

Parkinson's Disease Detection Using Dual-Branch WavLM Embeddings, Fuzzy Cellular Automata Statistics, and Mamba Sequence Modeling

Hema Sudha Rani ¹, Senthil Athithan ^{2,*}

¹Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Vaddeswaram, Guntur (Dt.), 522502 AP, India

²Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Vaddeswaram, Guntur (Dt.), 522502 AP, India

Abstract Parkinson's disease (PD) is a progressive neurodegenerative condition that can affect the muscular control of speech at an early stage. This work presents an FCA–Mamba framework for automatic Parkinson's disease detection from speech using contextual representation learning and uncertainty-aware spatial modeling. Preprocessed speech waveforms are processed by a pretrained, frozen WavLM encoder to obtain contextualized speech embeddings. The Mamba branch can selectively model long-distance temporal dependencies in contextual speech representations, while the FCA branch exploits fuzzy neighborhood evolution and standalone higher-order statistics for localized uncertainty-aware spatial relationships. These representations are then fused temporally and statistically using the two branches to produce final Parkinson speech classification. Experiments are conducted using stratified five-fold cross-validation on an established dataset of 655 speech samples from people with Parkinson's disease. The proposed FCA–Mamba framework achieves 98.68% accuracy, 98.58% precision, 99.49% recall, 99.03% F1-score, and 99.78% ROC-AUC, outperforming conventional machine learning and deep learning-based baseline models. Furthermore, FCA-based heatmap analysis provides clear temporal–spatial activation patterns of representation between healthy and Parkinsonian speech, enabling further interpretation. In conclusion, the results indicate that combining Mamba-based temporal modeling with uncertainty-aware FCA in a statistical learning architecture provides an accurate and interpretable system for automatic Parkinson's disease detection from speech signals.

Keywords Parkinson's Disease Detection, Fuzzy Cellular Automata, Mamba State-Space Model, WavLM, Speech Analysis, Deep Learning

DOI: 10.19139/soic-2310-5070-4726

1. Introduction

[1] Parkinson's disease (PD) is a progressive neurological disorder categorized by decreased motor function and disrupted neuromuscular coordination. In addition to movement-related symptoms such as rigidity, tremors, and bradykinesia, PD also causes significant vocal impairments due to dysfunction of the speech production system [2]. Parkinsonian speech commonly exhibits dysphonia-related abnormalities including vocal instability, reduced loudness, monotonic pitch, breathiness, articulation irregularities, and impaired vocal fold vibration [3]. Because speech abnormalities frequently manifest during the early stages of PD, Speech-based analysis has emerged as a promising non-invasive approach for the early detection and continuous monitoring of Parkinson's disease.

Get accurate and timely speech recognition using the latest machine learning techniques. One of the most effective methods for analysing Parkinsonian speech is automating it using current advancements in Machine Learning and Deep Learning [4,5]. Traditional methods, however, still mainly use manual acoustic features like

*Correspondence to: Senthil Athithan. Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Vaddeswaram, Guntur (Dt.), 522502 AP, India.

jitter, MFCCs, shimmer and HNR with classifiers such as SVM[11], Random Forests, and shallow neural networks. These methods achieve relatively high performance for classification problems, but do not work well in capturing non-linear temporal dependencies and relative relationships between the elements in speech sequences.

Several of these limitations have been addressed through the application of deep learning architectures, including Convolutional Neural Networks (CNNs) [10], Long Short-Term Memory (LSTM) networks, and Transformer-based models [12], for pathological speech analysis. CNN-based approaches learn spatial priors from spectrogram data, and more sequential models, such as LSTM and Transformer architectures, improve temporal representations. However, Transformer-based approaches are often computationally expensive due to quadratic self-attention computations and thus inefficient for long speech sequences.

Moreover, previous approaches do not account for (1) uncertainty-aware spatial interactions and (2) local activation variabilities within contextual embedding spaces[14]. In the recent years, state-space sequence modeling frameworks including Mamba have been shown to be effective for scalable long-range temporal dependency learning with linear time complexity. Mamba model [6]: Mamba is a class of models that propagates contextual information in input-dependent state transitions which are ideal for tasks like speech sequence analysis as pathological vocal abnormalities may occur in scattered temporal frames. However, current approaches for the detection of PD using speech tend to combine uncertainty-aware spatial statistical modeling with efficient sequence learning techniques. To address these challenges, this research presents a new FCA–Mamba dual-branch framework for Parkinsons disease detection based on voice signals. The framework consists of three blocks, WavLM based contextual embedding generation, Mamba selective state-space temporal sequence modeling and Auxiliary Fuzzy Cellular Automata (FCA) [7] statistical representations learning. The Mamba branch mines extended contextual temporal dependencies in the speech embeddings and the [19] FCA branch uses fuzzy neighborhood evolution (FN) and higher-order statistical feature extraction to represent localized uncertainty-aware spatial interactions. Complementary fusion of temporal uncertainty and spatial uncertainty-aware representations improves discriminative power for Parkinsonian speech analysis [8].

Moreover, the framework utilizes strong experimental paradigms including but not limited to fold-wise normalization, frozen pre-trained embeddings and stratified cross-validation designed to avoid data leakage and increase reproducibility.

The main contributions of this work are enumerated as follows:

1. A dual branch architecture is developed in which WavLM contextual embeddings are processed by Mamba for long range temporal modelling and FCA for uncertainty aware spatial representation. The complimentary temporal and spatial feature are fused for PD detection, with FCA modelling the embedding matrix as a fuzzy field using a 3×3 Moore neighbourhood and iterative fuzzy state evolution.
2. An FCA-based uncertainty representation mechanism is introduced for contextual speech embeddings, in which the WavLM embedding matrix is interpreted as a two-dimensional fuzzy field and evolved using a 3×3 Moore neighborhood, product t-norm interactions, and iterative fuzzy-state updates to characterize localized activation variability.
3. A compact FCA statistical representation is derived from the evolved fuzzy field to summarize activation distribution and uncertainty characteristics, providing complementary information to the temporal representation generated by the Mamba branch.
4. A temporal–spatial fusion strategy is implemented by combining the Mamba-derived temporal representation with FCA-derived statistical descriptors before classification, allowing the classifier to jointly exploit sequential speech dynamics and localized uncertainty information.
5. It includes an FCA-based method for visualizing model activations, which is used to visualize fuzzy activation fields evolved across generations of the GA as temporal–embedding heatmaps, so that comparisons can be made between activation propagation in Parkinsonian and healthy-control speech representations.
6. We quantitatively analyze the contribution of dual-branch design using 5-fold cross-validation and component-wise ablation; precision, accuracy, F1-score recall, and ROC-AUC are used as metrics in performance evaluation.

To highlight the methodological novelty of the proposed framework, as opposed to merely applying known techniques, we present a comparative novelty matrix in Table 1.

Table 1. Novelty Matrix Comparing the Proposed FCA-Mamba Framework with Existing Approaches

Approach	Contextual Speech Representation	Long- range Temporal Modeling	Uncertainty- aware Spatial Modeling	FCA Evolution	Temporal- Spatial Fusion	Contribution Status
Handcrafted acoustic features + ML	No	No	No	No	No	Existing
CNN-based approaches	Limited	Limited	No	No	No	Existing
LSTM/GRU-based approaches	No/Optional	Yes	No	No	No	Existing
Transformer/SSL approaches	Yes	Yes	No	No	No	Existing
WavLM-based representation	Yes	Contextual encoding	No	No	No	Existing component
Mamba-based modeling	Optional	Yes	No	No	No	Existing component
FCA-based biomedical modeling	No	No	Yes	Yes	No	Existing component
Proposed FCA-Mamba	Yes (WavLM)	Yes (Mamba)	Yes (FCA)	Yes	Yes	Proposed integration

The various components that are used for this study, including WavLM, Mamba and Fuzzy Cellular Automata (FCA) are established methods with no individual claims of being new. Their contributions to the proposed framework consist of method-specific dual-branch integration for Parkinsonian speech analysis. We propose to leverage the complementary representations contained in WavLM contextual embeddings by processing them through two branches: specifically, the Mamba branch captures long-range temporal dependencies while the auxiliary FCA branch models localized uncertainty via iterative neighborhood evolution seen as a fuzzy spatial field on top of observed contextual embedding matrix. These FCA-derived statistical descriptors are then of what is called, fused with the Mamba temporal representative for parkinsonian classification. Thus, the primary contribution is to incorporate temporal sequence modeling with uncertainty-aware spatial statistical modeling into a unified framework for speech-based screening of Parkinson's disease.

2. Literature Survey

2.1. Speech-Based Parkinson's disease Detection

This has led to considerable interest in using speech analysis to automatically detect Parkinson's, as it is non-invasive and audio features are known to appear early in the disease process. Handcrafted acoustic features, such as jitter, shimmer, pitch perturbation, harmonic-to-noise ratio (HNR), and Mel-frequency cepstral coefficients (MFCCs) [9], were mostly used in early studies. We used these features with traditional machine learning classifiers, including SVM, Random Forests and k-NN, to classify the Parkinsonian speech from healthy speech. While these methods performed reasonably well, they were often limited by hand-engineered features and lacked the ability to properly model long-range contextual speech dependencies.

In the past few years research has made great strides for conducting pathological speech analysis with deep learning [10]. CNNs have been used to learn spatial discriminative representation from spectrograms [13] and

recurrent neural networks like LSTM or Gated Recurrent Unit(GRU)[15] based models improved the temporal modelling of speech sequences. More recently, self-supervised speech representation models [20], such as HuBERT [14] and WavLM [21], have surpassed them by learning high-quality contextualised speech representations from large amounts of unlabeled acoustic corpora. These pretrained models find long-term temporal dependencies and some of the shallow vocal abnormalities that challenge standard hand-crafted acoustic features.

However, there are still some limitations that persist. Existing approaches mainly use traditional deep neural networks, Transformer-based architectures or attention mechanisms to exploit sequential order learning from signal or irregular embeddings without temporal representation learning. As a result, local uncertainty propagation and spatial activation features are not directly modeled in the contextual embedding space. Limited study on subject-level data leak prevention from existing studies during cross-validation raises concerns about the validity of reported performance [30]

Recent studies [22, 31] on automatic detection of idiopathic PD from dysphonia speech using HuBERT and WavLM embeddings have utilized CNN, Transformer or multilayer perceptron (MLP) classifiers. These techniques mainly utilize contextual speech embeddings for temporal representation learning without explicitly modeling uncertainty-aware spatial activity patterns in the embedding space they learn. [24, 25] Different from these approaches, our proposed framework has an auxiliary [29] FCA branch which applies fuzzy neighborhood evolution on our pretrained contextual embedding of WavLM before fusing it with the Mamba-based temporal sequence representation. As a result, the introduced framework fully benefits from both long-term temporal dependencies and uncertainty-aware spatial information to deliver complementary representations of PD speech classification.

To address these limitations, the proposed framework combines pretrained WavLM contextual embeddings with an auxiliary Fuzzy Cellular Automata (FCA) branch and a Mamba-based temporal sequence modeling branch. The FCA branch explicitly models localized uncertainty-aware spatial activation patterns through fuzzy neighborhood evolution and summarizes them using compact statistical descriptors, while the Mamba branch captures long-range temporal dependencies in contextual speech sequences. The complementary fusion of these temporal and uncertainty-aware spatial representations enables more comprehensive characterization of Parkinsonian speech than existing self-supervised learning approaches that rely primarily on temporal contextual embeddings alone.

2.2. State-Space Sequence Modeling

State-space sequence modelling is an effective alternative to recurrent and Transformer-based architectures for long-sequence learning tasks, which has been recently introduced. RNNs often face gradient instability and limited ability to learn long-range dependencies, whereas Transformer architectures need quadratic computational time complexity for self-attention computation. Structured State Space Models (S4) proposed a solution to these limitations by performing continuous-time sequence propagation with linear computational complexity. Expanding on this idea, the Mamba architecture performed state-space modeling only for select states with input-dependent state transitions. However, unlike fixed-transition state-space models, Mamba allows the sequence propagation behavior to change according to the current input representation of each particle [16, 17]. Additionally, this selective state-space mechanism is well suited for analyzing dysarthric speech in Parkinsonians, given that the pathological nature of certain vocal tremors manifests intermittently throughout continuous phonations. In addition, Mamba is less memory-consuming and more scalable than Transformer-based architectures and makes it very suitable for contextualized speech embedding analysis [18]. Mamba, a state-of-the-art temporal modelling framework, has been proposed to detect PD; however, Mamba-based frameworks rarely use it with uncertainty-aware spatial representation learning methods.

