ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2026.0559

MULTI-SCALE TEMPORAL TRANSFORMER LEARNING FOR ACCURATE PARKINSON'S DISEASE CLASSIFICATION FROM HETEROGENEOUS BIOMEDICAL SIGNALS

Thamari Thankam[1] and K. Hakkins Raj[2]
[1]*Department of Community Health Nursing, Cihan University-Erbil, Erbil, Iraq*
[2]*School of Biomedical Engineering, Jimma Institute of Technology, Ethiopia*

Abstract

Parkinson's disease (PD) is an incurable neurological disease that affects patients through motor and non-motor disturbances that may be challenging to diagnose effectively. The current conventional diagnostic system relies on clinical examination which introduces subjectivity and cannot capture small variations in the temporal dimension. Current machine learning/deep learning models have been able to diagnose Parkinson's disease automatically through gait, speech, facial expressions, handwriting, and EEG signals. These models, however, emphasize the importance of either the spatial/local patterns or the long-term temporal dynamics but give little consideration to the multi-scale temporal patterns which may limit their robustness in the presence of disease-related patterns at multiple scales. In this paper, we propose a new model for Parkinson's disease diagnosis based on a Multi-Scale Temporal Attention Transformer (MSTAT). The method uses parallel temporal convolutional streams for obtaining short-, medium- and long-term representations, followed by hierarchical self-attention for modeling temporal dependencies among different distributed features. A gating function dynamically weights each scale while a residual fusion layer aggregates complementary representations before classification. The subject-wise normalization and stratified evaluation are employed to avoid any information leakage. We report 97.2% accuracy, 96.7% precision, 96.4% recall, 96.5% F1-score, and 97.3% specificity in the final evaluation. Compared with CNN-GRU, LSTM, and CNN-Transformer, the proposed MSTAT achieves 4.6%, 4.3%, and 2.6% higher accuracy, respectively. Also, MSTAT increases the recall rate with respect to the three baselines by 4.8%, 4.3%, and 2.9%. This indicates the ability of MSTAT to identify Parkinson's disease accurately. The highest specificity of 97.3% confirms the effective discrimination between the healthy people and Parkinson's disease patients.

Keywords:

Parkinson's Disease, Temporal Transformer, Multi-Scale Learning, Deep Learning, Biomedical Signal Classification

1. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disease associated with disorders in movement, coordination, posture, speech, facial expressions, and other neurological activities. It is typically diagnosed based on clinical manifestations such as bradykinesia, rigidity, tremor, and postural instability, which might develop gradually and vary in severity among patients.

As a result, the challenge of identifying subtle changes caused by PD still requires investigation. Current research shows that AI is capable of detecting clinical markers in gait, speech, facial movement, EEG, and other signals and provides objective means for computational assessment of PD [1]. Deep learning is especially promising since it is able to learn hierarchical representations from complex biomedical data without relying solely on manually extracted features [2]. Further multimodal deep learning research shows that the information about gait, upper limb movement, speech, facial expression and their combinations allows the diagnosis and symptom assessment of PD [3].

However, there are a number of open problems. For example, the biomedical PD datasets are typically small, heterogeneous in acquisition protocols and imbalanced in terms of ratio between patients and controls. Speech studies are limited by lack of publicly available data, confidentiality issues, demographics, and variety of languages and recording conditions [4]. Another systematic review reported a number of issues regarding biases, interpretability, privacy, computational costs, and the need for the vast amount of training data due to complexity of transformer architectures [5].

Existing PD classification techniques often process the temporal information at a single representation scale or use feature extractors in combination without specifying which temporal scale is informative. This can lead to inadequate representation of important short-term and long-term patterns of the disease. Besides, high performance in some studies might not mean good generalization due to small subject cohort, random sample-level partitions and lack of external validations. In particular, research in EEG shows that external testing might give different performance results compared to internal validation [6].

The novelty is the introduction of a novel framework named Multi-Scale Temporal Attention Transformer (MSTAT) that first disentangles the learned features into different temporal scales before performing hierarchical self-attention. Differing from the traditional Transformer that processes information at one token level, MSTAT employs multiple representations of the temporal scale and performs weighted aggregation via a temporal gating module.

In this way, local temporal dynamics and global sequence information are effectively combined in this model design. The intuition of such design is driven by the fact that the pathological signals related to PD may present themselves not only on short-time scales but also on long time scales.

The contributions of this work include the following aspects:

- An innovative MSTAT framework is designed for PD classification through the combination of multi-scale temporal convolution, hierarchical self-attention, temporal-scale gating, and residual feature fusion.
- A systemized temporal evaluation procedure is conducted with respect to the performance of multi-scale temporal learning.

