applied sciences

MDPI

Article

Pathological Gait Classification Using Early and Late Fusion of Foot Pressure and Skeleton Data

Muhammad Tahir Naseem [1], Haneol Seo [1], Na-Hyun Kim [2] and Chan-Su Lee [2,*]

[1] Research Institute of Human Ecology, Yeungnam University, Gyeongsan 38541, Republic of Korea; nmtahir@yu.ac.kr (M.T.N.); haneol@yu.ac.kr (H.S.)
[2] Department of Electronic Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea; nh_kim@yu.ac.kr
* Correspondence: chansu@ynu.ac.kr; Tel.: +82-53-810-3527

Abstract: Classifying pathological gaits is crucial for identifying impairments in specific areas of the human body. Previous studies have extensively employed machine learning and deep learning (DL) methods, using various wearable (e.g., inertial sensors) and non-wearable (e.g., foot pressure plates and depth cameras) sensors. This study proposes early and late fusion methods through DL to categorize one normal and five abnormal (antalgic, lurch, steppage, stiff-legged, and Trendelenburg) pathological gaits. Initially, single-modal approaches were utilized: first, foot pressure data were augmented for transformer-based models; second, skeleton data were applied to a spatiotemporal graph convolutional network (ST-GCN). Subsequently, a multi-modal approach using early fusion by concatenating features from both the foot pressure and skeleton datasets was introduced. Finally, multi-modal fusions, applying early fusion to the feature vector and late fusion by merging outputs from both modalities with and without varying weights, were evaluated. The foot pressure-based and skeleton-based models achieved 99.04% and 78.24% accuracy, respectively. The proposed multi-modal approach using early fusion achieved 99.86% accuracy, whereas the late fusion method achieved 96.95% accuracy without weights and 99.17% accuracy with different weights. Thus, the proposed multi-modal models using early fusion methods demonstrated state-of-the-art performance on the GIST pathological gait database.

Keywords: deep learning; machine learning; early fusion; late fusion; foot pressure; skeleton; transformer; spatiotemporal graph convolutional network

Citation: Naseem, M.T.; Seo, H.; Kim, N.-H.; Lee, C.-S. Pathological Gait Classification Using Early and Late Fusion of Foot Pressure and Skeleton Data. *Appl. Sci.* **2024**, *14*, 558. https://doi.org/10.3390/app14020558

Academic Editors: João M. F. Rodrigues, Mukesh Prasad, Pang-jo Chun, Xian Tao and Ali Braytee

Received: 7 October 2023
Revised: 1 December 2023
Accepted: 5 January 2024
Published: 9 January 2024

1. Introduction

Gait defines a distinct pattern of human mobility, strongly influenced by individual biological characteristics such as height, weight, age, and gender. Therefore, it enables differentiation among individuals. Gait analysis has been extensively used to study walking behaviors in both healthy individuals [1] and those afflicted with neurological conditions such as Huntington's disease (HD), Amyotrophic Lateral Sclerosis (ALS), and Parkinson's disease (PD) [2]. Furthermore, it has been used beyond healthcare, extending to domains such as biometric security systems [3]. Researchers have devised various techniques for assessing gait, utilizing a range of sensors, including inertial sensors [4], foot pressure sensors [5], depth cameras [6], and motion capture systems [7].

Wearable and non-wearable sensors are the two primary categories of sensors employed for analyzing human gaits. A typical wearable sensor is the inertial sensor [8], which can be affixed to the hips, knees, and ankles. OptiTrack and Vicon are also wearable sensors that require users to attach the sensors to their bodies. Wearable sensors provide accurate data; however, they require users to attach the sensors to their bodies, resulting in discomfort during walking [9]. Another noteworthy sensor is the foot pressure sensor [10], which comprises an array of sensors that calculate cumulative pressure across the entire surface by measuring pressure on individual cells. With advancements in chip technology,

foot pressure sensors [11] are now integrated into shoe insoles, allowing for more efficient and comfortable data collection.

Recently developed depth cameras, such as Kinect v1, v2, and Azure Kinect by Microsoft (Redmond, DC, USA) [12], are unintrusive sensors capable of capturing 3D information about human joints. However, they offer lower accuracy. Additionally, one can only capture a few steps at a time while the participant is walking. Nonetheless, the reduced accuracy is generally deemed acceptable owing to the fact that 3D skeleton data are derived as secondary information from depth data. Furthermore, these cameras can cover long distances and are more cost-effective.

Marker-based gait quantification is considered a gold standard by the research and health communities. It reconstructs motion in 3D and provides parameters to measure gait. However, it is an expensive and intrusive technique, limited to soft tissue artifacts, prone to incorrect marker positioning, and associated with skin sensitivity problems. Ref. [13] illustrated a 3D gait motion analyzer employing impulse radio ultra-wideband (IR-UWB) wireless technology. The prototype can measure 3D motion and determine quantitative parameters considering anatomical reference planes. Knee angles have been calculated from the gait by applying vector algebra. Simultaneously, the model has been corroborated by the popular marker-less camera-based 3D motion-capturing system, the Kinect sensor. Three-dimensional shape information is a crucial clue to understanding the posture and shape of pedestrians. However, most existing person-Re-ID methods learn pedestrian feature representations from images, ignoring the real 3D human body structure and the spatial relationship between pedestrians and interferents. To address this problem, ref. [14] devised a new point cloud Re-ID network (PointReIDNet), designed to obtain 3D shape representations of pedestrians from point clouds of 3D scenes. The model consists of modules, namely the global semantic guidance and local feature extraction modules. The global semantic guidance module enhances feature representation and reduces the interference caused by 3D shape reconstruction or noise.

The authors in [15] simulated the depth-wise separable convolution calculation method in the point cloud and proposed a new type of convolution, namely dynamic cover convolution (DC-Conv), to aggregate local features. The core of DC-Conv is the space cover operator (SCOP), which constructs anisotropic spatial geometry in a local area to cover the local feature space to enhance the compactness of local features. DC-Conv achieves the capture of local shapes by dynamically combining multiple SCOPs in the local neighborhood. Among them, the attention coefficients of the SCOPs are adaptively learned from the point position in a data-driven manner. Experiments on the 3D point cloud shape recognition benchmark datasets ModelNet40, ModelNet10, and ScanObjectNN show that this method can effectively improve the performance of 3D point cloud shape recognition and robustness to sparse point clouds, even in the case of a single scale.

