Foot Pressure-Based Abnormal Gait Recognition With Multi-Scale Cross-Attention Fusion

Menghao Yuan, Yan Wang, *Member, IEEE*, Xiaohu Zhou, *Member, IEEE*, Meijiang Gui, Aihui Wang, *Member, IEEE*, Chen Wang, *Member, IEEE*, Guotao Li, *Member, IEEE*, Hongnian Yu, *Senior Member, IEEE*, Lin Meng, and Zengguang Hou, *Fellow, IEEE*

***Abstract*— Abnormal gait recognition plays a critical role in healthcare, particularly for the early diagnosis and continuous monitoring of neurological and musculoskeletal disorders, such as Parkinson's disease and orthopedic injuries. This study proposes MSCAF-Gait, a Multi-Scale Cross-Attention Fusion Network designed specifically for abnormal gait recognition using foot pressure sensors. MSCAF-Gait incorporates multi-scale convolutional modules with channel and spatial attention mechanisms to effectively capture features across temporal, channel, and spatial dimensions. A novel cross-attention fusion module further enhances feature representation, enabling precise recognition of diverse abnormal gait patterns. To facilitate this research, we introduce the Pressure-Insole Abnormal Gait (PIAG) dataset, comprising gait data associated with common neurological and musculoskeletal abnormalities. Extensive experiments on the publicly available Gait in Parkinson's Disease (GaitinPD) dataset and our self-constructed PIAG dataset validate the effectiveness of MSCAF-Gait. Specifically, the model achieves 99.61% accuracy in Parkinsonian gait recognition and 98.88% accuracy in Parkinson's severity classification. On the PIAG dataset, which includes multiple abnormal gait patterns, MSCAF-Gait attains a high accuracy of 99.42%. Notably, these results are obtained with a lightweight architecture characterized by reduced FLOPs and parameter count, demonstrating that MSCAF-Gait offers both high accuracy and computational efficiency, making it well-suited for real-time deployment on wearable platforms.**

***Index Terms*— Multi-scale convolution, self-attention, cross-attention, foot pressure sensors, gait recognition.**

Received 1 January 2025; revised 2 June 2025 and 27 July 2025; accepted 2 August 2025. Date of publication 11 August 2025; date of current version 14 August 2025. This work was supported in part by Henan Province Key International Science and Technology Cooperation Project under Grant 251111520400; in part by the National Key Research and Development Program of China under Grant 2023YFC2415100; in part by the Backbone Teacher Support Program under Grant GG202414; in part by the Zhongyuan University of Technology (ZUT) Graduate Research Innovation Program under Grant YKY20252K06; in part by Henan Province Key Research and Development Project under Grant 241111312000, Grant 25210221110, and Grant 252102320281; in part by the National Natural Science Foundation of China under Grant 62222316, Grant 62373351, Grant 82327801, Grant 62073325, and Grant 62303463; in part by Chinese Academy of Sciences Project for Young Scientists in Basic Research under Grant YSBR-104; in part by Beijing Natural Science Foundation under Grant F252068 and Grant 4254107; in part by Beijing Nova Program under Grant 20250484813; in part by the China Postdoctoral Science Foundation (CPSF) under Grant 2024M763535; in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20251170; and in part by the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS) under Grant 2023-I2M-C&T-B-017. *(Corresponding authors: Yan Wang; Xiaohu Zhou.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the School of Automation and Electrical Engineering, Zhongyuan University of Technology.

Menghao Yuan, Yan Wang, and Aihui Wang are with the School of Automation and Electrical Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China (e-mail: ywang@zut.edu.cn).

Xiaohu Zhou, Meijiang Gui, Chen Wang, Guotao Li, and Zengguang Hou are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xiaohu.zhou@ia.ac.cn).

Hongnian Yu is with the School of Computing, Engineering and Built Environment, Edinburgh Napier University, EH10 5DT Edinburgh, U.K.

Lin Meng is with the School of Science and Engineering, Ritsumeikan University, Kusatsu 525-8577, Japan.

Digital Object Identifier 10.1109/TNSRE.2025.3597639

I. Introduction

WALKING is a fundamental human activity, governed by the complex interplay between the nervous system and the musculoskeletal system. Neurological and orthopedic conditions, such as Parkinson's disease (PD) and lower-limb injuries, can disrupt this coordination, resulting in abnormal gait patterns that significantly impair mobility and daily life [1]. In PD, gait abnormalities can manifest early with symptoms like mild bradykinesia and reduced arm swing. As the disease progresses, patients exhibit shortened stride length, stooped posture, and episodes of freezing of gait (FoG). In advanced stages, stride length may become minimal, postural instability worsens, and FoG becomes more frequent [2]. Similarly, in orthopedic rehabilitation, quantitative gait assessment enables clinicians to monitor recovery and evaluate treatment efficacy [3], [4]. Thus, precise recognition of abnormal gait patterns is crucial for early diagnosis, treatment planning, and long-term monitoring [5], [6].

Foot pressure gait analysis has emerged as an effective method for identifying gait abnormalities, providing detailed insights into plantar force distribution. However, the inherent multi-dimensional nature of pressure data, including temporal trends, sensor-channel interactions, and spatial layouts, poses significant challenges for feature extraction and recognition.

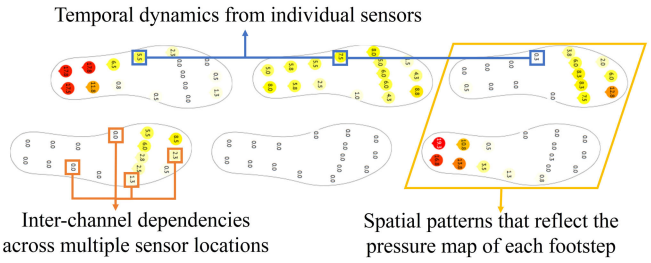


Fig. 1. Illustration of diverse features in foot pressure data across temporal, channel, and spatial domains.

As illustrated in Fig. 1, foot pressure gait data exhibit diverse and intricate characteristics: temporal dynamics from individual sensors, inter-channel dependencies across multiple sensor locations, and spatial patterns that reflect the pressure map of each footstep.

Despite recent advances in deep learning, many existing gait recognition models prioritize accuracy at the expense of computational cost, limiting their deployment on wearable or edge devices. Large model sizes and high Floating Point Operations (FLOPs) make it difficult to meet the real-time, low-power requirements of mobile health applications. To address these challenges, we propose MSCAF-Gait, a Multi-Scale Cross-Attention Fusion Network tailored for abnormal gait recognition using foot pressure data. The model integrates multi-scale convolution to capture both short and long-range temporal features, channel and spatial attention mechanisms to enhance salient features along each dimension, and a cross-attention fusion module that adaptively aggregates complementary information across temporal, spatial, and channel domains. The main contributions of this paper are summarized as follows:

1) A novel Multi-Scale Cross-Attention Fusion Network (MSCAF-Gait) is proposed, combining multi-scale convolution, channel and spatial attention, and cross-attention fusion for comprehensive feature extraction from foot pressure data.
2) A new dataset, PIAG, is introduced to capture diverse clinically relevant abnormal gait patterns and support rigorous model evaluation.
3) Our model achieves state-of-the-art accuracy on both the public GaitinPD dataset and the PIAG dataset, while maintaining low computational complexity suitable for deployment on resource-constrained devices.

The remainder of this paper is organized as follows: Section II reviews existing gait analysis techniques and deep learning models. Section III details the construction of the PIAG dataset and introduces the proposed MSCAF-Gait architecture. Section IV presents experimental settings, performance evaluations, and comparative analyses. Finally, Section V outlines potential directions for future research.

II. Related Work

Early studies on gait analysis primarily relied on traditional signal processing and machine learning techniques. These approaches focused on extracting handcrafted features such as stride length, gait speed, and gait cycle parameters [7]. For instance, Wu et al. [8] employed a 3D motion capture system to extract 36 spatiotemporal and kinematic features. Kernel Principal Component Analysis (KPCA) was used for feature correlation analysis, and Support Vector Machines (SVMs) were then applied for classification, enabling differentiation between young and elderly gait patterns. However, such methods are often constrained by their limited capacity to represent the complex and dynamic characteristics of high-dimensional gait data [9].

In recent years, deep learning has significantly advanced abnormal gait recognition by enabling end-to-end feature learning. Liu et al. [10] proposed a dual-branch model for Parkinson's disease diagnosis, which combines Convolutional Neural Networks (CNNs) for extracting spatial features and Bi-directional Long Short-Term Memory networks (Bi-LSTMs) for capturing temporal features. The model processes data from the left and right feet separately using CNNs, and utilizes the Bi-LSTM layer to integrate independent and joint features, thereby capturing the interaction between the left and right foot data. However, during the data preprocessing stage, a threshold-based method is employed to segment the data according to gait cycles, ensuring accurate division of the gait cycle and eliminating incomplete gait cycles. While this preprocessing approach improves model performance in ideal experimental conditions, it may limit the effectiveness of real-time deployment in practical applications. Furthermore, although LSTM networks can effectively capture temporal features, they still face limitations in handling long-range dependencies [11]. In high-noise environments, this could lead to the omission or over-smoothing of critical information, thereby affecting the accurate detection of subtle gait abnormalities.

To address this, attention mechanisms have been introduced to selectively focus on salient features [12]. For example, Nguyen et al. [13] leveraged a transformer encoder for 1D gait signals, capturing both temporal dependencies within individual channels and spatial correlations across channels. While such models yield promising results, they are typically resource-intensive and difficult to deploy on embedded platforms [14].

In response, lightweight attention mechanisms such as channel attention Squeeze-and-Excitation Network (SENet) [15] and Convolutional Block Attention Module (CBAM) [16] have been adopted to improve efficiency. For instance, Li et al. [17] integrated CBAM with a parallel MobileNetv2 [18] backbone for pathological gait recognition, achieving a favorable trade-off between accuracy and computational cost. However, many of these works still rely on basic fusion techniques like concatenation or summation, which fall short in modeling complex interactions across multiple feature domains.