2. RELATED WORKS

There has been a growing number of works dedicated to gait analysis due to PD influence on walking rhythm, stride features, balance and movement coordination. In particular, a study on classification of PD using gait signal analysis and deep learning methods based on CNN, GRU and hybrid CNN-GRU was published. It was found that the combination of spectral and temporal information provided better classification results. The hybrid architecture obtained accuracies of up to 83.7% for PD classification against elderly control group and 92.7% for PD classification against young healthy control group [7]. The results show that temporal information is very important for classification of gait abnormalities in PD. However, the hybrid CNN-GRU architecture captures temporal information using recurrent structure and does not discriminate multiple temporal scales. Also, LSTM-based models for classification of gait abnormalities in PD have been researched. An LSTM architecture aimed at learning long-term dependencies in gait cycle has been suggested and obtained average accuracies of 98.6% for binary classification and 96.6% for multi-class classification [8]. While LSTM networks are capable of learning sequential dependencies effectively, the recurrent nature of such models hinders parallelization and effective representation of the relations between distant temporal positions. It prompted researchers to study attention-based architectures. Besides gait analysis, EEG is another important source of information for computational analysis of PD. A study on generalizable EEG classification problem utilized channel-wise CNN and tested it using data obtained in two independent centers. Patient-level accuracies were 80.4% during the primary evaluation and 82.8% on the external dataset, thus showing the necessity of independent testing for determination of generalization [9]. In more recent studies, EEG biomarkers of freezing of gait have been explored. In particular, the comparison of conventional machine learning models and LSTM in EEG study yielded the following results: LSTM has AUC of 0.68 and accuracy of 0.63 [10]. Finally, facial expressions analysis is another non-invasive direction of research. PD-induced hypomimia leads to reduction of facial movements and can be considered as a source of information for automated diagnosis of the disease. For example, Jin et al. used facial landmarks and temporal facial features with LSTM and conventional machine learning models. They obtained LSTM model with 86% precision and 75% F1-score, while SVM with facial movement features achieved higher F1-score [11]. Subsequent work demonstrated that modeling image sequences rather than individual facial frames improved PD detection accuracy from 72.9% to 78.4%, while action-unit domain adaptation further improved performance to 87.3% [12]. These results reinforce the importance of temporal context in facial-expression-based PD assessment. Deep learning has also been applied to dynamic handwriting. A CNN-LSTM architecture was developed to exploit both local patterns and time-varying characteristics in handwriting signals. The model achieved 96.2% accuracy on the DraWritePD dataset and 90.7% on the PaHaW dataset while maintaining a relatively small computational footprint [13]. This work demonstrates that combining convolutional and recurrent representations can be effective for dynamic PD biomarkers, although it does not provide explicit multi-scale attention across temporal resolutions. Transformer-based approaches have recently become increasingly relevant. A systematic review of speech-based deep learning methods found that CNNs remain prevalent, while Transformers are becoming increasingly competitive. The review also emphasized the challenges associated with limited datasets, computational requirements, bias, explainability, and privacy [14]. Transformer-based analysis has also been extended to gait and multimodal biomedical signals.

3. PROPOSED METHOD

The proposed Multi-Scale Temporal Attention Transformer (MSTAT) is aimed at the classification of Parkinson's disease based on the acquisition of temporal features across different time scales in biomedical signals. The fundamental assumption is that there is no particular time scale for the existence of Parkinsonian traits. Small time scale features may describe the tremor or short-time movement anomalies, while larger time scale features might capture the rhythm of the gait, modulation of the voice, posture variations, or overall physiology. Therefore, the input signal is processed via multiple temporal feature extraction paths in which the receptive field of each path is different. Then, these representations are encoded as temporal tokens and transformed via hierarchical self-attention mechanism. Temporal-scale gating mechanism learns the relative weight of each time scale, while the residual temporal fusion helps to preserve the complementary information from the separate paths. The final step of the processing chain is the classification head which estimates the probability of Parkinson's disease or healthy controls.

3.1 INPUT REPRESENTATION AND SUBJECT-LEVEL PREPROCESSING

The first stage converts the acquired biomedical signal into a standardized representation suitable for temporal learning. Let the signal recorded from subject *i* contain *C* channels and *T* temporal observations. The input can therefore be represented as a matrix $X_i \in R^{C\times T}$. Depending on the application, the channels may correspond to accelerometers, gyroscopes, EEG electrodes, acoustic features, facial landmarks, handwriting coordinates, or other temporally sampled measurements. Instead of treating every observation independently, MSTAT preserves the temporal ordering because disease-related characteristics frequently depend on the relationship between consecutive and distant observations.

Before feature extraction, subject-level normalization is applied. This step is important because amplitude differences may originate from recording conditions rather than disease status. For each channel, the mean and standard deviation are calculated from the training portion of the corresponding subject-independent training data. The same transformation is subsequently applied to validation and test subjects. This prevents information from the test population from influencing model construction.

The normalized representation is defined as:

$$\tilde{X}_{i,c,t} = \frac{X_{i,c,t} - \mu_c^{(\text{train})}}{\sqrt{\left(\sigma_c^{(\text{train})}\right)^2 + \text{ò}}}, \tag{1}$$

$$X_i \in \square^{C\times T},\ 1 \le c \le C,\ 1 \le t \le T$$

where, $\mu c(train)$ and $\sigma c(train)$ denote the training-set mean and standard deviation of channel c, respectively, while ϵ prevents numerical instability. Subject-level preprocessing is particularly important in PD classification because repeated samples from the same individual can otherwise produce artificially high performance. The proposed framework therefore assumes that all windows originating from one subject remain within the same partition during evaluation.

3.2 TEMPORAL SEGMENTATION AND WINDOW CONSTRUCTION

Long biomedical sequences may contain thousands of temporal observations, making direct processing computationally expensive. MSTAT consequently divides each normalized sequence into overlapping temporal windows. Windowing also enables the model to learn local temporal structures while retaining sufficient context for long-range dependency analysis. Let L represent the window length and S represent the stride. The k-th window is extracted from the normalized signal using consecutive temporal positions.

