IEEE Open Journal of the Computer Society

Received 1 October 2024; revised 7 November 2024; accepted 20 November 2024. Date of publication 22 November 2024; date of current version 9 January 2025. The review of this article was arranged by Associate Editor Li Li.

Digital Object Identifier 10.1109/OJCS.2024.3504864

Time Series Classification of Raw Voice Waveforms for Parkinson's Disease Detection Using Generative Adversarial Network-Driven Data Augmentation

MARTA REY-PAREDES [1], CARLOS J. PÉREZ [2], AND ALFONSO MATEOS-CABALLERO [1]

[1]Departamento de Inteligencia Artificial, ETSIINF, Universidad Politécnica de Madrid, 28660 Madrid, Spain
[2]Departamento de Matemáticas, Universidad de Extremadura, 10003 Cáceres, Spain

CORRESPONDING AUTHOR: CARLOS J. PéREZ (e-mail: carper@unex.es).

This work was supported in part by the R&D&I projects under Grant PID2021-122209OB-C31 and Grant PID2021-122209OB-C32 and in part by the MICIU/AEI/10.13039/501100011033/ FEDER, UE.

ABSTRACT Parkinson's disease (PD) is a neurodegenerative disorder that affects more than 10 million people worldwide. Despite its prevalence, the detection of PD remains a complicated task, as no gold standard test has yet been developed to provide an accurate diagnosis. In this context, many recent studies have focused on the automatic detection and progression tracking of PD from voice-related characteristics, being feature engineering the most common approach. This work intends to address an existing research gap by introducing a novel strategy that analyzes raw voice waveforms. Despite recent advancements, one of the significant hurdles is still the lack of extensive and diverse datasets. This article also implements a data augmentation solution. Big Vocoder Slicing Adversarial Network (BigVSAN) is used to generate synthetic voice data that mimics the characteristics of real patients and healthy subjects. For the PD detection task, deep learning models such as ResNet, LSTM-FCN, InceptionTime, and CDIL-CNN are used. The experiments were performed using the speech task of sustained vowel /a/ in the PC-GITA database, which contains the recordings of healthy and PD subjects. CDIL-CNN achieves the best results, improving the accuracy by 15.87% (8.96%) compared to the model that does not use augmented data (from the best method found in the literature that uses voice waveforms). The results of this study indicate that models trained with raw waveforms showcase modest but promising performance, underlying the potential of audio analysis to improve the early detection of PD, providing a non-invasive and potentially remotely applicable method.

INDEX TERMS Deep learning, generative adversarial networks, Parkinson's disease, vocal signal analysis.

I. INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease. There are more than 10 million people estimated to live with it worldwide [1], and its prevalence is rising due to population aging [2].

PD is characterized by the gradual death of dopamine-producing neurons in the sustantia nigra pars compacta, a region of the brain that regulates motor coordination and movement control. Thus, dopamine deficiency leads to significant physical and neurological impairments [3]. Clinically, the disorder is identified by resting tremor, rigidity, bradykinesia, and gait impairments, and, in addition, it may manifest other features such as postural instability, speech difficulties, autonomic disturbances, sensory changes, mood disorders, sleep dysfunction, cognitive impairment, or dementia [3].

The neurodegenerative processes in PD are thought to start a few years before the motor symptoms that serve as the basis for the diagnosis. There is no gold standard test in the diagnosis of PD. As a result, detection is made through clinical evaluation, by history-taking, and physical examination of the signs and symptoms of the patient. This evaluation is

subjective and susceptible to human error and can take from months to years [4]. These difficulties in the diagnosis motivate the development of methods for early detection of the prodromal phase of the disease, which have gained particular interest in recent years [5].

Speech production is affected in PD subjects, so voice and speech articulation are thought to be suitable options for PD detection [6]. Even in the early stages of PD, patients may exhibit noticeable changes in vocal characteristics, collectively known as hypokinetic dysarthria. These changes can manifest in every domain of human speech production, with more emphasis on the areas of articulation, phonation, speech fluency, and facial movements. Observed impairments include decreased vocal intensity, harsh and breathy voice quality, heightened voice nasality, monotonous speech, reduction of loudness, involuntary pauses, imprecise consonant articulation, and disturbances in speech rate [7]. The presence of recognizable speech disorders in PD patients is driving the development of non-invasive voice-based assessment technologies that enable early diagnosis and remote monitoring of hypokinetic dysarthria [8], [9].

Numerous studies have been published recently that point to the possibility of using voice features as biomarkers in the creation of automated PD screening tools [10]. The acoustic analysis may be applied to the classification task to distinguish between speech recordings representing healthy controls (HC) or PD cases. Recent advancements in artificial intelligence have led to increased interest in applying diverse machine learning techniques to analyze speech data for PD detection, reducing diagnostic time and associated costs. Such techniques may include the application of Support Vector Machines, K-Nearest Neighbors, Naive Bayes classifiers, or Decision Trees [11], [12]. Of particular interest are Deep Neural Networks (DNNs), primarily employed due to their ability to model complex patterns and dependencies in data [13].

Whereas traditional approaches in voice-based PD detection have primarily focused on feature engineering, such as extracting spectrograms [14], [15] or other statistical features from voice recordings [16], [17], little research has been done on the direct analysis of raw Time Series (TS) data from speech recordings [18], which has created a significant research gap. TS analysis could potentially reveal more intricate patterns in the speech of PD patients that have not been captured by previously studied methods, which motivates the exploitation of raw speech data without relying on extensive feature extraction.

In spite of these recent advancements, one of the main impediments to the development of robust voice-based diagnostic tools for PD is still the scarcity of extensive and diverse datasets [19]. Most existing databases are constrained by the number of individuals, and subject to strict privacy regulations, which limit the variability of recording conditions, producing a lack of meaningful data capturing disease progression. These limitations restrict the ability of DNNs to generalize well beyond their training data, leading to models that may perform well in controlled experimental settings but fail in real-world applications [20].

Motivated by the challenge of limited labeled data, which can significantly degrade the performance of classification models, this work proposes using Generative Adversarial Networks (GANs) for data augmentation. Traditional augmentation methods are limited to producing lower-quality samples based on simple transformations. In contrast, GANs can generate synthetic voice data that mimics real patient recordings, providing a more diverse and richer dataset without relying on predetermined augmentation techniques. By enhancing the dataset with realistic, artificially created voice samples, the robustness and accuracy of PD detection models can be significantly improved.

The main contributions of this work can be summarized as follows:

- Provide an alternative approach for PD detection by using raw voice signals as biomarkers.
- Implement a novel pipeline that combines the generation of synthetic data using GANs with raw waveform classification.
- Demonstrate that the data generated mimic the vocal characteristics of patients with PD and can improve the generalizability of the models in real scenarios.
- Validate different DNN architectures for speech classification on this particular task.

The remainder of the article is organized as follows. Section II discusses the related work. Section III describes the data used, its preprocessing, and the proposed methodology both for data augmentation and PD detection. Section IV summarizes the results obtained throughout the work. Finally, the main conclusions extracted are in Section V.

II. RELATED WORK

The vast majority of strategies found so far extract temporal features from the raw audio signals or convert the temporal signals into a time-frequency domain by representing it with spectrograms or some variations such as mel-spectrograms, or even extract spectral characteristics, such as *Bark Spectrogram Cepstral Coefficients*, *Mel Spectrogram Cepstral Coefficients*, *Bark Frequency Cepstral Coefficients* or *Mel-Frequency Cepstral Coefficients* [21]. A hybrid U-lossian DNN was proposed in [22] for PD screening and detection, converting the voice signals into spectrograms, and then extracting the *Mel-Frequency Cepstral Coefficients*. Two datasets were employed, Italian PVS [23] and Lithuanian PD voice dataset. In [24], *Hilbert Cepstral Coefficients* were investigated as new features, and evaluated using vowels and words from the PC-GITA database, which includes a balanced number of recordings. The *Hilbert Cepstral Coefficients* were utilized for classification using a Multi-Layer Perceptron (MLP) model evaluated with the vowel /a/ the word /apto/.

