enhancement Using Graph Neural Networks* (IEEE, 2021), 3415–3419.

40. S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech Enhancement Generative Adversarial Network," preprint, arXiv, arXiv:170309452, 2017.

41. P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, *Wgansing: A Multi-Voice Singing Voice Synthesizer Based on the Wasserstein-Gan* (IEEE, 2019), 1–5.

42. R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, *A Study on Incorporating Whisper for Robust Speech Assessment* (IEEE, 2023), 1–6.

43. J. Prince, F. Andreotti, and M. De Vos, "Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data," *IEEE Transactions on Biomedical Engineering* 66, no. 5 (2019): 1402–1411, https://doi.org/10.1109/TBME.2018.2873252.

44. S. Park, C. No, S. Kim, et al., "A Multimodal Screening System for Elderly Neurological Diseases Based on Deep Learning," *Scientific Reports* 13, no. 1 (2023): 21013, https://doi.org/10.1038/s41598-023-48071-y.

45. A. Johnson, J. Bélisle-Pipon, D. Dorr, et al., "Data From: Bridge2AI-Voice: An Ethically-Sourced, Diverse Voice Dataset Linked to Health Information (version 1.1)," *PhysioNet* (2025), https://doi.org/10.13026/249v-w155.

46. W. Ji and Y. Li, "Energy-Based Feature Ranking for Assessing the Dysphonia Measurements in Parkinson Detection," *IET Signal Processing* 6, no. 4 (2012): 300–305.

47. A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease," *IEEE Transactions on Biomedical Engineering* 59, no. 5 (2012): 1264–1271, https://doi.org/10.1109/TBME.2012.2183367.

48. N. Erfanian Saeedi and F. Almasganj, "Wavelet Adaptation for Automatic Voice Disorders Sorting," *Computers in Biology and Medicine* 43, no. 6 (2013): 699–704, https://doi.org/10.1016/j.compbiomed.2013.03.006.

49. L. Verde, G. De Pietro, and G. Sannino, "Voice Disorder Identification by Using Machine Learning Techniques," *IEEE Access* 6 (2018): 16246–16255.

50. M. Yang, J. Ma, P. Wang, et al., "Hierarchical Boosting Dual-Stage Feature Reduction Ensemble Model for Parkinson's Disease Speech Data," *Diagnostics* 11, no. 12 (2021): 2312, https://doi.org/10.3390/diagnostics11122312.

51. E. Özbay, F. A. Özbay, N. Khodadadi, F. S. Gharehchopogh, and S. Mirjalili, "Multifeature Fusion Method With Metaheuristic Optimization for Automated Voice Pathology Detection," *Journal of Voice* (2024), https://doi.org/10.1016/j.jvoice.2024.08.018.

52. R. Lamba, T. Gulati, and A. Jain, "A Hybrid Feature Selection Approach for Parkinson's Detection Based on Mutual Information Gain and Recursive Feature Elimination," *Arabian Journal for Science and Engineering* 47, no. 8 (2022): 10263–10276.

53. C. C. Huang, Y. S. Leu, C. F. J. Kuo, W. L. Chu, Y. H. Chu, and H. C. Wu, "Automatic Recognizing of Vocal Fold Disorders From Glottis Images," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 228, no. 9 (2014): 952–961, https://doi.org/10.1177/0954411914551851.

54. N. Erfanian Saeedi, F. Almasganj, and F. Torabinejad, "Support Vector Wavelet Adaptation for Pathological Voice Assessment," *Computers in Biology and Medicine* 41, no. 9 (2011): 822–828, https://doi.org/10.1016/j.compbiomed.2011.06.019.

55. D. Montaña, Y. Campos-Roca, and C. J. Pérez, "A Diadochokinesis-Based Expert System Considering Articulatory Features of Plosive Consonants for Early Detection of Parkinson's Disease," *Computer Methods and Programs in Biomedicine* 154 (2018): 89–97, https://doi.org/10.1016/j.cmpb.2017.11.010.

56. S. Lahmiri and A. Shmuel, "Detection of Parkinson's Disease Based on Voice Patterns Ranking and Optimized Support Vector Machine," *Biomedical Signal Processing and Control* 49 (2019): 427–433.

57. K. Harimoorthy and M. Thangavelu, "Cloud-Assisted Parkinson Disease Identification System for Remote Patient Monitoring and Diagnosis in the Smart Healthcare Applications," *Concurrency and Computation: Practice and Experience* 33, no. 21 (2021): e6419.

58. S. R. Sharma, B. Singh, and M. Kaur, "Classification of Parkinson Disease Using Binary Rao Optimization Algorithms," *Expert Systems* 38, no. 4 (2021): e12674.

59. O. Oguz and H. Badem, "A New Metaheuristic Approach to Diagnosis of Parkinson's Disease Through Audio Signals," *Elektronika ir Elektrotechnika* 30, no. 4 (2024): 68–75.

60. M. Pramanik, R. Pradhan, P. Nandy, A. K. Bhoi, and P. Barsocchi, "Machine Learning Methods With Decision Forests for Parkinson's Detection," *Applied Sciences* 11, no. 2 (2021): 581.

61. A. Vaish, "A Machine Learning Approach for Early Identification of Prodromal Parkinson's Disease," *Cureus* 16, no. 6 (2024): e63240, https://doi.org/10.7759/cureus.63240.

62. S. Sheikhi and M. T. Kheirabadi, "An Efficient Rotation Forest-Based Ensemble Approach for Predicting Severity of Parkinson's Disease," *Journal of Healthcare Engineering* 2022 (2022): 5524852, https://doi.org/10.1155/2022/5524852.