[26] Existing Mamba-based speech classification frameworks primarily focus on modeling long-range temporal dependencies within speech sequences. However, they do not explicitly characterize uncertainty or localized spatial activation patterns within the learned embedding space. The proposed FCA branch complements the Mamba encoder by modeling uncertainty-aware neighborhood interactions through fuzzy cellular evolution and summarizing these interactions using compact statistical descriptors. Accordingly, the fused framework not only captures temporal dynamics through Mamba branch, spatial characteristics that are especially prone to uncertainty such as blurred classification boundaries via FCA branch, and also offers complimentary representations that

augment both discriminative power and model interpretability over simple temporal sequence modeling alone. Comparing to earlier methods for disordered speech, our framework combines Mamba-based temporal modeling of WavLM embeddings with FCA-based uncertainty-aware spatial modeling with their delegations which are fused for Parkinson's disease detection.

As such, this presents the dual-branch FCA-Mamba framework to overcome the shortcomings of previous Mamba-based approaches for detecting PD. This solution combines effective temporal sequence modeling with uncertainty-sensitive spatial representation learning in a single model architecture. This design characterizes Parkinsonian speech more holistically than temporal modeling alone, while maintain computational efficiency.

2.3. Fuzzy Cellular Automata in Biomedical Signal Analysis

Cellular Automata (CA) were first formalized by Von Neumann in 1948 and are a class of discrete abstract math models comprising cells which evolve according to simple local neighborhood interaction rules through discrete time dynamical systems [23]. While performing deterministic binary state transitions, Classical Cellular Automata relies on a cell's state; however, Fuzzy Cellular Automata (FCA) prototypes model this operation by representing the membership value of each cell (or neighborhood region during field iteration) as a real number between 0 and 1 in order to provide fuzzy coverage in between transitions [19,24–26]. This fuzzy nature allows FCA to have a better approach in modeling uncertainty and gradual state transitions than binary cellular automata. FCA include fuzzy membership functions and t-norm-based neighborhood interactions to propagate uncertainty over spatial activation fields in the frame-work of FCA.

Many other biomedical applications also successfully use FCA [27, 28] for medical image segmentation and analysis, EEG [29] seizure detection, and biosignal anomaly detection, where it is capable of modeling localized spatial interactions under uncertainty through propagation. However, such studies generally work on image data or handcrafted biomedical features and often integrate the evolved FCA representations with classical machine learning classifiers. Hence, how we apply FCA in modern deep contextual representation learning frameworks is less exposed.

The vocal characteristics in Parkinsonian speech analysis are classically described as heterogeneous and exhibit uncertain activation patterns across their contextual speech encodings. As such, FCA is a suitable approach to model the propagation of localized uncertainty as well as obtaining supplementary spatial statistical abstractions from the high dimensional embedding space. In contrast to the main body of previous FCA-based biomedical studies, which typically used either medical images, EEG signals, or handcrafted biomedical features and metrics, the proposed framework applies FCA directly on very high-dimensional contextual speech embeddings generated by a pre-trained WavLM model. Instead of direct classification with FCA, the evolved fuzzy activation field is compacted into simple statistical descriptors then incorporated with Mamba-based temporal sequence representations.

As a result, FCA works as an uncertainty-aware spatial representation learning module that makes Mamba more capable of modeling temporal dimensions. Although the Mamba branch is capable of accurately capturing long-range temporal dependencies and occasional speech abnormalities, the FCA branch learns localized uncertainty propagation and spatial activation distributions over contextual embedding spaces. The proposed FCA-Mamba framework helps to generalize the scope of FCA beyond traditional biomedical signal processing by leveraging multi-level temporal speech dynamics and uncertainty-aware spatial information, providing a unified deep contextual learning framework for PD detection from speech signals.

3. Methodology

3.1. Dataset Description

The proposed FCA-Mamba framework was evaluated using Italian Parkinson's Voice and Speech (ItalianPVS) dataset [8], available through IEEE DataPort. The corpus contains speech recordings from individuals with PD and HC participants and provides multiple speech-production tasks for investigating vocal and articulatory characteristics associated with Parkinsonian speech.

For the present study, 655 speech recordings were selected, comprising 405 PD recordings and 250 HC recordings, indicating a moderate class imbalance. The dataset includes several speech-production tasks, including text reading (B1 and B2), phrase reading (FB1), word reading (FB2), syllable repetition (D1 and D2), and sustained vowel phonation (/a/, /e/, /i/, /o/, and /u/). According to the published dataset documentation, the corpus consists of 28 participants with PD, 22 elderly healthy controls, and 28 younger healthy controls, with representation from both male and female speakers.

The published dataset documentation further reports that none of the participants with PD had speech or language disorders unrelated to PD. However, detailed participant-level information, including age distribution, disease duration, medication status, and other clinical characteristics, is not publicly available in the dataset documentation. Consequently, these variables could not be incorporated into the present study and are acknowledged as a limitation of the dataset rather than the proposed methodology.

For the preprocessing pipeline, each audio recording was converted to a mono waveform and resampled to 16 kHz. Recordings longer than 15 s were truncated to the first 15 s before WavLM embedding extraction. All 655 recordings were successfully processed by the acoustic-feature extraction and WavLM embedding pipelines, with no recordings excluded because of preprocessing or feature-extraction failure.

The resulting dataset exhibits a moderate class imbalance because PD recordings outnumber HC recordings. This imbalance was considered during the construction of the subject-independent five-fold cross-validation protocol while maintaining complete separation of participant identities between the training and validation folds. Details of the cross-validation strategy and data-leakage prevention procedure are presented in Section 3.8 (Training Protocol).

Although sustained vowel phonation is widely used for speech-based PD assessment, the ItalianPVS dataset employed in this study is not limited to vowel recordings. The dataset includes multiple speech-production tasks, including sustained vowels, text reading, phrase reading, word reading, and syllable repetition. Consequently, the proposed FCA–Mamba framework was evaluated using a diverse collection of speech tasks rather than sustained vowel phonations alone. Nevertheless, the present study was evaluated only on the ItalianPVS dataset, and external validation on independent PD speech datasets was not performed. Results from this study show strong performance on the ItalianPVS benchmark but should not be interpreted as evidence of widespread language or recording environment generalizability, nor clinical population transfer. In future work using the FCA–Mamba framework, more extensive cross-validation on larger sets of data from independent PD speech datasets will be conducted, allowing for the demonstration of these findings through use of standardized subject-independent evaluation protocols, in addition to multiple languages and continuous-speech corpora.

3.2. Overall Framework

A new framework for PD detection from speech signals using WavLM-based contextual speech representations, Mamba-based temporal sequence modeling and Auxiliary Fuzzy Cellular Automata (FCA) statistical analysis was proposed. The global goal of the framework is to model joint long-range temporal dependencies and uncertainty-aware spatial activation features from Parkinsonian speech embeddings.

The frozen WavLM encoder is used to retrieve contextual representations from the preprocessed speech recordings as raw audio waveforms. For traditional baseline analysis, traditional acoustic descriptors such as MFCC, jitter, shimmer, and HNR are extracted as handcrafted speech features. These manually created acoustic features are not immediately supplied as inputs to WavLM by the suggested FCA–Mamba structure. Rather, WavLM creates contextual frame-level embeddings by working directly on the preprocessed raw waveform. The Mamba and FCA branches then process these embeddings. In order to employ deep learning applications for Parkinson’s voice data, the embeddings generated preserve contextual acoustic properties and latent speech dynamics transferred from the attributes associated with one of its prevalent vocal deficits.

These contextual embedding sequences are then separated and routed through two parallel branches. The main branch uses a Mamba selective state-space model that takes advantage of fast, linear-time sequence modeling to capture long-range temporal dependencies in speech sequences across the entire repetition for larger data sets. Meanwhile, the auxiliary FCA branch employs such contextual embedding matrix as a fuzzy spatial field and implements uncertainty-aware neighbourhood evolution with fuzzy t-norm transition rules. Higher-order statistical

descriptors are then extracted from the developed fuzzy field to illustrate spatial entropy, activation dispersion, and local variability patterns.

Features from both branches are concatenated and directed into a fully connected classification head, and finally predicting PD. The dual-branch proposed strategy allows for necessary learning of temporally sequencing information and uncertainty-aware spatial statistical representations.

The framework uses fold-wise normalization, stratified cross-validation, and frozen pre-trained WavLM embeddings to prevent data leakage for confident evaluation. In Fig.1, we present the overall architecture of our proposed FCA-Mamba framework that consists of 4 modules: i) preprocessing; ii) dual-branch representation learning; iii) feature fusion and iv) classification.

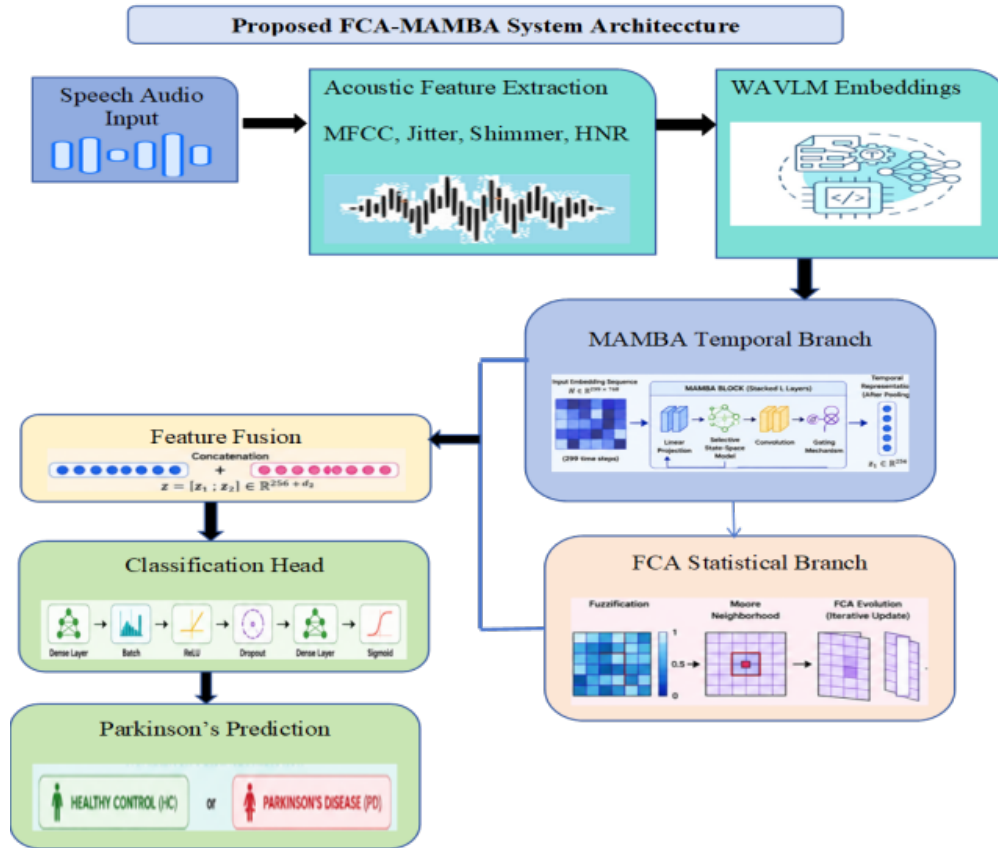


Figure 1. Simplified architecture of the proposed dual-branch WavLM-FCA-Mamba framework for PD detection from speech signals.

3.3. Acoustic Feature Extraction

In the proposed framework, speech recordings are first processed to extract acoustic features strongly associated with Parkinsonian dysphonia. PD affects neuromuscular control of the vocal system, leading to abnormalities in vocal fold vibration, pitch stability, and voice quality. To capture these pathological characteristics, four clinically significant acoustic features are extracted from each speech signal

- Mel-Frequency Cepstral Coefficients (MFCCs),
- jitter,
- shimmer,
- Harmonic-to-Noise Ratio (HNR).