Regarding the use of vocal features, a novel hybrid technique for early detection of PD was presented in [25]. Data was retrieved from 68 participants from the UCI Machine

Learning Repository [26]. Weights from an MLP were used for feature selection. Posteriorly, these selected features were input into a Lagrangian Support Vector Machine for classification. In [16], two approaches using Convolutional Neural Networks (CNNs) with 9 layers were presented. In the first approach, different vocal feature sets were combined before feeding them as input to the network. In the second, the feature sets were passed through parallel input layers connected to convolutional layers, and the obtained features from each branch were then combined. In [27], the goal was to extract dynamic time features from speech signals by using a bidirectional Long Short-Term Memory (LSTM) to improve PD detection accuracy compared to approaches using static features. A voice-based diagnosis model for PD was introduced in [28] using a Recurrent Neural Network (RNN) with 300 vocal features from 252 subjects of the UCI ML Repository as input.

Several studies have also explored the use of spectrograms for PD detection, by converting the audio signals into time-frequency representations. Spectrograms can be considered as images and used as input for popular image classification models. A novel method was presented in [14] that uses spectrogram representations as input for the ResNet architecture. The model was first pretrained on the ImageNet database and Saarbruecken Voice Database and the PD data was sourced from the PC-GITA database. A framework that utilizes time-distributed 2D-CNNs followed by a 1D-CNN was proposed in [29] to extract dynamic features from TS data and then capture dependencies between them. Data was extracted from sustained vowel /a/ sounds belonging to two databases, PC-GITA and another collected in GYENNO SCIENCE Parkinson's Disease Research Center. The work of [30] introduced a novel approach for detecting PD using mel-spectrograms derived from denoised speech signals of the PC-GITA database, processed with Variational Mode Decomposition. ResNet-18, ResNet-50, and ResNet-101 models were used as pretrained deep learning models to extract features from the mel-spectrogram representations, and then, those features were passed to the LSTM model for the final classification stage. In [31], a hybrid model that performed dynamic feature extraction from mel-spectrograms with a pretrained CNN model, ResNet-50 was used, and then, an LSTM network was applied for the final classification [31]. A CNN-LSTM hybrid model to facilitate the detection of PD was proposed in [32], extracting spectrogram representations from 22 HC and 28 PD patients of the Italian PVS database.

To date, the effectiveness of deep learning models for Time Series Classification (TSC) using raw audio signals (sustained vowels, words, or sentences) as input, and without performing any transformations on them, has not been properly studied. To the best of the author's knowledge, only one study has been reported that uses the raw speech signal as input for a model to detect PD [33]. This approach uses deep learning models trained on raw speech and raw voice source waveforms. Raw speech waveforms refer to recordings of patients with sustained phonations, reading words, short sentences, or monologues. On the other hand, the authors also utilized glottal flow waveforms extracted through different filtering methods that process the voice source waveforms. The proposed architecture combined CNNs and one MLP.

III. METHODOLOGY

In this section, the proposed methodological framework is explained, as depicted in Fig. 1. It includes three separate parts: first, the dataset used and the preprocessing techniques applied are explained; then, the methodology used for data augmentation through data synthesis is presented; finally, the approach used for PD detection is described. The proposed methodology is implemented in a computer with a Nvidia A100 GPU with 80GB of memory.

A. DATASET AND PREPROCESSING

1) DATASET DESCRIPTION

The PC-GITA [34] speech database was specifically designed to analyze the speech of individuals with PD. The database includes voice recordings in Spanish from 100 Colombian speakers, containing 50 PD patients (25 men, age $= 62.2 \pm 11.2$ years; 25 women, age $= 60.1 \pm 7.8$ years) and 50 HC speakers matched in age and gender (25 men, age $= 61.2 \pm 11.3$ years; 25 women, age $= 60.7 \pm 7.7$ years). All participants signed an informed consent.

The voice recordings were conducted under controlled noise conditions using a soundproof booth with a Shure SM63L microphone and a professional audio card. Data was obtained with a sampling frequency of 44.1 kHz and a resolution of 16 bits. Participants with PD were diagnosed by neurologists and then classified according to the Unified Parkinson's Disease Rating Scale, and subjects with HC showed no symptoms of PD or other neurological diseases.

2) PREPROCESSING

The sustained phonations of the vowel /a/ have been considered. The data consists of 300 instances, comprising the recordings of 100 participants who repeated the phonation task three times.

The first step includes reducing the sampling frequency of all recordings. The original recordings were obtained at a sampling frequency of 44.1 kHz, but the generative model works with 22 kHz or 24 kHz. Therefore, it is decided to use the maximum possible value, downsampling the signals from 44.1 kHz to 24 kHz. The second task involves ensuring that all recordings have the same duration, so the information obtained from them is comparable to each other. Hence, the shortest duration among all the audio recordings is sought and all the data are trimmed to that length. The shortest duration corresponds to 480 milliseconds, so, all 300 recordings are adjusted to that length, obtaining 11,520 timesteps per sequence.

A visual study of amplitude differences between HC and PD subjects is also conducted, and the mean of each amplitude distribution in every sample data is calculated. Fig. 2 illustrates in boxplots the distribution of these amplitudes,

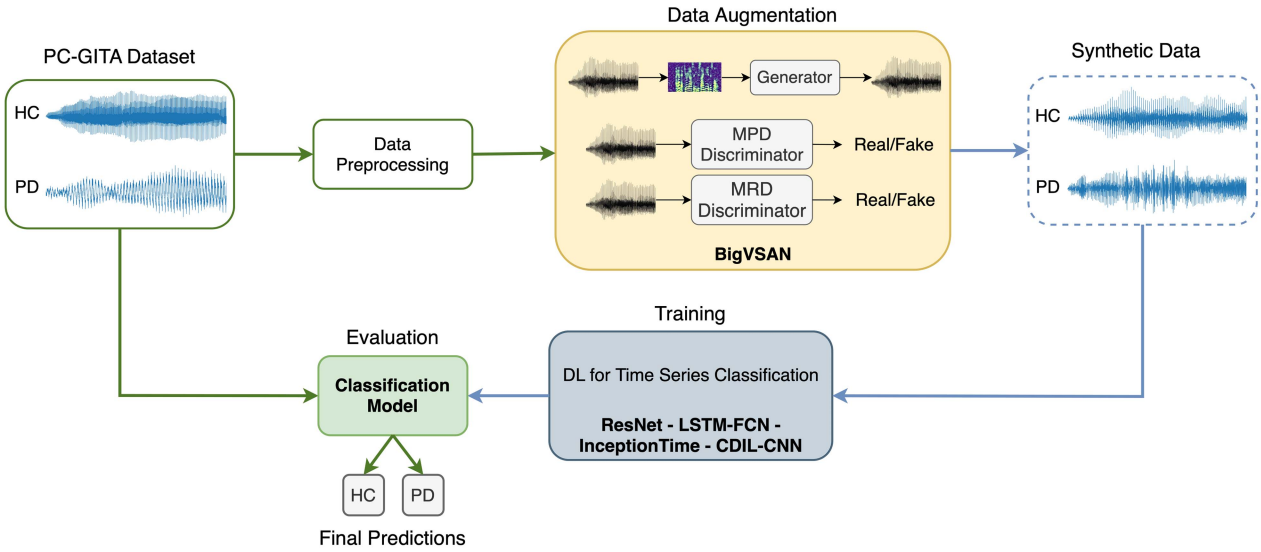

FIGURE 1. Methodological framework overview.