Ref. [16] addressed the problem of gait classification using different pre-trained single- and multi-modal foot pressure and skeleton datasets. This method uses conventional convolutional neural networks (CNNs) for single- and multi-modal methods. However, other methods, such as vision transformers and spatiotemporal graph convolutional network (ST-GCN), exist. Regarding fusion, the method uses only early fusion, which concatenates the best features from both modalities. Similarly, fusion can be performed in other ways, such as by combining the outputs from different modalities with and without weighted summation. Additionally, the accuracies for single-modals, foot pressure, and skeleton datasets were 68.82% and 93.40%, respectively, which are low. However, for the multimodal, the accuracy was 97.60%, which was somewhat adequate. To address these limitations, a classification model other than CNN for single-modal methods is required to effectively classify pathological gaits and achieve improved performance. Other methods for fusing the different multi-modalities are also required.

This study focuses on a more effective approach to classifying one normal gait and five abnormal pathological gaits (antalgic, lurching, steppage, stiff-legged, and Trendelenburg) using datasets from the GIST pathological gait database [16], which includes foot

pressure and skeleton datasets. Single-modal and multi-modal methods were proposed, employing early and late fusion techniques to assess gait classification accuracy. In the single-modal approaches, transfer-based models for the foot pressure dataset and the ST-GCN model for the skeleton dataset were utilized. In the multi-modal approach using early fusion, features from both modalities were concatenated, whereas in the multi-modal approach using late fusion, the outputs from the two modalities were combined in various ways, both with and without different weights. The choice of weights depended on the accuracy of the trained single models. Our early fusion method outperformed the late fusion method. When comparing our single-modal and multi-modal approaches, both in early and late fusion configurations, our models demonstrated state-of-the-art performance with the GIST pathological gait database.

The remainder of the paper is organized as follows: Section 2 summarizes related works, encompassing sensor-based, vision-based, and multimodal-based methods. Section 3 discusses the limitations of previous works and our contributions. Section 4 describes the details of the dataset utilized in this study. Section 5 first discusses the proposed single-modal models (MViTv2_base and MViTv2_small) and ST-GCN; second, the proposed multi-modal model using early fusion is discussed; and third, the proposed multi-modal using late fusion with and without different weights is discussed. Section 6 presents potential applications of gaits, conclusions, and some future research directions.

2. Related Works

Related works on machine learning (ML) and deep learning (DL) methods for gait recognition can be categorized into three distinct approaches: sensor-based, vision-based, and multimodal-based approaches.

2.1. Sensor-Based Approaches

Wearable sensors are widely employed in studying the impact and significance of descriptive statistical factors on osteopenia and sarcopenia using artificial intelligence (AI) [17]. ML techniques have been explored to address natural variations in gaits among different subjects using wearable sensors, with an accuracy rate of 81% [18]. Furthermore, a DL approach based on long short-term memory (LSTM) has been proposed for the automatic detection of initial contact (IC) and toe-off (TO) using foot-marker kinematics [19]. Another DL-based study has been reported on the accuracy of various neural networks in modeling lower body joint angles in the sagittal plane, utilizing kinematic records from a single inertial measurement unit (IMU) attached to the foot [20]. This model demonstrates superior performance compared to ML approaches. Additionally, a paper [21] delves into gait classification based on CNN using interferometric radar.

In [22], a smart insole equipped with various sensor arrays is discussed for gait classification based on DL. The results encompass seven types of gaits: walking, fast walking, running, stair climbing, stair descending, hill climbing, and hill descending. The method achieves a high classification rate of 90%. In [23], two models are presented, one hardware-based, consisting of multiple sensor modules (MSM), and another software-based, composed of a biomedical and inertial sensor algorithm, along with the leg health classification net (LCNet) model. These models achieve an impressive accuracy of 94.41%. Ref. [24] introduces a DL-based model that conducts a comprehensive examination of ground reaction force (GRF) patterns to detect normal gait and gait abnormalities. Another model, described in [25], employs five classifiers (K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and Stochastic Gradient Descent (SGD)) for the electromyography (EMG) dataset. The results demonstrate 99% accuracy using KNN and RF, whereas the DT classifier achieves 97% accuracy. In [26], a model is discussed for classifying numerous anatomical regions and their combinations using a vast and highly unbalanced dataset. Furthermore, ref. [27] presents a model that employs Shapley's additive explanations to select important parameters with the Support Vector

Machine (SVM), RF, and multilayer perceptron. The highest accuracy of 95% was achieved using an SVM classifier.

These sensor-based approaches sometimes achieve high accuracy in gait classification. However, they demand specialized hardware sensors and possess limitations in their general applicability, particularly for gait classification in cases of sarcopenia.

2.2. Vision-Based Approaches

In [28], the Kinect motion system is utilized to collect spatiotemporal gait data from seven healthy subjects across three walking trials: normal, pelvic-obliquity, and knee-hyperextension walking. Four classifiers—LSTM, SVM, KNN, and CNN—are employed. Notably, SVM and KNN perform well, achieving classification accuracies of 94.9% and 94.0%, respectively. Ref. [29] introduces a deep CNN that employs 3D convolutions for gait recognition from multiple viewpoints, capturing spatiotemporal features. This approach is evaluated across three different datasets, accounting for variations in clothing, walking speeds, and viewing angles. Additionally, ref. [30] proposes gait recognition using an enhanced CNN with the incorporation of a Gabor filter.

In [31], the researchers propose a model based on Gait Energy Images (GEIs) to automatically extract robust and discriminative spatial gait features for human identification. They leverage deep 3-dimensional CNNs to learn the temporal gait features, referred to as Convolutional 3D (C3D) representations. In [32], a classifier based on Gated Recurrent Units (GRUs) is proposed. It is used to classify six different gaits, including one normal and five pathological gaits (antalgic, stiff-legged, lurching, steppage, and Trendelenburg), achieving an accuracy of 93.67%.