63. P. Schlegel, A. M. Kist, M. Semmler, et al., "Determination of Clinical Parameters Sensitive to Functional Voice Disorders Applying Boosted Decision Stumps," *IEEE Journal of Translational Engineering in Health and Medicine* 8 (2020): 1–11, https://doi.org/10.1109/JTEHM.2020.2985026.

64. J. Skibińska and J. Hosek, "Computerized Analysis of Hypomimia and Hypokinetic Dysarthria for Improved Diagnosis of Parkinson's

Disease," *Heliyon* 9, no. 11 (2023): e21175, https://doi.org/10.1016/j.heliyon.2023.e21175.

65. N. A. N. Za'im, F. T. Al-Dhief, M. Azman, M. R. M. Alsemawi, N. Abdul Latiff, and M. Mat Baki, "The Accuracy of an Online Sequential Extreme Learning Machine in Detecting Voice Pathology Using the Malaysian Voice Pathology Database," *Journal of Otolaryngology - Head & Neck Surgery* 52, no. 1 (2023): 62, https://doi.org/10.1186/s40463-023-00661-6.

66. M. Pramanik, R. Pradhan, P. Nandy, S. M. Qaisar, and A. K. Bhoi, "Assessment of Acoustic Features and Machine Learning for Parkinson's Detection," *Journal of Healthcare Engineering* 2021 (2021): 9957132, https://doi.org/10.1155/2021/9957132.

67. D. Voigt, M. Döllinger, A. Yang, U. Eysholdt, and J. Lohscheller, "Automatic Diagnosis of Vocal Fold Paresis by Employing Phonovibrogram Features and Machine Learning Methods," *Computer Methods and Programs in Biomedicine* 99, no. 3 (2010): 275–288, https://doi.org/10.1016/j.cmpb.2010.01.004.

68. J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, et al., "Detecting Parkinson's Disease With Sustained Phonation and Speech Signals Using Machine Learning Techniques," *Pattern Recognition Letters* 125 (2019): 55–62.

69. L. Zhang, Y. Qu, B. Jin, L. Jing, Z. Gao, and Z. Liang, "An Intelligent Mobile-Enabled System for Diagnosing Parkinson Disease: Development and Validation of a Speech Impairment Detection System," *JMIR Medical Informatics* 8, no. 9 (2020): e18689, https://doi.org/10.2196/18689.

70. Z. Karapinar Senturk, "Early Diagnosis of Parkinson's Disease Using Machine Learning Algorithms," *Medical Hypotheses* 138 (2020): 109603, https://doi.org/10.1016/j.mehy.2020.109603.

71. S. Wan, Y. Liang, Y. Zhang, and M. Guizani, "Deep Multi-Layer Perceptron Classifier for Behavior Analysis to Estimate Parkinson's Disease Severity Using Smartphones," *IEEE Access* 6 (2018): 36825–36833.

72. S. H. Fang, Y. Tsao, M. J. Hsiao, et al., "Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach," *Journal of Voice* 33, no. 5 (2019): 634–641, https://doi.org/10.1016/j.jvoice.2018.02.003.

73. G. Pahuja and T. N. Nagabhushan, "A Comparative Study of Existing Machine Learning Approaches for Parkinson's Disease Detection," *IETE Journal of Research* 67, no. 1 (2021): 4–14.

74. T. P. R. Vital, J. Nayak, B. Naik, and D. Jayaram, "Probabilistic Neural Network-Based Model for Identification of Parkinson's Disease by Using Voice Profile and Personal Data," *Arabian Journal for Science and Engineering* 46, no. 4 (2021): 3383–3407.

75. M. Jyotiyana, N. Kesswani, and M. Kumar, "A Deep Learning Approach for Classification and Diagnosis of Parkinson's Disease," *Soft Computing* 26, no. 18 (2022): 9155–9165.

76. G. Muhammad, M. F. Alhamid, M. Alsulaiman, and B. Gupta, "Edge Computing With Cloud for Voice Disorder Assessment and Treatment," *IEEE Communications Magazine* 56, no. 4 (2018): 60–65.

77. M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Noth, "Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification," 2019 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2019, 717–720, https://doi.org/10.1109/EMBC.2019.8856972.

78. H. Gunduz, "Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets," *IEEE Access* 7 (2019): 115540–115551.

79. D. Valeriani and K. Simonyan, "A Microstructural Neural Network Biomarker for Dystonia Diagnosis Identified by a Dystonianet Deep Learning Platform," *Proceedings of the National Academy of Sciences* 117, no. 42 (2020): 26398–26405, https://doi.org/10.1073/pnas.2009165117.

80. L. Zahid, M. Maqsood, M. Y. Durrani, et al., "A Spectrogram-Based Deep Feature Assisted Computer-Aided Diagnostic System for Parkinson's Disease," *IEEE Access* 8 (2020): 35482–35495.

81. J. Lee and H.-J. Choi, "Deep Learning Approaches for Pathological Voice Detection Using Heterogeneous Parameters," *IEICE Transactions on Information and Systems* 103, no. 8 (2020): 1920–1923.

82. M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, et al., "Voice Pathology Detection and Classification Using Convolutional Neural Network Model," *Applied Sciences* 10, no. 11 (2020): 3723.

83. Z. Mou, W. Ye, C.-C. Chang, and Y. Mao, "The Application Analysis of Neural Network Techniques on Lexical Tone Rehabilitation of Mandarin-Speaking Patients With Post-Stroke Dysarthria," *IEEE Access* 8 (2020): 90709–90717.

84. M. A. Mohammed, M. Elhoseny, K. H. Abdulkareem, S. A. Mostafa, and M. S. Maashi, "A Multi-Agent Feature Selection and Hybrid Classification Model for Parkinson's Disease Diagnosis," *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, no. 2s (2021): 1–22.

85. H. C. Hu, S. Y. Chang, C. H. Wang, et al., "Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study," *Journal of Medical Internet Research* 23, no. 6 (2021): e25247, https://doi.org/10.2196/25247.