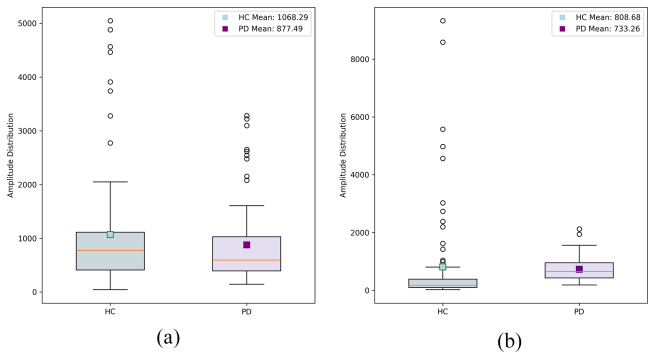

FIGURE 2. Boxplots representing amplitude distributions in (a) women and (b) men for HC and PD. Orange line represents the median; green and purple squares, the respective means; horizontal lines, the upper and lower quartiles.

segregated by patient type (HC or PD) and gender (female or male). PD patients show, in general, a lower amplitude range in both males and females. In the case of HC men, the majority of the recordings are observed to exhibit a low amplitude compared to PD. The remaining, in comparison, have a very high mean amplitude, which are regarded as outliers. Despite the visible variations in amplitude distributions, it can be inferred that these differences are influenced by the recording conditions and the inherent voice volume of each subject. Therefore, it is decided that each audio sample should be individually normalized using min-max normalization to avoid the influence of the voice volume of each subject.

Finally, given that the recordings were made in a controlled recording environment, it is assumed that a further preprocessing step to remove noise is not necessary, since external noise and environmental disturbances were already addressed. An analysis of the waveforms is conducted, where it is revealed that the normalized amplitude after preprocessing is still significantly higher for the HC cases, demonstrating vocal strength. A more complex and dense pattern is observed, with more variability, and a greater number of peaks, which may be indicative of a clearer vocalization. In contrast, PD waveforms showcase lower variability, resulting in a more monotonous sound and difficulties in vocalization, reflecting possible troubles in sound production. The mentioned observations are consistent across both genders, in line with studies on the impact of PD on voice production [35], [36].

B. DATA AUGMENTATION WITH GENERATIVE ADVERSARIAL NETWORK

1) BIGVSAN

BigVSAN is a generative model designed for high-fidelity audio synthesis [37]. The generator function, denoted as G: $\boldsymbol{S} \rightarrow \hat{\boldsymbol{X}}$ converts the mel-spectrogram $\boldsymbol{s} \in \boldsymbol{S}$ input into a waveform signal $\hat{\boldsymbol{x}} \in \hat{\boldsymbol{X}}$. BigVSAN generator is made up of a series of blocks to increase the time resolution of the mel-spectrogram and synthesize a high-quality audio signal. The main steps of the generator include:

1) *Periodic Activation:* The *Snake* periodic activation function is used, defined as:

$$f_\alpha(\boldsymbol{x}) = \boldsymbol{x} + \frac{1}{\alpha}\sin^2(\alpha \boldsymbol{x}), \tag{1}$$

where α is a trainable parameter that controls the frequency of the periodic component and $\boldsymbol{x}$ is the input to the activation function.

2) *Anti-aliased Multi-periodicity Composition (AMP):* This module aggregates various signal components using learnable periodicities and applies a low-pass filter to reduce high-frequency artifacts. It consists of residual dilated convolutional layers with Snake activations.

3) *Upsampling:* The generator increases the temporal resolution of the signal through 1D transposed convolution blocks followed by AMP layers.

The discriminator takes real $\boldsymbol{x}$ and generates $\hat{\boldsymbol{x}}$ input waveforms, transforms it into a mel-spectrogram representation $\boldsymbol{s}$ and outputs the waveform together with the probability distribution. It is composed of a combination of discriminators operating in different waveform resolution windows:

- *Multi-period discriminator (MPD):* Reorganizes the 1D signals into 2D representations with varying sizes to identify multiple periodic structures with 2D convolutions.
- *Multi-resolution discriminator (MRD):* Applies discriminators in the time-frequency domain using linear spectrograms of different Short-Time Fourier Transform (STFT) resolutions.

For an individual ground-truth waveform $\boldsymbol{x}$ and mel-spectrogram $\boldsymbol{s}$, the general objective functions are denoted as $\mathcal{L}_G$ for the generator, which is to be minimized, and $\mathcal{L}_D$ for the discriminator, which is to be maximized:

$$\mathcal{L}_G = \sum_{k=1}^{K} [\mathcal{L}_{\text{adv}}(G; D_k) + \lambda_{\text{fm}} \mathcal{L}_{\text{fm}}(G; D_k)] + \lambda_{\text{mel}} \mathcal{L}_{\text{mel}}(G), \tag{2}$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \mathcal{L}_{\text{adv}}(D_k; G), \tag{3}$$

where D_k denotes the k-th MPD or MRD submodules. The global generator loss is composed of three losses. $\mathcal{L}_{\text{fm}}$ is the feature matching loss, $\mathcal{L}_{\text{mel}}$ represents the mel-spectrogram loss and $\mathcal{L}_{\text{adv}}$ is the general adversarial loss. The scalar parameters λ_{fm} and λ_{mel} balance the GAN losses. The least-square GAN is included in the original adversarial loss $\mathcal{L}_{\text{adv}}$ as follows:

$$\mathcal{L}_{\text{adv}}(G; D_k) = \mathbb{E}_{\boldsymbol{s}} \left[(D_k(G(\boldsymbol{s})) - 1)^2 \right], \tag{4}$$

$$\mathcal{L}_{\text{adv}}(D_k; G) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})} \left[(D_k(\boldsymbol{x}) - 1)^2 + (D_k(G(\boldsymbol{s})))^2 \right]. \tag{5}$$

BigVSAN introduces a *soft monotonization* technique using the *softplus* function $\zeta(\cdot)$, i.e. $\zeta(a) = \log(1 + e^a)$, to convert least-squares GAN to least-squares Slicing Adversarial Network (SAN), enhancing the discriminative capacity of the model.

The final min-max objectives are constructed for least-squares SAN by adding *softplus* function to the adversarial losses, modifying (4) and (5) as:

$$\mathcal{L}_{\text{adv}}(G; D_k) = \mathbb{E}_{\boldsymbol{s}} \left[\zeta(D_k(G(\boldsymbol{s})) - 1)^2 \right], \tag{6}$$

$$\mathcal{L}_{\text{adv}}(D_k; G) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})} \left[\zeta(D_k(\boldsymbol{x}) - 1)^2 + \zeta(D_k(G(\boldsymbol{s})))^2 \right]. \tag{7}$$

2) GAN EVALUATION METRICS

The objective metrics for evaluating the BigVSAN model are designed to compare the generated data to the ground-truth audios across various types of distances, assessing the quality of the generated data. The five metrics proposed for the evaluation are the following:

- *Multi-resolution short-time Fourier Transform* (M-STFT) measures spectral distances across multiple resolutions, computing the STFT several times with varying parameters such as the window size or frameshift [38]:

$$L_{\text{M-STFT}}(G) = \frac{1}{N} \sum_{i=1}^{N} L_{\text{s}}^{(i)}(G), \tag{8}$$

  where L_{s} represents a single STFT loss in the generator and N is the number of calculated losses. M-STFT values are always positive and unbounded, with values closer to 0 indicating better performance.
- *Perceptual Evaluation of Speech Quality* (PESQ) provides an automated assessment of generated signal perceived quality with the original reference signal by modeling the perceptions of the human ear and brain [39]. It aligns the signals in time, transforms them into perceptual domains, calculates the disturbance parameters, and aggregates all parameters into a mean opinion score:

$$\text{PESQ MOS} = 4.5 - 0.1 \times d_{\text{SYM}} - 0.0309 \times d_{\text{ASYM}}, \tag{9}$$

  where d_{SYM} and d_{ASYM} are symmetric and asymmetric disturbances, respectively. It ranges from -0.5 to 4.5, and higher values indicate better-perceived voice quality.
- *Mel-Cepstral Distortion* (MCD) is a distance that measures the difference between the mel-cepstral coefficients extracted from the original and the generated audio [40].