Recently, cross-attention is increasingly recognized as a powerful solution for integrating heterogeneous feature sources. Unlike self-attention, which operates within a single feature domain, cross-attention allows for the adaptive alignment and fusion of distinct features, e.g., temporal vs. spatial, by computing attention weights across different representations [19]. This makes it particularly suitable for foot pressure

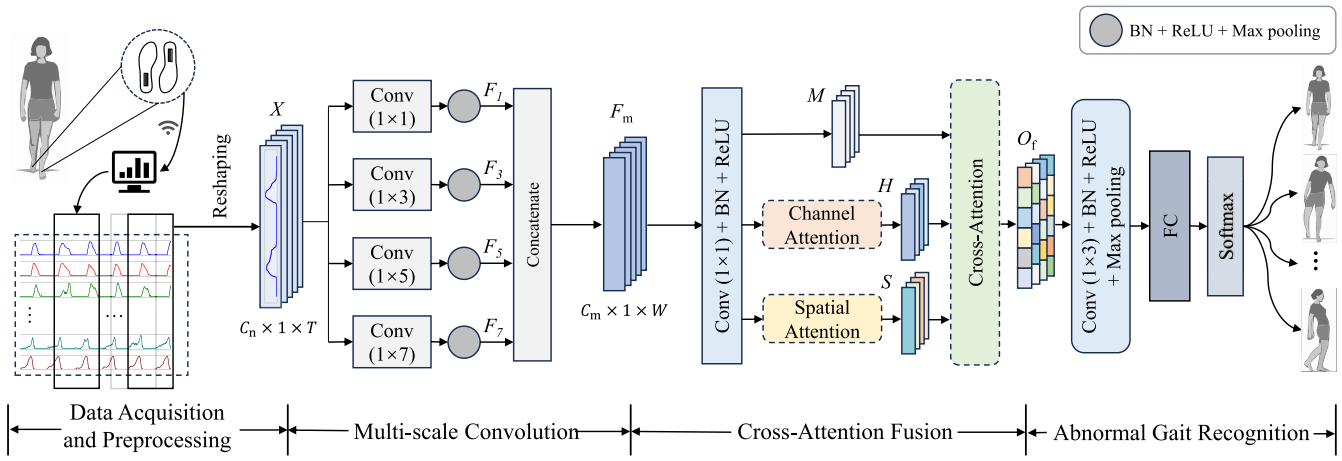


Fig. 2. Architecture of MSCAF-Gait. *M*: Multi-scale convolution output with pixel-level enhancement; *H*: Channel attention output; *S*: Spatial attention output. O_f: Final fused feature from *M*, *H*, and *S* via cross-attention.

data, where rich multidimensional signals coexist and interact dynamically across time, space, and sensor channels.

Building on these insights, we propose MSCAF-Gait, a Multi-Scale Cross-Attention Fusion Network designed for abnormal gait recognition. The model addresses the aforementioned challenges from three key perspectives: (1) it employs multi-scale convolution to capture temporal dynamics at varying resolutions; (2) it incorporates lightweight channel and spatial attention modules to enhance intra-dimensional feature saliency; (3) it integrates a cross-attention mechanism to adaptively fuse complementary features across dimensions. To ensure deployability, MSCAF-Gait is implemented using multi-channel 1D inputs, enabling low-latency, high-resolution processing suitable for real-time applications on edge devices. The technical details of the proposed framework are elaborated in Section III.

III. Proposed Approach

In this section, we introduce MSCAF-Gait, a Multi-Scale Cross-Attention Fusion Network designed specifically for abnormal gait recognition using foot pressure data, as illustrated in Fig. 2. MSCAF-Gait first employs a multi-scale convolutional module with varying kernel sizes to extract temporal features, enabling the model to capture gait patterns over diverse time scales. Subsequently, channel attention and spatial attention mechanisms are employed to adaptively recalibrate the feature representations along their respective dimensions, enhancing the model's focus on the most informative signals. Finally, a cross-attention module integrates these features, enriching the feature representation and enhancing the network's performance in abnormal gait recognition.

A. Data Acquisition and Preprocessing

1) Datasets: In this study, we utilized two datasets: the publicly available Gait in Parkinson's Disease (GaitinPD) dataset from Physionet and our self-constructed Pressure-Insole Abnormal Gait (PIAG) dataset, collected via pressure insoles to capture diverse abnormal gait patterns.

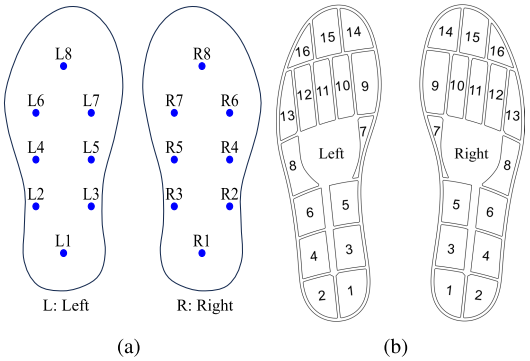


Fig. 3. Layout of pressure sensors embedded in the insoles. (a) GaitinPD dataset with 8 sensors in each insole. (b) PIAG dataset with 16 sensors in each insole.

The GaitinPD dataset, collected by Yogev et al. [20], Hausdorff et al. [21], and Frenkel-Toledo et al. [22], includes gait data from 93 patients diagnosed with idiopathic PD and 73 healthy controls. During data acquisition, each participant walked at a self-selected, comfortable pace on a level surface for approximately two minutes. Vertical ground reaction forces were recorded using eight pressure sensors embedded in each insole, with a sampling rate of 100 Hz. The sensor layout is shown in Fig. 3 (a), where L1–L8 and R1–R8 represent the sensors in the left and right insoles, respectively.

In addition to raw pressure data, the dataset includes demographic and clinical information, notably disease severity labeled using the Hoehn & Yahr (H&Y) staging scale, a widely accepted clinical standard for assessing PD progression from stage 1 (unilateral symptoms) to stage 5 (wheelchair user or bedridden) [23]. In this study, we focus on PD subjects within stages 2, 2.5, and 3, which represent mild-to-moderate disease severity. Stage 2 indicates bilateral symptoms without balance impairment. Stage 2.5 involves mild bilateral involvement with intact postural stability (e.g., recovery on pull test). Stage 3 reflects moderate bilateral symptoms with postural instability, though patients remain physically independent. These H&Y stage labels are used in our experiments as labels for PD severity recognition.

We collected the PIAG dataset using the OpenGo smart insole system developed by Moticon. Each insole is equipped with 16 capacitive pressure sensors and a six-axis inertial measurement unit, with data sampled at 100 Hz. In this study, we exclusively analyzed the plantar pressure data. The layout of the 16 sensors is illustrated in Fig. 3 (b), where sensors are evenly distributed across the toes, metatarsal heads, arch, and heel regions, ensuring detailed representation of plantar force distribution, which supports interpretable biomechanical analysis of different gait patterns [24], [25].

The PIAG dataset was collected from 12 healthy adult participants (7 males and 5 females, aged 22–26), all of whom had no history of neurological or musculoskeletal disorders. Data acquisition was conducted in a 60-meter-long flat corridor with a hard surface. To reduce variability due to foot size, all participants wore standardized insoles tailored to their foot dimensions. The dataset comprises 11 distinct gait categories, including four normal gait types, i.e. **Slow walking**, **Normal walking**, **Fast walking**, and **Running**, and seven clinically inspired abnormal gait patterns, each characterized by specific biomechanical signatures, as detailed below.

1) **In-toeing gait**: Characterized by inward foot rotation with toes pointing toward the body's midline and a negative progression angle, often due to excessive internal rotation of the femur or tibia.
2) **Out-toeing gait**: Defined by outward foot rotation with lateral toe deviation and a progression angle exceeding 10°, possibly caused by femoral/tibial torsion or foot overpronation [26].
3) **Right leg pain gait**: A protective pattern with shortened stance duration on the right leg, reduced step length, and compensatory weight shifting to the contralateral side.
4) **Left leg pain gait**: Mirror pattern of the right leg pain gait, resulting in asymmetrical loading and disrupted coordination.
5) **Magnetic gait**: Characterized by a dragging foot motion with small, shuffling steps due to impaired foot lift, commonly seen in neurological conditions such as normal pressure hydrocephalus [27].
6) **Steppage gait**: Involves exaggerated hip/knee flexion to compensate for foot drop, typically linked to peroneal nerve injury or peripheral neuropathy [28].
7) **Gluteus medius gait**: Marked by pelvic drop and torso lean due to weakness in the gluteus medius muscle, producing a side-to-side sway [29].

To ensure accurate gait simulation, all participants were trained using standardized instructional videos derived from clinical sources. Each gait type was rehearsed before recording, and participants performed continuous walking for approximately 2 minutes per condition. Rest intervals were provided between trials to prevent fatigue and maintain data quality. Each trial was stored as a separate CSV file, containing raw plantar pressure data from 32 sensors (16 per foot), IMU readings, and computed gait-related parameters. The corresponding gait label was appended to the final column. This well-structured acquisition protocol ensures that the PIAG dataset is reliable, reproducible, and suitable as a benchmark for abnormal gait recognition research.

TABLE I
SAMPLE COUNTS PER SUBJECT AND GAIT CATEGORY IN PIAG DATASET WITH DATASET SPLIT SUMMARY

Subject	Sample Count	Gait Category	Sample Count
subject 0	1,407	Slow Walking	1,516
subject 1	1,434	Normal Walking	1,517
subject 2	1,432	Fast Walking	1,534
subject 3	1,397	Running	1,545
subject 4	1,418	In-toeing gait	1,529
subject 5	1,410	Out-toeing gait	1,594
subject 6	1,477	Right leg pain gait	1,560
subject 7	1,459	Left leg pain gait	1,575
subject 8	1,383	Magnetic gait	1,634
subject 9	1,452	Steppage gait	1,491
subject 10	1,384	Gluteus medius gait	1,585
subject 11	1,427	–	–
Total			**17,080**
Train : Validation : Test = 10,248 : 3,416 : 3,416 (6:2:2)			

Note: A sliding window of 2000 ms with a stride of 1000 ms was applied during preprocessing.