86. E. Majda-Zdancewicz, A. Potulska-Chromik, J. Jakubowski, M. Nojszewska, and A. Kostera-Pruszczyk, "Deep Learning vs Feature Engineering in the Assessment of Voice Signals for Diagnosis in Parkinson's Disease," *Bulletin of the Polish Academy of Sciences, Technical Sciences* 69, no. 3 (2021): 137347.

87. I. Kodrasi, "Temporal Envelope and Fine Structure Cues for Dysarthric Speech Detection Using Cnns," *IEEE Signal Processing Letters* 28 (2021): 1853–1857.

88. S. Gupta, A. T. Patil, M. Purohit, et al., "Residual Neural Network Precisely Quantifies Dysarthria Severity-Level Based on Short-Duration Speech Segments," *Neural Networks* 139 (2021): 105–117, https://doi.org/10.1016/j.neunet.2021.02.008.

89. B. Sonawane and P. Sharma, "Speech-Based Solution to Parkinson's Disease Management," *Multimedia Tools and Applications* 80, no. 19 (2021): 29437–29451.

90. L. Verde, N. Brancati, G. De Pietro, M. Frucci, and G. Sannino, "A Deep Learning Approach for Voice Disorder Detection for Smart Connected Living Environments," *ACM Transactions on Internet Technology* 22, no. 1 (2021): 1–16.

91. T. S. Mian, "An Unsupervised Neural Network Feature Selection and 1D Convolution Neural Network Classification for Screening of Parkinsonism," *Diagnostics* 12, no. 8 (2022): 1796, https://doi.org/10.3390/diagnostics12081796.

92. S. Fujimura, T. Kojima, Y. Okanoue, et al., "Classification of Voice Disorders Using a One-Dimensional Convolutional Neural Network," *Journal of Voice* 36, no. 1 (2022): 15–20, https://doi.org/10.1016/j.jvoice.2020.02.009.

93. S. Hidaka, Y. Lee, M. Nakanishi, K. Wakamiya, T. Nakagawa, and T. Kaburagi, "Automatic GRBAS Scoring of Pathological Voices Using Deep Learning and a Small Set of Labeled Voice Data," *Journal of Voice* (2022), https://doi.org/10.1016/j.jvoice.2022.10.020.

94. J. Reid, P. Parmar, T. Lund, D. K. Aalto, and C. C. Jeffery, "Development of a Machine-Learning Based Voice Disorder Screening Tool," *American Journal of Otolaryngology* 43, no. 2 (2022): 103327, https://doi.org/10.1016/j.amjoto.2021.103327.

95. D. Fu, X. Zhang, D. Chen, and W. Hu, "Pathological Voice Detection Based on Phase Reconstitution and Convolutional Neural Network," *Journal of Voice* 39 (2022): 2, https://doi.org/10.1016/j.jvoice.2022.08.028.

96. M. Hireš, M. Gazda, P. Drotár, N. D. Pah, M. A. Motin, and D. K. Kumar, "Convolutional Neural Network Ensemble for Parkinson's Disease Detection From Voice Recordings," *Computers in Biology and Medicine* 141 (2022): 105021, https://doi.org/10.1016/j.compbiomed.2021.105021.

97. G. Mary and N. Suganthi, "Detection of Parkinson's Disease With Multiple Feature Extraction Models and Darknet CNN Classification," *Computer Systems Science & Engineering* 43, no. 1 (2022): 333–345.

98. C. Quan, K. Ren, Z. Luo, Z. Chen, and Y. Ling, "End-to-End Deep Learning Approach for Parkinson's Disease Detection From Speech Signals," *Biocybernetics and Biomedical Engineering* 42, no. 2 (2022): 556–574.

99. C. H. Hung, S. S. Wang, C. T. Wang, and S. H. Fang, "Using SincNet for Learning Pathological Voice Disorders," *Sensors* 22, no. 17 (2022): 6634, https://doi.org/10.3390/s22176634.

100. A. Iyer, A. Kemp, Y. Rahmatallah, et al., "A Machine Learning Method to Process Voice Samples for Identification of Parkinson's Disease," *Scientific Reports* 13, no. 1 (2023): 20615, https://doi.org/10.1038/s41598-023-47568-w.

101. D. Aziz and S. Dávid, "Multitask and Transfer Learning Approach for Joint Classification and Severity Estimation of Dysphonia," *IEEE Journal of Translational Engineering in Health and Medicine* 12 (2024): 233–244, https://doi.org/10.1109/JTEHM.2023.3340345.

102. V. Skaramagkas, A. Pentari, D. I. Fotiadis, and M. Tsiknakis, "Using the Recurrence Plots as Indicators for the Recognition of Parkinson's Disease Through Phonemes Assessment," 2023 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2023, 1–4, https://doi.org/10.1109/EMBC40787.2023.10340177.

103. R. Kumari and P. Ramachandran, "Deep Convolution Neural Network Based Parkinson's Disease Detection Using Line Spectral Frequency Spectrum of Running Speech," *Journal of Intelligent & Fuzzy Systems* 45, no. 3 (2023): 4599–4615.

104. G. S. Liu, J. M. Hodges, J. Yu, C. K. Sung, E. Erickson-DiRenzo, and P. C. Doyle, "End-to-End Deep Learning Classification of Vocal Pathology Using Stacked Vowels," *Laryngoscope Investigative Otolaryngology* 8, no. 5 (2023): 1312–1318, https://doi.org/10.1002/lio2.1144.

105. D. Escobar-Grisales, C. D. Ríos-Urrego, and J. R. Orozco-Arroyave, "Deep Learning and Artificial Intelligence Applied to Model Speech and Language in Parkinson's Disease," *Diagnostics* 13, no. 13 (2023): 2163, https://doi.org/10.3390/diagnostics13132163.