The present study focused on a compact set of clinically established acoustic descriptors, namely MFCC, jitter, shimmer, and HNR, because these features have been widely validated for speech-based PD assessment and capture complementary spectral and phonatory characteristics. Rather than incorporating a large number of handcrafted acoustic features, the proposed framework primarily relies on the contextual representations learned by the pretrained WavLM model, which implicitly encode rich spectral, prosodic, and temporal speech information from large-scale speech corpora. Consequently, additional handcrafted features such as pitch period entropy, glottal flow parameters, correlation dimension, and Lyapunov exponents were not included in the present study to avoid feature redundancy and unnecessary model complexity. Future work will investigate the complementary contribution of these advanced nonlinear and physiological speech descriptors for further improving PD detection performance.

3.3.1. MFCC Feature Extraction: Since MFCC extraction follows a well-established signal processing pipeline, only the key computational steps are presented here. The novelty of the proposed work lies in the subsequent WavLM-based contextual representation learning and the proposed FCA-Mamba architecture

Following noise reduction, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to represent the short-term spectral characteristics of speech. Standard MFCC extraction was performed using 40 coefficients, 26 Mel filter banks, a 25 ms frame length, a 10 ms frame shift, and a Hamming window. Framing splits the speech signal into short overlapping frames typically 20–40 ms because speech is non-stationary but approximately quasi-stationary over short intervals, and windowing then applies a Hamming window to each frame to reduce spectral leakage, it is expressed in equation (1)

$$x_w[n] = x[n]w[n] \quad (1)$$

Where

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{(N-1)}\right) \quad 0 \leq n \leq N \text{ is the Hamming Window Function}$$

Where $x[n]$ denotes the speech samples within a frame, $w[n]$ represents the Hamming window, $x_w[n]$ is the windowed speech frame, n is the sample index, and N is the number of samples in the frame. Each windowed frame is then transformed from the time domain to the frequency domain using the Fast Fourier Transform (FFT), as described in Equation (2).

$$X[k] = \sum_{n=0}^{N-1} x_w[n] e^{-j2\pi kn/N}, \quad k = 0, 1, \dots, N-1 \quad (2)$$

where $X[k]$ represents the frequency-domain spectrum of the windowed speech frame, $x_w[n]$ is the windowed speech signal, N is the number of samples in the frame, n is the time-domain sample index, k is the frequency-bin index, and $j = \sqrt{-1}$.

$$P[k] = \frac{1}{N} |X[k]|^2, \quad k = 0, 1, \dots, N-1 \quad (3)$$

where $P[k]$ represents the power spectrum at frequency bin k , $X[k]$ is the Fourier transform obtained from Equation (2), and N denotes the number of samples in the frame.

It is expressed in equation (4).

$$L(m) = \log \left(\sum_{k=0}^{N-1} P[k] H_m \right), \quad m = 1, 2, \dots, M \quad (4)$$

where $L(m)$ is the log Mel filter-bank energy of the m^{th} filter, $P[k]$ is the power spectrum obtained from Equation (3), $H_m[k]$ is the response of the m^{th} Mel filter at frequency bin k , and M is the total number of Mel filters.

$$C[r] = \sum_m^M L[m] \cos \left[\frac{\pi r}{M} \left(m - \frac{1}{2} \right) \right], \quad r = 0, 1, \dots, C-1 \quad (5)$$

where $C[r]$ represents the r^{th} MFCC coefficient, $L[m]$ is the log energy of the m^{th} Mel filter obtained from Equation (4), M is the total number of Mel filters, and C is the number of retained MFCC coefficients.

3.3.2. Jitter Extraction Jitter measures the cycle-to-cycle variations in the fundamental frequency of speech, as high jitter means more disrupted vibration of vocal-folds which is a prominent feature found in Parkinsonian speech.

Jitter is computed as in equation (6)

$$Jitter = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| \frac{T_i - T_{i+1}}{T_i} \right| \quad (6)$$

Where:

- T_i represents the i^{th} pitch period,
- N Denotes the total number of pitch cycles.

Higher jitter values indicate irregular vocal frequency perturbations caused by impaired neuromuscular control.

3.3.3. Shimmer Extraction Shimmer measures variations in speech amplitude between consecutive vocal cycles. Parkinsonian patients frequently exhibit unstable amplitude modulation and breathy voice characteristics, leading to elevated shimmer values.

Shimmer is calculated as in equation (7):

$$Shimmer = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| \frac{A_i - A_{i+1}}{A_i} \right| \quad (7)$$

Where:

- A_i denotes the amplitude of the i^{th} speech cycle,
- N Represents the number of vocal cycles.

Shimmer reflects vocal intensity instability associated with Parkinsonian dysphonia.

3.3.4. Harmonic-to-Noise Ratio (HNR) HNR measures the proportion of both periodic harmonic components and noise components in the speech signal. Incomplete closure of the vocal folds during phonation can lead to increased breathiness and turbulent noise, both characteristics of Parkinsonian speech.

HNR is computed as:

$$HNR = 10 \log_{10} \left(\frac{P_{harmonic}}{P_{noise}} \right) \quad (8)$$

Where:

- $P_{harmonic}$ denotes harmonic signal power,
- P_{noise} Represents noise power.

Lower HNR values indicate degraded voice quality and increased vocal noise.

After feature extraction, each speech sample is represented as a sequential acoustic feature matrix containing MFCC, jitter, shimmer, and HNR descriptors across multiple temporal frames. The recovered acoustic features are kept for baseline analysis, and the preprocessed raw speech waveform is then fed to the WavLM encoder to produce contextual speech embeddings.

3.4. WavLM-Based Contextual Embedding Generation

Each speech recording is converted to a mono waveform and resampled to 16 kHz before WavLM encoding. Audio recordings longer than 15 s are truncated to the first 15 s (240,000 samples). The raw one-dimensional waveform is processed directly by the Hugging Face Wav2Vec2 feature extractor and provided as input to the pretrained WavLM-base encoder (microsoft/wavlm-base). No handcrafted acoustic features, spectrograms, or MFCCs are supplied to WavLM. No manual amplitude normalization is performed because the Wav2Vec2 feature extractor internally handles waveform preprocessing and padding. Variable-length recordings are represented using an attention mask to distinguish valid speech samples from padded samples. The frozen WavLM-base encoder generates contextual frame-level embeddings with a hidden dimension of 768, which are converted into a fixed-dimensional utterance representation using temporal mean pooling.

Let the preprocessed waveform be represented as

$$X \in \mathbb{R}^L \quad (9)$$

where L denotes the number of waveform samples after preprocessing. The WavLM encoder generates contextual hidden representations as

$$H = \text{WavLM}(X), \quad H \in \mathbb{R}^{T' \times 768} \quad (10)$$

where T' denotes the output temporal sequence length and 768 is the hidden dimension of WavLM-base. Temporal mean pooling is subsequently applied to obtain a fixed-dimensional representation:

$$e = \frac{1}{T'} \sum_{t=1}^{T'} H_t, \quad e \in \mathbb{R}^{768} \quad (11)$$

where H_t denotes the contextual representation at temporal step t , T' represents the WavLM output sequence length, and e denotes the resulting 768-dimensional utterance-level embedding.

The generated contextual representations encode temporal and acoustic information learned by the pre-trained WavLM model. The WavLM-base encoder was initialized using the publicly available pre-trained microsoft/wavlm-base checkpoint. The encoder was not fine-tuned using the PD dataset employed in this study. During embedding extraction, WavLM was operated in evaluation mode with gradient computation disabled, and all pre-trained parameters remained frozen. Therefore, neither class labels nor information from the validation folds was used to update the WavLM parameters. Since WavLM was pre-trained on external speech corpora, potential overlap between its original pre-training data and the downstream speech dataset cannot be completely excluded; this issue is acknowledged as a limitation of the present study.

3.5. Mamba-Based Temporal Sequence Modeling Branch

The primary branch of the proposed framework employs a Mamba selective state-space model to capture long-range temporal dependencies within contextual speech embedding sequences. Parkinsonian speech abnormalities often occur intermittently across temporal frames; therefore, effective sequence modeling is essential for identifying pathological vocal dynamics.

Unlike Transformer architectures that rely on quadratic self-attention mechanisms, Mamba performs efficient linear-complexity sequence modeling using selective state-space transitions. This enables scalable processing of long contextual embedding sequences while maintaining computational efficiency.

Given the contextual embedding sequence:

$$H \in \mathbb{R}^{T \times D}$$

The Mamba branch processes the sequence through selective state-space evolution.

The hidden state update is represented as:

$$h_t = Ah_{t-1} + Bx_t \quad (12)$$

Where:

- h_t denotes the hidden state at time step t ,
- x_t represents the contextual embedding input,
- A And B are learnable state-transition parameters.

The output sequence representation is computed as:

$$y_t = Ch_t + Dx_t \quad (13)$$

Where:

- C And D are learnable projection parameters.

Mamba introduces selective state updates in which transition parameters dynamically depend on the current input representation. This provides the model with a way to remember relevant temporal speech patterns and ignore uninformative activations. The temporal feature representation extracted by the Mamba is represented as:

$$z_1 \in \mathbb{R}^{256}$$

The Mamba branch effectively captures:

- long-range temporal dependencies,
- contextual sequence relationships,
- And Parkinsonian temporal vocal variations.

These temporal representations are later fused with uncertainty-aware FCA statistical descriptors for final classification.

3.6. Auxiliary FCA Statistical Branch

The Auxiliary Fuzzy Cellular Automata (FCA) branch aims at modelling the uncertainty-driven spatial relations between speech embedding manifolds in a contextual space. Unlike the Mamba branch dedicated to temporal sequence learning only, FCA captures localized spatial activation patterns as well as uncertainty propagation between neighborhoods at the same level.

First, the contextual embedding matrix is interpreted as a p-dimensional space-time field:

$$H \in \mathbb{R}^{T \times D}$$

Where:

- rows correspond to temporal frames,
- Columns correspond to contextual embedding dimensions.

Step 1: Fuzzification

The contextual embedding values are transformed into fuzzy membership representations using a sigmoid membership function:

$$\mu(t, d) = \frac{1}{1 + e^{-H(t, d)}} \quad (14)$$

Where:

- $\mu(t, d)$ Represents the fuzzy membership value.

The sigmoid membership function was chosen because it smoothly and continuously differentiates the mapping of contextual embedding values into $[0, 1]$, carrying over gradual activation transitions and allowing for stable uncertainty propagation in FCA evolution. Sigmoid function can provide smoother activation transitions and better

numerical stability compared with triangular or trapezoidal membership functions for high-dimensional speech embeddings, which is continuous value.

Step 2: Moore Neighborhood Construction

For each fuzzy cell, local interactions are modeled using a 3×3 Moore neighborhood:

$$N(t, d) = \{(t + i, d + j) \mid i, j \in [-1, 0, 1]\} \quad (15)$$

This neighborhood captures local spatial dependencies across nearby temporal and embedding regions.

Step 3: FCA Evolution

The FCA evolution parameters were determined through sensitivity analysis. The number of evolution iterations was evaluated for $K=\{1,3,5,7,10\}$ and $K=3$ achieved the highest mean classification accuracy. The evolution rate was evaluated for $\alpha=\{0.1,0.3,0.5,0.7,0.9\}$ with $\alpha=0.5$ providing the best performance. The convergence threshold was evaluated over $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$. Since the maximum iteration limit was reached before the convergence criterion for all tested thresholds, performance was insensitive to ϵ . Accordingly, $K=3$, $\alpha=0.5$, and $\epsilon=10^{-4}$ were adopted in the subsequent experiments.

Neighborhood interactions are computed using the fuzzy product t-norm:

$$T(a, b) = a \cdot b \quad (16)$$

The neighborhood aggregation is calculated as:

$$f_k(t, d) = \frac{1}{|N(t, d)|} \sum_{(i,j) \in N(t,d)} T(\mu_k(t, d), \mu_k(i, j)) \quad (17)$$

The fuzzy field is iteratively updated as:

$$\mu_{k+1}(t, d) = (1 - \alpha)\mu_k(t, d) + \alpha f_k(t, d) \quad (18)$$

Where $\alpha=0.5$ denotes the FCA evolution rate selected through sensitivity analysis.