2) Raw Data Preprocessing: In the domain of gait analysis utilizing plantar pressure sensors, raw data are typically collected from multiple sensors distributed across different regions of the foot sole. In our framework, plantar pressure data from both the left and right feet are utilized simultaneously. Each OpenGo insole contains 16 pressure sensors, resulting in a total of 32 pressure channels per trial. To facilitate efficient processing, we employ a sliding window technique to segment the raw temporal data. Specifically, a sliding window of size T with a 50% overlap is applied along the time axis to capture temporal dependencies within the gait cycle. The data stream from each sensor is treated as a distinct input channel, and the data from the 32 sensors are organized into 32 one-dimensional sequences of length T.

As a result, each input sample X is structured as a three-dimensional tensor with shape $C_{\mathrm{n}} \times 1 \times T$, where $C_{\mathrm{n}} = 32$ corresponds to the number of channels capturing plantar pressure signals from both feet, and T is the temporal length of each segment. This representation enables the model to jointly learn spatial-temporal patterns across all sensors, effectively capturing both inter-foot and intra-foot pressure dynamics. To accelerate training convergence and enhance feature extraction, Min-Max normalization is applied to the raw pressure data before segmentation. This preprocessing pipeline ensures consistency across subjects and gait types, and improves the model's ability to accurately detect abnormal gait patterns from subtle temporal variations. A detailed summary of the resulting sample counts per subject and gait category in the PIAG dataset, along with the dataset split statistics, is presented in Table I.

B. Multi-Scale Convolution

The multi-scale convolutional module is specifically designed to extract local temporal features from foot pressure sensor data using convolutional kernels of varying sizes. These kernels provide different temporal receptive fields, enabling the network to capture both short-term and long-term

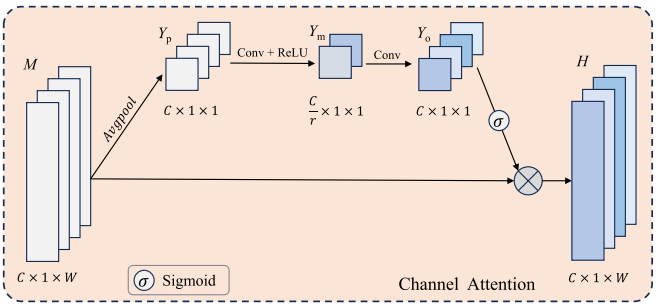


Fig. 4. The structure of channel attention module.

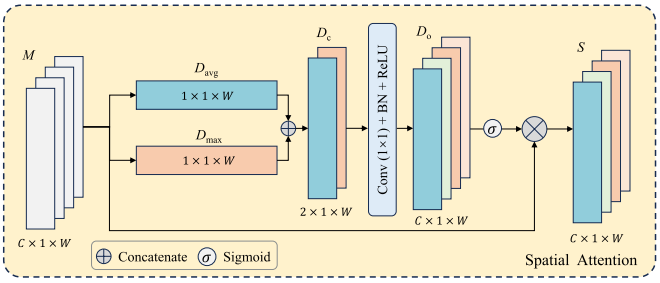


Fig. 5. The structure of spatial attention module.

dependencies, which are essential for identifying diverse gait dynamics. Specifically, we employ convolutional filters with kernel sizes k_i, where $k_i \in \{1, 3, 5, 7\}$, to process the input data $X \in \mathbb{R}^{C_n \times 1 \times T}$. Each convolutional branch applies a $1 \times k_i$ kernel followed by batch normalization, ReLU activation, and max pooling, resulting in a feature map F_i in Eq. (1).

$$F_i = \text{Maxpooling}\Big\{\text{ReLU}\Big[\text{BN}\big(\text{Conv2d}(X, W_{k_i})\big)\Big]\Big\} \quad (1)$$

where W_{k_i} represents the convolutional kernel of size $1 \times k_i$. To ensure that the output feature maps from different branches (F_1, F_3, F_5, F_7) can be directly concatenated, we apply appropriate zero-padding in the temporal dimension for each convolution, such that all output feature maps maintain the same temporal resolution. This design guarantees consistent alignment across branches, regardless of kernel size. The resulting feature maps from each branch are then concatenated along the channel dimension to form the final multi-scale representation in Eq. (2).

$$F_m = \text{Concat}(F_1, F_3, F_5, F_7) \quad (2)$$

This combined feature map is denoted as $F_m \in \mathbb{R}^{C_m \times 1 \times W}$, where C_m represents the aggregated output channels from each convolutional block. The resulting architecture leverages multi-scale convolution operations to construct a rich temporal feature representation, effectively capturing local patterns across varying time scales and thereby improving the recognition of diverse gait patterns.

C. Cross-Attention Fusion

Following the multi-scale convolution process, we apply a pixel-level convolution layer to further integrate the extracted features, generating an intermediate feature map M from F_m.

1) Channel Attention: The channel attention is strategically designed to highlight important channels in the input feature map, enabling the network to emphasize critical information across different feature channels. As shown in Fig. 4, channel attention mechanism emphasizes the most important channels in the feature map $M \in \mathbb{R}^{C \times 1 \times W}$. First, a global average pooling operation is applied across the width dimension of M, resulting in a channel-wise descriptor $Y_p \in \mathbb{R}^{C \times 1 \times 1}$. This operation condenses temporal information by computing the average value of each channel. It is then passed through a 1×1 convolution layer to reduce the channel dimension, producing an intermediate feature map $Y_m \in \mathbb{R}^{\frac{C}{r} \times 1 \times 1}$, followed by ReLU activation. A second 1×1 convolution restores the original channel size, yielding $Y_o \in \mathbb{R}^{C \times 1 \times 1}$, which is processed through a sigmoid function to generate the attention weights. Finally, the original feature map M is reweighted via element-wise multiplication with these attention weights, generating the output $H \in \mathbb{R}^{C \times 1 \times W}$. The process is encapsulated in Eqs. (3), (4) and (5).

$$Y_p = \frac{1}{W} \sum_{i=1}^{W} M_{c,1,i} \quad (3)$$

$$Y_o = \text{Conv2d}\Big\{\text{ReLU}\Big[\text{Conv2d}\big(Y_p\big)\Big]\Big\} \quad (4)$$

$$H = \sigma\left(Y_o\right) \odot M \quad (5)$$

2) Spatial Attention: The spatial attention mechanism focuses on identifying significant spatial regions within the feature map by analyzing the distribution of pressure across the foot. As shown in Fig. 5, this is achieved through pooling operations that summarize the input features across the channel dimension. Specifically, global average pooling and max pooling are used to capture essential spatial information from the input feature map.

The Spatial Attention mechanism identifies important spatial regions in the feature map $M \in \mathbb{R}^{C \times 1 \times W}$. First, global average pooling and max pooling are performed along the channel dimension, resulting in two spatial descriptors, D_{avg} and D_{max}, where $D_{avg}, D_{max} \in \mathbb{R}^{1 \times 1 \times W}$. These descriptors are concatenated and passed through a 1×1 convolution, followed by batch normalization and ReLU activation, generating the intermediate feature map $D_o \in \mathbb{R}^{C \times 1 \times W}$. A sigmoid function is then applied to produce the spatial attention weights. Finally, the original feature map M is reweighted by element-wise multiplication with these attention weights, producing the output $S \in \mathbb{R}^{C \times 1 \times W}$. The process is summarized in Eqs. (6), (7) and (8).

$$D_{avg} = \frac{1}{C} \sum_{i=1}^{C} M_i, \quad D_{max} = \max_{i=1}^{C} M_i \quad (6)$$

$$D_o = \text{ReLU}\Big[\text{BN}\Big\{\text{Conv2d}\big(\text{Concat}(D_{avg}, D_{max})\big)\Big\}\Big] \quad (7)$$

$$S = \sigma(D_o) \odot M \quad (8)$$

3) Cross-Attention: The cross-attention module, shown in Fig. 6, is designed to enhance the fusion of temporal and spatial information while preserving the structural characteristics of the original feature map. The module takes three inputs:

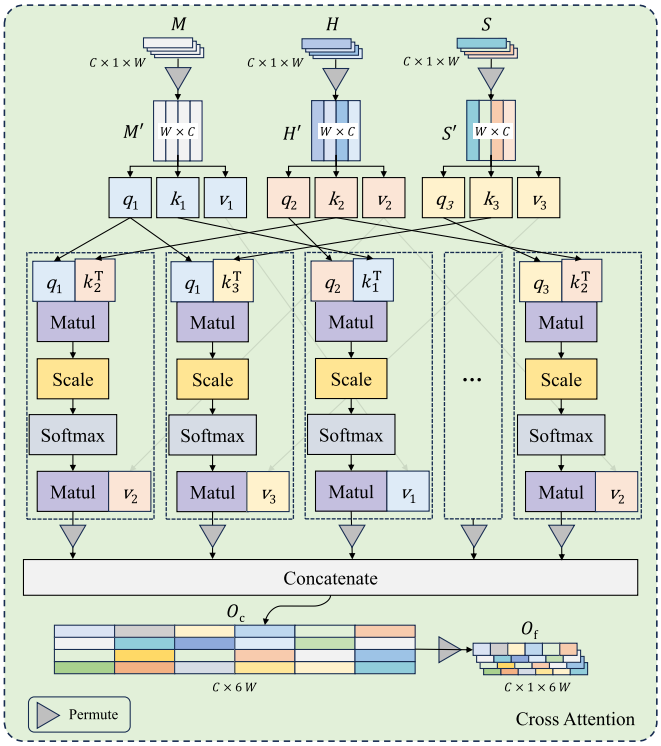


Fig. 6. The structure of the cross-attention module.

the output of the previous convolutional layer M, the channel attention output H, and the spatial attention output S.

As detailed in Algorithm 1, the cross-attention module combines temporal and spatial information from the input feature maps M, H, S. Initially, the input feature maps M, H, and S (each of size $C \times 1 \times W$) are permuted to produce M', H', and S', each reshaped to $W \times C$. Linear transformations are then applied to extract the queries (q), keys (k), and values (v) from each input, according to the formulas $q = XW_Q, k = XW_K, v = XW_V$, where W_Q, W_K, W_V are the learned weight matrices. In each branch, the query from one feature map interacts with the keys from the other feature maps, and attention scores are computed using the scaled dot-product attention mechanism, as expressed in Eq. (9).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)V \tag{9}$$

where d_{k} is the embedding dimension of the key vectors, used to scale the dot product and stabilize the softmax function when handling high-dimensional keys.