$$\text{MCD} = \frac{10}{\ln 10} \sqrt{2 \sum_{i=1}^{N} (c_i - \hat{c}_i)^2}, \tag{10}$$

  where c_i and $\hat{c}_i$ represent the i-th coefficients of the reference and generated signals respectively. Positive values with no upper bound are obtained, seeking smaller values.
- *Periodicity error:* The periodicity of a signal represents the regularity in the repetition of patterns. Thus, the periodicity error measures the differences between the periodicity of the original speech signal and the generated one [41]:

$$\text{Periodicity Error} = \frac{1}{N} \sum_{i=1}^{N} |p_i - \hat{p}_i|, \tag{11}$$

  where p_i and $\hat{p}_i$ represent the periodicity measures of the original and generated signals respectively, and N, the number of measures taken. Periodicity error takes positive, unbounded values, seeking values closer to 0.
- *F1-score of Voiced/Unvoiced Classification* (V/UV F1) measures the effectiveness of the model in distinguishing voiced sounds (containing a clear fundamental frequency) or unvoiced sounds (with less regular patterns,

and less clear fundamental frequency) [41]. V/UV F1 values range from 0 to 1, where higher results are sought.

3) EXPERIMENTAL SETUP

In traditional GANs, the input consists of a noise vector that is modeled by the generator to output the desired vector. However, in BigVSAN, real data has a twofold function since it is used as input for the generator and also as a reference for the discriminator. The input for the generator is a mel-spectrogram representation, transformed from the input audio. This transformation is straightforward. However, the challenge of the generator lies in learning to reconstruct the audio sample from the mel-spectrogram provided. This is not a simple task, since the time-frequency representation of sounds lacks certain key aspects needed to reconstruct high-quality audio. The generative model is used independently to generate the HC and PD data.

For this study, a pretrained BigVSAN model is used to generate new audio samples of the HC and PD subjects. BigVSAN model was trained in the work of [37] for 10 million steps on the LibriTTS dataset [42] at a sampling rate of 24 kHz taking a full 100-band mel-spectrogram as input.

During the synthesis process, a mel-spectrogram is calculated from each true audio of the loaded PC-GITA dataset. The generated mel-spectrograms are then fed into the BigVSAN model, which processes them and reconstructs the audio samples.

Since the pretrained BigVSAN model is already providing consistent results, only an inference process is conducted in this study, with no fine-tuning of the model on the PC-GITA dataset required. Conversely, when the same checkpoints are loaded and the same input audio files are used, the generated audio recordings will be identical, as different noise vectors cannot be introduced at each iteration. Therefore, the GAN can only generate a number of synthetic samples equal to that of the input data. To address this problem, two approaches are proposed:

- *Randomization factor:* A randomization factor (μ) that modifies the pretrained weights has been introduced:

$$\boldsymbol{W}' = \boldsymbol{W} + \mu \cdot \mathcal{N}(0, 1), \tag{12}$$

  where $\boldsymbol{W}$ represents the original weights of the model, $\mathcal{N}(0, 1)$ is the Gaussian noise and $\boldsymbol{W}'$ denotes the new adjusted weights after noise is added. These small, controlled perturbations are temporary and do not permanently alter the weights learned in the training process, but only affect their use during the inference process. This approach allows repeating the inference process as many times as needed, generating slightly different versions of each input audio file. Multiple synthetic audio samples can be generated from the same input dataset, enriching the dataset available for the posterior model training and evaluation. This method increases the diversity of the dataset but the randomization factor has to be precisely adjusted.
- *Sequential generation:* All the real audio files of a certain group are introduced in the first round. Subsequently, each batch of generated audio serves as input for generating the next batch. This strategy introduces controlled variability and ensures a gradual diversification of the dataset.

C. PARKINSON'S DISEASE DETECTION

1) RESNET

ResNet network was first introduced for image recognition tasks, but posteriorly adapted for TSC [43]. Its architecture contains skip connections that appear between consecutive convolutional layers, facilitating the gradient flow to be transmitted directly through these connections, aimed at reducing the vanishing gradient problem, and enhancing feature extraction capabilities for complex TS patterns. ResNet model is composed of three residual blocks, each of which contains three convolutional layers with varying units. These blocks are then followed by a Global Average Pooling layer and a final softmax classifier.

2) LSTM-FCN

The LSTM-FCN architecture combines two main blocks: a Fully Convolutional Network (FCN) with an LSTM network [44]. The FCN block treats the TS of length T as a univariate series with T timesteps. In contrast, when the LSTM block receives the univariate TS with T timesteps, its performance decreases considerably due to fast overfitting when dealing with short sequences and difficulty in learning long-term dependencies on datasets with long sequences. Thus, in this architecture, it receives the input TS as a multivariate TS with one timestep, employing a dimension-shifting layer, which transposes the time dimension of the sequence. This way, a univariate series of length T, after the transformation, can be interpreted as a multivariate TS with T variables with a single timestep.

3) INCEPTIONTIME

InceptionTime model is inspired by the Inception architecture [45], initially developed for image classification [46]. InceptionTime consists of a set of five Inception networks, to reduce the variability in performance and improve the accuracy of the model. The architecture of all the networks is identical, differing only in the initialization of their weights. The output prediction of each of the five networks is given the same importance, with all predictions being averaged to produce the final prediction of the classifier. Every Inception network classifier contains two residual blocks, each containing three Inception modules. Every module applies multiple filters of varying lengths simultaneously to the input TS. Additionally, every module includes a bottleneck layer that reduces the dimensionality of the input sequence.

4) CDIL-CNN

The Circular Dilated Convolutional Neural Network (CDIL-CNN) incorporates symmetric dilated convolutions, that allow the reception of any information extracted by previous layers [47]. The dilation sizes increase exponentially with the depth of the network, allowing the receptive field to expand rapidly and the network to scale to very long sequences. CDIL-CNN applies circular mixing, which allows a signal at one end to be mixed with signals at the other end, making the model more robust to changes in information position, thus preventing it from focusing only on local information. Each block contains mainly a circular dilated convolutional layer and a residual connection. The number of total convolutional layers L in the network depends on the length of the input sequence. Ensemble learning is applied to all positions in the final convolutional layer to achieve better performance.

5) CLASSIFICATION EVALUATION METRICS

The metrics used to evaluate the performance of the classification models are accuracy, sensitivity, specificity, F1-score, and AUC. The associated formulas are given below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{13}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{14}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{15}$$

$$\text{F1-score} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}, \tag{16}$$

where TP is the number of True Positives, TN is the number of True Negatives, FP is the number of False Positives, and FN is the number of False Negatives.

Since the data used in the study belongs to a restricted set of subjects, the evaluation metrics can vary significantly between the different folds. Therefore, using the Coefficient of variation (CV) gives a better idea of the consistency of the results obtained, measuring the relative variability in performance metrics. The CV is a standardized, unitless metric that measures the variability of the results [48]. A low CV result indicates that the performance metrics are more evenly distributed near the average, and therefore, are more consistent. For a specific performance metric, the CV is calculated as the ratio of the standard deviation (σ) to the mean (μ):

$$\text{CV}(\%) = \frac{\sigma}{\mu} \cdot 100 \tag{17}$$

6) EXPERIMENTAL SETUP

For the experiments conducted in this study, a speaker-independent stratified 5-fold cross-validation strategy has been followed, ensuring that there was no overlap of speakers between folds. The original dataset with 100 speakers has been divided into these 5 folds, with three separate recordings, resulting in a total support of 300. Since there are only 50 patients in each group, a 60-20-20 split has been chosen, allowing 10 patients from each group to be evaluated in the final test.