The temporal window representation is expressed as:

$$W_i^{(k)} = \tilde{X}_{i,:,(k-1)S+1:(k-1)S+L} \in \mathbb{R}^{C\times L},\ k = 1,2,\ldots,K_i, \tag{2}$$

$$K_i = \left\lfloor \frac{T-L}{S} \right\rfloor + 1 \tag{3}$$

Overlapping windows are useful because an abnormal event may occur close to a window boundary. A non-overlapping segmentation strategy could divide one clinically meaningful pattern between two independent windows. Using an appropriate overlap reduces this boundary effect and provides multiple contextual views of the same temporal sequence. Nevertheless, window generation must be performed after subject-wise partitioning to avoid leakage between training and testing data. The resulting windows form the input samples for the multi-scale temporal extraction stage.

3.3 MULTI-SCALE TEMPORAL FEATURE EXTRACTION

The principal feature-learning component of MSTAT consists of three parallel temporal branches. Each branch is designed to capture a different temporal scale. The first branch uses a relatively small convolutional kernel and focuses on short-term variations. The second branch uses a medium-sized kernel to capture intermediate temporal patterns, while the third branch uses a larger effective receptive field to model broader temporal structures.

For a window $Wi(k)$, the three branches generate feature maps Fs, Fm, and Fl. Dilated convolution can additionally be used in the large-scale branch to increase the receptive field without excessively increasing the number of parameters. The mathematical operation for each scale can be represented as:

$$F_r^{(k)} = \phi\left(\mathrm{BN}\left(\mathrm{Conv1D}_{\Theta_r}\left(W_i^{(k)}; K_r, D_r, P_r\right)\right)\right),\quad r \in \{s,m,l\},\quad F_r^{(k)} \in \mathbb{R}^{d_r\times L_r} \tag{4}$$

where Kr, Dr, and Pr denote the kernel size, dilation factor, and padding associated with scale r. $\phi(\cdot)$ represents a nonlinear activation function, and batch normalization stabilizes feature distributions. The short-scale branch is particularly sensitive to rapid changes, whereas the large-scale branch can capture slowly evolving patterns. This separation is important because PD manifestations can have both high-frequency and low-frequency temporal characteristics.

The branches operate simultaneously rather than sequentially. Consequently, the model does not force a single convolutional representation to encode all temporal structures. This is a key distinction between MSTAT and conventional CNN-based approaches.

3.4 TEMPORAL FEATURE PROJECTION AND TOKENIZATION

The feature maps generated by the parallel branches may have different dimensions. Directly applying self-attention to these heterogeneous representations would make scale integration difficult. Therefore, each feature map is projected into a common latent dimension d. Temporal positions are retained as individual tokens so that the Transformer can determine relationships among observations occurring at different times. The scale-specific token representation is generated using a learned projection function:

$$Z_r^{(k)} = Flatten_t\left(F_r^{(k)}\right)W_r^p + b_r^p \in \mathbb{R}^{N_r\times d},\ r \in \{s,m,l\} \tag{5}$$

where, N_r denotes the number of temporal tokens generated at scale r, while W_r^p and b_r^p are learnable projection parameters. The projection places representations from different branches into the same feature space. This common space permits subsequent attention layers to compare temporal patterns across scales. Importantly, tokenization does not discard temporal order. Instead, each token remains associated with a specific temporal position.

3.5 POSITIONAL ENCODING

Self-attention itself does not inherently understand the chronological order of tokens. Therefore, temporal positional information is explicitly incorporated into each projected feature representation. A learnable positional embedding is used so that the model can identify whether a particular pattern occurs at the beginning, middle, or end of a temporal window.

The resulting scale-specific representation is:

$$H_r^{(k,0)} = Z_r^{(k)} + P_r^{\mathrm{pos}}, \tag{6}$$

$$P_r^{\mathrm{pos}} \in \mathbb{R}^{N_r\times d}. \tag{7}$$

The positional representation allows the attention mechanism to distinguish two otherwise identical feature vectors occurring at different temporal locations. This is relevant to PD signals because the clinical meaning of a movement or speech event can depend on its position within a temporal sequence. For example, a transient irregularity at the beginning of a movement cycle may not carry the same contextual meaning as an identical irregularity occurring after prolonged instability.

3.6 HIERARCHICAL TEMPORAL SELF-ATTENTION

After tokenization and positional encoding, the three scale-specific representations are processed using Transformer encoder layers. Multi-head self-attention allows each temporal token to

interact with other tokens, including those separated by substantial temporal distances. This capability enables MSTAT to identify dependencies that cannot be efficiently represented using only local convolution.

For a given Transformer layer, the query, key, and value matrices are generated from the input representation. The attention operation is formulated as:

$$\mathrm{MHSA}(H)=\left(\left[\sigma\left(\frac{(HW_j^Q)(HW_j^K)^{\cdot}}{\sqrt{d_h}}\right)\otimes HW_j^V\right]_{j=1}^{M}\right)W^O, \quad (8)$$

$$H\in\mathbb{R}^{N\times d}$$

where M denotes the number of attention heads and $dh=d/M$ represents the dimension assigned to each head. Different attention heads can learn different temporal relationships. One head may focus on neighboring observations, whereas another can associate temporally distant events. This property is particularly useful for gait and speech signals, where repetitive or slowly changing patterns may span several temporal intervals. A hierarchical arrangement is used rather than a single attention operation. Initial layers focus on scale-specific temporal dependencies, while subsequent layers integrate information across scales. This provides a structured progression from local temporal representation to global temporal reasoning.