The FCA evolution process continues for a fixed number of iterations or until convergence is achieved:

$$|\mu_{k+1} - \mu_k| F < \epsilon \quad (19)$$

Where $\epsilon=10^{-4}$ denotes a small convergence threshold. In the selected configuration ($K=3$), the maximum iteration limit was reached before this threshold; therefore, ϵ acts as an auxiliary stopping criterion rather than the primary termination condition. The product t-norm was selected because it preserves continuous fuzzy interactions while avoiding abrupt activation changes during neighborhood evolution. Table 2 shows that, as expected, between min and Lukasiewicz t-norms, the product t-norm leads to smoother uncertainty propagation taking into account continuous embedding representations and greater numerical stability. In addition, the 3×3 Moore neighborhood could facilitate multidirectional contextual interaction among neighboring cells while also preserving computational simplicity than its simpler neighborhood structures (e.g., the 2D four-neighbour).

3.6.1. Statistical Feature Extraction After completion of the FCA evolution process, the final evolved fuzzy membership field contains uncertainty-aware spatial activation patterns learned from the contextual speech embeddings. However, the evolved field itself is high-dimensional and cannot be directly used for classification. Therefore, statistical feature extraction is performed to summarize the spatial characteristics of the evolved fuzzy field into a compact discriminative representation.

Let the final evolved fuzzy field be represented as:

$$\mu_K \in [0, 1]^{T \times D} \quad (20)$$

Where:

- K denotes the final FCA iteration,

- T represents temporal frames,
- D denotes embedding dimensions.

Unlike the preliminary version of the framework, the revised FCA branch employs only four complementary statistical descriptors (mean activation, standard deviation, spatial entropy, and energy). This compact representation was adopted to reduce feature redundancy and computational complexity while preserving the most informative uncertainty-aware statistical characteristics of the evolved fuzzy field.

The FCA statistical representation was intentionally limited to four complementary descriptors, namely mean activation, standard deviation, spatial entropy, and energy, because these descriptors summarize the global activation distribution, variability, uncertainty, and overall activation strength of the evolved fuzzy field while maintaining a compact and computationally efficient representation. More complex texture descriptors, including gray-level co-occurrence matrix (GLCM)-based measures such as contrast, homogeneity, and correlation, as well as other higher-order statistical descriptors, were not incorporated in the present study to avoid increasing feature dimensionality, computational complexity, and redundancy with the contextual information already captured by the pretrained WavLM embeddings. As the main purpose of the FCA branch is to complement global statistical descriptors instead of detailed texture descriptions, four selected descriptors offer an ideal tradeoff between discriminative power and computational efficiency.

3.6.2. Mean Activation Mean activation measures the average fuzzy activation intensity across the entire evolved field.

$$S_1 = \frac{1}{TD} \sum_{t=1}^T \sum_{d=1}^D \mu_k(t, d) \quad (21)$$

A higher mean value indicates stronger overall activation patterns within Parkinsonian speech embeddings.

3.6.3. Standard Deviation: Standard deviation measures the dispersion of fuzzy activations around the mean.

$$S_2 = \sqrt{\frac{1}{TD} \sum_{t=1}^T \sum_{d=1}^D (\mu_k(t, d) - S_1)^2} \quad (22)$$

where S_1 denotes the mean activation defined in Equation (22). The standard deviation S_2 measures the dispersion of fuzzy activation values around the mean, reflecting variability within the evolved contextual speech representation.

This descriptor captures variability and instability within contextual speech representations.

3.6.4. Spatial Entropy Spatial entropy measures uncertainty and randomness within the evolved fuzzy field.

$$P_{t,d} = \frac{\mu_k(t, d)}{\sum_{i=1}^T \sum_{j=1}^D \mu_k(i, j) + \epsilon} \quad (23)$$

The spatial entropy is then calculated as:

$$S_3 = \frac{1}{TD} \sum_{t=1}^T \sum_{d=1}^D P_{t,d} \log_2 (P_{t,d} + \epsilon) \quad (24)$$

where ϵ is a small positive constant used to prevent numerical instability. Higher entropy indicates a more dispersed or uncertain fuzzy activation distribution.

3.6.5. Energy Energy measures the overall strength of fuzzy activations.

$$s_4 = \frac{1}{TD} \sum_{t=1}^T \sum_{d=1}^D [\mu_k(t, d)]^2 \quad (25)$$

where s_4 represents the average energy of the evolved fuzzy membership field. Higher energy values indicate stronger and more concentrated fuzzy activation patterns.

Higher energy values indicate stronger and more concentrated activation responses.

Finally, all descriptors are concatenated to form the FCA statistical feature vector:

$$z_2 = [s_1, s_2, s_3, s_4] \in \mathbb{R}^4 \quad (26)$$

where z_2 denotes the four-dimensional FCA statistical feature vector comprising mean activation, standard deviation, spatial entropy, and energy. These descriptors provide complementary uncertainty-aware statistical information to the temporal representation learned by the Mamba branch.

These uncertainty-aware statistical descriptors provide complementary spatial information to the temporal representations learned by the Mamba branch.

3.7. Feature Fusion and Classification

The temporal feature representation generated by the Mamba branch and the uncertainty-aware statistical representation generated by the FCA branch are combined through feature fusion.

The fused representation is expressed as:

$$z_1 \in \mathbb{R}^{256} \quad (27)$$

and the FCA statistical representation as

$$z_2 \in \mathbb{R}^4$$

The fused representation is defined as

$$z = [z_1 : z_2] \in \mathbb{R}^{260}$$

where z_1 represents the temporal features learned by the Mamba branch and z_2 represents the FCA-derived statistical features.

The Mamba representation is transformed to 256 dimensions to provide a compact temporal representation while maintaining sufficient information for classification. On the other hand, the FCA branch is more focused in nature, which summarizes the updated fuzzy membership field into four complementary statistical descriptors, such as mean activation, std. Thus, the functional nature of the two branches is reflected by their dimensional difference in this proposed framework.

Feature-level concatenation maintains both branch representations without the need for extra fusion parameters or manually determined weighting factors; it was chosen. The relative contributions of the Mamba and FCA features are learned by the fully connected layers that follow. While weighted fusion and attention-based fusion are potential substitutes, they were not explored in this study and are reserved for future research. The fully connected classification layers can retrain the relative contribution of Mamba temporal features and FCA-derived statistical descriptors in a supervised fashion. The high dimensionality of the concatenation can, therefore, contain richer information while retaining the original feature distributions compared to other fusion methodologies such as weighted fusion and attention-based fusion which incur additional computational complexity and parameter overhead. Thus, it offers a good trade-off between information preservation, computation and implementation efficiency for the proposed framework. This 260-dimensional fused representation is further processed by fully connected layers in a classification head, containing dense layers with batch normalization, followed by ReLU activation and dropout regularization ($p = 0.5$). The last output layer uses a sigmoid to predict the likelihood of having Parkinson's:

$$\hat{y} = \sigma(Wz + b) \quad \text{Equation (28)}$$

where $\hat{y}$ denotes the predicted probability of PD, W and b represent the learnable weight matrix and bias term, respectively, and $\sigma(\cdot)$ denotes the sigmoid activation function. The resulting probability is used to assign each speech sample to either the healthy-control or PD class.

Classification Head Architecture: Fully connected layers goes over the 260-dimensional fused representation which is concatenated through Mamba features of dimension 256 with FCA statistical features of dimension 4,

yielding a totally concatenated output of (260). This is done with batch normalization, ReLU activation and dropout ($p = 0.5$) to improve generalisation. Finally, the output corresponds to a single neuron designed with a sigmoid function estimating Parkinsons disease (PD) probability. The entire architecture of classification-head is shown in Figure 2.

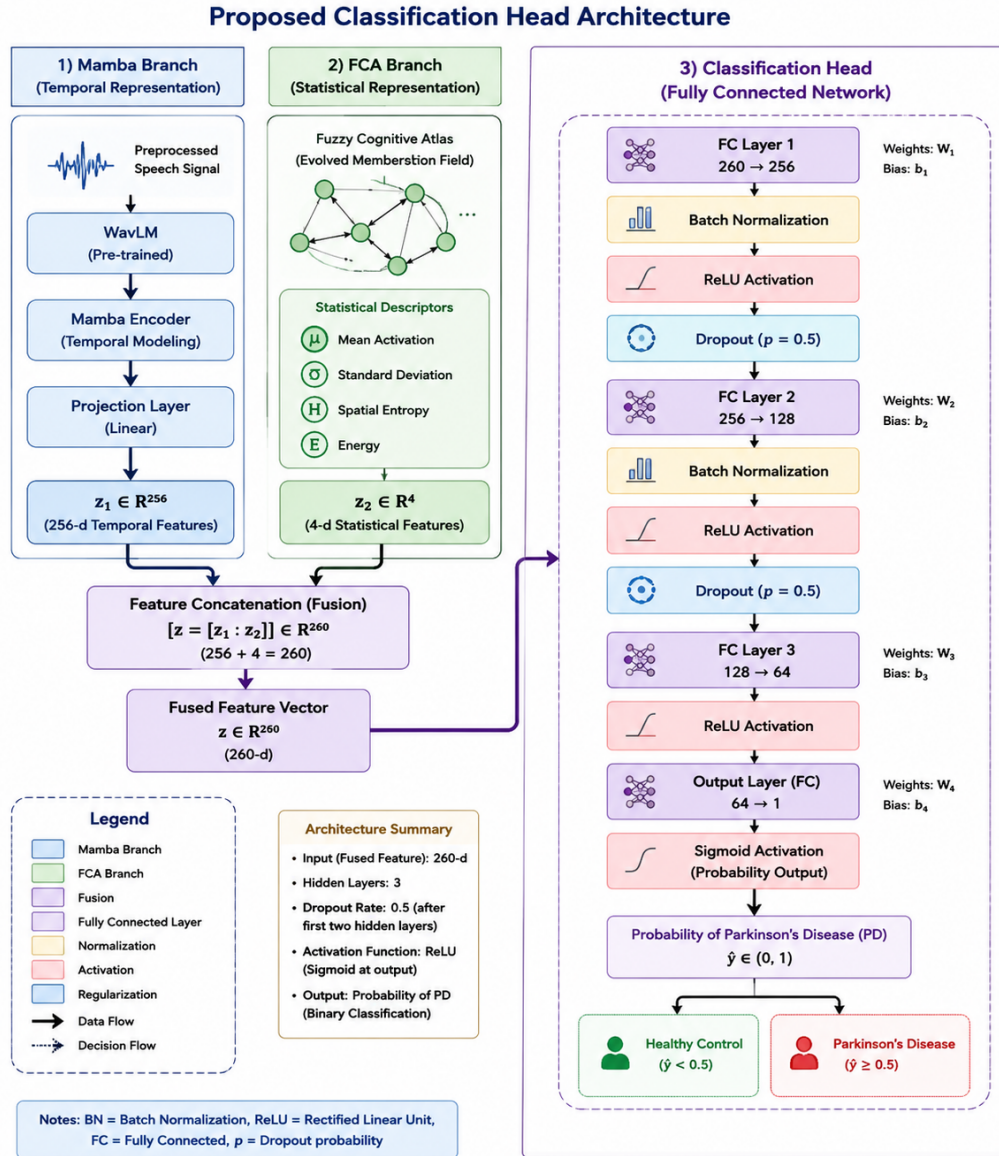


Figure 2. Architecture of the proposed classification head.

3.8. Training Protocol

This study evaluated the FCA-Mamba framework with subject independent five-fold cross-validation: folds were allocated at the participant-level to prevent subject overlap in training and validation sets. Folds were generated with subject IDs from the ItalianPVS metadata, to ensure a somewhat balanced class distribution across folds. Non-overlap where subject-level separation was confirmed for each fold;

Preprocessing parameters such as feature normalization were estimated on the training data only so as to prevent information leakage and applied to the corresponding validation data. Third, we maintained the pretrained WavLM encoder frozen while extracting features and training the model. For the optimizer, AdamW with cosine-annealing learning-rate scheduling was used. To increase generalization and training stability, batch normalization, dropout ($p = 0.5$), and gradient clipping were employed. Training was carried out with a batch size of 8 using a CUDA-enabled GPU for 20 epochs.

Table 2. Subject-independence verification across the five cross-validation folds

Fold	Train recordings	Validation recordings	Subject overlap
1	569	86	0
2	543	112	0
3	511	144	0
4	505	150	0
5	492	163	0

Training/validation subject identifier intersection was zero for each fold as shown in Table 2, confirming that reported evaluation was without subject-level overlap.