A study detailed in [33] that introduces an RNN-based model for the classification of pathological gaits based on skeleton data has garnered significant attention. Utilizing bidirectional LSTM, the model achieves an accuracy of 88.90%. It categorizes gaits into six categories: normal, in-toeing, out-toeing, drop foot, pronation, and supination. Furthermore, in [34], features extracted by an LSTM-based autoencoder are employed in a GRU classifier. This approach accurately distinguishes between normal, limping, and knee-stiff gaits across various levels, achieving an impressive accuracy rate of 95.90%.

2.3. Multimodal-Based Approaches

In [35], an automated knee osteoarthritis (KOA) classification algorithm based on the Kellgren–Lawrence (KL) grading system is examined utilizing radiographic imaging and gait analysis data. The model achieves an F1-score, sensitivity, and precision of 0.70, 0.76, and 0.71, respectively.

In [16], a DL-based multimodal model is presented for classifying six gait types, including one normal and five pathological (antalgic, lurch, steppage, stiff-legged, and Trendelenburg) gaits. This model uses both skeletal and foot pressure data. To efficiently extract features from these two different data types for classification, the model inputs sequential skeletal data into recurrent neural network (RNN)-based encoding layers and average foot pressure data into a CNN. The classification accuracy for the pressure-based and skeleton-based single models is 68.82% and 93.40%, respectively. However, the hybrid model outperforms them with an accuracy of 95.66%. Subsequently, the model underwent improvement, achieving an accuracy of 97.60% through a three-step training process.

In [36], gait freezing episodes were classified with an accuracy of 98.1% through the data fusion of scalograms from Wi-Fi sensing and spectrograms from radar sensors. Additionally, ref. [37] explores sensor fusion using data from ambulatory inertial sensors (AISs) and plastic optical fiber-based floor sensors (FSs). The model attains an accuracy of 88% with ANN and 91% with CNN, as it learns optimal data representations from both sensor modalities.

3. Limitations of Previous Works and Contributions

Table 1 outlines the issues with existing works, which often exhibit one or more of the following limitations:

- The system outlined in [16] is intricate, despite its high accuracy, as it involves processing double instances of each dataset.
- The model presented in [21] boasts good accuracy but offers fewer classes and limited datasets from various sources.
- In contrast to the work described in [24], which achieves high accuracy but lacks diversified datasets, the approach addressed in [19] only achieves 81%, a comparatively lower result.
- Whereas the model mentioned in [17] provides good accuracy, its applicability is restricted to specific age groups. Moreover, it is gender-specific and suitable only for slow speeds.
- The model in [25] conducted experiments with fewer classes.

Table 1. Comparison and shortcomings of earlier works.

References	Method	Dataset	Accuracy (%)	Limitations
Jun et al. [16]	RNN and CNN	Foot pressure dataset 3D Skeleton dataset	97.60	Complicated Uses double instances of datasets
Hassan et al. [21]	CNN	Micro-Doppler and Interferometric Spectrogram Image	99.5	Small number of classes Lack of diverse datasets
Wolf et al. [29]	CNN	Casia-B	100	Overfitting Lack of diverse datasets
Youn et al. [18]	SVM, RF, Logistic Regression	Open dataset based on inertial sensor-based system	81	Low accuracy Lack of diverse datasets Handcrafted features
Kim et al. [17]	XGBoost, ResNet	Open dataset based on inertial sensor-based system	93.75	Restricted age group Gender discrimination Restricted speed Lack of diverse datasets Handcrafted features
Naji et al. [25]	KNN, RF, DT, LR, SGD	EMG dataset	99	Less number of classes Handcrafted features
Pandey et al. [24]	GaitRec-Net	GRF measurements from different patients	91.6	Lack of diverse datasets

Achieving a deep understanding of image properties is imperative when relying on a fixed set of handcrafted features [17,18,25]. These methods rely on texture analysis, where classifiers like RF are fed a limited number of locally computed descriptors from the image. Whereas some studies have demonstrated the high accuracy of these strategies, they are constrained in terms of generalizability and inter-dataset variability.

Some of the aforementioned models have demonstrated exceptional accuracy in their respective work. However, they are often tested in a limited number of classes. Others may have a sufficient number of classes but are limited to specific age or gender groups and can only operate within certain speed limits.

Additionally, some researchers have employed diverse datasets, but they often belong to the same modality. In light of the challenges outlined in Table 1, this study makes the following significant contributions:

- Transformer-based single-modal models, MViTv2_base and MViTv2_small, were introduced using the original instances of the foot pressure dataset. These models exhibit enhanced performance in gait classification from foot pressure data.

- A single-modal approach, the ST-GCN model, utilizing the original instances of the skeletal dataset, was presented.
- To enhance the accuracy of the foot pressure dataset, various augmentations were introduced.
- The proposed single-modal models outperform the baseline single-modal models.
- A multi-modal approach employing early fusion was proposed, which utilizes the original instances of both foot pressure and skeletal data simultaneously.
- Additionally multi-modal methods utilizing late fusion were introduced, where the outputs from both modalities were combined, both without and with varying weights.
- The proposed multi-modal method using early fusion performed better than our proposed late fusion methods.
- Our proposed multi-modal models show state-of-the-art performance on the GIST pathological database.

4. Dataset Used

In this study, a publicly available foot pressure and skeleton dataset was utilized from [16]. This dataset comprises one normal gait and five abnormal pathological gaits (antalgic, lurching, steppage, stiff-legged, and Trendelenburg). Twelve healthy males participated in the data collection. All of them were laboratory staff members and fully understood the data collection system. They had watched videos of each pathological gait and trained until they became familiar with simulating them. Data collection was conducted under strict supervision. The data were acquired using a recently released single-depth camera (Azure Kinect, Microsoft, USA), along with foot pressure data collected from a pressure plate (GW1100, GHiWell, Yangju-si, Republic of Korea). Although the size of the dataset was sufficient, it had certain limitations. For example, for the DL models, the size is low. Additionally, the dataset was created for only one gender. Moreover, there are no details about the dataset regarding which age groups participated, and the dataset was not created by real patients.