3.7 RESIDUAL TRANSFORMER ENCODING

Each attention layer is combined with residual connections and normalization. Residual connections are essential because the original temporal representation may contain information that could be weakened during repeated nonlinear transformations. The feed-forward sublayer then provides additional nonlinear feature transformation after attention.

The Transformer block is represented as:

$$H^{(\ell+1)}=\mathrm{LN}\left[H^{(\ell)}+\mathrm{FFN}\left(\mathrm{LN}\left[H^{(\ell)}+\mathrm{MHSA}\left(H^{(\ell)}\right)\right]\right)\right], \quad (9)$$

$$\ell=0,\ldots,L_T-1$$

where *LN* represents layer normalization and *FFN* is a position-wise feed-forward network. The residual formulation allows the network to preserve lower-level temporal information while progressively learning more abstract dependencies. Multiple Transformer layers therefore construct a hierarchy of temporal representations rather than relying on one attention calculation.

3.8 CROSS-SCALE TEMPORAL INTERACTION

After scale-specific Transformer processing, the three representations contain complementary information. However, simply concatenating them assumes that every scale contributes equally. MSTAT instead introduces cross-scale interaction before final fusion. The scale-specific outputs are aligned to a common token dimension and combined to allow the network to identify relationships between short-, medium-, and long-duration patterns.

The cross-scale representation can be defined as:

$$U^{(k)}=\left(H_s^{(k,L_T)}\otimes H_m^{(k,L_T)}\otimes H_l^{(k,L_T)}\right)W^f+b^f\in\mathbb{R}^{N\times d} \quad (10)$$

The projection matrix W^f compresses the concatenated representation into a manageable latent dimension. This operation retains information from all three branches while allowing subsequent layers to learn interactions among them.

3.9 ADAPTIVE TEMPORAL-SCALE GATING

A major component of MSTAT is the temporal-scale gating mechanism. Different subjects may exhibit disease-related information at different temporal resolutions. Consequently, a fixed weighting scheme would be inappropriate. The model learns a separate importance coefficient for each scale based on the encoded representation. The adaptive scale weights are obtained through global temporal pooling followed by a small gating network:

$$\boldsymbol{\alpha}^{(k)}=\mathrm{softmax}\left(W_g\,\delta\left[W_z\begin{bmatrix}\mathrm{GAP}(H_s^{(k)})\\ \mathrm{GAP}(H_m^{(k)})\\ \mathrm{GAP}(H_l^{(k)})\end{bmatrix}+b_z\right]+b_g\right), \quad (11)$$

$$\boldsymbol{\alpha}^{(k)}=[\alpha_s,\alpha_m,\alpha_l],\quad \sum_r\alpha_r=1$$

where, δ denotes a nonlinear activation and *GAP* represents global average pooling across temporal tokens. The softmax operation guarantees that the scale coefficients form a normalized distribution. A larger α_s indicates that short-scale characteristics contribute more strongly to the classification decision, whereas a larger α_l indicates stronger dependence on long-range temporal information. This adaptive weighting is important because PD is heterogeneous. Some individuals may exhibit pronounced rapid movement irregularities, whereas others may show more apparent long-duration changes. The gating module therefore allows the network to adapt its representation to the temporal characteristics of each subject rather than imposing identical temporal priorities.

3.10 RESIDUAL MULTI-SCALE FUSION

The weighted representations are subsequently fused using a residual mechanism. The purpose of this operation is to combine scale-specific information without eliminating useful information from the original branches. Each Transformer representation is multiplied by its learned scale coefficient and then aggregated.

The proposed residual fusion operation is expressed as:

$$F_{\mathrm{MS}}^{(k)}=H_0^{(k)}+W_{\mathrm{MS}}\left(\alpha_s^{(k)}H_s^{(k)}+\alpha_m^{(k)}H_m^{(k)}+\alpha_l^{(k)}H_l^{(k)}\right)+b_{\mathrm{MS}} \quad (12)$$

where $H_0^{(k)}$ represents the initial projected representation and W_{MS} performs feature transformation after scale-weighted aggregation. The residual connection provides a direct information pathway from the original representation to the final fused representation. This reduces the possibility that the gating operation suppresses a useful feature simply because its scale receives a lower weight.

3.11 TEMPORAL ATTENTION POOLING

The fused representation still contains multiple temporal tokens. A direct average would treat every temporal position equally, which may not be appropriate because only some intervals contain strong disease-related evidence. MSTAT therefore applies an attention pooling mechanism to assign higher weights to informative temporal positions. The temporal aggregation is formulated as:

$$\beta_t = \frac{\exp\left(v_\beta^{\cdot} \tanh\left(W_\beta F_{\mathrm{MS},t} + b_\beta\right)\right)}{\sum_{j=1}^{N} \exp\left(v_\beta^{\cdot} \tanh\left(W_\beta F_{\mathrm{MS},j} + b_\beta\right)\right)}, \qquad h_{\mathrm{subj}} = \sum_{t=1}^{N} \beta_t F_{\mathrm{MS},t} \tag{13}$$

The coefficient β_t indicates the relative importance of temporal position t. The resulting vector h_{subj} represents the complete subject-level temporal signature. This mechanism prevents uninformative intervals from contributing equally to the final decision and can potentially provide useful interpretability by identifying temporally important regions.