The training hyperparameters were chosen based on initial empirical experiments as well as standard parameter settings for fine-tuning of deep neural networks with frozen pre-trained feature representation from a large corpus. For stable optimization that did not have oscillation during training, a learning rate of 1×10^{-3} was used, while the AdamW optimizer with cosine annealing scheduling allowed for gradual reduction in learning-rate over time, and better convergence stability. Since preliminary empirical experiments confirmed that both the training and validation performance stabilized between epochs during training, we chose to randomly stop training after 20 epochs, as further increments in random seed selection above this epoch number had no meaningful impact on mean validation performance. Thus, 20 epochs were chosen to reach a sweet spot across convergence, computational efficiency and overfitting avoidance. To stabilize and generalize albeit training, Dropout regularization, Batch Normalization, and gradient clipping were added.

Although early stopping was considered, it was not incorporated because preliminary experiments indicated that validation performance stabilized within the selected 20 training epochs without noticeable overfitting. Nevertheless, adaptive early stopping will be investigated in future work to further enhance training efficiency and generalization.

Although the dataset exhibits a moderate class imbalance (405 PD recordings and 250 healthy control recordings), additional imbalance-handling techniques such as class-weighted loss functions, oversampling, or data augmentation were not employed. Instead, the proposed framework addressed class imbalance through subject-independent stratified five-fold cross-validation while maintaining complete subject separation between the training and validation partitions. Given the moderate imbalance ratio and the consistent performance observed across all cross-validation folds, additional resampling or augmentation strategies were not considered necessary in the present study. Nevertheless, future work will investigate class-weighted optimization, synthetic data augmentation, and advanced resampling techniques for larger or more severely imbalanced PD speech datasets.

4. Results

4.1. Experimental Setup

Using a subject-level five-fold cross-validation framework, the FCA-Mamba model was evaluated on the Italian Parkinsons speech dataset containing 655 recordings. The data were split by participant so that recordings from a given subject were not present in both the training and validation subsets. To avoid bias during the supervised models training and testing on the proposed model, in parallel with all comparative approaches, the same folds

were predefined. In each execution $\lfloor N \rfloor$ folds were for training and one fold was used for validation, that is, all $\lfloor k = 5 \rfloor$ folds served as a validation set once. Likewise, feature normalization was performed separately for each training partition, and the corresponding parameters were only applied to this validation partition. This ensured that each subject evaluation was independent and free of leakage for all experiments.

Speech recordings were processed using the pretrained WavLM model to extract contextual speech representations. Each recording produced a sequence of 749 temporal frames, where every frame was represented by a 768-dimensional embedding vector. The pretrained WavLM encoder was used as a frozen feature extractor, and its parameters remained unchanged throughout model training. The encoder was intentionally kept frozen because the Italian PD speech dataset contains only 655 recordings, which is relatively small for fine-tuning a large self-supervised speech model. Fine-tuning even the final transformer layers under such limited data conditions may substantially increase the risk of overfitting and reduce the generalization capability of the learned representations. By freezing the pretrained WavLM parameters, the proposed framework leverages robust contextual speech representations learned from large-scale speech corpora while allowing the proposed Mamba encoder and FCA branch to learn task-specific discriminative patterns. The extracted sequence representations were subsequently processed by the proposed Mamba encoder to capture long-range temporal dependencies. In parallel, four statistical descriptors (mean, standard deviation, entropy, and energy) were extracted from the FCA-evolved representations to characterize the spatial activation patterns. The temporal representation learned by the Mamba encoder and the FCA statistical descriptors were concatenated and supplied to the final multilayer perceptron classifier for PD classification.

FCA-Mamba has been implemented in PyTorch and trained with AdamW (learning rate: 1×10^{-3} and a weight decay of 1×10^{-4}) under cosine-annealing scheduling. Training was performed for 20 epochs, using a batch size of 8 and the best validation checkpoint from each fold was retained. FCA-Mamba has been implemented in PyTorch and trained with AdamW (learning rate: 1×10^{-3} and a weight decay of 1×10^{-4}) under cosine-annealing scheduling. Training was performed for 20 epochs, using a batch size of 8 and the best validation checkpoint from each fold was retained.

4.2. Overall Classification Performance

Table 3. Overall classification performance of the proposed FCA-Mamba framework using subject-independent five-fold cross-validation.

Metric	Mean (%)	Std (%)
Accuracy	98.68	1.58
Precision	98.58	1.49
Recall	99.49	1.15
F1-Score	99.03	1.12
ROC-AUC	99.78	0.28

Table 3 depicts the overall classification performance of the proposed FCA-Mamba framework using subject-independent five-fold cross-validation. Average accuracy of $98.68 \pm 1.58\%$ indicates the great discriminative abilities to separate between PD from Healthy Control speech sound samples. It achieved a precision of $98.58 \pm 1.49\%$, recall of $99.49 \pm 1.15\%$ and F1-score of $99.03 \pm 1.12\%$ on the validation folds, demonstrating even classification performance across all folds which is an indicator for reliability in decision making (Figure [f], Methods). The proposed framework also obtained a $99.78 \pm 0.28\%$ average ROC-AUC score, demonstrating its excellent capability to differentiate between PD and healthy speech at various decision thresholds. Note that the standard deviations were low for all the evaluation metrics, indicating robust and reliable performance of our model over cross-validation.

Table 4. Class-wise Performance

Class	Metric	Value (%)
Parkinson's Disease	Sensitivity (Recall)	99.49
Healthy Control	Specificity	98.00

Table 4 reports the class-wise performance of the proposed FCA–Mamba framework. Sensitivity (recall) for the PD class reached 99.49%, indicating that the proposed framework correctly identified nearly all PD speech recordings. The specificity for the Healthy Control class was 98.00%, demonstrating a low false-positive rate and effective discrimination between healthy and Parkinsonian speech.

Although the overall classification performance is high, several factors provide evidence that the reported performance is unlikely to result from data leakage. The evaluation employed strict subject-independent five-fold cross-validation with explicit verification that no participant appeared in both the training and validation sets. Furthermore, all preprocessing operations were performed independently within each training fold, and the WavLM encoder remained frozen throughout model training. The consistency of the fold-wise performance, together with the training and validation learning curves and the fold-wise confusion matrices, indicates that the reported performance reflects good generalization rather than memorization.

4.3. Fold-wise Performance

Table 5. Fold-wise performance of the proposed FCA–Mamba framework under subject independent five-fold cross-validation

Fold	Accuracy	Precision	Recall	F1	ROC-AUC
1	97.67	96.72	100.00	98.33	99.56
2	96.43	97.44	97.44	97.44	99.40
3	99.31	98.77	100.00	99.38	99.94
4	100.00	100.00	100.00	100.00	100.00
5	100.00	100.00	100.00	100.00	100.00

Table 5 presents the Performance of the FCA–Mamba over other methods with subject-independent five-fold cross-validation (average values per fold). The performance of the model was stable and only slightly varying during all validation folds (classifications accuracies from 96.43% to 100.00%, mean accuracy overall = $98.68 \pm 1.58\%$).

The error rates of the five folds are similar, with Fold 3 and Fold 5 giving best classification performance who is perfectly classified in both cases while Fold 1 and Fold 2 produced lower accuracies. Such variations, however, are to be expected in subject-independent evaluation (as each fold is comprised of entirely different subjects not seen before) and reflect the normal heterogeneity present in Parkinsonian speech samples rather than model instability. Additionally, since the ItalianPVS dataset does not have subject-level clinical characteristics or recording-condition metadata [33], we could not perform a more detail-oriented analysis of why models performed better in some folds than others. However, the small standard deviation and tight confidence intervals confirm that the proposed FCA–Mamba framework generalizes well to new subjects.

4.4. Confusion Matrix Analysis

Figure 3 shows the aggregated out-of-fold predictions through the subject-independent five-fold cross-validation scheme in a confusion matrix format. Combined the predictions from all five validation folds to display a single confusion matrix for compact presentation. In contrast, the performance metrics reported in table 5 were calculated separately for each validation fold and then averaged across all folds. Therefore, small numbers differences between the pooled confusion matrix metrics and the fold-wise average evaluation metrics are normal.

The proposed FCA–Mamba framework classified 245 Healthy Control (HC) speech recordings and 403 PD speech recordings correctly, while 5 HC were misclassified as PD samples by false positives and two PD

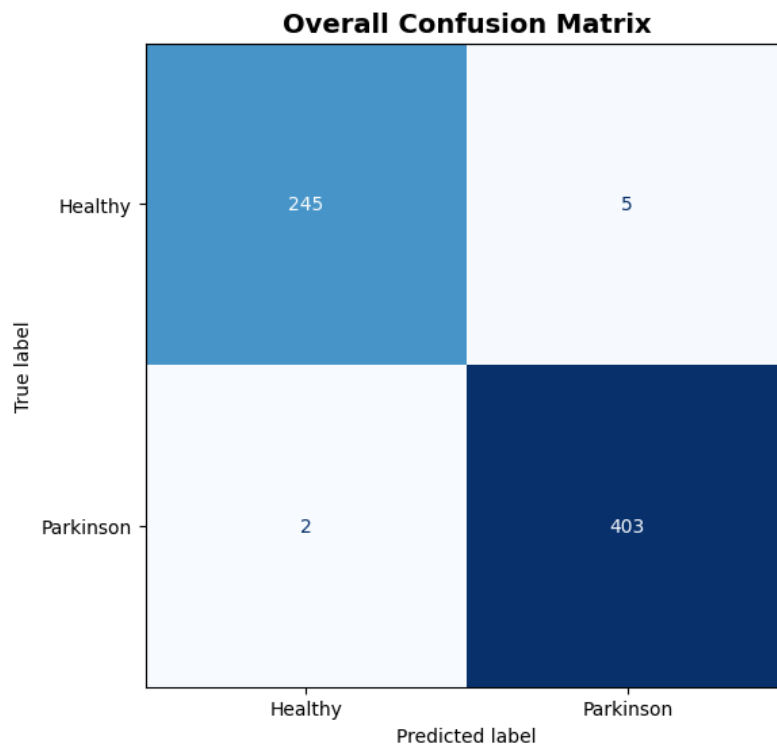


Figure 3. Confusion matrix of the proposed FCA–Mamba framework obtained using five-fold cross-validation.

misclassified HC data by false negatives as shown in figure-3. So only 7 versus 655 and the system fails to classify them properly showing the very strong discriminative nature of the proposed whole-frame based framework.

The low number of false-positive predictions indicates that the proposed framework can prevent healthy individuals from being misdiagnosed with PD, and the very small number of false-negative predictions demonstrates its high sensitivity in accurately diagnosing recordings of speech belonging to people with Parkinson's. This balanced classification performance is highly needed for robust computer-assisted screening of PD, where missed diagnoses must be minimized.

The evaluation metrics reported in Table 5 were calculated separately for each validation fold using the corresponding fold-wise confusion matrix and then averaged across the five validation folds. Therefore, these averaged metrics are not expected to be numerically identical to the values derived directly from the pooled confusion matrix shown in Figure 3. Overall, the confusion matrix confirms that integrating WavLM-based contextual speech representations with FCA-derived statistical descriptors enables accurate discrimination between PD and healthy control speech recordings.

4.5. Error Analysis

The proposed FCA–Mamba framework achieved a very low number of misclassifications, as reflected by the confusion matrices presented in the previous sections. Most validation folds contained either no errors or only a small number of false-positive and false-negative predictions. These misclassified samples are likely associated with speech recordings exhibiting characteristics that overlap between the PD and healthy control classes, thereby reducing class separability. Since the ItalianPVS dataset does not provide detailed clinical severity scores, disease progression information, or recording-quality annotations, a more detailed analysis of the underlying causes of these misclassifications could not be performed. However, the low steady state error level across all validation folds suggests that the framework generalizes well whilst maintaining strong discrimination ability between healthy and

Parkinsonian speech. Future work will explore explainable AI techniques, as well as clinical metadata, to better characterize challenging samples and to identify potential failure modes.

4.6. Receiver Operating Characteristic (ROC) Curve Analysis

The ROC curves obtained with the proposed FCA–Mamba framework are illustrated in Fig. 4 for each of the five subject-independent validation folds.