In Step 3 of Algorithm 1, cross-attention is computed across each pair of branches. For example, q_1 from M' interacts with k_2 from H' and k_3 from S', producing the cross-attention outputs O_{c1} and O_{c2}, respectively. This process is repeated for all branch pairs, resulting in six cross-attention outputs O_{c1} to O_{c6}. These outputs are then concatenated along the temporal dimension to form $O_{\text{c}} \in \mathbb{R}^{C\times 6W}$, which is subsequently permuted to produce the final output $O_{\text{f}} \in \mathbb{R}^{C\times 1\times 6W}$. This final output integrates enriched temporal, spatial, and channel-wise information, leveraging cross-attention across multiple branches.

Algorithm 1 Cross-Attention Mechanism

Input: $M, H, S \in \mathbb{R}^{C\times 1\times W}$
Output: $O_{\text{f}} \in \mathbb{R}^{C\times 1\times 6W}$
Step 1: Permute input feature maps

$$M', H', S' \in \mathbb{R}^{W\times C} \leftarrow M, H, S \in \mathbb{R}^{C\times 1\times W}$$

Step 2: Extract queries, keys, and values

$$q_1, k_1, v_1 \leftarrow M'; \quad q_2, k_2, v_2 \leftarrow H'; \quad q_3, k_3, v_3 \leftarrow S'$$

Step 3: Compute cross-attention using attention mechanism

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)V$$

$$O_{\text{c1}} \leftarrow \text{Attention}(q_1, k_2, v_2), \quad O_{\text{c2}} \leftarrow \text{Attention}(q_1, k_3, v_3);$$

$$O_{\text{c3}} \leftarrow \text{Attention}(q_2, k_1, v_1), \quad O_{\text{c4}} \leftarrow \text{Attention}(q_2, k_3, v_3);$$

$$O_{\text{c5}} \leftarrow \text{Attention}(q_3, k_1, v_1), \quad O_{\text{c6}} \leftarrow \text{Attention}(q_3, k_2, v_2)$$

Step 4: Concatenate all cross-attention outputs

$$O_{\text{c}} \in \mathbb{R}^{C\times 6W} \leftarrow \text{Concatenate}(O_{\text{c1}}^{\text{T}}, O_{\text{c2}}^{\text{T}}, O_{\text{c3}}^{\text{T}}, O_{\text{c4}}^{\text{T}}, O_{\text{c5}}^{\text{T}}, O_{\text{c6}}^{\text{T}})$$

Step 5: Permute concatenated output

$$O_{\text{f}} \in \mathbb{R}^{C\times 1\times 6W} \leftarrow O_{\text{c}} \in \mathbb{R}^{C\times 6W}$$

Return: O_{f}

As shown in Fig. 2, the fused output O_{f} is passed through a convolutional layer with BN and ReLU activation for feature refinement. A fully connected (FC) layer followed by a softmax layer generates the final classification scores to classify different gait patterns for abnormal gait recognition.

IV. Experimental Results and Discussions

In this section, we present the experimental setup and evaluate the proposed method. Our evaluation includes experiments conducted on the publicly available GaitinPD dataset and our newly collected PIAG dataset. The experimental results provide insights into the effectiveness of the method, including quantitative performance metrics, qualitative visualizations, and deployment evaluations.

A. Experimental Setup

1) Experimental Details: All experiments were conducted on a laptop equipped with an NVIDIA GTX1650 GPU using the PyTorch framework. The dataset was randomly split into training, validation, and testing sets in a 6:2:2 ratio. The proposed network is trained using the Adam optimizer and the cross-entropy loss function, with the primary objective of optimizing the model by minimizing multi-class classification error. The cross-entropy loss function performs effectively in multi-class classification tasks, as it penalizes incorrect classifications by comparing the predicted probability distribution with the true labels, thereby guiding the optimization process. The Adam optimizer adapts the learning rate by computing

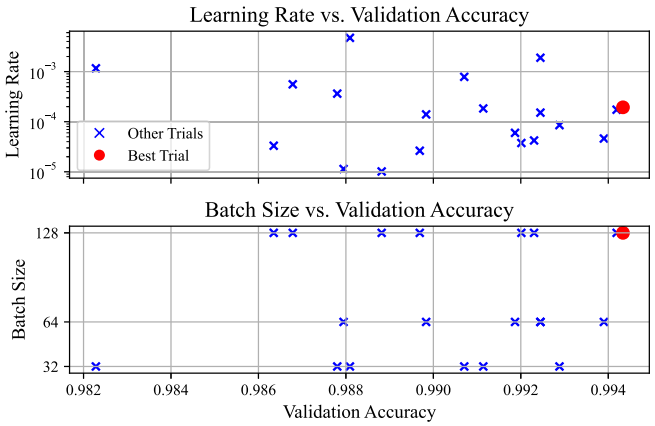
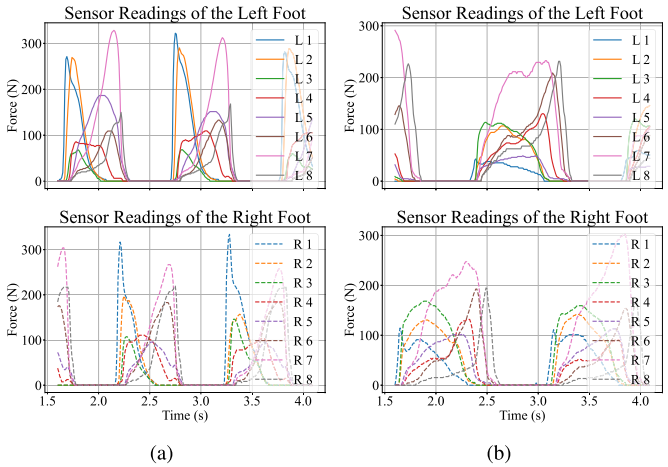


Fig. 7. Visualization of hyperparameter tuning using Optuna.



(a) (b)

Fig. 8. Vertical ground reaction force curves in GaitinPD dataset. (a) Healthy. (b) PD gait.

first and second moment estimates of the gradients, ensuring stable and efficient convergence during training. The model is trained for a maximum of 100 epochs.

To enhance model performance, we employ the Optuna framework [30] for automated hyperparameter tuning. Optuna adopts a Bayesian optimization strategy and utilizes the Tree-structured Parzen Estimator (TPE) as its core algorithm. By modeling the relationship between past hyperparameter trials and their corresponding performance, TPE guides the search toward promising regions in the parameter space, thereby achieving competitive performance with fewer trials. The objective is to maximize recognition accuracy on the validation set, promoting robust generalization beyond the training data. The optimization focuses on two key hyperparameters: the learning rate and batch size. The learning rate is sampled from a logarithmic scale within the range $[10^{-5}, 10^{-2}]$ to allow for fine-grained exploration of smaller values, while the batch size is selected from the discrete set 32, 64, 128, covering commonly used configurations in deep learning. As shown in Fig. 7, we conducted 20 optimization trials. The top subplot illustrates the relationship between learning rate and validation accuracy, while the bottom subplot shows the impact of batch size. The red dots indicate the configurations that produced the best results, a learning rate of 1.9×10^{-4} and a batch size of 128, which yielded high accuracy and stable convergence.

2) Evaluation Metrics: To rigorously evaluate the performance of the networks, we utilized four commonly used metrics: Precision (Pr.), Recall (Re.), F_1-score, and Accuracy (Acc.). The formulas for these metrics are defined as Eqs. (10)–(13).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

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

where TP, FN, FP, and TN represent true positives, false negatives, false positives, and true negatives, respectively. To further measure the computational efficiency, we also consider the number of parameters and the computational cost, measured in Floating Point Operations (FLOPs), as key evaluation metrics.

B. Experiment Results and Performance Comparison

1) Determining Sliding Window Size Based on Gait Cycle Duration: The duration of a complete gait cycle varies significantly across individuals and gait types due to differences in walking speed, stride length, and neuromuscular conditions. To ensure sufficient temporal coverage for capturing essential gait characteristics, it is essential to select a sliding window size T that approximates the average gait cycle duration. An appropriately sized window enables the model to retain key temporal features, such as stride timing and phase transitions, thereby enhancing recognition accuracy.

As shown in Fig. 8, notable differences in plantar pressure variation are observed between healthy individuals and patients with PD. In healthy gait, sensors L1 and R1 exhibit pronounced pressure fluctuations, reflecting rapid shifts in the center of gravity and dynamic force transitions. In contrast, PD patients present with lower amplitude and slower pressure changes, indicative of prolonged gait cycles and reduced mobility. To accommodate this variability, we employ a multi-scale convolutional module that captures hierarchical temporal dependencies. However, selecting a well-matched input window remains critical for maximizing feature completeness.

Based on prior analysis of the GaitinPD dataset [10], where typical gait cycles ranged between 1.0 and 1.6 seconds, we selected a window size of 1600 ms at a 100 Hz sampling rate, providing full-cycle coverage.

For the PIAG dataset, we conducted a detailed analysis across all 11 gait patterns to determine average gait cycle durations (see Fig. 9). A range of window sizes from 800 ms to 3200 ms was tested, and the recognition performance under each setting is summarized in Table II. The highest

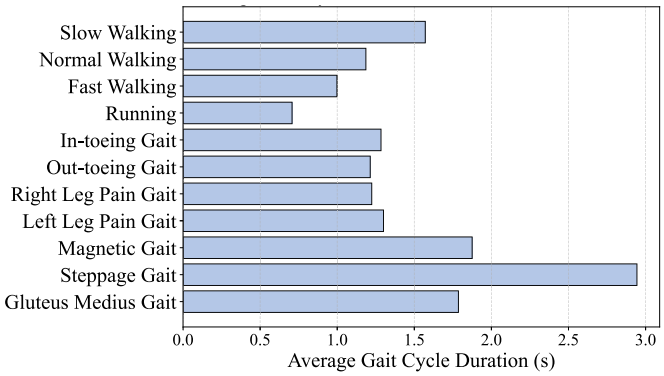


Fig. 9. Average gait cycle duration for different gait patterns in PIAG dataset.