4.1. Gait Types

The publicly accessible foot pressure and 3D skeleton datasets used in this study include normal, antalgic, lurch, steppage, and stiff-legged gait types. In a normal gait, the spine and pelvis play crucial roles. Antalgic gait is characterized by pain resulting from a specific disease or leg injury. Steppage gait occurs when the toes of the foot avoid contact with the ground due to muscle and motor nerve abnormalities in the front shin. Lurch gait is induced by hip area abnormalities, such as weakness or paralysis of the gluteus maximus. Stiff-legged gait, also known as stiff-knee gait, is often due to quadriceps weakness, caused by joint abnormalities in the knee region. In contrast, Trendelenburg gait results from weakness or paralysis of the blunt middle force, often due to a weak gluteus medius muscle, causing the torso to tilt in the direction of the symptoms during walking [32].

4.2. Foot Pressure Dataset

The GW1100 pressure plate, capable of measuring pressure up to 100 N/cm^2, is equipped with 6144 high-voltage matrix sensors and is 1080 mm $\times$ 480 mm in size [12]. Typically, this sensor can capture data for two walking steps. To create the dataset utilized in this study, which measures average foot pressure, the planar foot pressures from all time sequences were averaged. The foot pressure dataset, encompassing normal, antalgic, lurch, steppage, and Trendelenburg gaits, is depicted in Figure 1. A single-channel image suffices to represent the average foot pressure. For data collection, gait datasets were compiled from 12 subjects, encompassing six different gait types and 20 trials, resulting in a total of $12 \times 6 \times 20 = 1440$ foot pressure samples.

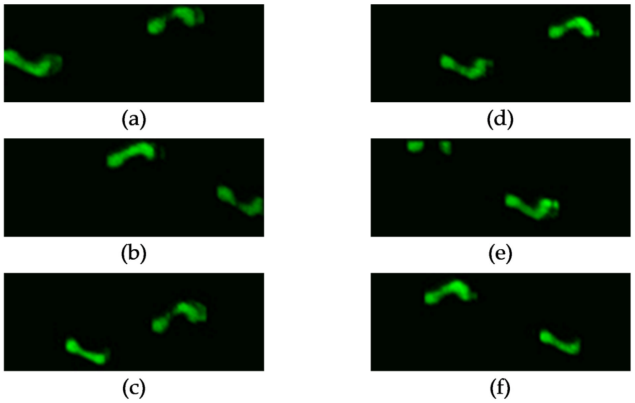

Figure 1. Foot pressure dataset: (**a**) normal gait, (**b**) antalgic gait, (**c**) lurch gait, (**d**) steppage gait, (**e**) stiff-legged gait, and (**f**) Trendelenburg gait.

4.3. The 3D Skeleton Dataset

The most recent Kinect sensor available is the Azure Kinect. In [16], researchers utilized the Azure Kinect sensor along with the associated Microsoft software development kit (SDK) to collect skeleton data. This enabled them to capture the 3D XYZ coordinates for 32 joints, including the hips, knees, ankles, feet, nose, clavicles, shoulders, elbows, wrists, hands, hand tips, and thumbs. Figure 2 illustrates the skeleton dataset encompassing normal, antalgic, lurch, steppage, stiff-legged, and Trendelenburg gaits. The skeletal gaits of the walkers were recorded on a 4 m boardwalk. Similar to the foot-pressure dataset, gait datasets were collected for 12 subjects, comprising six different gait types and 20 trials, resulting in a total of $12 \times 6 \times 20 = 1440$ skeleton data samples.

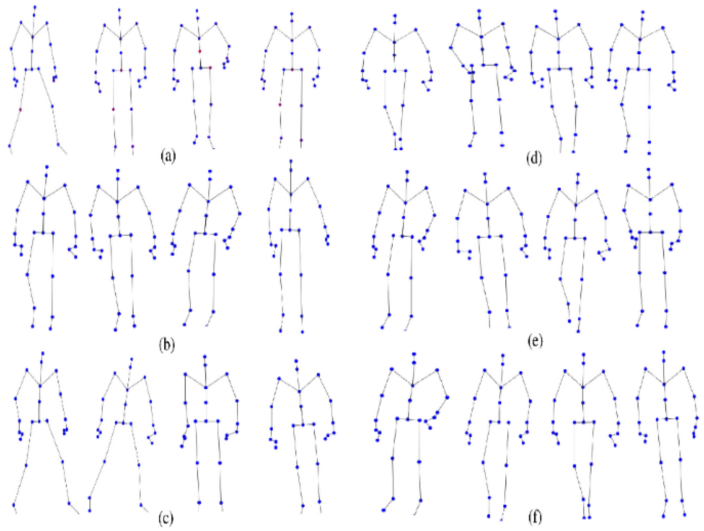

Figure 2. The 3D skeleton dataset: (**a**) normal gait, (**b**) antalgic gait, (**c**) lurch gait, (**d**) steppage gait, (**e**) stiff-legged gait, and (**f**) Trendelenburg gait.

5. Proposed Models

In this section, various aspects of our approach were covered. First, two single-modal models, MViTv2_base, and MViTv2_small, both of which are transformer-based and utilize the foot pressure dataset were introduced. Following that, another single-modal model, ST-GCN, tailored for the skeleton dataset was discussed. Moving into multi-modal methods, the early fusion approach, which concatenates features from both the foot pressure and the skeleton datasets, was explained. Finally, the multi-modal technique was explored using late fusion, both with and without different weights, where the outputs from both modalities are combined.