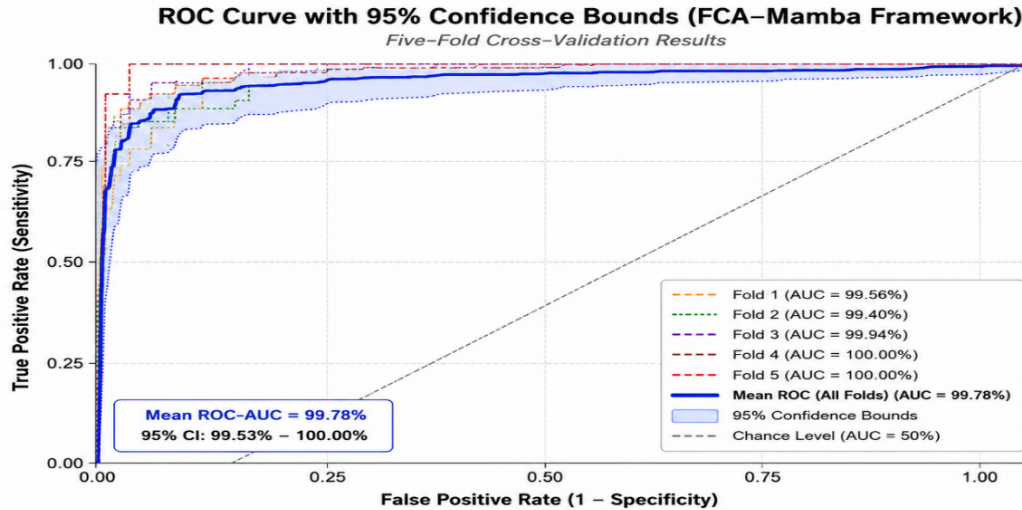


Figure 4. Receiver Operating Characteristic (ROC) curves of the proposed FCA–Mamba framework obtained using subject-independent five-fold cross-validation.

The mean ROC-AUC obtained from the proposed FCA–Mamba framework was $99.78 \pm 0.28\%$, equivalent to a 95% confidence interval of 99.53–100.00%, indicating that following feature concatenation, Mamba is able to discriminate between PD and healthy control speech recordings with excellent accuracy. ROC curves for all validation folds were close to the upper-left corner of the ROC space, showing consistently high sensitivity and specificity throughout cross-validation.

The high ROC-AUC values obtained across all validation folds demonstrate that the proposed framework maintains reliable classification performance over a wide range of decision thresholds rather than relying on a single operating point. These results further confirm the effectiveness of combining WavLM-based temporal representations with FCA-derived statistical descriptors for robust PD speech classification.

Although the mean ROC-AUC and its 95% confidence interval demonstrate excellent overall discriminative performance, additional operating-point analyses such as false-positive rate (FPR) at predefined clinically relevant sensitivity thresholds were not evaluated in the present study. Such threshold-specific analysis depends on the intended clinical application and decision threshold. Future work will investigate clinically optimized operating points together with FPR, sensitivity, specificity, and decision-threshold calibration using larger multicenter PD speech datasets.

4.7. Curve Analysis

Figure 5 illustrates the training and validation learning curves of the proposed FCA–Mamba framework over 20 training epochs. **Figure 5(a)** presents the training and validation accuracy, whereas **Figure 5(b)** depicts the corresponding training and validation loss.

As shown in **Figure 5(a)**, both the training and validation accuracies increased steadily throughout the training process, demonstrating effective learning of discriminative speech representations. The training accuracy improved from approximately 76.7% in the first epoch to nearly 99.3% at the final epoch. Likewise, the increase in the

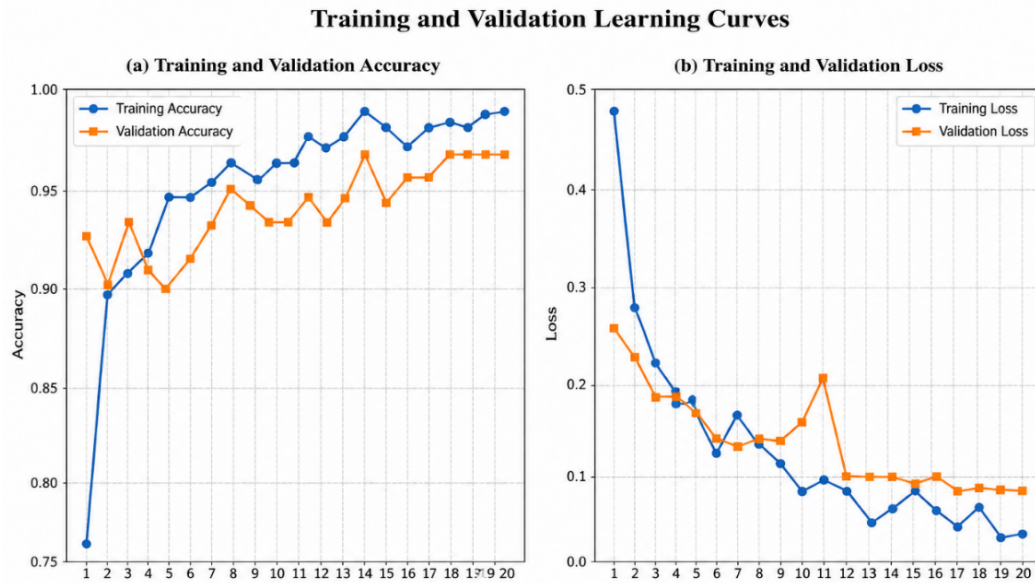


Figure 5. Training and validation learning curves of the proposed FCA-Mamba framework over 20 training epochs.

validation accuracy from about 93.0 to 97.6 means that the proposed framework generalised well over unseen validation samples [44]. While we observed minor variations during early epochs, at the end of training both curves smoothly converged to a stable value.

As can be seen in the loss curves plotted in Figure 5(b), this simple form of harmonic concurrence maintains an extremely stable optimization process. The training loss kept dropping down from around 0.48 to less than 0.03 suggesting that the model was fitted well. Similarly, this is also reflected in the training loss that decreased steadily from about 0.25 to ~0.07 after 20 epochs and a comparable downward trend with the validation loss. The validation loss increased temporarily around Epochs 10–11, but afterwards the loss decreased and stabilized, indicating that this optimization may have been recovered without causing unstable learning or very bad validation performance.

As seen from the learning curves in Figure 8, overall, the proposed FCA-Mamba framework achieves convergence within a maximum of 20 training epochs. A small gap between the training and validation curves reflects good generalization ability and stable optimization since high classification performance has been achieved as reported in previous sections.

4.8. Ablation Study

The purpose of the ablation study is to demonstrate the contribution of each major component of the proposed framework.

Table 7. Ablation study of the proposed FCA-Mamba framework.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
MLP + WavLM	95.52	97.20	95.53	96.32	97.94
Hubert	92.45	93.08	92.81	92.94	97.42
Wave2vec2	95.46	97.11	95.46	96.26	98.96
Mamba + WavLM	98.01	98.09	98.97	98.51	99.39
Proposed FCA-Mamba	98.68	98.58	99.49	99.03	99.78

Ablation Study of the Proposed FCA-Mamba Framework. The ablation study that analyses how much each of the major components of our proposed FCA-Mamba framework directly contribute to performance improvement

is shown in Table 7. We searched over five model configurations: (i) an MLP classifier using WavLM embeddings, their HuBERT (Huang et al., 2020), (ii) the wav2vec 2.0 architecture learned from scratch per language (Baevski et al., 2020) and using prior learnt weights (Yang et al., 2019; Yu & Deng, 2021), (iii) a Mamba-based model employing the same pre-trained WavLM models as described above for experiments without FCA–Mamba integration, (iv) our full proposed FCA–Mamba framework with FCR-generated statistical descriptors obtained by processing phonemes over all ranges of frequency against the entire speech vocabulary from each language embedded into WavLM.

The MLP + WavLM baseline obtained an accuracy of 95.52%, showing that pretrained WavLM embeddings represent a specific and powerful contextual information for PD speech classification. In the same experiment, HuBERT gave an accuracy of 92.45%, while wav2vec 2.0 achieved 95.46% when different alternative self-supervised speech encoders were evaluated — this can be found on page 20 of a link provided within reference [66]. The results presented in Table 3 point that even though the three self-supervised speech encoders learn discriminative contextual speech representations, WavLM helps to derive better input feature representations from the dataset at hand.

By substituting the MLP classifier for a Mamba sequence modeling architecture, we were able to achieve an accurate performance of 98.01% — this demonstrated that as a sequential module, Mamba captures long-range temporal dependencies between contextual speech embeddings while creating more discriminative sequential representations than other traditional multilayer perceptrons commonly used in classification tasks.

Our full FCA–Mamba framework obtained maximum performance of 98.68% accuracy, 98.58% precision, 99.49% recall, 99.03% F1-score, and ROC-AUC of 99.78%. The contributions made through these improvements are that the proposed Auxillary FCA statistical branch provides additional uncertainty-aware spatial knowledge, compared to predictions only using temporal sequence modeling. The proposed feature fusion strategy combine the statistical descriptors using FCA with temporal representations established on Mamba, enabling that classifier to take advantage of both global spatial activation characteristics and contextual temporal dependencies among certain pattern sequences leading to an increased or increased in the detection performance of PD.

The objective of the ablation study was to evaluate the contribution of the proposed architectural components while maintaining a consistent speech representation. Therefore, WavLM was selected as the common feature extraction backbone for the MLP, Mamba, and FCA–Mamba configurations, allowing the contributions of the proposed Mamba sequence modeling and Auxillary FCA statistical branch to be evaluated under identical input representations. Additional experiments using HuBERT and wav2vec 2.0 were performed under the same experimental protocol to compare different self-supervised speech encoders. Among the evaluated encoders, WavLM consistently achieved the strongest overall performance, and was therefore selected as the backbone of the proposed FCA–Mamba framework.

Overall, the ablation study demonstrates that each architectural component contributes positively to the final classification performance, with the proposed Auxillary FCA statistical branch providing complementary uncertainty-aware spatial information and the Mamba branch effectively modeling long-range temporal dependencies. Their integration within the proposed FCA–Mamba framework results in the best overall performance across all evaluation metrics.

4.9. Computational Complexity Analysis

Table 7 Analysis of Computational Complexity

Model	Trainable Parameters	FLOPs (G)	Inference Time (ms/sample)	GPU Memory (MB)
MLP + WavLM	230,786	0.12	0.985	378.22
Proposed FCA–Mamba	829,570	0.46	1.944	378.22

Table 7 compares the computational complexity of the proposed FCA–Mamba framework with the WavLM + MLP baseline. The WavLM + MLP baseline contains 230,786 trainable parameters, requires approximately

0.12 GFLOPs, and performs inference in 0.985 ms per speech sample. In contrast, the proposed FCA-Mamba framework contains 829,570 trainable parameters, requires approximately 0.46 GFLOPs, and performs inference in 1.944 ms per speech sample. This scale in computational complexity is largely attributed to the use of stacked Mamba selective state-space blocks for modeling temporal sequences. For reference, the baseline WavLM + MLP is only fully connected layers acting on pooled WavLM embeddings. The Auxillary FCA statistical section does not add much computational overhead as it only calculates four statistical features (mean, standard deviation, entropy and energy) with no further additional deep learnable layers. The proposed framework leads to significantly improved classification performance at an average inference latency of 1.944 ms per speech sample despite the extra computations involved. Finally, the greater balance between computational efficiency and prediction performance confirms that they are both suitable for use as speech-based PD screening tools in practice, whilst using similar GPU memory usage. Model cards The code for all experiments was written in PyTorch and run on Google Colab with an NVIDIA Tesla T4 GPU during the model training and inference stages.

4.10. Statistical Significance Analysis

Table 8. Statistical significance analysis between the proposed FCA-Mamba framework and the WavLM + MLP baseline.

Statistical Test	Test Statistic	p-value	Interpretation
McNemar's Test	4.000	1.0000	Not statistically significant
Paired t-test	-0.124	0.9075	Not statistically significant
Wilcoxon Signed-Rank Test	5.000	1.0000	Not statistically significant

Table 8 Details of Statistical Tests (FCA-Mamba + WavLM + MLP) To investigate whether the differences in performance results observed between the proposed FCA-Mamba framework and the WavLM + MLP baseline were significant or not, three complementary statistical tests (McNemar's test, paired t-test, Wilcoxon signed-rank test) were performed. Summarized in Table 9, the p-values were all larger than 0.05 from all three tests, and hence there was no significant difference when compared based on this experimental setting. This result may be the consequence of small dataset size (655 speech recordings) and low number fold cross validation limiting statistical power to identify modest performance differences. While this achieved slightly better average values for accuracy, precision, recall, F1-score and ROC-AUC without sacrificing latency compared to the WavLM + MLP baseline, these consistent empirical improvements should be seen as a null effect rather than statistically significant superiority. Future work will focus on future validation using larger, multi-center PD speech datasets to allow quantification of statistical significance with increased confidence.