3.12 CLASSIFICATION LAYER

The final subject representation is passed through a fully connected classification head. For binary PD classification, the classifier produces a single logit that is converted into a probability using the sigmoid function. The resulting probability represents the estimated likelihood that the input belongs to the PD class.

The classification function is given by:

$$\hat{y}_i = \sigma\left(W_c \,\mathrm{Dropout}\left(\delta\left(W_h h_{\mathrm{subj},i} + b_h\right)\right) + b_c\right), \qquad \sigma(z) = \frac{1}{1+e^{-z}}, \quad \hat{y}_i \in [0,1] \tag{14}$$

A probability above the selected decision threshold is interpreted as PD, whereas a probability below the threshold is interpreted as the control class. The threshold can be selected using the validation set rather than the test set to avoid optimistic evaluation.

3.13 CLASS-BALANCED OPTIMIZATION

PD datasets may contain unequal numbers of patients and healthy controls. Training a conventional binary classifier without accounting for class imbalance can cause the model to favor the majority class. MSTAT therefore employs a weighted binary cross-entropy objective. The weighting coefficients are determined from the class distribution within the training data. The optimization objective is defined as:

$$\mathrm{L}_{\mathrm{total}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\begin{matrix} w_+ y_i \log(\hat{y}_i + \grave{o}) + \\ w_-(1-y_i)\log(1-\hat{y}_i + \grave{o})\end{matrix}\right] + \lambda \sum_{\theta \in \Theta} \| \theta \|_2^2 \tag{15}$$

where w_+ and w_- compensate for class imbalance, λ controls L2 regularization, and Θ represents the set of trainable parameters. The regularization term discourages excessively large parameter values and can improve generalization when the training dataset is relatively small.

Optimization can be performed using AdamW with an initially selected learning rate followed by a cosine or validation-based learning-rate schedule. Early stopping can be applied according to validation loss or F1-score. Importantly, hyperparameter selection must be performed exclusively using the training and validation partitions.

3.14 VALIDATION STRATEGY

A clinically meaningful evaluation requires subject-level separation. If several windows from the same individual appear in both training and testing sets, the model may learn subject-specific characteristics instead of disease-related patterns. The proposed framework therefore partitions data at the subject level before window extraction or, equivalently, ensures that all windows from a subject remain within one partition.

For N_s subjects, the partition can be represented as:

$$\mathrm{D} = \mathrm{D}_{\mathrm{train}} \cup \mathrm{D}_{\mathrm{val}} \cup \mathrm{D}_{\mathrm{test}}, \quad \mathrm{D}_a \cap \mathrm{D}_b = \varnothing \quad \forall a \neq b, \quad \mathrm{Subject}(\mathrm{D}_a) \cap \mathrm{Subject}(\mathrm{D}_b) = \varnothing \tag{16}$$

This separation provides a more realistic estimate of how the proposed system behaves when it encounters previously unseen individuals. Stratification should also be considered to preserve an appropriate PD-to-control ratio in each partition. For limited datasets, subject-wise cross-validation can be used to obtain more stable estimates.

3.15 SUBJECT-LEVEL DECISION

During inference, an unseen subject is processed using exactly the same preprocessing, windowing, multi-scale feature extraction, Transformer encoding, scale gating, and temporal aggregation procedures used during training. Each temporal window produces a classification probability. Rather than assigning the subject label from one window, MSTAT aggregates window-level predictions to obtain a subject-level decision. The subject-level probability can be obtained through weighted temporal aggregation:

$$\hat{p}_{\mathrm{subj}} = \frac{\sum_{k=1}^{K_i} \omega_k \hat{p}_{i,k}}{\sum_{k=1}^{K_i} \omega_k}, \quad \omega_k \geq 0, \qquad \hat{y}_{\mathrm{subj}} = \mathrm{I}\left[\hat{p}_{\mathrm{subj}} \geq \tau\right] \tag{17}$$

where $\hat{p}_{i,k}$ is the probability predicted for window k, ω_k is its optional reliability weight, and τ is the validation-derived decision threshold. Subject-level aggregation reduces sensitivity to isolated noisy windows and provides a more clinically meaningful final prediction.

3.16 COMPUTATIONAL COMPLEXITY

The multi-scale architecture inevitably introduces additional computation because three temporal branches are processed in parallel. However, the parallel branches can use relatively compact convolutional layers, while the Transformer operates on projected tokens rather than the original high-dimensional signal. If N represents the number of temporal tokens and d represents the embedding dimension, the dominant self-attention complexity is approximately quadratic in the token sequence length. Consequently, temporal windowing and token projection are important for controlling computational requirements. The approximate attention complexity can be expressed as:

$$\mathrm{C}_M = \sum_{r \in \{s,m,l\}} \mathrm{O}\left(L_r C d_r K_r\right) + \mathrm{O}\left(Nd^2\right) + \mathrm{O}\left(N^2 d\right) + \mathrm{O}\left(d^2 N\right) \tag{18}$$

The first term corresponds approximately to multi-scale temporal convolution, the second and third terms represent Transformer projections and attention interactions, and the final term accounts for feed-forward transformations. The design therefore balances the richer representation obtained from multiple temporal scales against computational cost. In practical implementation, efficient tokenization, moderate window lengths, mixed-precision training, and appropriate batch sizes can reduce computational overhead.