106. M. Hireš, P. Drotár, N. D. Pah, Q. C. Ngo, and D. K. Kumar, "On the Inter-Dataset Generalization of Machine Learning Approaches to Parkinson's Disease Detection From Voice," *International Journal of Medical Informatics* 179 (2023): 105237, https://doi.org/10.1016/j.ijmedinf.2023.105237.

107. X. Xie, H. Cai, C. Li, Y. Wu, and F. Ding, "A Voice Disease Detection Method Based on MFCCs and Shallow CNN," *Journal of Voice* (2023), https://doi.org/10.1016/j.jvoice.2023.09.024.

108. Z. Chen, P. Zhu, W. Qiu, J. Guo, and Y. Li, "Deep Learning in Automatic Detection of Dysphonia: Comparing Acoustic Features and Developing a Generalizable Framework," *International Journal of Language & Communication Disorders* 58, no. 2 (2023): 279–294, https://doi.org/10.1111/1460-6984.12783.

109. G. Costantini, V. Cesarini, P. Di Leo, et al., "Artificial Intelligence-Based Voice Assessment of Patients With Parkinson's Disease Off and On Treatment: Machine vs. Deep-Learning Comparison," *Sensors* 23, no. 4 (2023): 2293, https://doi.org/10.3390/s23042293.

110. E. J. Ibarra, J. D. Arias-Londoño, M. Zañartu, and J. I. Godino-Llorente, "Towards a Corpus (and Language)-Independent Screening of Parkinson's Disease From Voice and Speech Through Domain Adaptation," *Bioengineering* 10, no. 11 (2023): 1316, https://doi.org/10.3390/bioengineering10111316.

111. P. Shanmugapriya and V. Mohan, "Comparative Analysis of Deep Learning Models for Dysarthric Speech Detection," *Soft Computing* 28, no. 6 (2024): 5683–5698.

112. İ. Cantürk and O. Günay, "Investigation of Scalograms With a Deep Feature Fusion Approach for Detection of Parkinson's Disease," *Cognitive Computation* 16 (2024): 1198–1209.

113. H. S. Malekroodi, N. Madusanka, B. Lee, and M. Yi, "Leveraging Deep Learning for Fine-Grained Categorization of Parkinson's Disease Progression Levels Through Analysis of Vocal Acoustic Patterns," *Bioengineering* 11, no. 3 (2024): 295, https://doi.org/10.3390/bioengineering11030295.

114. J. B. Lee and H. G. Lee, "Quantitative Analysis of Automatic Voice Disorder Detection Studies for Hybrid Feature and Classifier Selection," *Biomedical Signal Processing and Control* 91 (2024): 106014.

115. G. DİŞKen, "Multi-Label Voice Disorder Classification Using Raw Waveforms," *Turkish Journal of Electrical Engineering and Computer Sciences* 32, no. 4 (2024): 590–604.

116. M. Sha and M. P. Rahamathulla, "Quantum Deep Learning in Parkinson's Disease Prediction Using Hybrid Quantum–Classical Convolution Neural Network," *Quantum Information Processing* 23, no. 12 (2024): 383.

117. H. Majidinia, F. Khatib, S. J. Seyyed Mahdavi Chabok, H. R. Kobravi, and F. Rezaeitalab, "Diagnosis of Parkinson's Disease Using Convolutional Neural Network-Based Audio Signal Processing on FPGA," *Circuits, Systems, and Signal Processing* 43 (2024): 4221–4238.

118. B. Akila and J. J. V. Nayahi, "Parkinson Classification Neural Network With Mass Algorithm for Processing Speech Signals," *Neural Computing and Applications* 36 (2024): 10165–10181.

119. V. Guedes, F. Teixeira, A. Oliveira, et al., "Transfer Learning With Audioset to Voice Pathologies Identification in Continuous Speech," *Procedia Computer Science* 164 (2019): 662–669.

120. A. Lauraitis, R. Maskeliunas, R. Damasevicius, and T. Krilavicius, "Detection of Speech Impairments Using Cepstrum, Auditory Spectrogram and Wavelet Time Scattering Domain Features," *IEEE Access* 8 (2020): 96162–96172.

121. C. Quan, K. Ren, and Z. Luo, "A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech," *IEEE Access* 9 (2021): 10239–10252.

122. F. Demir, A. Sengur, A. Ari, K. Siddique, and M. Alswaitti, "Feature Mapping and Deep Long Short Term Memory Network-Based Efficient Approach for Parkinson's Disease Diagnosis," *IEEE Access* 9 (2021): 149456–149464.

123. M. B. Er, E. Isik, and I. Isik, "Parkinson's Detection Based on Combined CNN and LSTM Using Enhanced Speech Signals With Variational Mode Decomposition," *Biomedical Signal Processing and Control* 70 (2021): 103006.

124. T. Fujita, Z. Luo, C. Quan, K. Mori, and S. Cao, "Performance Evaluation of RNN With Hyperbolic Secant in Gate Structure Through Application of Parkinson's Disease Detection," *Applied Sciences* 11, no. 10 (2021): 4361.

125. N. Chintalapudi, G. Battineni, M. A. Hossain, and F. Amenta, "Cascaded Deep Learning Frameworks in Contribution to the Detection of Parkinson's Disease," *Bioengineering* 9, no. 3 (2022): 116, https://doi.org/10.3390/bioengineering9030116.

126. S. S. Wang, C. T. Wang, C. C. Lai, Y. Tsao, and S. H. Fang, "Continuous Speech for Improved Learning Pathological Voice Disorders," *IEEE Open Journal of Engineering in Medicine and Biology* 3 (2022): 25–33, https://doi.org/10.1109/OJEMB.2022.3151233.