TABLE II
PERFORMANCE OF MSCAF-GAIT ON PIAG OVER DIFFERENT WINDOW SIZES

Window Size	Acc. (%)	Pr. (%)	Re. (%)	F_1-score (%)
3200 ms	99.03	99.03	99.02	99.03
2800 ms	99.31	99.31	99.31	99.31
2400 ms	99.09	99.10	99.09	99.09
2000 ms	**99.42**	**99.42**	**99.42**	**99.42**
1600 ms	99.05	99.05	99.05	99.05
1200 ms	98.69	98.70	98.69	98.69
800 ms	98.27	98.27	98.27	98.27

accuracy (99.42%) was achieved with a 2000 ms window, suggesting this size offers optimal temporal-spatial resolution for discriminating between fine-grained gait patterns. These findings confirm that aligning the input window size with actual gait cycle durations substantially improves recognition performance, especially when combined with multi-scale temporal modeling.

2) Experimental Results and Performance Comparison on the GaitinPD Dataset: In the experiments conducted on the GaitinPD dataset, we aimed to evaluate the proposed method in both binary and four-class classification tasks. The first task was a binary classification to distinguish between the control group with healthy gait (Co) and PD gait. The second task involved a four-class classification based on the H&Y scale, intended to assess PD severity levels for clinical decision support. As summarized in Table III, we compared the performance of related work using the GaitinPD dataset for both tasks. Daliri [31] employed an SVM model, achieving an accuracy of 91.2% for PD gait diagnosis. Açıci et al. [32] utilized a random forest (RF) model, obtaining an accuracy of 98.04%. Veeraragavan et al. [33] applied an artificial neural network (ANN), reaching 97.4% accuracy for PD diagnosis and 87.1% for H&Y stage classification. Xia et al. [34] adopted a CNN-LSTM hybrid model, achieving a PD diagnosis accuracy of 99.07% and a H&Y stage classification accuracy of 98.03%. Liu et al. [10] used the CNN-BiLSTM method, achieving a PD gait recognition accuracy of 99.22%. However, both the methods by Xia et al. and Liu et al.

TABLE III
PERFORMANCE COMPARISON BETWEEN MSCAF-GAIT AND THE RELATED WORK ON GAITINPD DATASET

Related Work	Model	PD diagnosis Acc. (%)	H&Y stage Acc. (%)
Daliri M.R. [31]	SVM	91.2	-
Açıci et al. [32]	RF	98.04	-
Veeraragavan et al. [33]	ANN	97.4	87.1
Xia et al. [34]	CNN-LSTM	99.07	98.03
Liu et al. [10]	CNN-BiLSTM	99.22	-
Nguyen et al. [13]	Transformers	95.2	-
Naimi et al. [35]	HCT	97	87
Ours	MSCAF-Gait	**99.61**	**98.88**

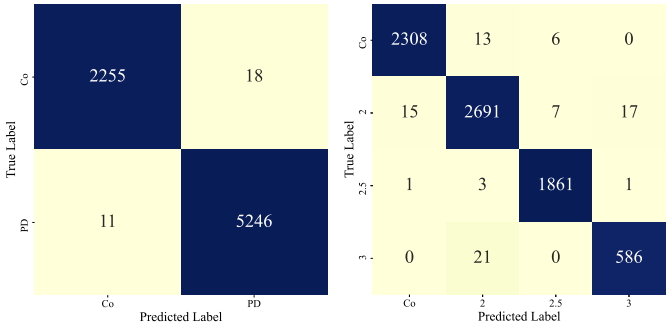


Fig. 10. Confusion matrix of MSCAF-Gait on the GaitinPD dataset. (a) Co-PD. (b) H&Y stage.

depend on extracting gait cycles from raw data for recognition. Naimi et al. [35] developed a hybrid ConvNet-Transformer (HCT) model that achieved 97% accuracy in PD diagnosis and 87% in H&Y stage classification. MSCAF-Gait demonstrated superior performance compared to existing methods in both tasks. For the binary classification task, the proposed model achieved an accuracy of 99.61%, outperforming the highest accuracy reported in Table III. For the four-class classification task assessing the severity of PD, MSCAF-Gait achieved a recognition accuracy of 98.88%.

Fig. 10 presents the confusion matrices of the proposed MSCAF-Gait network for both tasks on the GaitinPD dataset. As shown in Fig. 10 (b), which illustrates MSCAF-Gait's performance in the H&Y stage grading task, the classification accuracy shows a slight decline when distinguishing between H&Y 2 and H&Y 3. The number of test samples for each class was as follows: 2327 for Control (Co), 2730 for H&Y 2, 1866 for H&Y 2.5, and 607 for H&Y 3. The significantly smaller sample size for H&Y 3 compared to H&Y 2 is likely a contributing factor to the reduced classification performance in this category.

3) Experimental Results and Performance Comparison on the PIAG Dataset: In the experiments conducted on the PIAG dataset, we evaluated several CNN-based network variants, including CNN-LSTM, CNN-SENet, CNN-CBAM, CNN-Self-Attention, and the proposed MSCAF-Gait network. CNN-LSTM consists of two convolutional layers followed by an LSTM layer. CNN-SENet and CNN-CBAM augment the same backbone with SENet and CBAM modules, respectively. CNN-Self-Attention combines a single convolutional

TABLE IV
PERFORMANCE COMPARISON BETWEEN MSCAF-GAIT AND OTHER CNN-BASED MODELS ON PIAG

Model	Acc. (%)	Pr. (%)	Re. (%)	F_1-score (%)
CNN-LSTM	95.90	95.97	95.90	95.92
CNN-SENet	96.46	96.53	96.46	96.46
CNN-CBAM	97.07	97.10	97.07	97.07
CNN-Self-attention	97.44	97.48	97.44	97.43
MSCAF-Gait	**99.42**	**99.42**	**99.42**	**99.42**

layer with a self-attention mechanism. As shown in Table IV, models incorporating attention mechanisms consistently outperformed the CNN-LSTM, which achieved an accuracy of 95.90%, with improvements ranging from 0.56% to 3.52%, depending on the specific architecture. CBAM improved performance over SENet by 0.25%, while CNN-Self-Attention outperformed CBAM by 0.37%. The proposed MSCAF-Gait, integrating channel and spatial attention in a parallel structure with cross-attention fusion, achieved the highest accuracy of 99.42%. Since the PIAG dataset was balanced across different activity classes during data collection, the four evaluation metrics (Acc., Pr., Re., and F_1-score) exhibited minimal variation.

4) Leave-One-Subject-Out Evaluation and Model Efficiency Analysis: To evaluate model generalization under inter-subject variability typically encountered in real-world applications, we employed the Leave-One-Subject-Out (LOSO) cross-validation strategy. In each fold, one subject is held out for testing, while the remaining subjects are used for training. This procedure is repeated until every subject has been used as the test set once. Final evaluation metrics are obtained by averaging the results across all folds, providing a robust measure of model performance. This approach provides a realistic and unbiased estimate of model performance on unseen individuals, which is critical for practical deployment.

Table V summarizes the LOSO evaluation results and computational efficiency of various models on the PIAG and GaitinPD datasets. The baseline CNN-LSTM model achieves recognition accuracies of 72.05% on PIAG and 71.41% on GaitinPD but suffers from high computational cost, requiring over 43M and 33M FLOPs on the PIAG and GaitinPD datasets, respectively, and approximately 0.45M parameters. The CNN-SENet and CNN-CBAM models, which incorporate channel and spatial attention mechanisms, respectively, achieve higher accuracies than CNN-LSTM, with moderate performance gains. They achieve 74.04% (PIAG) and 76.03% (GaitinPD) accuracy, while reducing FLOPs to 6M-9M, significantly lower than the CNN-LSTM baseline. However, their parameter counts increase to 2.48M and 3.14M, respectively. On the other end of the spectrum, the CNN-Self-Attention model demonstrates extremely low computational cost of 0.53M FLOPs and 0.08M parameters on GaitinPD, but this efficiency comes at the expense of accuracy, achieving only 67.39% (PIAG) and 71.28% (GaitinPD).

The proposed MSCAF-Gait model achieves the highest accuracy of 77.62% on PIAG and 82.31% on GaitinPD, while maintaining a lightweight architecture with 15.83M / 9.96M FLOPs, and 0.22M / 0.08M parameters, respectively. This indicates that MSCAF-Gait offers a favorable trade-off between recognition performance and computational cost. Such a balance in conjunction with its robust generalization demonstrated under LOSO evaluation underscores the suitability of MSCAF-Gait for deployment in wearable gait monitoring systems and intelligent rehabilitation platforms.

C. Ablation Study

1) Ablation of Key Components: To access the contribution of each individual component to the overall performance of the proposed MSCAF-Gait network, we conducted a comprehensive ablation study, as summarized in Table VI. The experiments were carried out on two tasks: PD diagnosis using the GaitinPD dataset and multi-class activity recognition using the PIAG dataset. Key modules, including multi-scale convolution (MS), channel attention, spatial attention, and cross-attention, were selectively enabled or disabled, while retaining the CNN backbone architecture. A total of seven experimental configurations (Exp I to VII) were designed to systematically investigate both individual and joint contributions of these components on classification performance, measured in terms of accuracy, precision, recall, and F_1-score.

As shown in Table VI, the comparison from Exp I to Exp III reveals that incorporating either channel attention or spatial attention enhances network performance, with channel attention slightly outperforming spatial attention. Comparing Exp IV with the complete network (Exp VII) highlights that the inclusion of cross-attention significantly improves the overall performance. Similarly, comparing Exp V with the complete network reveals that adding multi-scale convolution increases accuracy on the GaitinPD and PIAG datasets by 1.07% and 0.96%, respectively. While multi-scale convolution enhances temporal feature extraction, it inevitable introduces additional computational overhead. Furthermore, the comparison between Exp VI and Exp VII demonstrates that the parallel structure of channel and spatial attention further enhances network performance. The complete network (Exp VII), incorporating all components, achieves the highest performance metrics on both datasets—99.61% accuracy on the GaitinPD dataset and 99.42% accuracy on the PIAG dataset. These findings highlight the importance of multi-scale convolution and attention mechanisms—especially the parallel structure of channel and spatial attention combined with cross-attention fusion—in achieving optimal accuracy and model efficiency.