For training and validation, synthetic data generated with BigVSAN have been used. The training process used 3,600 recordings from each group (equivalent to 30 speakers), and the validation set consisted of 1,200 recordings from each group (10 speakers). For the final test, only the real data have been used, providing an accurate estimate of the overall performance of the models.

Before incorporating the synthetic data, a preliminary experiment is performed with all four models following the splitting strategy described above but considering only the 300 samples that correspond to the real data. It is intended that these tests serve as a reference to analyze the improvement of each model when incorporating the data augmentation strategy. Then, an ablation test is also performed on all models, where the data augmentation strategy is incorporated, and the same hyperparameter settings selected in the preliminary experiment are used, thus providing a fair comparison. Finally, Experiments 1 to 3 consist of several tests in which the model configuration is adjusted for better performance.

IV. RESULTS AND DISCUSSION

This section discusses the experimental results obtained throughout the study, evaluating the audio generation strategy and conducting comparative evaluations of the PD detection models.

A. AUDIO SYNTHESIS PERFORMANCE

To evaluate the performance of the BigVSAN model, initial results are obtained without any iterative processes, generating only 150 recordings for HC and 150 for PD. For the HC (PD) group, the M-STFT is 0.72 (0.73), the PESQ is 4.21 (4.17), the MCD is 0.38 (0.27), the Periodicity is 0.09 (0.14), and the V/UV F1 score is 0.98 (0.97).

The following experiments involve 40 iterations, generating 6,000 synthetic samples that represent HC samples and as many for PD subjects.

The optimization of the randomization factor, μ, proceeds as follows: first, various pretrained weight instances are analyzed to determine their approximate order of magnitude, which is found to be around 10^{-2}. Based on this information, several randomization factors are selected for evaluation to analyze their respective impacts. Table 1 summarizes the performance comparison of different randomization factors for distorting the learned parameters, along with the sequential approach.

The results obtained in Table 1 indicate that the addition of controlled randomness, at small levels such as 10^{-5} factor can generate beneficial variability in the training data while maintaining high quality in the generated voice samples. However, although an even smaller factor, 10^{-6}, preserves the quality of the data, the provided variability is insufficient. On the other hand, higher levels of randomization, such as 10^{-3}, introduce excessive noise, significantly compromising audio quality. Experiments using the sequential approach have

TABLE 1. Objective Evaluations of BigVSAN Using the Randomization Factors and Sequential Approach for HC and PD Groups

Approach	Group	M-STFT ↓	PESQ ↑	MCD ↓	Periodicity ↓	V/UV F1 ↑
Factor 10^{-3}	HC	55.66	1.31	8.87	0.56	0.47
	PD	20.38	1.28	6.51	0.62	0.49
Factor 10^{-4}	HC	4.30	3.08	2.26	0.17	0.96
	PD	1.46	3.94	2.03	0.23	0.94
Factor 10^{-5}	HC	0.91	4.16	0.73	0.10	0.98
	PD	0.82	4.14	0.58	0.16	0.97
Factor 10^{-6}	HC	0.72	4.19	0.39	0.09	0.98
	PD	0.72	4.17	0.28	0.14	0.97
Sequential	HC	1.54	2.18	2.06	0.28	0.89
	PD	1.50	2.11	1.63	0.37	0.86

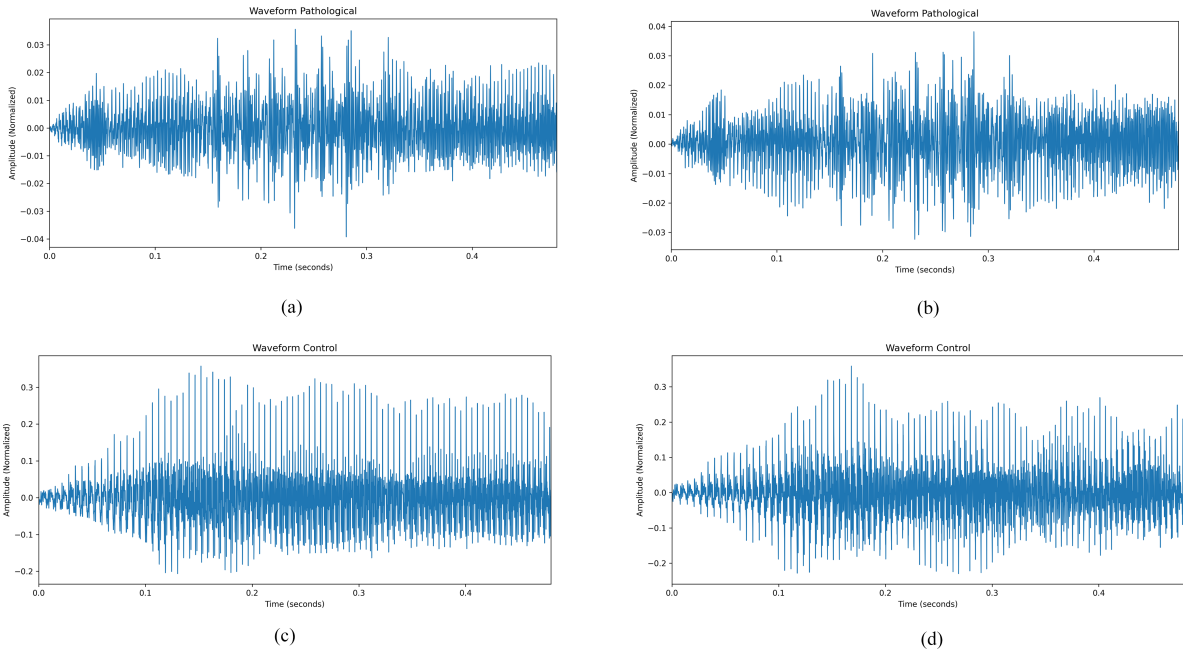

FIGURE 3. Waveforms comparison. (a) Ground-truth example for PD. (b) Corresponding PD-generated audio with the selected configuration. (c) Ground-truth example for HC. (d) Corresponding HC-generated audio with the selected configuration.

demonstrated that as the iterations progress, the quality of the audio deteriorates. This is because the model attempts to replicate as closely as possible the input provided, but over time, small imperfections and distortions are accumulated, resulting in generated samples that are increasingly different from their corresponding original.

The final selected configuration implements a randomization factor of 10^{-5} since it is considered to provide an appropriate trade-off between maintaining high quality in the generated data and introducing minimal perturbations to promote sufficient variability, giving rise to a more diverse dataset that improves the posterior generalization capacity of the classification model.

A visual comparison of the generated waveforms against the real ones is depicted in Fig. 3, including an example both for PD and HC cases. The generated waveforms closely resemble the real ones, showing comparable amplitude patterns and peaks. This similarity demonstrates the effectiveness of the BigVSAN model in generating high-quality synthetic recordings that maintain the fundamental features characterizing the sounds of PD patients and healthy individuals.

B. CLASSIFICATION PERFORMANCE

The performance of the PD detection models has been evaluated in different experiments using varying configurations for their architecture, summarized in Table 2. In all experiments, the models have been trained for 500 epochs, using the Reduce LR OnPlateau technique with a 0.5-factor decrease in 30 epochs and with an early stopping at 50 epochs, and monitoring both the decrease in validation loss and the improvement in validation accuracy. The associated results can be found in Table 3.