127. D.-H. Shih, C.-H. Liao, T.-W. Wu, X.-Y. Xu, and M.-H. Shih, *Dysarthria Speech Detection Using Convolutional Neural Networks With Gated Recurrent Unit* (MDPI, 2022), 1956.

128. A. A. Bahaddad, M. Ragab, E. B. Ashary, and E. M. Khalil, "Metaheuristics With Deep Learning-Enabled Parkinson's Disease Diagnosis and Classification Model," *Journal of Healthcare Engineering* 2022 (2022): 9276579, https://doi.org/10.1155/2022/9276579.

129. M. Chaiani, S. A. Selouani, M. Boudraa, and M. Sidi Yakoub, "Voice Disorder Classification Using Speech Enhancement and Deep Learning Models," *Biocybernetics and Biomedical Engineering* 42, no. 2 (2022): 463–480.

130. F. Javanmardi, S. R. Kadiri, and P. Alku, "A Comparison of Data Augmentation Methods in Voice Pathology Detection," *Computer Speech & Language* 83 (2024): 101552.

131. J. Y. Han, C. J. Hsiao, W. Z. Zheng, et al., "Enhancing the Performance of Pathological Voice Quality Assessment System Through the Attention-Mechanism Based Neural Network," *Journal of Voice* (2023), https://doi.org/10.1016/j.jvoice.2022.12.026.

132. U. K. Lilhore, S. Dalal, N. Faujdar, et al., "Retracted Article: Hybrid CNN-LSTM Model With Efficient Hyperparameter Tuning for Prediction of Parkinson's Disease," *Scientific Reports* 13, no. 1 (2023): 14605, https://doi.org/10.1038/s41598-023-41314-y.

133. A. Rehman, T. Saba, M. Mujahid, F. S. Alamri, and N. ElHakim, "Parkinson's Disease Detection Using Hybrid LSTM-GRU Deep Learning Model," *Electronics* 12, no. 13 (2023): 2856.

134. A. Ksibi, N. A. Hakami, N. Alturki, M. M. Asiri, M. Zakariah, and M. Ayadi, "Voice Pathology Detection Using a Two-Level Classifier Based on Combined CNN–RNN Architecture," *Sustainability* 15, no. 4 (2023): 3204.

135. S. Tayebi Arasteh, T. Weise, M. Schuster, E. Noeth, A. Maier, and S. H. Yang, "The Effect of Speech Pathology on Automatic Speaker Verification: A Large-Scale Study," *Scientific Reports* 13, no. 1 (2023): 20476, https://doi.org/10.1038/s41598-023-47711-7.

136. T. D. Pham, S. B. Holmes, L. Zou, M. Patel, and P. Coulthard, "Diagnosis of Pathological Speech With Streamlined Features for Long Short-Term Memory Learning," *Computers in Biology and Medicine* 170 (2024): 107976, https://doi.org/10.1016/j.compbiomed.2024.107976.

137. R. F. Mansour, "Quantum Mayfly Optimization Based Feature Subset Selection With Hybrid CNN for Biomedical Parkinson's Disease Diagnosis," *Neural Computing and Applications* 36, no. 15 (2024): 8383–8396.

138. Y. N. Zhang, "Can a Smartphone Diagnose Parkinson Disease? A Deep Neural Network Method and Telediagnosis System Implementation," *Parkinson's Disease* 2017 (2017): 6209703, https://doi.org/10.1155/2017/6209703.

139. M. Masud, P. Singh, G. S. Gaba, et al., "CROWD: Crow Search and Deep Learning Based Feature Extractor for Classification of Parkinson's Disease," *ACM Transactions on Internet Technology* 21, no. 3 (2021): 1–18.

140. A. S. Almasoud, T. Abdalla Elfadil Eisa, F. N. Al-Wesabi, et al., "Parkinson's Detection Using RNN-Graph-LSTM With Optimization Based on Speech Signals," *Computers, Materials & Continua* 72 (2022): 871–886.

141. R. Khaskhoussy and Y. B. Ayed, "Speech Processing for Early Parkinson's Disease Diagnosis: Machine Learning and Deep Learning-Based Approach," *Social Network Analysis and Mining* 12, no. 1 (2022): 73.

142. M. T. García-Ordás, J. A. Benítez-Andrades, J. Aveleira-Mata, J.-M. Alija-Pérez, and C. Benavides, "Determining the Severity of Parkinson's Disease in Patients Using a Multi Task Neural Network," *Multimedia Tools and Applications* 83, no. 2 (2024): 6077–6092.

143. K. T. Chui, M. D. Lytras, and P. Vasant, "Combined Generative Adversarial Network and Fuzzy C-Means Clustering for Multi-Class Voice Disorder Detection With an Imbalanced Dataset," *Applied Sciences* 10, no. 13 (2020): 4571.

144. Z.-J. Xu, R.-F. Wang, J. Wang, and D.-H. Yu, "Parkinson's Disease Detection Based on Spectrogram-Deep Convolutional Generative Adversarial Network Sample Augmentation," *IEEE Access* 8 (2020): 206888–206900.

145. T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," preprint, arXiv, arXiv:180205957, 2018.

146. S. Zhao, G. Dai, J. Li, et al., "An Interpretable Model Based on Graph Learning for Diagnosis of Parkinson's Disease With Voice-Related EEG," *npj Digital Medicine* 7, no. 1 (2024): 3, https://doi.org/10.1038/s41746-023-00983-9.

147. A. Vaswani, "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30 (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

148. D. Ribas, M. A. Pastor, A. Miguel, D. Martinez, A. Ortega, and E. Lleida, "Automatic Voice Disorder Detection Using Self-Supervised Representations," *IEEE Access* 11 (2023): 14915–14927.