2) Ablation Study of Multi-Scale Convolution Kernels: To quantify the trade-off between recognition performance and computational cost introduced by different kernel configurations, we conducted an ablation study using different kernel configurations within the multi-scale convolutional module. As shown in Table VII, we compare models using both single-kernel (e.g., 1×1, 1×3, 1×5, or 1×7) and multi-kernel combinations (e.g., $1\times1+1\times3$, $1\times1+1\times3+1\times5$, up to $1\times1+1\times3+1\times5+1\times7$). The results indicate that larger kernel sizes, such as 1×5 and 1×7, contribute to improved performance by providing wider receptive fields, which better capture long-term temporal dependencies in gait sequences. Notably, the combination of $1\times1+1\times5$ achieves 99.33% accuracy with relatively low computational overhead, demonstrating a favorable balance between efficiency and recognition performance.

TABLE V
LOSO-BASED PERFORMANCE COMPARISON ON PIAG AND GAITINPD DATASETS

Model	PIAG			GaitinPD		
	Acc. (%)	FLOPs (M)	Params (M)	PD diagnosis Acc. (%)	FLOPs (M)	Params (M)
CNN-LSTM	72.05	43.04	0.45	71.41	33.10	0.44
CNN-SENet	73.66	8.97	3.14	74.20	6.28	2.48
CNN-CBAM	74.04	8.99	3.14	76.03	6.29	2.48
CNN-Self-attention	67.39	1.38	0.19	71.28	0.53	0.08
MSCAF-Gait	**77.62**	15.83	0.22	**82.31**	9.96	0.08

TABLE VI
ABLATION STUDY OF MSCAF-GAIT ON THE GAITINPD AND PIAG DATASETS

Dataset	Exp	Main Components				Metrics (%)			
		Multi-Scale Convolution	Channel Attention	Spatial Attention	Cross-Attention	Acc.	Pr.	Re.	F_1-score
GaitinPD	I	✓				98.98	98.98	98.98	98.98
	II	✓	✓			99.40	99.40	99.40	99.40
	III	✓		✓		99.16	99.16	99.16	99.16
	IV	✓	✓	✓		99.00	99.00	99.00	99.00
	V		✓	✓	✓	98.54	98.55	98.54	98.54
	VI	✓			✓	99.16	99.16	99.16	99.16
	VII	✓	✓	✓	✓	**99.61**	**99.61**	**99.61**	**99.61**
PIAG	I	✓				98.93	98.93	98.93	98.93
	II	✓	✓			99.22	99.22	99.22	99.22
	III	✓		✓		99.05	99.05	99.05	99.05
	IV	✓	✓	✓		99.01	99.02	99.01	99.01
	V		✓	✓	✓	98.66	98.67	98.66	98.67
	VI	✓			✓	99.36	99.36	99.36	99.36
	VII	✓	✓	✓	✓	**99.42**	**99.42**	**99.42**	**99.42**

TABLE VII
ABLATION STUDY OF MULTI-SCALE CONVOLUTION KERNEL COMBINATIONS

1×1	1×3	1×5	1×7	Acc. (%)	FLOPs	Params
✓				98.66	8.42 M	0.20 M
	✓			98.77	9.24 M	0.20 M
		✓		98.79	10.06 M	0.20 M
			✓	98.91	10.88 M	0.20 M
✓	✓			99.19	10.07 M	0.20 M
✓		✓		99.33	10.89 M	0.21 M
✓			✓	98.98	11.71 M	0.21 M
	✓	✓		99.06	11.71 M	0.21 M
	✓		✓	99.13	12.53 M	0.21 M
		✓	✓	98.92	13.35 M	0.21 M
✓	✓	✓		99.10	12.54 M	0.21 M
✓	✓		✓	99.36	13.36 M	0.21 M
✓		✓	✓	99.36	14.18 M	0.21 M
	✓	✓	✓	99.33	15.00 M	0.21 M
✓	✓	✓	✓	**99.42**	**15.83 M**	**0.22 M**

When combining all four kernel sizes (1×1, 1×3, 1×5, and 1×7), the model achieves the highest accuracy of 99.42%, with consistent performance across multiple runs. Based on this analysis, the full multi-scale configuration was adopted as the default design to maximize temporal feature representation across different time scales. Additionally, lighter alternatives, such as the 1×1+1×5 setup, offer viable options for resource-constrained environments, providing flexibility in deployment without significant performance loss.

D. Visualization

Interpretability remains an important challenge in deep learning, particularly in understanding model decision-making processes and the contribution of learned features. To explore the contributions of different components, such as multi-scale convolution and attention mechanisms, we conducted a comprehensive visualization analysis of internal feature activations and attention responses.

Firstly, to reveal the contribution of each component in feature extraction, we performed layer-by-layer visualization of intermediate feature maps, as shown in Fig. 11. This figure illustrates the evolution of the transformation of feature representations derived from plantar pressure data during running activity, starting from the initial input and progressing through multi-scale convolution, CNN modules, channel attention, spatial attention, and cross-attention, ultimately forming more discriminative features.

From Fig. 11 (a) to Fig. 11 (b), multi-scale convolution refines the input gait features, enhancing their detail and representation. Comparing Fig. 11 (c) and Fig. 11 (d), channel attention significantly emphasizes critical channels, such as Channel 34, which corresponds to a specific plantar region, highlighting its importance in feature extraction. In contrast, the spatial attention output in Fig. 11 (e) reveals a distinct weight distribution across the feature map. The cross-attention feature map in Fig. 11 (f) illustrates the integration of channel and spatial features, further enhancing feature representation. Finally, after additional processing

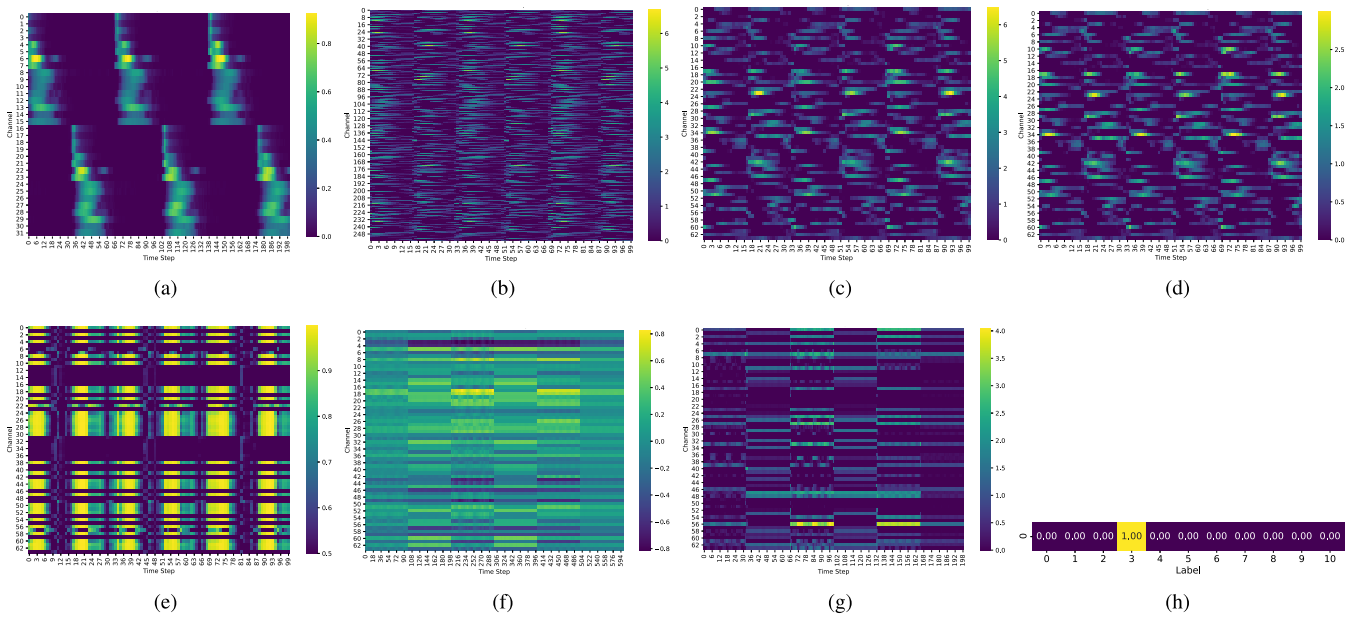


Fig. 11. Visualization of MSCAF-Gait's processing of pressure data from the PIAG dataset during running. (a) Input. (b) Multi-scale convolution output (F_{m}). (c) Pixel-level refined output (M). (d) Channel attention output (H). (e) Spatial attention output (S). (f) Cross-attention output (O_{f}). (g) Final output before FC layer. (h) Softmax output.

through the CNN module, the final softmax output in Fig. 11 (h) assigns the label 3, which corresponds to the running activity in the classification task.

To evaluate the feature representation capabilities of different models, t-distributed Stochastic Neighbor Embedding (t-SNE) is employed to visualize intermediate features extracted from the penultimate layer. As shown in Fig. 12., 2,000 samples are randomly selected from the training set of the PIAG dataset, and their high-dimensional features are projected into a two-dimensional space using t-SNE for visualization. To visualize and assess the clustering structure in the learned feature space, we applied K-Means clustering [36] on the t-SNE-reduced representations. The Homogeneity Score and Completeness Score are then computed to quantify the consistency between the clustering assignments and the ground-truth labels.

Although the CNN-CBAM model achieves a high classification accuracy of 97.07%, its Homogeneity and Completeness Scores are only 0.775 and 0.779, respectively, substantially below those of other models. While CBAM enhances local saliency through channel and spatial attention, this selective focus on localized features may inadvertently increase intra-class dispersion in the latent space, thus degrading clustering quality.

The proposed MSCAF-Gait model integrates multi-scale attention mechanisms and cross-branch feature enhancement, enabling it to learn more structured and discriminative representations. As shown in Fig. 12, MSCAF-Gait produces well-separated inter-class clusters and compact intra-class distributions in the t-SNE embedding space. It achieves Homogeneity and Completeness Scores of 0.996 each, substantially outperforming all baseline models by more than 20% relative to the best-performing alternative. These results confirm the model's strong capacity to learn robust and discriminative features, directly contributing to its high recognition performance.