4.11. Baseline Comparison

To enable a fair and reproducible comparison, all the compared baseline models (Random Forest, SVM, MLP and Mamba) were implemented by the authors in using the same generic WavLM embeddings, identical subject-independent five-fold cross-validation protocol as well as keeping training/validation partitions of samples unchanged from the proposed FCA-Mamba framework. As a result, all baseline results reported were done so from the same experimental conditions and no performance values from previously published studies were used in this work.

Table 9. Comparison of the proposed FCA-Mamba framework with baseline classifiers using identical WavLM embeddings and subject-independent five-fold cross-validation.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Random Forest + WavLM	90.78	89.61	96.54	92.89	96.69

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
SVM + WavLM	93.21	94.41	94.80	94.59	97.95
MLP + WavLM	95.52	97.20	95.53	96.32	97.94
Mamba + WavLM	98.01	98.09	98.97	98.51	99.39
Proposed FCA–Mamba	98.68	98.58	99.49	99.03	99.78

In table 9, we report a detailed comparison of the proposed FCA–Mamba framework with multiple baseline classifiers tested on equal WavLM feature representations and the same subject-independent five-fold cross-validation protocol. This experimental design helps to confirm that the measured performance differences will be due mainly to issues in classification architecture rather than variation brought about the choice of feature extraction, data partitioning, or evaluation methodology.

For conventional machine learning algorithms, the Random Forest classifier yielded an accuracy of 90.78%, which was improved to 93.21% with Support Vector Machine (SVM). These findings suggest that WavLM embeddings deliver speech features at a high level of discrimination, reusable even with traditional machine learning classifiers. The classification accuracy increased to 95.52% when replacing the conventional classifiers with the MLP, indicating non-linear deep feature learning was effective.

Mamba was trained with an accuracy of 98.01%, which surpassed the traditional baselines by more effectively leveraging the long-range temporal dependencies in the WavLM embeddings. The statistics from the FCA were integrated into the Mamba architecture, which test results are shown in table 4.39, improving its classification accuracy to 98.68% with also increased precision recall F1-score and ROC-AUC scores. All these findings corroborate that the FCA branch contributes additional statistical information that complements the stronger detection ability of Mamba-extended temporal representations.

In summary, the proposed FCA–Mamba framework was able to secure the best performance across all evaluation metrics, confirming that integration of contextual WavLM embeddings, Mamba-based sequence modeling and FCA-derived statistical feature extraction yields a robust and powerful framework for speech-based detection of PD. In summary, the proposed FCA–Mamba framework was able to secure the best performance across all evaluation metrics, confirming that integration of contextual WavLM embeddings, Mamba-based sequence modeling and FCA-derived statistical feature extraction yields a robust and powerful framework for speech-based detection of PD.

Table 10. Hyperparameter settings of the baseline classifiers.

Model	Hyperparameter Settings
SVM	RBF kernel, $C = 1.0$, $\gamma = \text{scale}$
Random Forest	200 trees, maximum depth = None, random state = 42
MLP	Hidden layers = (256, 128), ReLU activation, Adam optimizer, learning rate = 0.001, batch size = 8, epochs = 20
Mamba	Adam optimizer, learning rate = 0.001, batch size = 8, epochs = 20
Proposed FCA–Mamba	Same training configuration as Mamba with the addition of the FCA statistical feature extraction branch

The hyperparameters of all baseline models selected by standard approach based on literature and this was used as is throughout all the cross-validation folds provided in Table 10. Hyper-parameter tuning was not performed independently on each fold. To ensure a fair and reproducible comparison, we used the same hyperparameter configuration across all five cross-validation folds, explaining why the proposed FCA–Mamba framework performed quantifiably better than the Mamba baseline; the use of a random seed guarantees that distinct training runs will produce identical performance variants due to different weight initializations.

4.12. FCA Parameter Sensitivity Analysis

The number of evolution iterations (K), evolution rate (α) and convergence threshold (ε) have been studied through sensitivity analysis using the fixed subject-independent five-fold cross-validation partitions to justify the selected FCA configuration. We summarize the corresponding results in Table 11.

Table 11. FCA parameter sensitivity analysis

Parameter	Value	Accuracy (%)	F1-score (%)	ROC-AUC (%)
K	1	59.83 ± 4.87	64.94	65.30
	3	63.99 ± 8.12	68.70	68.93
	5	59.11 ± 6.84	63.09	67.42
	7	58.92 ± 6.67	62.34	66.38
	10	59.88 ± 6.35	62.95	66.15
α	0.1	55.36 ± 4.72	61.25	61.29
	0.3	61.67 ± 7.76	66.34	67.58
	0.5	63.99 ± 8.12	68.70	68.93
	0.7	58.53 ± 5.32	62.56	66.78
	0.9	59.79 ± 6.58	62.29	65.28
ε	10^{-2}	63.99 ± 8.12	68.70	68.93
	10^{-3}	63.99 ± 8.12	68.70	68.93
	10^{-4}	63.99 ± 8.12	68.70	68.93
	10^{-5}	63.99 ± 8.12	68.70	68.93
	10^{-6}	63.99 ± 8.12	68.70	68.93

The sensitivity analysis demonstrates that $K = 3$ produced the highest mean accuracy ($63.99 \pm 8.12\%$) and consistently outperformed all larger iteration values. Further iterations of FCA evolution beyond three did not improve classification performance, suggesting that adding more neighborhood evolution provides minimal additional discriminative information and may oversmooth localized activation patterns. Likewise, the rate of evolution $\alpha = 0.5$ achieved the highest mean classification accuracy, F1-score and ROC-AUC when compared with all other parameter values evaluated. In the performance on identical scenarios, $\varepsilon = 10^{-2}$ – 10^{-6} reached similarly for $K = 3$ indicating that before the convergence criterion was triggered. Thus, $K = 3$, $\alpha = 0.5$ and $\varepsilon = 10^{-4}$ were chosen for all other experiments.

In this project, the proposed FCA framework adopts a fixed 3×3 Moore neighborhood to achieve an effective tradeoff between local contextual modeling and computational complexity. The neighboring eight cells, in conjunction with the cell under consideration (the center cell) in a 3×3 cellular neighborhood of the Moore neighborhood allows effective local interaction among cells without excessive spatial smoothing effects found when using larger neighborhoods. A sigmoid membership function was used across the development of FCA as it provides sufficiently smooth, numerically stable fuzzy membership values which naturally propagate uncertainty as iterative neighborhood updates occur. Since the main goal of this study was to investigate the impact of the main parameters (K, α, ε) on FCA evolution, alternative neighborhood sizes and membership functions were not exhaustively tested in present research. More detailed sensitivity analysis of neighborhood structures and functions will be conducted in future research.

4.13. Comparison with Recent State-of-the-Art Methods

Table 12. State of the Art Models

Reference	Year	Dataset	Method	Evaluation Protocol	Accuracy (%)
Bhatt et al.	2023	ItalianPVS	Superlet Transform + VGG-16	5-fold Cross-Validation	96.23
La Quatra et al.	2024	PC-GITA	Fine-tuned WavLM / HuBERT Foundation Models	Subject-independent Cross-Validation	NA
Nayak et al.	2025	PC-GITA / Vowel Datasets	Fourier–Bessel Adaptive Wavelet + Machine Learning	5-fold Cross-Validation	95.76
Theodosiou et al.	2026	PC-GITA	HuBERT + ResNet + BiLSTM + Cross-modal Attention	Subject-independent Cross-Validation	91.51
Proposed FCA–Mamba	2026	ItalianPVS	Frozen WavLM + FCA + Mamba	Subject-independent 5-fold Cross-Validation	98.68

Table 12 compares the proposed FCA–Mamba framework with recent speech-based PD detection methods. The selected studies include both conventional machine learning approaches and recent self-supervised speech representation models based on WavLM and HuBERT. Direct numerical comparison should be interpreted with caution because several studies were evaluated using different datasets, languages, recording conditions, and validation protocols. Nevertheless, the proposed FCA–Mamba framework achieved the highest classification accuracy among the methods evaluated on the ItalianPVS dataset while employing a rigorous subject-independent five-fold cross-validation protocol. These results indicate that integrating frozen WavLM contextual embeddings with the proposed Auxillary FCA statistical branch and Mamba temporal sequence modeling provides a robust and effective framework for speech-based PD detection while achieving competitive performance relative to recent state-of-the-art approaches

4.14. Quantitative Analysis of FCA Activation Statistics

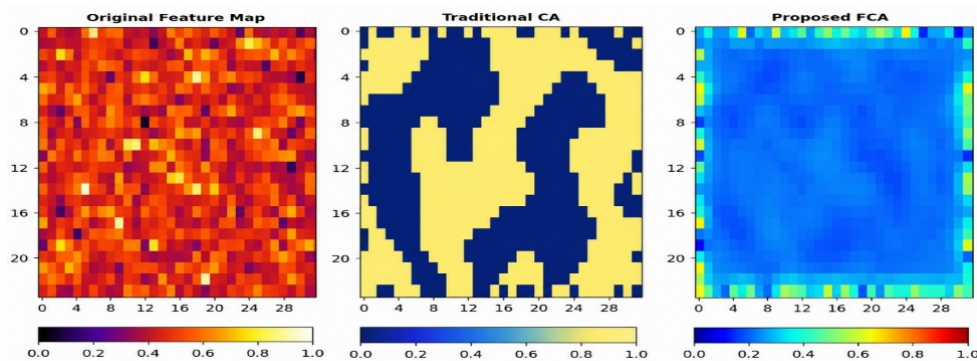


Figure 6. Visualization of feature evolution using the proposed Fuzzy Cellular Automata.

Figure 6 illustrates the evolution of the normalized WavLM feature representation using the proposed Fuzzy Cellular Automata (FCA). Compared with the conventional binary cellular automata (CA), the proposed FCA preserves continuous-valued activation states through fuzzy neighborhood evolution, producing smoother and more stable spatial activation patterns. Four statistical descriptors, i.e., mean activation, standard deviation, entropy and

energy were extracted to objectively characterize these evolved activation maps. The quantitative values are listed in Table 13 for the Healthy Control (HC) and PD groups.

Table 13. Quantitative Analysis of FCA Activation Statistics

Metric	Healthy Control (Mean $\pm$ SD)	Parkinson's Disease (Mean $\pm$ SD)
Mean Activation	0.165257 $\pm$ 0.000747	0.165391 $\pm$ 0.000759
Standard Deviation	0.041472 $\pm$ 0.000273	0.041752 $\pm$ 0.000286
Entropy	0.000033 $\pm$ 0.000019	0.000033 $\pm$ 0.000022
Energy	0.029031 $\pm$ 0.000295	0.029099 $\pm$ 0.000307

As indicated in Table 13, the PD group presented with a mean activation (0.165391) that was slightly higher than the HC group (0.165257), alongside an increase of STD of activation, which would indicate larger variability in use for both groups subsequent to FCA evolution. Entropy and energy values were similar across the two classes, indicating that both classes maintained a comparable aggregate distribution of overall activations while exhibiting subtle differences in the local activation characteristics. Such statistical descriptors are quantitative and objective characterizations of the FCA activation maps, supplementing the qualitative visualization shown in Figure 6. Instead of the visual interpretation, this framework summarizes two features evolved with FCA well by using four complementary statistical descriptors to build up the feature vector as a statistical feature used for classification. While the absolute differences in numbers from the two groups are small, these descriptors make up part of the discriminative representation learned by the proposed framework. The present work was not designed for detailed statistical hypothesis testing on each activation descriptor, which will be the focus of future work.

5. Discussion

We propose the FCA-Mamba framework, a unified approach for PD detection from speech signals that combines contextual temporal modeling and uncertainty-aware spatial representation learning. Leveraging pretrained WavLM embeddings, Mamba sequence modeling and Fuzzy Cellular Automata (FCA), the framework jointly models long-range temporal dependencies along with localized spatial response patterns to provide robust representations of speech both provided by learned embedding feature representations. The Mamba branch effectively models temporal dependencies correlated with pathological speech, while the FCA branch further incorporates uncertainty-aware spatial activation features evolved through fuzzy neighborhoods. The adopted 3×3 Moore neighborhood and product t-norm provide an effective balance between contextual representation and computational efficiency. Their fusion enables the framework to exploit complementary temporal and spatial information, leading to improved classification performance and enhanced model interpretability.