149. O. Klempíř, D. Příhoda, and R. Krupička, "Evaluating the Performance of wav2vec Embedding for Parkinson's Disease Detection," *Measurement Science Review* 23, no. 6 (2023): 260–267.

150. S. Tirronen, S. R. Kadiri, and P. Alku, "Hierarchical Multi-Class Classification of Voice Disorders Using Self-Supervised Models and Glottal Features," *IEEE Open Journal of Signal Processing* 4 (2023): 80–88.

151. D. Hemmerling, M. Wodzinski, J. R. Orozco-Arroyave, et al., "Vision Transformer for Parkinson's Disease Classification Using Multilingual Sustained Vowel Recordings," 2023 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2023, 1–4, https://doi.org/10.1109/EMBC40787.2023.10340478.

152. R. Nijhawan, M. Kumar, S. Arya, et al., "A Novel Artificial-Intelligence-Based Approach for Classification of Parkinson's Disease Using Complex and Large Vocal Features," *Biomimetics* 8, no. 4 (2023): 351, https://doi.org/10.3390/biomimetics8040351.

153. A. Zhao, N. Wang, X. Niu, M. Chen, and H. Wu, "A Triplet Multimodel Transfer Learning Network for Speech Disorder Screening of Parkinson's Disease," *International Journal of Intelligent Systems* 2024, no. 1 (2024): 8890592.

154. O. Klempíř and R. Krupička, "Analyzing Wav2Vec 1.0 Embeddings for Cross-Database Parkinson's Disease Detection and Speech Features Extraction," *Sensors* 24, no. 17 (2024): 5520, https://doi.org/10.3390/s24175520.

155. I. Tougui, M. Zakroum, O. Karrakchou, and M. Ghogho, "Transformer-Based Transfer Learning on Self-Reported Voice Recordings for Parkinson's Disease Diagnosis," *Scientific Reports* 14, no. 1 (2024): 30131, https://doi.org/10.1038/s41598-024-81824-x.

156. U. Irshad, R. Mahum, I. Ganiyu, et al., "UTran-DSR: A Novel Transformer-Based Model Using Feature Enhancement for Dysarthric Speech Recognition," *EURASIP Journal on Audio, Speech, and Music Processing* 2024, no. 1 (2024): 54.

157. N. Madusanka and B. Lee, "Vocal Biomarkers for Parkinson's Disease Classification Using Audio Spectrogram Transformers," *Journal of Voice* (2024), https://doi.org/10.1016/j.jvoice.2024.11.008.

158. A. D. Halai, A. M. Woollams, and M. A. Lambon Ralph, "Investigating the Effect of Changing Parameters When Building Prediction Models for Post-Stroke Aphasia," *Nature Human Behaviour* 4, no. 7 (2020): 725–735, https://doi.org/10.1038/s41562-020-0854-5.

159. A. Tankus, L. Solomon, Y. Aharony, A. Faust-Socher, and I. Strauss, "Machine Learning Algorithm for Decoding Multiple Subthalamic Spike Trains for Speech Brain-Machine Interfaces," *Journal of Neural Engineering* 18, no. 6 (2021): 066021, https://doi.org/10.1088/1741-2552/ac3315.

160. A. Suppa, F. Asci, G. Costantini, et al., "Effects of Deep Brain Stimulation of the Subthalamic Nucleus on Patients With Parkinson's Disease: A Machine-Learning Voice Analysis," *Frontiers in Neurology* 14 (2023): 1267360, https://doi.org/10.3389/fneur.2023.1267360.

161. A. Suppa, F. Asci, G. Saggio, et al., "Voice Analysis in Adductor Spasmodic Dysphonia: Objective Diagnosis and Response to Botulinum Toxin," *Parkinsonism & Related Disorders* 73 (2020): 23–30, https://doi.org/10.1016/j.parkreldis.2020.03.012.

162. D. S. Barbera, M. Huckvale, V. Fleming, et al., "NUVA: A Naming Utterance Verifier for Aphasia Treatment," *Computer Speech & Language* 69 (2021): 101221, https://doi.org/10.1016/j.csl.2021.101221.

163. A. Jain, K. Abedinpour, O. Polat, et al., "Voice Analysis to Differentiate the Dopaminergic Response in People With Parkinson's Disease," *Frontiers in Human Neuroscience* 15 (2021): 667997, https://doi.org/10.3389/fnhum.2021.667997.

164. D. Yao, L. C. O'Flynn, and K. Simonyan, "DystoniaBoTXNet: Novel Neural Network Biomarker of Botulinum Toxin Efficacy in Isolated Dystonia," *Annals of Neurology* 93, no. 3 (2023): 460–471, https://doi.org/10.1002/ana.26558.

165. M. Raza, M. Awais, N. Singh, M. Imran, and S. Hussain, "Intelligent IoT Framework for Indoor Healthcare Monitoring of Parkinson's Disease Patient," *IEEE Journal on Selected Areas in Communications* 39, no. 2 (2020): 593–602.

166. D. Mulfari, D. La Placa, C. Rovito, A. Celesti, and M. Villari, "Deep Learning Applications in Telerehabilitation Speech Therapy Scenarios," *Computers in Biology and Medicine* 148 (2022): 105864, https://doi.org/10.1016/j.compbiomed.2022.105864.

167. W. Rahman, A. Abdelkader, S. Lee, et al., "A User-Centered Framework to Empower People With Parkinson's Disease," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, no. 4 (2024): 1–29.

168. I. Mandal and N. Sairam, "Accurate Telemonitoring of Parkinson's Disease Diagnosis Using Robust Inference System," *International Journal of Medical Informatics* 82, no. 5 (2013): 359–377, https://doi.org/10.1016/j.ijmedinf.2012.10.006.

169. H. Ozkan, "A Comparison of Classification Methods for Telediagnosis of Parkinson's Disease," *Entropy* 18, no. 4 (2016): 115.