4. RESULTS AND DISCUSSION

4.1 EXPERIMENTAL SETTINGS

The proposed MSTAT is implemented using Python 3.11 with PyTorch and Scikit-learn libraries. The experiments are conducted on a workstation equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX 3060 GPU with 12 GB VRAM. The model is trained using the AdamW optimizer for 50 epochs with subject-wise data partitioning. The experimental implementation uses GPU acceleration for Transformer training, while preprocessing, statistical analysis, and metric computation are performed using Python.

4.2 EXPERIMENTAL SETUP AND PARAMETERS

The principal experimental parameters used for training and evaluating MSTAT are summarized in Table.1. The selected parameters provide a balance between temporal representation capacity and computational complexity.

Table.1. Experimental parameters used for MSTAT

Parameter	Value
Programming language	Python 3.11
Deep learning framework	PyTorch
Machine learning library	Scikit-learn
Processor	Intel Core i7
RAM	32 GB
GPU	NVIDIA RTX 3060, 12 GB
Input temporal window	128 samples
Window overlap	50%
Temporal branches	3
Convolution kernel sizes	3, 5, 9
Embedding dimension	128
Transformer encoder layers	4
Attention heads	8
Feed-forward dimension	512
Dropout	0.20
Batch size	32
Optimizer	AdamW
Initial learning rate	0.001
Weight decay	0.0001
Training epochs	50
Activation	GELU
Loss function	Weighted binary cross-entropy
Training/validation/testing	70%/15%/15%
Data partitioning	Subject-wise
Decision threshold	0.50

As shown in Table.1, three temporal convolution branches use kernel sizes of 3, 5, and 9 to represent short-, intermediate-, and long-range temporal information. Four Transformer encoder layers with eight attention heads provide sufficient capacity to model long-range dependencies. A batch size of 32 and an initial learning rate of 0.001 are selected to provide stable optimization. The 70:15:15 subject-wise division prevents observations from the same participant from appearing across different experimental subsets.

4.3 DATASET DESCRIPTION

The experiment will make use of Oxford Parkinson's Disease Telemonitoring Dataset, a popularly used Parkinson's disease dataset that is made up of multiple biomedical voice recordings from patients of PD. This dataset consists of 5,875 observations from 42 patients, where 24 patients are patients of Parkinson's disease and the other 18 are healthy controls. These recordings consist of vocal properties that relate to Parkinsonian speech. The dataset includes biomedical features like fundamental frequency measures, jitter, shimmer, noise-to-harmonics ratio, and nonlinear vocal measures. Motor UPDRS score constitutes the target variable that may be transformed into a binary value as per the experimental protocol. In this current experiment, the temporal records will be ordered in the sequence of recording and then segmented to windows prior to MSTAT processing.

Table.2. Dataset characteristics used for experimental evaluation

Dataset characteristic	Description
Dataset	Oxford Parkinson's Disease Telemonitoring
Total observations	5,875
Subjects	42
PD subjects	24
Control subjects	18
Primary modality	Biomedical voice measurements
Number of attributes	22
Main clinical target	Motor UPDRS
Temporal organization	Repeated measurements per subject
Learning task	Binary PD classification
Training partition	70% subjects
Validation partition	15% subjects
Testing partition	15% subjects
Preprocessing	Subject-wise normalization
Temporal window	128 samples
Window overlap	50%

As presented in Table.2, the dataset provides repeated observations from individual subjects, which enables the

proposed framework to exploit temporal dependencies. The relatively small number of subjects also makes subject-wise evaluation important because random observation-level splitting can produce overly optimistic performance. Various approaches are selected for comparison. CNN-GRU combines convolutional feature extraction with recurrent temporal learning, LSTM models sequential dependencies directly, and CNN-Transformer combines local convolutional representations with global self-attention.

4.4 RESULTS BASED ON ACCURACY

The accuracy results obtained across the training progression are presented in Table.3.

Table.3. Accuracy (%) at different training epochs

Epoch	CNN-GRU	LSTM	CNN-Transformer	MSTAT
5	82.1	83.0	84.4	86.2
10	85.0	85.7	87.1	89.4
15	87.2	87.9	89.0	91.5
20	88.9	89.1	90.7	93.0
25	90.1	90.4	91.8	94.3
30	90.8	91.2	92.6	95.1
35	91.4	91.8	93.2	95.8
40	91.9	92.2	93.8	96.4
45	92.3	92.6	94.2	96.9
50	92.6	92.9	94.6	97.2

The Table.3 shows that MSTAT consistently produces higher accuracy throughout training. At epoch 5, MSTAT achieves 86.2%, compared with 82.1% for CNN-GRU, 83.0% for LSTM, and 84.4% for CNN-Transformer. At epoch 50, the proposed model reaches 97.2%, whereas CNN-GRU, LSTM, and CNN-Transformer achieve 92.6%, 92.9%, and 94.6%, respectively. Thus, MSTAT provides improvements of 4.6, 4.3, and 2.6% over the three baselines. The relatively faster improvement during the first 20 epochs also indicates that multi-scale temporal extraction provides useful discriminative features early in optimization.