5.1. Single-Modals (MViTv2_Base and MViTv2_Small) for Foot Pressure Dataset

Recent advancements in image and video classification have been propelled by models based on vision transformers [38]. This unified architecture caters to image and video classification, along with object recognition, demonstrating remarkable performance in image processing and classification tasks. For our purposes, the pre-trained MViTv2_base and MViTv2_small models from MViTv2, which are already trained on ImageNet [39], were employed. Both the pre-trained models were fine-tuned according to a new dataset. The model's evaluation included ImageNet classification and Kinetics video recognition, incorporating decomposed relative positional embeddings and residual pooling connections. MViTv2 stands at the forefront in three key areas: image classification (achieving 88.8% accuracy), COCO object detection (achieving 58.7% accuracy), and video classification (achieving 86.1% accuracy).

Figure 3a illustrates the block diagram for the single-modal models, MViTv2_base and MViTv2_small, employed with the foot pressure dataset after undergoing data augmentation, as detailed in Section 5.1.1. These models were employed for the classification of six pathological gaits, including one normal gait and five abnormal ones (antalgic, lurch, steppage, stiff-legged, and Trendelenburg).

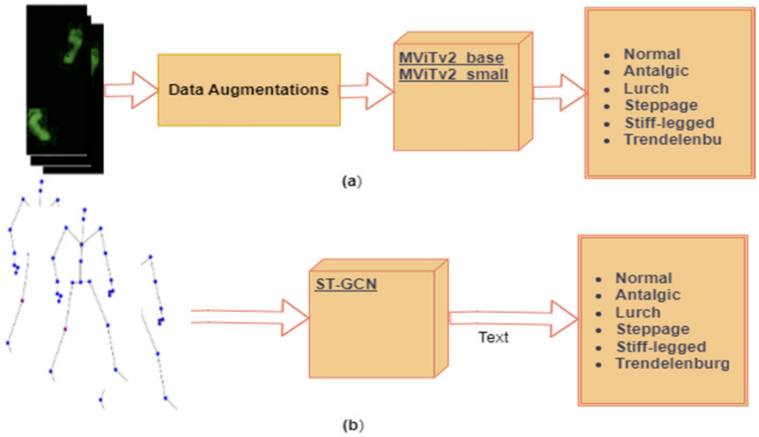

Figure 3. Proposed single-modals (**a**) MViTv2_base and MViTv2_small for foot pressure dataset, and (**b**) single-modal ST-GCN [40] for the skeleton dataset.

5.1.1. Data Augmentations

Insufficient training data can lead to improper training or overfitting problems. Data augmentation, achieved by generating modified data from the original dataset, can effectively augment the dataset's size to mitigate the issue of data scarcity. The foot pressure dataset utilized in our experiment also possesses a relatively limited amount of data (14 to 40 samples). For each foot pressure image, the following successive transforms were performed:

- Flip/mirror the image with a probability of 0.5.
- Rotate the image with a random angle from 15 to 45 (in degrees).

For MViTv2_base and MViTv2_small, the results both without and with augmentation, specifically flipping and rotating images, were evaluated as detailed in Table 2. Without any augmentation, an accuracy of 71.74% for MViTv2_base and 70.83% for MViTv2_small were achieved. For flipping, random horizontal flips with a probability of 0.5 were employed, resulting in an accuracy of 72.43% for MViTv2_base and 73.47% for MViTv2_small. Additionally, random rotations were applied at angles of 15, 30, and 45 degrees, respectively. With a probability of 0.5, these rotations were applied, and the highest accuracy was achieved with a 30-degree rotation: 76.46% for MViTv2_base and 78.24% for MViTv2_small. Notably, applying the 30-degree rotation augmentation led to a substantial accuracy increase of 7.41% (from 70.83% to 78.24%) for the MViTv2_small model.

Table 2. Validation accuracies of MViTv2_base and MViTv2_small with and without augmentations.

With and without Augmentations	MViTv2_Base	MViTv2_Small
Without augmentation	71.74%	70.83%
Flipping (0.5)	72.43%	73.47%
15° rotations	75.62%	76.32%
30° rotations	76.46%	**78.24%**
45° rotations	74.65%	75.56%

5.2. Single-Modal ST-GCN for Skeleton Data

Skeleton data consists of sequential time series data. Notably, the field of skeleton-based action detection has witnessed a surge in the adoption of graph convolutional networks (GCNs) [40,41]. The ST-GCN-based model enhances expressive power and exhibits stronger generalization capabilities. Recently, pathological gait categorization has employed ST-GCN, incorporating attention mechanisms with 3D skeletal data [42]. This innovative technique introduces an attention mechanism for spatiotemporal GCNs, enabling a focus on crucial joints within the current gait. There are two types of edges, namely the spatial edges that conform to the natural connectivity of joints and the temporal edges that connect the same joints across consecutive time steps. Multiple layers of ST-GCN are constructed thereon, which allow information to be integrated along both the graph and the temporal dimension. This is achieved after initially extracting spatiotemporal features from 3D skeletal data through joint linkages and subsequently applying these features to GCNs. The ST-GCN attention mechanism facilitates the concentration on significant joints during pathological gait classification based on skeleton data (refer to Figure 3b for the block diagram illustrating a single-modal ST-GCN). Here, our previously proposed ST-GCN model from [43] was utilized, which uses an attention technique applied to pathological gait classification from the skeleton information. Our focus was twofold. The first objective was to extract spatiotemporal features from skeletal information presented by joint connections and to apply these features to graph convolutional neural networks. The second objective was to develop an attention mechanism for spatiotemporal graph convolutional neural networks to focus on important joints in the current gait. This model establishes a pathological gait classification system for diagnosing sarcopenia.

As detailed in Section 4.3, Azure Kinect was employed to capture the 3D skeleton dataset, which encompasses 32 joints representing the human skeleton [44]. However, in this study, joints were selectively utilized as input data for the ST-GCN model. Specifically, joints such as the nose, left eye, right eye, left clavicle, and right clavicle were excluded. Instead, only 25 joints (as shown in Figure 4) sourced from Azure Kinect, which align with the same 25 joints found in Kinect v2, were used.