170. B. Erdogdu Sakar, G. Serbes, and C. O. Sakar, "Analyzing the Effectiveness of Vocal Features in Early Telediagnosis of Parkinson's Disease," *PLoS One* 12, no. 8 (2017): e0182428, https://doi.org/10.1371/journal.pone.0182428.

171. M. Alhussein and G. Muhammad, "Voice Pathology Detection Using Deep Learning on Mobile Healthcare Framework," *IEEE Access* 6 (2018): 41034–41041.

172. U. Cesari, G. De Pietro, E. Marciano, C. Niri, G. Sannino, and L. Verde, "Voice Disorder Detection via an m-Health System: Design and Results of a Clinical Study to Evaluate Vox4Health," *BioMed Research International* 2018 (2018): 8193694, https://doi.org/10.1155/2018/8193694.

173. O. Y. Chen, F. Lipsmeier, H. Phan, et al., "Building a Machine-Learning Framework to Remotely Assess Parkinson's Disease Using Smartphones," *IEEE Transactions on Biomedical Engineering* 67, no. 12 (2020): 3491–3500, https://doi.org/10.1109/TBME.2020.2988942.

174. M. S. R. Sajal, M. T. Ehsan, R. Vaidyanathan, S. Wang, T. Aziz, and K. A. A. Mamun, "Telemonitoring Parkinson's Disease Using Machine Learning by Combining Tremor and Voice Analysis," *Brain Informatics* 7, no. 1 (2020): 12, https://doi.org/10.1186/s40708-020-00113-1.

175. I. Tougui, A. Jilbab, and J. E. Mhamdi, "Analysis of Smartphone Recordings in Time, Frequency, and Cepstral Domains to Classify Parkinson's Disease," *Healthcare Informatics Research* 26, no. 4 (2020): 274–283, https://doi.org/10.4258/hir.2020.26.4.274.

176. N. Nonavinakere Prabhakera and P. Alku, "Automatic Assessment of Intelligibility in Speakers With Dysarthria From Coded Telephone Speech Using Glottal Features," *Computer Speech & Language* 65 (2021): 101117.

177. N. D. Pah, M. A. Motin, and D. K. Kumar, "Phonemes Based Detection of Parkinson's Disease for Telehealth Applications," *Scientific Reports* 12, no. 1 (2022): 9687, https://doi.org/10.1038/s41598-022-13865-z.

178. J. Carrón, Y. Campos-Roca, M. Madruga, and C. J. Pérez, "A Mobile-Assisted Voice Condition Analysis System for Parkinson's Disease: Assessment of Usability Conditions," *BioMedical Engineering OnLine* 20, no. 1 (2021): 114, https://doi.org/10.1186/s12938-021-00951-y.

179. S. Arora, C. Lo, M. Hu, and A. Tsanas, "Smartphone Speech Testing for Symptom Assessment in Rapid Eye Movement Sleep Behavior Disorder and Parkinson's Disease," *IEEE Access* 9 (2021): 44813–44824.

180. M. A. Motin, N. D. Pah, S. Raghav, and D. K. Kumar, "Parkinson's Disease Detection Using Smartphone Recorded Phonemes in Real World Conditions," *IEEE Access* 10 (2022): 97600–97609.

181. D. Worasawate, W. Asawaponwiput, N. Yoshimura, A. Intarapanich, and D. Surangsrirat, "Classification of Parkinson's Disease From Smartphone Recording Data Using Time-Frequency Analysis and Convolutional Neural Network," *Technology and Health Care* 31, no. 2 (2023): 705–718, https://doi.org/10.3233/THC-220386.

182. S. Mishra, L. Jena, N. Mishra, and H. T. Chang, "PD-Detector: A Sustainable and Computationally Intelligent Mobile Application Model for Parkinson's Disease Severity Assessment," *Heliyon* 10, no. 14 (2024): e34593, https://doi.org/10.1016/j.heliyon.2024.e34593.

183. T. He, J. Chen, X. Xu, and W. Wang, "Exploiting Smartphone Voice Recording as a Digital Biomarker for Parkinson's Disease Diagnosis," *IEEE Transactions on Instrumentation and Measurement* 73 (2024): 1–12.

184. S. Iliya and F. Neri, "Towards Artificial Speech Therapy: A Neural System for Impaired Speech Segmentation," *International Journal of Neural Systems* 26, no. 6 (2016): 1650023, https://doi.org/10.1142/S0129065716500234.

185. S. Chandrakala, S. Malini, and S. V. Veni, "Histogram of States Based Assistive System for Speech Impairment Due to Neurological Disorders," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021): 2425–2434, https://doi.org/10.1109/TNSRE.2021.3125314.

186. M. Chu, M. Yang, C. Xu, et al., "E-DGAN: An Encoder-Decoder Generative Adversarial Network Based Method for Pathological to Normal Voice Conversion," *IEEE Journal of Biomedical and Health Informatics* 27, no. 5 (2023): 2489–2500, https://doi.org/10.1109/JBHI.2023.3239551.

187. X. Pan, T. Feng, and N. Zhang, "PVGAN: A Pathological Voice Generation Model Incorporating a Progressive Nesting Strategy," *Journal of Voice* (2023), https://doi.org/10.1016/j.jvoice.2023.10.006.

188. Q. W. Oung, H. Muthusamy, S. N. Basah, H. Lee, and V. Vijean, "Empirical Wavelet Transform Based Features for Classification of Parkinson's Disease Severity," *Journal of Medical Systems* 42, no. 2 (2017): 29, https://doi.org/10.1007/s10916-017-0877-2.

189. C. Lo, S. Arora, F. Baig, et al., "Predicting Motor, Cognitive & Functional Impairment in Parkinson's," *Annals of Clinical and Translational Neurology* 6, no. 8 (2019): 1498–1509, https://doi.org/10.1002/acn3.50853.