For the ResNet model, when data augmentation is applied, it is revealed a trend of improvement in the classification metrics as the complexity of the experiments increases, increasing the baseline accuracy result by 12.90%. In general, improvements are observed in the means of accuracy, specificity, and AUC with each successive experiment,

TABLE 2. Selected Hyperparameters for the Different Experiments Using the PC-GITA Dataset

Model	Batch Size	LR	Configuration
ResNet Baseline	64	Adam 10^{-5}	units: 64,128,128; filter size: 8,5,3
ResNet Experiment 1	64	Adam 10^{-5}	units: 32,64,64; filter size: 8,5,3
ResNet Experiment 2	64	Adam 10^{-5}	units: 64,128,128; filter size: 8,5,3
ResNet Experiment 3	64	Adam 10^{-6}	units: 128,256,256; filter size: 8,5,3
LSTM-FCN Baseline	128	Adam 10^{-3}	2 LSTM modules: 4 units; dropout: 0.8. FCN: units: 128,64,128; filter size: 8,5,3
LSTM-FCN Experiment 1	128	Adam 10^{-4}	2 LSTM modules: 4 units; dropout: 0.9. FCN: units: 128,64,128; filter size: 8,5,3
LSTM-FCN Experiment 2	128	Adam 10^{-4}	1 LSTM module: 8 units; dropout: 0.9. FCN: units: 128,64,128 filter size: 8,5,3
LSTM-FCN Experiment 3	128	Adam 10^{-4}	2 LSTM modules: 8 units; dropout: 0.9. FCN: units: 128,64,128; filter size: 8,5,3
Inception-Time Baseline	64	Adam 10^{-4}	6 modules: units: 32,32,32; filter size: 20,10,5; bottleneck size: 32
Inception-Time Experiment 1	64	Adam 10^{-4}	6 modules: units: 32,32,32; filter size: 20,10,5; bottleneck size: 32
Inception-Time Experiment 2	64	Adam 10^{-6}	6 modules: units: 32,32,32; filter size: 40,20,10; bottleneck size: 32
Inception-Time Experiment 3	64	Adam 10^{-5}	6 modules: units: 32,32,32; filter size: 20,10,5; bottleneck size: 16
CDIL-CNN Baseline	128	Adam 10^{-3}	units: 32; filter size: 3
CDIL-CNN Experiment 1	128	Adam 10^{-5}	units: 16; filter size: 3
CDIL-CNN Experiment 2	128	Adam 10^{-5}	units: 64; filter size: 5
CDIL-CNN Experiment 3	128	Adam 10^{-5}	units: 32; filter size: 3

although sensitivity does not follow this trend and shows a decrease in the most complex experiment. In addition, the CV also improves, especially for the accuracy and AUC values, indicating that the obtained results are more consistent and robust across the different folds of the training process.

When data augmentation is applied in the LSTM-FCN model, results reveal a general improvement in classification metrics as the complexity of the model increases, enhancing the baseline accuracy by 16.12%. In the first and second experiments, a notable discrepancy between sensitivity and specificity is observed. However, in the following experiments, this imbalance is progressively reduced. CV values do not have such a clear trend. The values for accuracy and AUC reflect greater consistency in the results. However, other metrics suggest less robustness, reflected by slightly higher CV values.

The InceptionTime model also improves its performance when adding the data augmentation strategy, reaching a 15% increase in accuracy. The trend of improvement in the classification metrics corresponds with the decrease in the complexity of the model, employing a smaller bottleneck and using smaller filters. In the first, second, and third experiments, a notable mismatch is observed between sensitivity and specificity, and, in the case of the first experiment, a high variability in specificity is also observed (56%), indicating inconsistent performance in identifying the HC cases. In the last experiment, this imbalance is reduced, as both metrics are more balanced.

The CDIL-CNN model obtained a 15.87% improvement over the baseline model due to the data augmentation technique. The best result is obtained by finding a balance in the complexity of its architecture, using 32 units in each convolutional layer. In all experiments, a notable mismatch between sensitivity and specificity is observed, indicating a high performance in the detection of PD individuals compared to the HC cases. Even so, throughout the experiments, this difference is slightly mitigated. Additionally, the CV remains at acceptable values, very similar throughout the different tests.

The best overall results among all models are obtained using CDIL-CNN, obtaining an accuracy of 0.73. It is also important to note that the LSTM-FCN model takes up to five times less time to train than CDIL-CNN, due to its simpler architecture, but still achieves very similar results, reaching an accuracy of 0.72.

It can be confirmed that the use of the data augmentation technique is beneficial, significantly enhancing the performance of all models when trained with a larger number of data compared to the baseline models, which demonstrate limited learning capacity using only the 300 real recordings. The overall accuracy increased by approximately 15.87%, sensitivity by 6.49%, and specificity had the largest improvement with a 28.57% increase. In addition, baseline models showed greater variability between folds, suggesting less consistency and reliability in their performance compared to those using data augmentation.

TABLE 3. Mean (CV%) Results for the Different Experiments Following a 5-Fold Cross-Validation With the PC-GITA Dataset

Model	Accuracy	Sensitivity	Specificity	F1-Score	AUC	Time (mins)
ResNet Baseline	0.62 (16%)	0.60 (50%)	0.63 (28%)	0.56 (45%)	0.62 (16%)	8
ResNet Experiment 1	0.66 (15%)	0.77 (13%)	0.55 (18%)	0.69 (12%)	0.66 (15%)	115
ResNet Experiment 2	0.68 (16%)	0.77 (13%)	0.59 (19%)	0.71 (13%)	0.68 (16%)	155
ResNet Experiment 3	0.70 (7%)	0.63 (11%)	0.76 (12%)	0.67 (7%)	0.70 (5%)	380
LSTM-FCN Baseline	0.62 (18%)	0.71 (22%)	0.52 (12%)	0.65 (18%)	0.62 (19%)	2
LSTM-FCN Experiment 1	0.67 (9%)	0.88 (7%)	0.46 (36%)	0.73 (5%)	0.67 (9%)	17
LSTM-FCN Experiment 2	0.69 (13%)	0.74 (13%)	0.64 (30%)	0.70 (9%)	0.69 (13%)	21
LSTM-FCN Experiment 3	0.72 (8%)	0.78 (21%)	0.67 (18%)	0.73 (11%)	0.72 (8%)	25
InceptionTime Baseline	0.60 (12%)	0.74 (41%)	0.46 (57%)	0.63 (20%)	0.60 (12%)	5
InceptionTime Experiment 1	0.67 (17%)	0.74 (22%)	0.60 (56%)	0.69 (9%)	0.67 (17%)	94
InceptionTime Experiment 2	0.68 (10%)	0.75 (17%)	0.60 (12%)	0.70 (11%)	0.68 (9%)	280
InceptionTime Experiment 3	0.69 (18%)	0.71 (19%)	0.67 (16%)	0.69 (19%)	0.69 (19%)	150
CDIL-CNN Baseline	0.63 (14%)	0.77 (27%)	0.49 (21%)	0.66 (19%)	0.63 (15%)	7
CDIL-CNN Experiment 1	0.70 (12%)	0.83 (4%)	0.57 (27%)	0.75 (8%)	0.70 (15%)	120
CDIL-CNN Experiment 2	0.71 (13%)	0.83 (10%)	0.60 (22%)	0.74 (10%)	0.74 (13%)	130
CDIL-CNN Experiment 3	0.73 (13%)	0.82 (17%)	0.63 (20%)	0.75 (13%)	0.73 (13%)	125

Classification models for TS have significant limitations when dealing with sequences that contain a large number of timesteps. Another limiting factor in the experiments is the nature and amount of data available. The limited dataset size, with only 50 patients per group, restricts the diversity of speech disturbances associated with PD. This limitation may prevent the model from learning less common vocal features, reducing its robustness. While data augmentation with BigVSAN has effectively increased the sample count, the high similarity between the synthetic and original samples could constrain the learning capacity of the classifiers. This lack of diversity may result in models with limited generalizability, making them less adaptive and robust in practical applications.