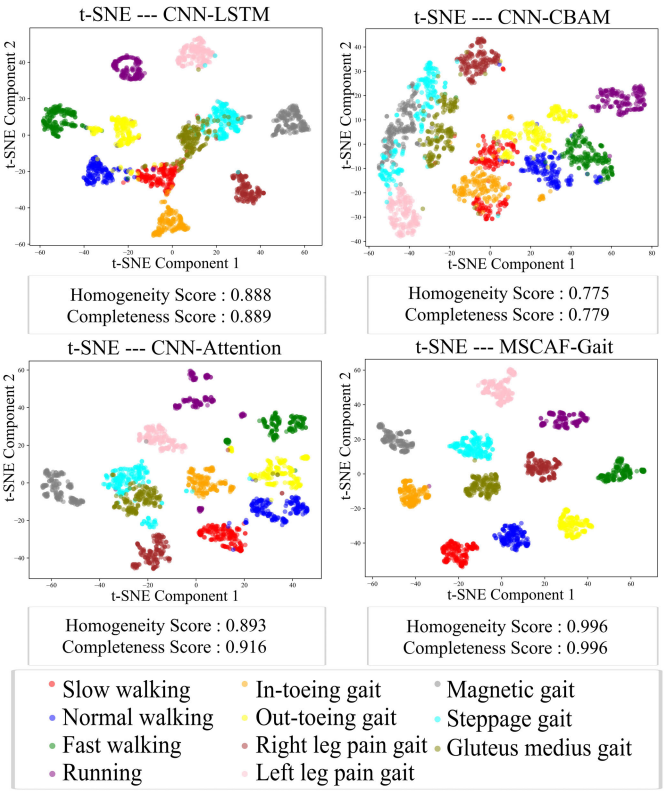


Fig. 12. t-SNE visualizations of intermediate feature representations learned by different models on the PIAG dataset. Homogeneity and Completeness Scores are reported below each subfigure to quantitatively assess the quality of feature clustering.

E. Model Deployment

To evaluate the real-time deployment performance of the MSCAF-Gait model, experiments were conducted on the Raspberry Pi 4B platform. The model was first trained on the PIAG dataset and subsequently exported to the edge

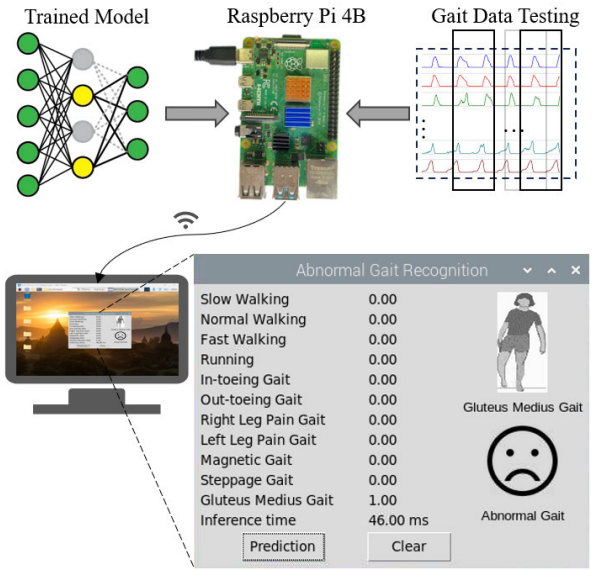


Fig. 13. Deployment of MSCAF-Gait on Raspberry Pi 4B for real-time abnormal gait recognition.

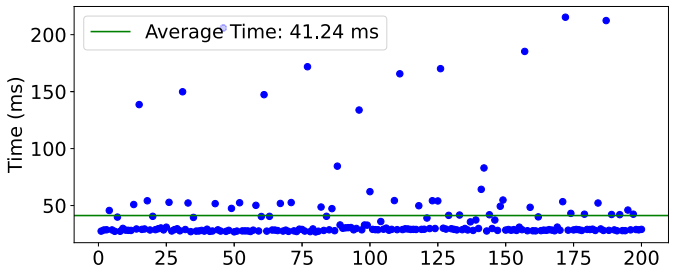


Fig. 14. Inference time of MSCAF-Gait tested 200 times on Raspberry Pi.

device for on-device inference. MSCAF-Gait was designed for computational efficiency, with approximately 0.22 million parameters and 15.83 million FLOPs, making it well-suited for resource-constrained environments. As shown in Fig. 13, a lightweight interface was developed for real-time visualization of the model's recognition probabilities, enabling gait type recognition and abnormal gait alerts directly on the edge device.

To assess latency, inference times were measured for 200 randomly selected test samples on the Raspberry Pi 4B. The results, visualized in Fig. 14, show that the model achieves an average inference time of 41.24 ms, with over 95% of samples processed within 100 ms. Most samples were completed within 50 ms, satisfying the latency requirements of real-time applications such as fall risk detection and rehabilitation monitoring. A small fraction of samples exhibited longer inference times (>120 ms), likely due to runtime system-level variations such as CPU frequency scaling, memory access latency, and background task interference. These outliers do not reflect the model's computational efficiency but are instead attributed to dynamic resource scheduling behavior typical of embedded platforms. Overall, the MSCAF-Gait model demonstrates robust low-latency performance and practical deployability on edge devices, validating its suitability for on-device gait recognition and health monitoring applications.

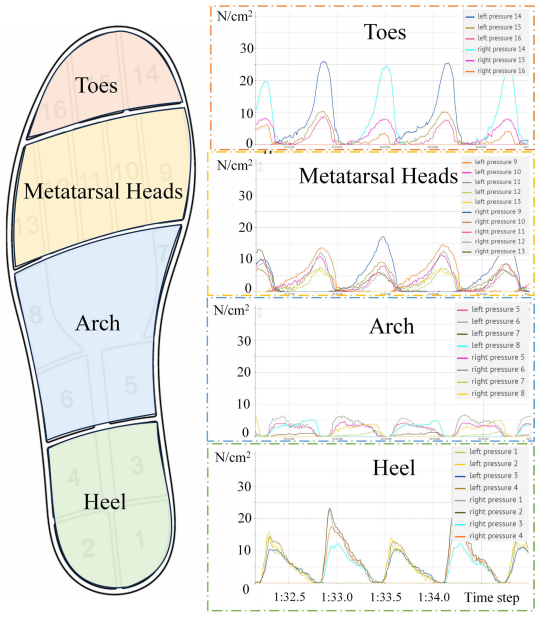


Fig. 15. Visualization of foot pressure distribution across four regions and raw data during normal walking in the PIAG.

F. Discussion

1) Impact of Foot Pressure Sensor Placement on Gait Recognition: To investigate the physiological impact of sensor placement on abnormal gait recognition, we examine differences in pressure sensor distributions across the two datasets. The plantar surface of the foot is biomechanically divided into four regions based on pressure distribution: toes, metatarsal heads, arch, and heel. As shown in Fig. 15, the PIAG dataset captures plantar pressure variations across these regions during normal walking, with the arch region showing relatively minimal pressure fluctuation. To further investigate regional contributions, Table VIII compares performance using sensors from individual regions and excluding specific regions. Using only the heel region results in a drop of just 0.86% in accuracy, the smallest decline among the four regions, suggesting that the heel region alone contributes substantially to the overall accuracy. Conversely, excluding the heel region leads to a 0.44% reduction in accuracy, further underscoring its importance in optimal sensor configuration.

Fig. 16 visualizes classification precision across different gait patterns for each region. Heel exhibits the most consistent performance, with precision approaching 1.0 across all gait patterns, particularly excelling in Fast Walking and pathological gaits like Magnetic Gait. Similarly, the Metatarsal heads maintain high precision, especially in Normal Walking, Right Leg Pain and Left Leg Pain, highlighting the region's sensitivity to pressure variations during gait cycles. In contrast, the Arch region shows slightly lower precision than other regions, with noticeable fluctuations in longer-cycle gaits, such as Slow Walking and Gluteus Medius Gait. However, it performs well in gaits directly influenced by the arch, such as In-toeing Gait. The Toes region exhibits a precision trend similar to Metatarsal Heads, though with overall lower precision, suggesting limited capability in capturing distinct gait characteristics. Overall, the Heel and Metatarsal Heads regions are key contributors to

TABLE VIII
ACCURACY AND DIFFERENCES IN FOOT PRESSURE REGIONS AND COMBINATIONS

Foot Pressure Distribution	Acc. (%)	Difference (%)
Toes	96.50	-2.92
Metatarsal Heads	97.71	-1.71
Arch	96.41	-3.01
Heel	**98.56**	**-0.86**
Metatarsal Heads, Arch, Heel	99.36	-0.06
Toes, Arch, Heel	99.39	-0.03
Toes, Metatarsal Heads, Heel	99.02	-0.40
Toes, Metatarsal Heads, Arch	**98.98**	**-0.44**
Toes, Metatarsal Heads, Arch, Heel	**99.42**	**0.00**

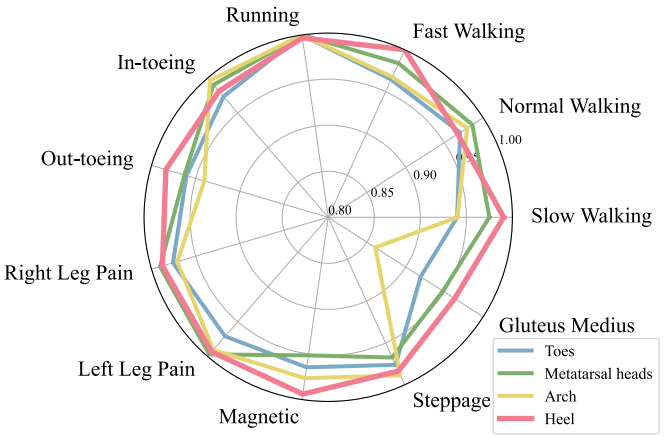


Fig. 16. Classification precision across foot regions for various gait patterns.

gait recognition accuracy, showcasing superior performance compared to other regions.