190. J. C. Vasquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Noth, "Multimodal Assessment of Parkinson's Disease: A Deep Learning Approach," *IEEE Journal of Biomedical and Health Informatics* 23, no. 4 (2019): 1618–1630, https://doi.org/10.1109/JBHI.2018.2866873.

191. N. R. Yousif, H. M. Balaha, A. Y. Haikal, and E. M. El-Gendy, "A Generic Optimization and Learning Framework for Parkinson Disease via Speech and Handwritten Records," *Journal of Ambient Intelligence and Humanized Computing* 14 (2022): 10673–10693, https://doi.org/10.1007/s12652-022-04342-6.

192. W. S. Lim, S. I. Chiu, M. C. Wu, et al., "An Integrated Biometric Voice and Facial Features for Early Detection of Parkinson's Disease," *npj Parkinson's Disease* 8, no. 1 (2022): 145, https://doi.org/10.1038/s41531-022-00414-8.

193. M. Goni, S. B. Eickhoff, M. S. Far, K. R. Patil, and J. Dukart, "Smartphone-Based Digital Biomarkers for Parkinson's Disease in a Remotely-Administered Setting," *IEEE Access* 10 (2022): 28361–28384.

194. D. Dotov, V. Cochen de Cock, V. Driss, B. Bardy, and S. Dalla Bella, "Coordination Rigidity in the Gait, Posture, and Speech of Persons

With Parkinson's Disease," *Journal of Motor Behavior* 55, no. 4 (2023): 394–409, https://doi.org/10.1080/00222895.2023.2217100.

195. R. Indu, S. C. Dimri, and P. Malik, "A Modified kNN Algorithm to Detect Parkinson's Disease," *Network Modeling Analysis in Health Informatics and Bioinformatics* 12, no. 1 (2023): 24.

196. M. Sivakumar and K. Devaki, "Improved Glowworm Swarm Optimization for Parkinson's Disease Prediction Based on Radial Basis Functions Networks," *Information Technology and Control* 53, no. 2 (2024): 342–354.

197. D. Kumar, U. Satija, and P. Kumar, "Pathological Speech and Electroglottography Signals Analysis Using Invariance Scattering Network," *Circuits, Systems, and Signal Processing* (2024): 1–18.

198. E. H. Shortliffe, "The Adolescence of AI in Medicine: Will the Field Come of Age in the '90s," *Artificial Intelligence in Medicine* 5, no. 2 (1993): 93–106, https://doi.org/10.1016/0933-3657(93)90011-q.

199. International Organization for S. ISO/IEC 42001:2023—Information Technology—Artificial Intelligence—Management System, 2023.

200. Administration USFaD, "Performance Evaluation Methods for Evolving Artificial Intelligence (AI)-Enabled Medical Devices," 2023, https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/performance-evaluation-methods-evolving-artificial-intelligence-ai-enabled-medical-devices.

201. B. Murdoch, "Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era," *BMC Medical Ethics* 22, no. 1 (2021): 122, https://doi.org/10.1186/s12910-021-00687-3.

202. J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Speech Sanitizer: Speech Content Desensitization and Voice Anonymization," *IEEE Transactions on Dependable and Secure Computing* 18, no. 6 (2021): 2631–2642.

203. J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, "V-Cloak: Intelligibility-, Naturalness- & Timbre-Preserving Real-Time Voice Anonymization," 5181–5198.

204. Administration USFaD, "Artificial Intelligence and Machine Learning Software as a Medical Device Action Plan," 2021, https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

205. Regulation (EU) 2024 /1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, 2024, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

206. P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable Artificial Intelligence: An Analytical Review," *WIREs Data Mining and Knowledge Discovery* 11, no. 5 (2021): e1424.

207. S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. R. Müller, "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn," *Nature Communications* 10, no. 1 (2019): 1096, https://doi.org/10.1038/s41467-019-08987-4.

208. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study," *PLoS Medicine* 15, no. 11 (2018): e1002683, https://doi.org/10.1371/journal.pmed.1002683.

209. D. Saraswat, P. Bhattacharya, A. Verma, et al., "Explainable AI for Healthcare 5.0: Opportunities and Challenges," *IEEE Access* 10 (2022): 84486–84517.

210. M. Shen, P. Mortezaagha, and A. Rahgozar, "Explainable Artificial Intelligence to Diagnose Early Parkinson′ s Disease via Voice Analysis," *medRxiv*, 2024, 2024-09.

211. P. R. Magesh, R. D. Myloth, and R. J. Tom, "An Explainable Machine Learning Model for Early Detection of Parkinson's Disease Using LIME on DaTSCAN Imagery," *Computers in Biology and Medicine* 126 (2020): 104041, https://doi.org/10.1016/j.compbiomed.2020.104041.

212. T. Pianpanit, S. Lolak, P. Sawangjai, T. Sudhawiyangkul, and T. Wilaiprasitporn, "Parkinson's Disease Recognition Using SPECT Image and Interpretable AI: A Tutorial," *IEEE Sensors Journal* 21, no. 20 (2021): 22304–22316.

213. E. Mancini, F. Paissan, P. Torroni, M. Ravanelli, and C. Subakan, "Investigating the Effectiveness of Explainability Methods in Parkinson's Detection From Speech," preprint, arXiv, arXiv:241108013, 2024.

214. W. M. Kouw and M. Loog, "An Introduction to Domain Adaptation and Transfer Learning," preprint, arXiv, arXiv:181211806, 2018.

215. J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The Practical Implementation of Artificial Intelligence Technologies in Medicine," *Nature Medicine* 25, no. 1 (2019): 30–36, https://doi.org/10.1038/s41591-018-0307-0.