Future proposals to enrich the dataset could involve mixed data augmentation techniques, combining GAN-generated samples with traditionally augmented data through transformations like filtering or noise addition.

Additionally, incorporating other databases that use the sustained vowel /a/ protocol alongside PC-GITA could further enhance dataset diversity. Integrating multiple databases that use the same phonation protocol, would allow for the generation of synthetic data that capture specific variations to each database, enabling the model to learn broader patterns. Expanding the dataset with samples from diverse populations and varying recording conditions would further increase variability without losing the distinctive features between PD and HC subjects. By training the classifier on synthetic data from a wide range of sources, the model gains a richer representation of vocal features, enhancing its generalizability and robustness for real-world applications.

C. ABLATION EXPERIMENTS

To validate the effectiveness of the proposed data augmentation technique using GANs, a set of ablation experiments are carried out, shown in Table 3. In these two sets of experiments, referred to as Baseline with and without Data Augmentation, respectively, the results obtained using only real data are compared to those obtained with synthetic data under the same conditions, presented in Table 2.

By analyzing each of the four models individually, it can be seen that they all show improvements after incorporating the data augmentation strategy, highlighting the positive impact of using synthetic data on PD detection: ResNet improved by 9.67%, LSTM-FCN had a 4.83% increase, InceptionTime showed an improvement of 11.66%, and CDIL-CNN increased its performance by 7.93%.

This additional study proposed to validate the GANs-based approach showed that the use of synthetic data improves

the performance of all models under the same experimental conditions, underlying the potential of this technique in overcoming the limitations of models when dealing with small datasets, promoting better generalization and preventing overfitting, thus improving their ability to detect PD from speech waveforms.

D. COMPUTATIONAL TIME

To evaluate the practical feasibility of the proposed approach, the entire performance of the pipeline was assessed by estimating the total time required for generating and classifying the audio samples.

The BigVSAN inference process synthesizes 6,000 samples for both PD and HC groups, with an average generation time of 50 milliseconds per audio, demonstrating the capability of the model to produce large data volumes efficiently.

Since this study leverages a pretrained model, assessing its full training time is unnecessary. However, if the entire model were to be trained from scratch, it would take approximately 15 days on an Nvidia A100 GPU, underscoring the computational benefits of using a pretrained model and enhancing the feasibility of the approach in clinical environments.

Table 3 presents the training times for each of the four classifiers, focusing solely on training duration, as the prediction time is negligible.

For the complete pipeline, which includes using BigVSAN to generate 12,000 audio samples, training the CDIL-CNN on this data, and subsequently evaluating 300 real samples, the total estimated time is approximately 10 minutes for data generation with BigVSAN and 120 minutes for training the CDIL-CNN classifier. This aggregate time of 130 minutes highlights the practicality of the pipeline for real-world applications.

E. COMPARISON

Despite the increasing attention to the use of vocal features for PD detection, there is still a scarcity of comparable studies in the literature that analyze raw voice signals with an approach comparable to this work.

Consequently, the obtained results can be solely compared to those found in [33]. This approach uses as input the raw speech signal waveforms from the PC-GITA database but the analyzed information comes from words, short sentences, or monologues, instead of sustained vowels. Since sequences are longer, they are segmented into 250 milliseconds frames with 50 milliseconds shifts. The reported results indicate an accuracy of nearly 0.67, with 0.58 sensitivity. Compared to them, the experiments carried out in this work achieve higher performance, with an 8.96% improvement in accuracy and 41.38% in sensitivity, showcasing the power of more complex models also focused on TS and the benefit of data augmentation. This improvement could be further enhanced by incorporating various phonation protocols based on words or phrases, as they can bring additional information to that provided by the sustained vowel /a/.

V. CONCLUSION

In this research, an alternative approach for PD detection that uses voice as a biomarker was presented. It provides an innovative pipeline that combines GANs as a data augmentation technique and DNNs as TS classifiers, an integration that has so far been underexplored in this task. The conducted study has focused on exploiting intrinsic voice features present in their waveform representation, through the temporal analysis of raw audio signals, addressing an existing gap in current research.

GANs have demonstrated to be a robust alternative for audio-type data generation for PD detection purposes, exhibiting strong performance in reconstructing voice samples from their input spectrogram representation. It has also been shown that using synthetic data generated by GANs can improve the robustness and generalization of deep learning models, allowing their application in real-world scenarios. The best overall accuracy was obtained with the CDIL-CNN model, while LSTM-CNN followed closely, and required five times less time to train.

Although models trained with waveforms have shown promising performance, they still do not reach the accuracy obtained with spectrogram representations due to the inherent complexity of audio TS data. However, it is worth exploring this approach, which with further development of TSC models could lead to better results. The findings underline the potential of raw audio waveform analysis to improve early detection of PD, opening up new lines for research and clinical applications. Additionally, this strategy provides an objective, non-invasive, low-cost, and potentially remotely applicable method that could have a significant impact on improving the diagnosis and monitoring of the disease.

ACKNOWLEDGMENT

The authors are thankful to Prof. Orozco Arroyave and the GITA research group of the University of Antioquia (Colombia) for allowing us to use the PC-GITA database.

REFERENCES

[1] Parkinson's Foundation, "Parkinson's statistics," early access: Jun. 30, 2024. [Online]. Available: https://www.parkinson.org/understanding-parkinsons/statistics

[2] Y. Ben-Shlomo, S. Darweesh, J. Llibre-Guerra, C. Marras, M. San Luciano, and C. Tanner, "The epidemiology of Parkinson's disease," *Lancet*, vol. 403, no. 10423, pp. 283–292, 2024.

[3] S. Hauser and S. A. Josephson, *Harrison's Neurology in Clinical Medicine*, 4th ed. New York, NY USA: McGraw-Hill, 2017.

[4] O. Trifonova et al., "Parkinson's disease: Available clinical and promising omics tests for diagnostics, disease risk assessment, and pharmacotherapy personalization," *Diagnostics*, vol. 10, no. 5, 2020, Art. no. 339.

[5] A. Ibrahim and M. A. Mohammed, "A comprehensive review on advancements in artificial intelligence approaches and future perspectives for early diagnosis of Parkinson's disease," *Int. J. Math. Statist. Comput. Sci.*, vol. 2, pp. 173–182, 2024.

[6] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in Parkinson's disease: Early diagnostics and effects of medication and brain stimulation," *J. Neural Transm.*, vol. 124, no. 3, pp. 303–334, 2017.

[7] L. O. Ramig, C. Fox, and S. Sapir, "Speech Disorders in Parkinson's disease and the Effects of Pharmacological, Surgical and Speech Treatment With Emphasis on Lee Silverman Voice Treatment (LSVT)," in *Parkinson's Disease and Related Disorders* vol. 83, pp. 385–399, 2007.

[8] E. V. Altay and B. Alatas, "Association analysis of Parkinson disease with vocal change characteristics using multi-objective metaheuristic optimization," *Med. Hypotheses*, vol. 141, 2020, Art. no. 109722.

[9] J. Skibińska and J. Hosek, "Computerized analysis of hypomimia and hypokinetic dysarthria for improved diagnosis of Parkinson's disease," *Heliyon*, vol. 9, no. 11, 2023, Art. no. e21175.

[10] F. Amato, G. Saggio, V. Cesarini, G. Olmo, and G. Costantini, "Machine learning-and statistical-based voice analysis of Parkinson's disease patients: A survey," *Exp. Syst. Appl.*, vol. 219, 2023, Art. no. 119651.

[11] S. Lahmiri, D. A. Dawson, and A. Shmuel, "Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures," *Biomed. Eng. Lett.*, vol. 8, no. 1, pp. 29–39, 2018.