5.3. Multi-Modal Using Early Fusion

In this section, the multi-modal approach, employing early fusion by concatenating features from the foot pressure and 3D skeleton data, was introduced as illustrated in Figure 5. These features were extracted from the second-to-last fully connected (fc) layers of both MViTv2_small and ST-GCN. For the foot pressure data, MViTv2_small was used due to its superior performance. Feature extraction plays a pivotal role in uncovering valuable information, which is subsequently harnessed for image classification tasks. To illustrate this, feature extraction allows us to identify facial features such as the eyes, nose, and mouth when presented with an image of a human face. Through concatenation, the performance of these combined models was fused, thereby enhancing the accuracy of individual models.

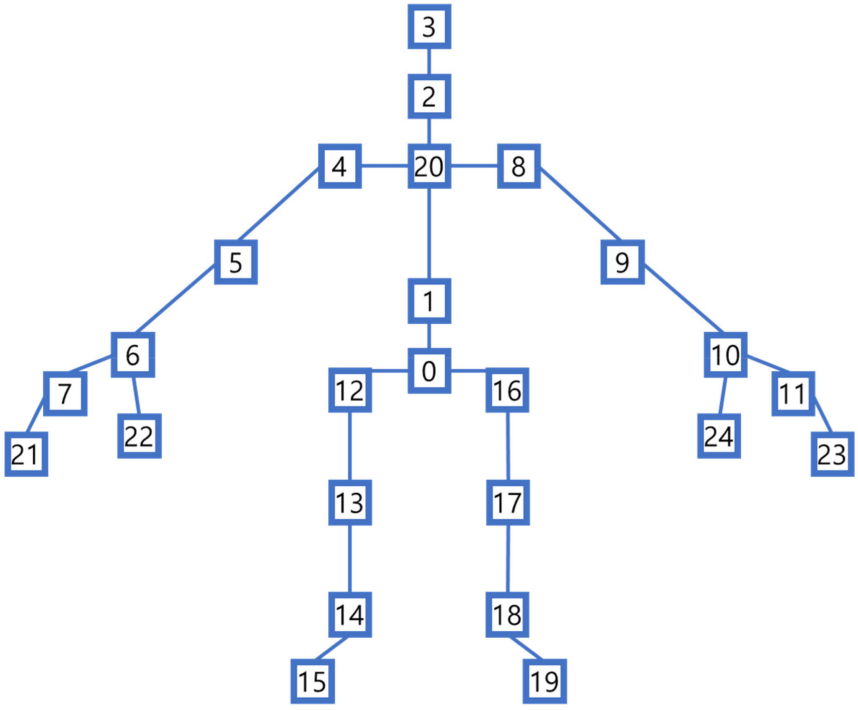

Figure 4. Twenty-five 3D Skeleton joints used in the experiments.

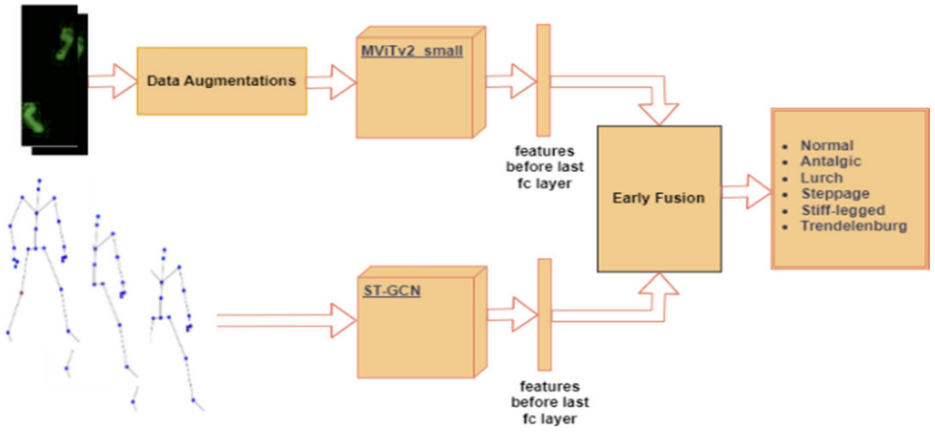

Figure 5. Proposed multi-modal using early fusion by concatenating features from foot pressure and ST-GCN [40] for the skeleton datasets.

5.4. Multi-Modal Using Late Fusion with and without Different Weights

In this section, the multi-modal approach employing late fusion, which involves the combination of both the skeleton and foot pressure datasets, was introduced by utilizing various weight configurations. Data fusion techniques have been extensively explored across diverse domains, including medical applications, as multimodal data fusion holds the potential to enhance the performance of ML models [45,46]. Once again, the MViTv2_small model for the foot pressure dataset was used due to its superior performance.

Similarly, for ST-GCN, our previously proposed model was employed as detailed in [43]. Figure 6 illustrates the block diagram of the late fused method, both without and with varying weights, showcasing the combination of outputs from the two single-modal models, MViTv2_small and ST-GCN, utilizing diverse weighting schemes. These weight selections were made in accordance with the accuracy levels achieved by the already-trained single-modal models, wherein higher accuracy corresponded to greater weights and vice versa. During the evaluation process, different late fusion techniques, including maximum, average, and multiplication, were employed both with and without varying weights.

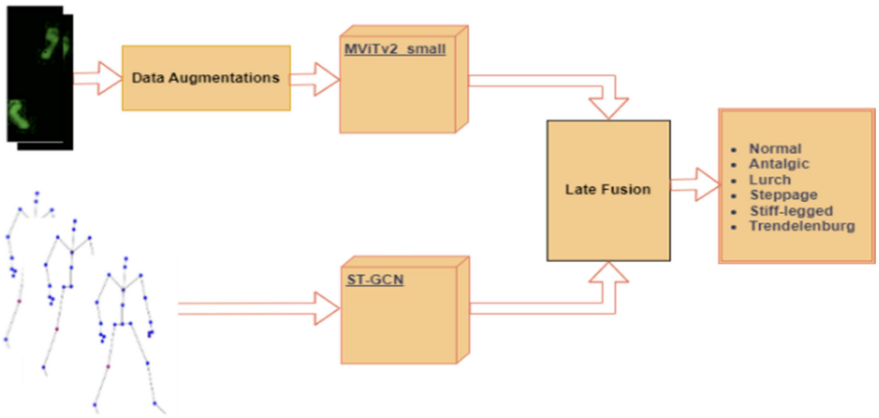

Figure 6. Proposed multi-modal using late fusion by combining the outputs from foot pressure and ST-GCN [40] for the skeleton datasets with and without different weights.