2) Practical Applicability and Deployment Scenarios: In practical gait recognition scenarios, while model accuracy remains essential, computational complexity and data processing efficiency are equally critical for real-world deployment. The MSCAF-Gait model strikes a balance between high recognition accuracy and computational complexity, while minimizing preprocessing steps. Unlike some existing methods that rely on gait cycle extraction to improve accuracy, which often require complex preprocessing and may hinder deployment, our model avoids excessive processing of raw data and uses a time-window segmentation approach for gait recognition. This ensures high precision while enhancing real-time analysis efficiency. The simplified data processing pipeline and lightweight design enable the model to operate efficiently on resource-constrained devices, meeting the real-time gait recognition and analysis requirements for practical deployment. Based on these considerations, the MSCAF-Gait model demonstrates significant potential in the following key application domains:

1) **Wearable gait monitoring systems and home-based rehabilitation:** Thanks to its lightweight architecture and low computational cost, MSCAF-Gait enables real-time inference on edge devices such as Raspberry Pi 4B, with an average latency of 41.24 ms. This low-latency performance facilitates seamless integration into wearable systems for continuous gait monitoring and remote rehabilitation management. Additionally, its ability to distinguish multiple abnormal gait patterns supports clinical demands for personalized and fine-grained rehabilitation guidance.
2) **Early warning systems for neurodegenerative diseases:** The model's sensitivity to subtle gait abnormalities renders it suitable for early screening and longitudinal monitoring of conditions such as Parkinson's and Alzheimer's diseases. Continuous foot pressure monitoring via wearable sensors may enable pre-symptomatic risk assessment, providing clinicians with an extended time window for early intervention and disease management.

Furthermore, the regional pressure analysis presented in this study offers actionable insights into optimizing sensor placement, which could reduce hardware complexity and improve user comfort while maintaining high recognition accuracy. Such optimizations are critical for enhancing user comfort and promoting device adoption in real-world scenarios.

V. CONCLUSION

This work presents MSCAF-Gait, a novel multi-scale cross-attention fusion network for abnormal gait recognition based on foot pressure data. The proposed model achieves over 99.4% accuracy in both general abnormal gait recognition and PD diagnosis tasks, demonstrating strong potential for real-time gait analysis. Through extensive experiments, including ablation studies, feature visualization, and LOSO evaluation, we demonstrate the model's robustness in capturing temporal-spatial gait dynamics while maintaining computational efficiency. However, the performance decline under LOSO validation highlights the challenge of generalizing across subjects with diverse gait characteristics. Additionally, class imbalance in PD severity recognition and dependence on specific gait patterns may restrict model generalizability across broader populations. Future work may explore improved generalization through enhanced regularization, domain adaptation, and meta-learning techniques, as well as exploring sensor placement optimization tailored to application-specific constraints. Overall, MSCAF-Gait offers a robust, efficient, and deployable solution for abnormal gait recognition with promising applications in rehabilitation monitoring, clinical screening, and wearable healthcare systems.

REFERENCES

[1] J. Baker, "Gait disorders," *Amer. J. Med.*, vol. 131, no. 6, pp. 602–607, 2017.
[2] A. P. J. Zanardi et al., "Gait parameters of Parkinson's disease compared with healthy controls: A systematic review and meta-analysis," *Sci. Rep.*, vol. 11, no. 1, p. 752, Jan. 2021.
[3] X. Wang, H. Yu, S. Kold, O. Rahbek, and S. Bai, "Wearable sensors for activity monitoring and motion control: A review," *Biomimetic Intell. Robot.*, vol. 3, no. 1, Mar. 2023, Art. no. 100089.
[4] V. Skaramagkas, A. Pentari, Z. Kefalopoulou, and M. Tsiknakis, "Multi-modal deep learning diagnosis of Parkinson's disease—A systematic review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2399–2423, 2023.

[5] J. A. Ramirez-Bautista, J. A. Huerta-Ruelas, S. L. Chaparro-Cárdenas, and A. Hernández-Zavala, "A review in detection and monitoring gait disorders using in-shoe plantar measurement systems," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 299–309, 2017.

[6] F. Liang et al., "Interlimb and intralimb synergy modeling for lower limb assistive devices: Modeling methods and feature selection," *Cyborg Bionic Syst.*, vol. 5, p. 122, Jan. 2024.

[7] T.-Y. Xiang et al., "Quantitative movement analysis using scaled information implied in monocular videos," *IEEE Trans. Med. Robot. Bionics*, vol. 5, no. 1, pp. 88–99, Feb. 2023.

[8] J. Wu, J. Wang, and L. Liu, "Feature extraction via KPCA for classification of gait patterns," *Hum. Movement Sci.*, vol. 26, no. 3, pp. 393–411, Jun. 2007.

[9] S.-I. Sakamoto, Y. Hutabarat, D. Owaki, and M. Hayashibe, "Ground reaction force and moment estimation through EMG sensing using long short-term memory network during posture coordination," *Cyborg Bionic Syst.*, vol. 4, p. 16, Jan. 2023.

[10] X. Liu, W. Li, Z. Liu, F. Du, and Q. Zou, "A dual-branch model for diagnosis of Parkinson's disease based on the independent and joint features of the left and right gait," *Int. J. Speech Technol.*, vol. 51, no. 10, pp. 7221–7232, Mar. 2021.

[11] T.-Y. Xiang et al., "Upper limb motor sequence analysis: From isolated to sequential," *IEEE Trans. Ind. Informat.*, vol. 21, no. 7, pp. 5093–5103, Jul. 2025.

[12] Z. Wu and Y. Cui, "GaitFFDA: Feature fusion and dual attention gait recognition model," *Tsinghua Sci. Technol.*, vol. 30, no. 1, pp. 345–356, Feb. 2025.

[13] D. M. D. Nguyen, M. Miah, G.-A. Bilodeau, and W. Bouachir, "Transformers for 1D signals in Parkinson's disease detection from gait," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 5089–5095.

[14] X. Wu, S. Bai, and L. O'Sullivan, "Editorial for the special issue on wearable robots and intelligent device," *Biomimetic Intell. Robot.*, vol. 3, no. 2, Jun. 2023, Art. no. 100102.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[17] C. Li, B. Wang, Y. Li, and B. Liu, "A lightweight pathological gait recognition approach based on a new gait template in side-view and improved attention mechanism," *Sensors*, vol. 24, no. 17, p. 5574, Aug. 2024.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[19] L. Tang et al., "A multimodal fusion network based on a cross-attention mechanism for the classification of parkinsonian tremor and essential tremor," *Sci. Rep.*, vol. 14, no. 1, p. 28050, Nov. 2024.

[20] G. Yogev, N. Giladi, C. Peretz, S. Springer, E. S. Simon, and J. M. Hausdorff, "Dual tasking, gait rhythmicity, and Parkinson's disease: Which aspects of gait are attention demanding?" *Eur. J. Neurosci.*, vol. 22, no. 5, pp. 1248–1256, Sep. 2005.

[21] J. M. Hausdorff, J. Lowenthal, T. Herman, L. Gruendlinger, C. Peretz, and N. Giladi, "Rhythmic auditory stimulation modulates gait variability in Parkinson's disease," *Eur. J. Neurosci.*, vol. 26, no. 8, pp. 2369–2375, Oct. 2007.

[22] S. Frenkel-Toledo, N. Giladi, C. Peretz, T. Herman, L. Gruendlinger, and J. M. Hausdorff, "Treadmill walking as an external pacemaker to improve gait rhythm and stability in Parkinson's disease," *Movement Disorders*, vol. 20, no. 9, pp. 1109–1114, Sep. 2005.

[23] C. G. Goetz et al., "Movement disorder society task force report on the hoehn and yahr staging scale: Status and recommendations the movement disorder society task force on rating scales for Parkinson's disease," *Movement Disorders*, vol. 19, no. 9, pp. 1020–1028, Sep. 2004.

[24] M.-J. Gui, X.-H. Zhou, X.-L. Xie, S.-Q. Liu, Z.-Q. Feng, and Z.-G. Hou, "Soft magnetic Skin's deformation analysis for tactile perception," *IEEE Trans. Ind. Electron.*, vol. 70, no. 12, pp. 12883–12893, Dec. 2023.

[25] M.-J. Gui et al., "Highly interpretable representation for multidimensional tactile perception," *IEEE Trans. Med. Robot. Bionics*, vol. 6, no. 1, pp. 340–350, Feb. 2024.

[26] D. Rosenbaum, "Foot loading patterns can be changed by deliberately walking with in-toeing or out-toeing gait modifications," *Gait Posture*, vol. 38, no. 4, pp. 1067–1069, Sep. 2013.

[27] G. R. Finney, "Normal pressure hydrocephalus," *Int. Rev. Neurobiol.*, vol. 84, pp. 263–281, Jan. 2009.

[28] N. Jamshidi, M. Rostami, S. Najarian, M. B. Menhaj, M. Saadatnia, and F. Salami, "Differences in center of pressure trajectory between normal and steppage gait," *J. Res. Med. Sci., Off. J. Isfahan Univ. Med. Sci.*, vol. 15, no. 1, p. 33, 2010.

[29] A. I. Semciw, T. Pizzari, G. S. Murley, and R. A. Green, "Gluteus medius: An intramuscular EMG investigation of anterior, middle and posterior segments during gait," *J. Electromyogr. Kinesiol.*, vol. 23, no. 4, pp. 858–864, Aug. 2013.

[30] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2019, pp. 2623–2631.

[31] M. R. Daliri, "Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease," *Biomed. Signal Process. Control*, vol. 8, no. 1, pp. 66–70, Jan. 2013.

[32] K. Açıcı, Ç. B. Erdaş, T. Aşuroğlu, M. K. Toprak, H. Erdem, and H. Oğul, "A random forest method to detect Parkinson's disease via gait analysis," in *Proc. 18th Int. Conf. Eng. Appl. Neural Netw.*, Athens, Greece. Cham, Switzerland: Springer, 2017, pp. 609–619.

[33] S. Veeraragavan, A. A. Gopalai, D. Gouwanda, and S. A. Ahmad, "Parkinson's disease diagnosis and severity assessment using ground reaction forces and neural networks," *Frontiers Physiol.*, vol. 11, Nov. 2020, Art. no. 587057.

[34] Y. Xia, Z. Yao, Q. Ye, and N. Cheng, "A dual-modal attention-enhanced deep learning network for quantification of Parkinson's disease characteristics," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 42–51, Jan. 2020.

[35] S. Naimi, W. Bouachir, and G.-A. Bilodeau, "HCT: Hybrid convnet-transformer for Parkinson's disease detection and severity prediction from gait," in *Proc. Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2023, pp. 814–819.

[36] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in K-means clustering," *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.