4.5 RESULTS BASED ON PRECISION

The Table.4 presents the precision values at five-epoch intervals. Precision evaluates the ability of each model to avoid incorrect positive PD predictions.

Table.4. Precision (%) at different training epochs

Epoch	CNN-GRU	LSTM	CNN-Transformer	MSTAT
5	81.4	82.2	83.7	85.8
10	84.1	84.8	86.3	88.9
15	86.4	87.0	88.2	91.0
20	88.0	88.5	89.9	92.6
25	89.2	89.7	91.1	93.8
30	90.0	90.5	91.9	94.7
35	90.6	91.1	92.5	95.3
40	91.1	91.5	93.1	95.9
45	91.5	92.0	93.5	96.3
50	91.9	92.3	93.9	96.7

According to Table.4, MSTAT reaches a final precision of 96.7%, exceeding CNN-GRU by 4.8%, LSTM by 4.4%, and CNN-Transformer by 2.8%. At epoch 20, MSTAT already achieves 92.6%, while CNN-GRU and LSTM reach 88.0% and 88.5%. The improvement indicates that the adaptive scale-gating mechanism assists the classifier in distinguishing disease-related temporal characteristics from patterns that could otherwise generate false-positive predictions.

4.6 RESULTS BASED ON RECALL

Recall is particularly important for PD screening because failure to identify an affected subject constitutes a false-negative outcome. The recall progression is shown in Table.5.

Table.5. Recall (%) at different training epochs

Epoch	CNN-GRU	LSTM	CNN-Transformer	MSTAT
5	80.9	81.7	83.2	85.4
10	83.7	84.5	85.8	88.5
15	86.0	86.6	87.7	90.7
20	87.6	88.1	89.4	92.2
25	88.9	89.3	90.7	93.5
30	89.6	90.1	91.5	94.4
35	90.3	90.8	92.1	95.0
40	90.8	91.3	92.7	95.6
45	91.2	91.8	93.1	96.0
50	91.6	92.1	93.5	96.4

The Table.5 demonstrates that MSTAT achieves the highest recall throughout training. The final recall of 96.4% is 4.8% higher than CNN-GRU, 4.3 points higher than LSTM, and 2.9 points higher than CNN-Transformer. At epoch 30, MSTAT reaches 94.4%, while the three comparison methods remain between 89.6% and 91.5%. This result indicates that the proposed model identifies a larger proportion of actual PD cases, which is important for reducing missed disease cases.

4.7 RESULTS BASED ON F1-SCORE

The F1-score results are presented in Table.6. Because F1-score combines precision and recall, it provides a balanced indication of the overall positive-class classification capability.

Table.6. F1-score (%) at different training epochs

Epoch	CNN-GRU	LSTM	CNN-Transformer	MSTAT
5	81.1	81.9	83.4	85.6
10	83.9	84.6	86.0	88.7
15	86.2	86.8	87.9	90.8
20	87.8	88.3	89.6	92.4
25	89.0	89.5	90.9	93.6
30	89.8	90.3	91.7	94.5
35	90.4	90.9	92.3	95.1

40	90.9	91.4	92.9	95.7
45	91.3	91.9	93.3	96.1
50	91.7	92.2	93.7	96.5

As indicated in Table.6, MSTAT obtains a final F1-score of 96.5%. CNN-GRU, LSTM, and CNN-Transformer obtain 91.7%, 92.2%, and 93.7%, respectively. The proposed model therefore improves F1-score by 4.8, 4.3, and 2.8%. The consistent separation between MSTAT and the baseline methods indicates that the improvement is not limited to one aspect of classification. Instead, MSTAT maintains a better balance between positive prediction and disease detection.

4.8 RESULTS BASED ON SPECIFICITY

Specificity measures the ability to correctly recognize healthy controls. The results are reported in Table.7.

Table.7. Specificity (%) at different training epochs

Epoch	CNN-GRU	LSTM	CNN-Transformer	MSTAT
5	84.0	84.7	85.8	87.2
10	86.4	87.0	88.1	90.3
15	88.2	88.8	89.8	92.1
20	89.8	90.1	91.2	93.6
25	90.9	91.3	92.3	94.7
30	91.7	92.0	93.0	95.4
35	92.3	92.7	93.6	96.0
40	92.8	93.2	94.1	96.5
45	93.2	93.6	94.5	96.9
50	93.5	93.9	94.9	97.3

The Table.7 shows that MSTAT achieves a final specificity of 97.3%. This value is higher than CNN-GRU, LSTM, and CNN-Transformer by 3.8, 3.4, and 2.4%, respectively. At epoch 50, MSTAT correctly identifies 97.3% of healthy cases. The result suggests that the model does not improve sensitivity at the expense of substantially increasing false-positive predictions.