6. Results and Discussions

In this section, a comprehensive evaluation and discussion of our approach are presented. First, the performance of our single-modal models, namely MViTv2_base and MViTv2_small, is assessed and discussed using the foot pressure dataset. Second, the ST-GCN architecture, focusing on the skeleton dataset, is evaluated and discussed. Third, the outcomes of our proposed multi-modal approach are scrutinized and discussed, employing early fusion. Lastly, the results of our proposed multi-modal technique, which employed late fusion, considering various weight configurations, are examined and discussed.

To assess the performance of our proposed single-modal and multi-modal classification models, a leave-one-subject-out cross-validation approach was employed. This entailed using data from one subject as the validation set, whereas the remaining data served as the training set. Consequently, the training procedure was performed 12 times, and the resultant average validation accuracy was calculated. The training process for both the foot-pressure and ST-GCN models was conducted independently for 200 epochs. For the foot-pressure model, the AdamW (Adam with Weight decay) optimizer was employed, whereas the ST-GCN model was optimized using the Adam optimizer. Additionally, in the case of the transformer-based model, the learning rate was set to 0.00006 with a weight decay of 0.01, and the batch size was set to eight. For ST-GCN, the learning rate was set at 0.1, with a weight decay of 0.0001, and a batch size of 16. The cross-entropy loss function was used.

The system specifications utilized for the evaluation in our study consist of an Intel (Santa Clara, CA, USA) CPU with 32 GB of RAM and an NVIDIA (Santa Clara, CA, USA) GeForce RTX 3060 GPU. The implementation of the models in this study was executed using PyTorch 2.1. For the 118 frames of the skeleton dataset, the system takes 11.31 s to evaluate, which processes ten frames per second, so it is applicable in the real world.

6.1. Performance of Single-Modals (MViTv2_Base and MViTv2_Small) for Foot Pressure Dataset

In this context, the foot pressure dataset to train both the MViTv2_base and MViTv2_small models was utilized. Validation accuracies for each individual subject were computed and subsequently calculated the average validation accuracy across all subjects. Figure 7 displays subject-specific validation accuracies when training the foot pressure data with MViTv2_base and MViTv2_small, employing a 30° rotation augmentation. For subject 1, both models achieved an accuracy of 77.50%, whereas for subject 2, MViTv2_base reached 76.67%, while MViTv2 achieved a higher accuracy of 80.83%. Notably, subjects 3, 4, 7, 8, 9, 10, and 11 exhibited improved accuracy when using MViTv2_small.

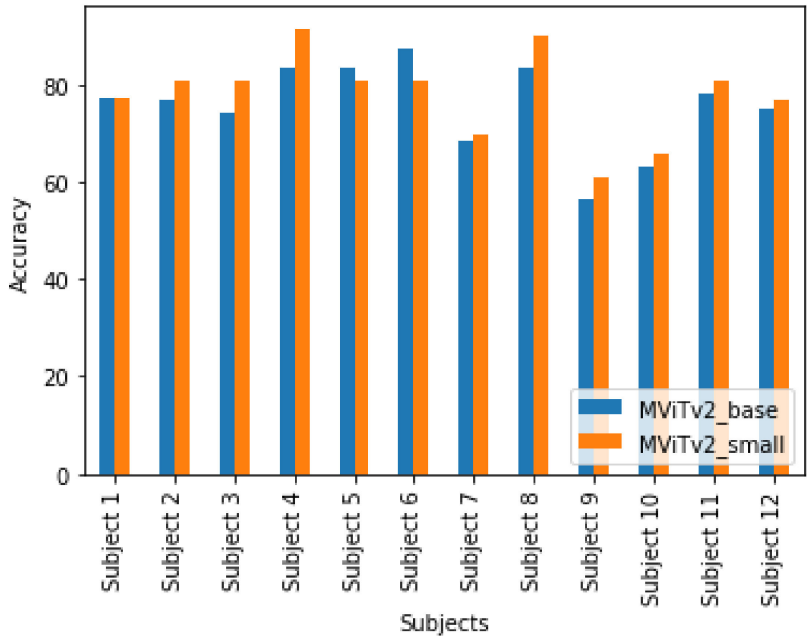

Figure 7. Subject-wise validation accuracies using MViTv2_base and MViTv2_small.

On average, the validation accuracy for MViTv2_base was 76.46%, whereas MViTv2_small achieved a higher average validation accuracy of 78.24%. In summary, our proposed single-modal models outperformed the baseline single-modal models, with one of our proposed single-modal models demonstrating superior performance over the other.

6.2. Performance of Single-Modal ST-GCN for Skeleton Dataset

Here, the model was trained using ST-GCN and subsequently evaluated the validation accuracies for each subject before averaging them. Experiments were conducted involving different numbers of frames for all subjects and then averaged the validation accuracies. Initially, only the last 50 alternate frames were utilized to test our ST-GCN model, and, subsequently, the middle 100 frames were selected to assess its performance. Next, the model's effectiveness was evaluated using the last 100 frames. Furthermore, the model's performance was examined with the minimum number of common frames (118). In the final step, our model was tested by employing all available frames (509).

Table 3 shows the subject-wise validation accuracies using various frame combinations for the single-modal ST-GCN model. When the last 50 frames were used, all the subjects except subjects 6 and 7 achieved an accuracy of 100%, and the average validation accuracy was 97.85%. When the middle 100 frames were used, the accuracy for all the subjects dropped a little except for subject 9, and the average validation accuracy was 93.76%. When the last 100 frames were used, the accuracy for all the subjects was 100% except for subjects 6 and 9. The average validation accuracy was 98.90%. There was an increase in accuracy from the middle 100 frames to the last 100 frames. Similarly, when the minimum number of common frames was considered, an accuracy of 100% for all the subjects except subjects 6 and 9 was achieved. The average validation accuracy was 99.04%. Again, there was an increase in accuracy for subjects 6 and 9 when the last 100 frames were chosen from a minimum number of common frames. Finally, there was a decrease in accuracy for subjects 6 and 9 when all the frames were chosen. The accuracy for all other subjects was 100%. Here, the average validation accuracy was 98.68%.