[12] A. Ouhmida, A. Raihani, B. Cherradi, and Y. Lamalem, "Parkinson's disease classification using machine learning algorithms: Performance analysis and comparison," in *Proc. 2nd IEEE Int. Conf. Inn. Res. Appl. Sci. Eng. Tech.*, 2022, pp. 1–6.

[13] M. S. Alzubaidi et al., "The role of neural network for the detection of Parkinson's disease: A scoping review," *Healthcare*, vol. 9, no. 6, 2021, Art. no. 740.

[14] 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," in *Proc. IEEE 41st Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2019, pp. 717–720.

[15] 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*, vol. 8, pp. 206888–206900, 2020.

[16] H. Gunduz, "Deep learning-based Parkinson's disease classification using vocal feature sets," *IEEE Access*, vol. 7, pp. 115540–115551, 2019.

[17] S. V. T. Dao, Z. Yu, L. V. Tran, P. Phan, T. Huynh, and T. Le, "An analysis of vocal features for Parkinson's disease classification using evolutionary algorithms," *Diagnostics*, vol. 12, no. 8, 2022, Art. no. 1980.

[18] J. Mallela et al., "Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM," in *Proc. Interspeech* 2020, pp. 4586–4590.

[19] O. Abayomi-Alli, R. Damaševičius, R. Maskeliūnas, and A. Abayomi-Alli, "BiLSTM with data augmentation using interpolation methods to improve early detection of Parkinson disease," in *Proc. IEEE 15th Conf. Comput. Sci. Inf. Syst.*, 2020, pp. 371–380.

[20] 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," *Biomed. Eng. Online*, vol. 20, no. 1, Art. no. 114, 2021.

[21] M. Mounia, B. Nouhaila, N. Benayad, and B. D. Taoufiq, "Use of ANN, LSTM and CNN classifiers for the new MSCC and BSCC methods in the detection of Parkinson's disease by voice analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 12, 2023, pp. 560–567.

[22] R. Maskeliūnas, R. Damaševičius, A. Kulikajevas, E. Padervinskis, K. Pribuišis, and V. Uloza, "A hybrid U-lossian deep learning network for screening and evaluating Parkinson's disease," *Appl. Sci.*, vol. 12, no. 22, 2022, Art. no. 11601.

[23] G. Dimauro and F. Girardi, "Italian Parkinson's voice and speech," 2019. early access: Jun. 30, 2024. [Online]. Available: https://doi.org/10.21227/AW6B-TG17

[24] B. Karan and S. Sekhar Sahu, "An improved framework for Parkinson's disease prediction using variational mode decomposition-hilbert spectrum of speech signal," *Biocybern. Biomed. Eng.*, vol. 41, no. 2, pp. 717–732, 2021.

[25] L. Parisi, N. RaviChandran, and M. L. Manaog, "Feature-driven machine learning to improve early diagnosis of Parkinson's disease," *Expert Syst. Appl.*, vol. 110, pp. 182–190, 2018.

[26] O. Kursun, B. Sakar, M. Isenkul, C. Sakar, A. Sertbas, and F. Gurgen, "Parkinson's speech with multiple types of sound recordings [dataset]," UCI Mach. Learn. Repository, 2013, doi: 10.24432/C5NC8M.

[27] C. Quan, K. Ren, and Z. Luo, "A deep learning based method for Parkinson's disease detection using dynamic features of speech," *IEEE Access*, vol. 9, pp. 10239–10252, 2021.

[28] Z. K. Senturk, "Layer recurrent neural network-based diagnosis of Parkinson's disease using voice features," *Biomedizinische Technik*, vol. 67, no. 4, pp. 249–266, 2022.

[29] 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," *Biocybern. Biomed. Eng.*, vol. 42, no. 2, pp. 556–574, 2022.

[30] 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," *Biomed. Signal Process. Control*, vol. 70, 2021, Art. no. 103006.

[31] U. K. Lilhore et al., "Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 14605.

[32] V. K. Pandey, S. S. Sahu, B. Karan, and S. K. Mishra, "Parkinson disease prediction using CNN-LSTM model from voice signal," *SN Comput. Sci.*, vol. 5, no. 4, 2024, Art. no. 381.

[33] N. P. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1925–1936, 2021.

[34] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. 9th Int. Conf. Lang. Res. Eval.*, 2014, pp. 342–347.

[35] A. Ma, K. K. Lau, and D. Thyagarajan, "Voice changes in Parkinson's disease: What are they telling us?," *J. Clin. Neurosci.*, vol. 72, pp. 1–7, 2020.

[36] R. Chiaramonte and M. Bonfiglio, "Acoustic analysis of voice in Parkinson's disease: A systematic review of voice disability and meta-analysis of studies," *Revista de Neurología*, vol. 70, no. 11, pp. 393–405, 2020.

[37] T. Shibuya, Y. Takida, and Y. Mitsufuji, "BigVSAN: Enhancing GAN-based neural vocoders with slicing adversarial network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, 2024, pp. 10121–10125.

[38] R. Yamamoto, E. Song, and J. M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, 2020, pp. 6199–6203.

[39] A. Rix, J. Beerends, M. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2001, pp. 749–752.

[40] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, 1993, pp. 125–128.

[41] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive GAN for conditional waveform synthesis," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–19.

[42] H. Zen et al., "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech* 2019, pp. 1526–1530.

[43] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Jt Conf. Neural Netw.*, 2017, pp. 1578–1585.

[44] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[46] H. I. Fawaz et al., "InceptionTime: Finding AlexNet for time series classification," *Data Min. Knowl. Disc.*, vol. 34, no. 6, pp. 1936–1962, 2020.

[47] L. Cheng, R. Khalitov, T. Yu, J. Zhang, and Z. Yang, "Classification of long sequential data using circular dilated convolutional neural networks," *Neurocomputing*, vol. 518, pp. 50–59, 2023.

[48] K. Pearson, "VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia," *Philos. Trans. R. Soc. A*, vol. 187, pp. 253–318, 1896.

MARTA REY-PAREDES received the B.Sc. degree in biomedical engineering from the Rey Juan Carlos University, Madrid, Spain, in 2023 and the M.Sc. degree in artificial intelligence from the Polytechnic University of Madrid, Madrid, Spain. She has coauthored a conference paper for the 2023 Computing in Cardiology. Her research interest includes signal processing for biomedical applications.

CARLOS J. PÉREZ received the M.Sc. degree in mathematical science and the Ph.D. degree in mathematics from the University of Málaga, Spain, in 1996 and 2003, respectively. He is currently a Full Professor of Statistics with the Department of Mathematics, University of Extremadura, Badajoz, Spain. He has authored or coauthored more than 80 JCR-indexed journal papers about statistical methodology and applications in diverse knowledge fields, including computer-aided diagnosis systems. He has participated in more than 30 research projects from competitive calls and contracts. He also has been a Reviewer for journals, such as *Expert Systems with Applications, Reliability Engineering and Safety Systems, Journal of Applied Statistics, and Annals of Applied Statistics*. His research interests include in the areas of Bayesian statistical inference and classification.

ALFONSO MATEOS-CABALLERO received his M.Sc. degree in mathematical science from the University of Extremadura, Badajoz, Spain, in 1991 and the Ph.D. degree in informatics from the Polytechnic University of Madrid, Madrid, Spain. He is currently a Full Professor with the Artificial Intelligence Department, Polytechnic University of Madrid. All of the professional life he has been related to Operations Research, Statistics, Artificial Intelligence, Decision Analysis and Decision Support Systems. He has written more than 100 papers in Spanish and international journals (44 of them listed in JCR), such as *EJOR, Computational Optimization and Applications, Annals of Operations Research, JORS, Reliability Engineering and System Safety, DSS, GDN, and Computers and Operations Research*. He has participated in more than 64 projects (4 of them are European Projects), has coauthored five books and was reader of a paper more than 180 times in conferences.