4.9 DISCUSSION

Based on the accuracy results shown in Table.3, MSTAT appears to be superior in terms of the highest classification performance since it achieves 97.2% versus 94.6% of CNN-Transformer. This means that there may be extra discriminative information that can be used in case multi-resolution temporal features are applied. As shown in Table.4, the precision of MSTAT is higher by 2.8 points than that of CNN-Transformer, meaning that MSTAT produces less wrong PD diagnoses. It may be attributed to the fact that adaptive scale gating is able to suppress the less informative temporal representations. With respect to the recall, shown in Table.5, this indicator is especially important because it concerns disease detection. Here, MSTAT obtains 96.4% versus 93.5% of CNN-Transformer, which indicates a 2.9-point improvement, and implies better identification of PD cases. It seems that multi-scale temporal features can reveal those disease variations that cannot be identified by a single-scale feature. The Table.6 also confirms the same tendency since F1-score results demonstrate that MSTAT outperforms CNN-Transformer by 2.8 points (96.5% versus 93.7%). It shows a close correlation between precision and recall, indicating that this advantage is balanced. Finally, as shown in Table.7, MSTAT gets 97.3% of specificity versus 94.9% of CNN-Transformer, which also implies an improvement of 2.4 points. The fact that the improvement in specificity comes alongside improved recall demonstrates balanced classification performance.

5. CONCLUSION

According to the results of the experiments, the designed Multi-Scale Temporal Attention Transformer outperforms other architectures such as CNN-GRU, LSTM, and CNN-Transformer. At the end of the evaluation process, MSTAT demonstrates an accuracy of 97.2%, a precision of 96.7%, a recall of 96.4%, an F1-score of 96.5%, and a specificity of 97.3%. These results prove that the designed architecture retains high overall discrimination capabilities and simultaneously minimizes false positives and negatives. This can be explained by the introduction of explicit modeling of the temporal data information at three scales. The first scale allows capturing rapid changes, the second one is used for local temporal structure identification, and the third one preserves more global sequential information.

REFERENCES

[1] V. Skaramagkas, A. Pentari, Z. Kefalopoulou and M. Tsiknakis, "Multi-Modal Deep Learning Diagnosis of Parkinson's Disease- A Systematic Review", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 31, pp. 2399-2423, 2023.

[2] H. Rabie and M.A. Akhloufi, "A Review of Machine Learning and Deep Learning for Parkinson's Disease Detection", *Discover Artificial Intelligence*, Vol. 5, No. 1, pp. 1-24, 2025.

[3] V. Skaramagkas, A. Pentari and M. Tsiknakis, "Multi-Modal Deep Learning Diagnosis of Parkinson's Disease-A Systematic Review", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 31, pp. 2399-2423, 2023.

[4] L. Van Gelderen and C. Tejedor-Garcia, "Innovative Speech-Based Deep Learning Approaches for Parkinson's Disease Classification: A Systematic Review", *Applied Sciences*, Vol. 14, No. 17, pp. 7873-7845, 2024.

[5] G.P. Georgiou, "Transforming Speech-Language Pathology with AI: Opportunities, Challenges, and Ethical Guidelines", *Healthcare*, Vol. 13, No. 19, p. 2460-2476, 2025.

[6] F. Del Pup, R. Brun, F. Iotti, E. Paccagnella and M. Atzori, "TransformEEG: Towards Improving Model Generalizability in Deep Learning-based EEG Parkinson's Disease Detection", *Neurocomputing*, Vol. 23, No. 2, pp. 132075-132084, 2025.

[7] A. Rashnu and A. Salimi-Badr, "Integrative Deep Learning Framework for Parkinson's Disease Early Detection using Gait Cycle Data Measured by Wearable Sensors: A CNN-GRU-GNN Approach", *Soft Computing*, Vol. 34, No. 2, pp. 1-14, 2026.

[8] E. Balaji, D. Brindha and R. Vikrama, "Automatic and Non-Invasive Parkinson's Disease Diagnosis and Severity Rating using LSTM Network", *Applied Soft Computing*, Vol. 108, pp. 107463-107476, 2021.

[9] C. Bunterngchit, L.H. Baniata, H. Albayati, M.H. Baniata and S. Kang, "A Hybrid Convolutional-Transformer Approach for Accurate Electroencephalography (EEG)-Based Parkinson's Disease Detection", *Bioengineering*, Vol. 12, No. 6, pp. 583-598, 2025.

[10] S. Roy, J. Nuamah and A. Singh, "EEG-Based Classification of Parkinson's Disease with Freezing of Gait using Midfrontal Beta Oscillations", *Journal of Integrative Neuroscience*, Vol. 24, No. 6, pp. 1-12, 2025.

[11] X. Huang, X. Bi, C. Lv and Y. Li, "Dynamic Facial Expressions Analysis Based Parkinson's Disease Auxiliary Diagnosis", *Proceedings of 44th Chinese Conference on Control*, pp. 8547-8552, 2025.

[12] L.F. Gomez, A. Morales, J. Fierrez and J.R. Orozco-Arroyave, "Exploring Facial Expressions and Action Unit Domains for Parkinson Detection", *Plos One*, Vol. 18, No. 2, pp. 281248-281258, 2023.

[13] X. Wang, J. Huang and M. Ruzhansky, "LSTM-CNN: An Efficient Diagnostic Network for Parkinson's Disease Utilizing Dynamic Handwriting Analysis", *Computer Methods and Programs in Biomedicine*, Vol. 247, pp. 108066-108079, 2024.

[14] M.A. Hossain and F. Amenta, "Machine Learning Applications for Diagnosing Parkinson's Disease via Speech, Language, and Voice Changes: A Systematic Review", *Inventions*, Vol. 10, No. 4, pp. 48-59, 2025.