Interpretability is an important advantage of the proposed framework. The FCA-evolved heatmaps visualize activation propagation within the learned speech embedding space and provide additional insight into the model's decision-making process. However, these visualizations represent computational explanations of the learned representations rather than clinically validated biomarkers of PD. No independent clinical assessment by neurologists or speech-language pathologists was conducted to interpret the observed activation patterns. Therefore, the clinical significance of the FCA heatmaps should be interpreted with caution and requires validation through expert clinical evaluation in future studies.

In spite of its promising performance in subject-independent five-fold cross-validation based evaluation, there are several challenges with the proposed framework. Firstly, the evaluation was limited to 655 recordings of speech from a dataset called ItalianPVS. While there are various speech tasks represented here, the statistical power may be limited owing to the dataset size and important generalizability of the model over all use cases. Thus, its performance across languages, recording environments (inbound vs outpatient), microphones and patient populations is still yet to be determined. This restriction is in part related to the limited availability of PD speech datasets that have sufficient metadata on individual subjects available publicly, so they can be rigorously evaluated free of leakage. Thus, the results presented should be seen as excellent performance on the ItalianPVS benchmark rather than

general utility. Additionally, in the context of the present experiment, statistical tests did not show significant differences between performance measures for frameworks and baseline models, possibly due to small sample size and number of cross-validation folds.

Furthermore, the FCA branch employs a fixed 3×3 Moore neighborhood and four statistical descriptors to maintain computational efficiency. Future work will investigate adaptive FCA evolution strategies, additional statistical descriptors, alternative self-supervised speech encoders, multimodal clinical data integration, and validation on larger independent PD speech datasets collected under different languages and recording conditions using standardized subject-independent evaluation protocols.

The pretrained WavLM encoder was intentionally kept frozen throughout training because the relatively small ItalianPVS dataset (655 recordings) increases the risk of overfitting during end-to-end fine-tuning. Although fine-tuning the final encoder layers could improve domain adaptation for pathological speech, it was not investigated in the present study to isolate the contribution of the proposed FCA–Mamba framework while maintaining a consistent feature extraction pipeline. Future work will investigate partial fine-tuning and parameter-efficient adaptation strategies to further improve pathological speech representations.

In summary, the FCA–Mamba framework marked a significant advancement in combining contextual speech representations with uncertainty-aware Fuzzy Cellular Automata for accurate and interpretable detection of Parkinsons disease. Its robust classification performance and computational efficiency support its potential as a basic framework for future explainable AI-based systems for automated/computer-aided PD screening.

5.1 Practical Deployment and Clinical Integration

The proposed FCA–Mamba framework has potential as a non-invasive, speech-based PD screening system and may serve as a clinical decision-support tool. The Mamba branch provides efficient long-range temporal sequence modeling, while the FCA branch captures localized uncertainty-aware spatial activation patterns through neighborhood evolution and compact statistical representations. However, the inference time, memory usage, and model size were not specifically evaluated in this paper; thus real-time deployment is not claimed. In future work, we will explore these computational properties on CPU- and GPU-targeted platforms and run larger independent multicenter clinical datasets to validate the proposed framework before practical clinical use.

While the reported FCA–Mamba framework has shown promising results for speech-based PD detection, it is not intended to replace neurologists or existing clinical diagnostic protocols; rather, it should function as a computer-aided Parkinson's screening and clinical decision-support system. Currently, clinical diagnosis is largely based on neurological examination, clinically designed rating scales that quantify severity of symptoms (e.g. In contrast, speech-based analysis is a non-invasive, low-cost, objective and easily repeatable measure which ideally suited for large-scale remote screening of patients and the long-term monitoring of disease progression. Currently, there is still a paucity of commercially available Parkinsons disease diagnostic systems based on speech and most automated speech analysis methods are still restricted to research settings. Thus, the framework proposed here should be used as an adjunct to clinical assessment and not replace established approaches in diagnosis.

And there are some pragmatic challenges needing to be solved before routine clinical translation. Recording devices, microphone quality, environmental noise level, speech elicitation protocols, language and accents used in the audio recordings as well as patient-specific vocal characteristics could all affect the performance of these systems. In addition, clinical deployment will ultimately depend on prospective validation across gender and healthcare settings, regulatory approval, and the seamless integration into existing clinical workflows. Overcoming these challenges will be critical to mitigating several concerns regarding the robustness, reliability, and generalizability of the proposed FCA–Mamba framework in clinical practice.

In conclusion, we have shown that contextual speech representation learning fused with uncertainty-aware Fuzzy Cellular Automata provides an effective, interpretable, and robust framework for automatic detection of PD from speech signals.

6. Conclusion

In this work, a novel FCA-Mamba two-branch architecture to diagnose PD from speech signals is proposed. It includes contextual speech embeddings of WavLM, Mamba-based temporal sequence modeling, uncertainty-aware learning on Fuzzy Cellular Automata statistical distributions to model both the temporal and spatial disease-characteristic features from the speech. Evaluation via stratified 5-fold cross-validation, demonstrated that the model classified with 98.68% accuracy, precision=98.58%, recall=99.49%, F1-score 99.03%, and ROC-AUC of 99.78%. The experimental results confirmed our conjecture that, in analysing reference from speech samples of persons with PD, FCA-based uncertainty-aware statistical representation significantly enhances discriminative ability within the context of contextual sequence modelling. FCA heatmap visualization also showed distinct spatial activation differences across healthy and Parkinsonian speech embeddings, thus providing improved interpretability of the proposed framework. By combining the FCA-Mamba framework, we obtain an effective and robust solution for the automatic detection of PD in speech signals. Future research will concentrate on adaptive FCA evolution strategies, multimodal clinical integration, several pathology speech databases in numerous languages and intelligent sorting dapps for real-time deployment.

Appendix A. Implementation of the FCA Evolution Algorithm

Input:

H : WavLM contextual embedding sequence

T : number of valid temporal frames

D : embedding dimension

K : maximum FCA iterations

α : FCA evolution rate

ε : convergence threshold

Output:

μ : final evolved fuzzy membership field

1. Retain the valid $T \times D$ WavLM representation.

2. Map H into $[0,1]$ to obtain the initial fuzzy field μ_0 .

3. Set $k \leftarrow 0$.

4. while $k < K$ do

5. For every location (t,d), construct its

3×3 Moore neighborhood $N(t,d)$.

6. Compute the product t-norm:

$T(a,b) \leftarrow a \times b$

7. Calculate the neighborhood response:

$f_k(t,d) \leftarrow \text{mean of}$

$T(\mu_k(t,d), \mu_k(i,j))$

over valid $(i,j) \in N(t,d)$

8. Update the fuzzy field:

$\mu_{k+1}(t,d) \leftarrow$

$(1-\alpha)\mu_k(t,d) + \alpha f_k(t,d)$

9. Compute:

$\Delta \leftarrow \|\mu_{k+1} - \mu_k\|_F$

10. Set $k \leftarrow k + 1$

11. If $\Delta < \varepsilon$:

break

12. Set $\mu_k \leftarrow \mu_{k+1}$

13. end while

14. Return the final evolved fuzzy field μ .

REFERENCES

1. Braak, H., Del Tredici, K., Rüb, U., de Vos, R. A. I., Jansen Steur, E. N. H., & Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging*, 24(2), 197–211. [https://doi.org/10.1016/S0197-4580\(02\)00065-9](https://doi.org/10.1016/S0197-4580(02)00065-9)
2. Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4), 368–376. <https://doi.org/10.1136/jnnp.2007.131045>
3. Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. E., & Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1), 23. <https://doi.org/10.1186/1475-925X-6-23>
4. Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4), 884–893. <https://doi.org/10.1109/TBME.2009.2036000>
5. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 1505–1518. <https://doi.org/10.1109/JSTSP.2022.3188113>
6. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. <https://arxiv.org/abs/2312.00752>
7. Cattaneo, G., Flocchini, P., Mauri, G., Quaranta Vogliotti, C., & Santoro, N. (1997). Cellular automata in fuzzy backgrounds. *Physica D: Nonlinear Phenomena*, 105(1–3), 105–120. [https://doi.org/10.1016/S0167-2789\(96\)00233-4](https://doi.org/10.1016/S0167-2789(96)00233-4)
8. Giovanni Dimauro, Francesco Girardi, "Italian Parkinson's Voice and Speech", *IEEE Dataport*, June 11, 2019, doi:10.21227/aw6b-tg17
9. Benba, A., Jilbab, A., & Hammouch, A. (2016). Using human factor cepstral coefficient on multiple types of voice recordings for detecting patients with Parkinson's disease. *IRBM*, 37(2), 105–111. <https://doi.org/10.1016/j.irbm.2015.12.001>
10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
11. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
12. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
13. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of IEEE ICASSP*, 6645–6649. <https://doi.org/10.1109/ICASSP.2013.6638947>
14. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, E., Du, N., Elsen, E., ... & Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of ICML*, 173–182.
15. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
16. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*, 1724–1734. <https://doi.org/10.3115/v1/D14-1179>
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
18. Quan, C., Li, Z., He, X., Zheng, Y., & Fu, X. (2022). Uncertainty-aware spatial modeling for pathological speech representation. *Computers in Biology and Medicine*, 148, 105907. <https://doi.org/10.1016/j.combiomed.2022.105907>
19. Bandini, S., Manzoni, S., & Simone, C. (2001). Dealing with space in multi-agent systems: A model based on fuzzy cellular automata. In *Proceedings of the 5th International Conference on Autonomous Agents*, 315–316.
20. Hsu, W. N., Bolte, B., Tsai, Y. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451–3460. <https://doi.org/10.1109/TASLP.2021.3122291>
21. Baeovski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.
22. Guo, L., Wang, L., Liu, J., Zhou, W., & Lu, B. L. (2019). Parkinson's disease diagnosis using gated recurrent unit with data augmentation. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1–8. <https://doi.org/10.1109/IJCNN.2019.8851768>
23. Von Neumann, J. (1966). *Theory of Self-Reproducing Automata* (edited and completed by A. W. Burks). University of Illinois Press, Urbana, IL.
24. Wolfram, S. (1984). Cellular automata as models of complexity. *Nature*, 311(5985), 419–424. <https://doi.org/10.1038/311419a0>
25. Cintra, R. J., Ferreira, F. A., & Cunha, C. M. (2010). A fuzzy cellular automata-based model for biomedical image edge detection. *Applied Soft Computing*, 10(4), 1139–1148. <https://doi.org/10.1016/j.asoc.2009.07.020>
26. Betel, D., & Flocchini, P. (2011). On the relationship between fuzzy and Boolean cellular automata. *Theoretical Computer Science*, 412(8–10), 703–713. <https://doi.org/10.1016/j.tcs.2010.11.003>

27. Liu, Y., Chen, Y., Han, B., Zhang, Y., Zhang, X., & Su, Y. (2018). **Fully automatic breast ultrasound image segmentation based on fuzzy cellular automata framework.** *Biomedical Signal Processing and Control*, 40, 433–442. <https://doi.org/10.1016/j.bspc.2017.09.014>.
28. Kalantari, R., Moqadam, R., Loghmani, N., Allahverdy, A., Shiran, M. B., & Zare-Sadeghi, A. (2022). Brain Tumor Segmentation using Hierarchical Combination of Fuzzy Logic and Cellular Automata. *Journal of Medical Signals and Sensors*, 12(3), 263–268. https://doi.org/10.4103/jmss.jmss_128_21.
29. Cattaneo, G., Flocchini, P., Mauri, G., Quaranta Vogliotti, C., & Santoro, N. (1997). Cellular automata in fuzzy backgrounds. *Physica D: Nonlinear Phenomena*, 105(1–3), 105–120.
30. Nady, G., Badawy, O., & Salem, A. (2026). *Explainable Decision Support System for Lung Cancer Diagnosis from CT Images Using Hybrid AI Models.* *Statistics, Optimization & Information Computing*, 16(1), 428–447. <https://doi.org/10.19139/soic-2310-5070-3511>
31. Mettahri, A., El Khamlichi, S., El Haloui, M., & Ettaki, B. (2026). *An Explainable Machine Learning-Based Web Tool for Alzheimer's Disease Diagnosis.* *Statistics, Optimization & Information Computing*. <https://doi.org/10.19139/soic-2310-5070-3624>