As utilizing the minimum number of common frames resulted in the highest average validation accuracy of 99.04% compared to other combinations, the minimum frame count of 118 was employed for ST-GCN in both the early fusion and late fusion (without and with different weights) methods.

Table 3. Subject-wise validation accuracies using different combinations for ST-GCN.

Subjects	ST-GCN (Last 50 Alternate Frames)	ST-GCN (Middle 100 Frames)	ST-GCN (Last 100 Frames)	ST-GCN (Min No. of Common Frames)	ST-GCN (All Frames)
Subject 1	100%	99.50%	100%	100%	100%
Subject 2	100%	93.67%	100%	100%	100%
Subject 3	100%	100%	100%	100%	100%
Subject 4	100%	97.67	100%	100%	100%
Subject 5	100%	95.83%	100%	100%	100%
Subject 6	89.50%	79.34%	90.50%	91.17%	87.29%
Subject 7	100%	97.00%	100%	100%	100%
Subject 8	100%	97.17%	100%	100%	100%
Subject 9	84.67%	91.67%	96.33%	97.33%	96.88%
Subject 10	100%	98.33%	100%	100%	100%
Subject 11	100%	80.83%	100%	100%	100%
Subject 12	100%	94.17%	100%	100%	100%
Average	97.85%	93.76%	98.90%	99.04%	98.68%

6.3. Performance of Multi-Modal Using Early Fusion

In this subsection, the features from both modalities were combined to determine the ultimate classification accuracy for all subjects. Table 4 illustrates the subject-wise validation accuracies for the multi-modal approach utilizing early fusion. Remarkably, 100% accuracy for all subjects except for subjects 6 and 9 was achieved. The overall average classification accuracy across all subjects was an impressive 99.86%. Our proposed multi-modal technique using early fusion outperformed the baseline for the multi-modal approach.

Table 4. Subject-wise validation accuracies of multi-modal using early fusion.

Subjects	Accuracy
Subject 1	100%
Subject 2	100%
Subject 3	100%
Subject 4	100%
Subject 5	100%
Subject 6	99.17%
Subject 7	100%
Subject 8	100%
Subject 9	99.17%
Subject 10	100%
Subject 11	100%
Subject 12	100%
Average	99.86%

6.4. Performance of Multi-Modal Using Late Fusion without and with Different Weights

As discussed in [40], various strategies have been proposed and modified for sensor fusion. Here, the previously trained single-modals, ST-GCN and MViTv2_small, which correspond to the foot pressure and skeleton data, respectively, were combined to evaluate the classification performance of our proposed multi-modal late fusion method, both with

and without applying different weights. The results were computed using the following procedures:

- A single vector of the same size by selecting the maximum values from the two input vectors (element-wise) was created.
- Two input vectors of the same size (element-wise) were averaged.
- Two input vectors of the same size (element-wise) were multiplied.

The validation accuracies for our multi-modal method using late fusion with and without different weights, and using the three methods described above, are presented in Table 5 and Figure 8. The accuracy using the maximum without weights is 96.95%, whereas the average accuracy with weights is 96.81%. When the outputs of two models were multiplied without weights, the accuracy was 96.67%. In the late fusion method with weight, the outputs with different weights were combined. A weight of 0.9 was assigned to the skeleton dataset due to its high accuracy and 0.1 to the foot pressure dataset due to its low accuracy. In the case of the maximum, an accuracy of 99.10% was achieved, whereas in the case of averaging, a 99.17% accuracy with different weights was achieved. Similarly, for multiplication, an accuracy of 96.67% was achieved. Our proposed multi-modal using late average fusion with different weights shows the highest performance among multi-modal late fusion methods.

Table 5. Validation accuracies of multi-modal using late fusion with and without different weights.

Different Fusion Methods	Accuracy without Weights	Accuracy with Different Weights
Maximum	**96.95%**	99.10%
Average	96.81%	**99.17%**
Multiplication	96.67%	96.67%

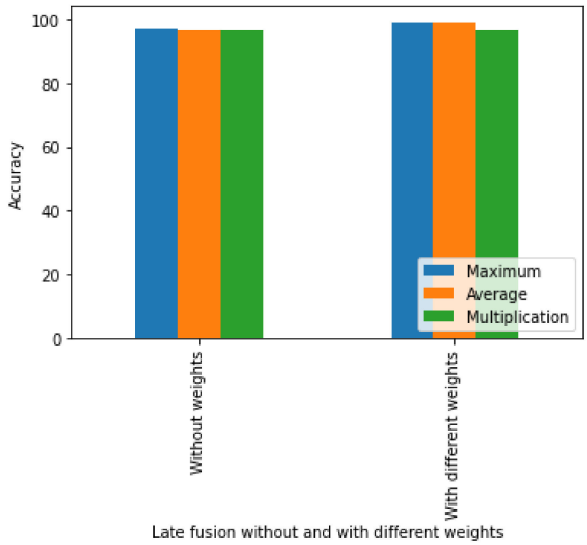

Figure 8. Validation accuracies of our multi-modal using late fusion without and with different weights.

A comparison of the suggested models with the baseline is presented in Table 6. Jun et al. [16] introduced single-modals using foot pressure and skeleton datasets, achieving accuracies of 68.82% and 93.40%, respectively. In contrast, our proposed single-modals achieved higher accuracies, reaching 78.2% for the foot pressure dataset and an impressive 99.04% for the skeleton dataset. In the realm of multi-modal approaches, Jun et al. reported an accuracy of 97.60%, whereas our proposed multi-modal utilizing early fusion attained an accuracy of 99.86%. Likewise, our proposed multi-modal using late fusion, both with and without different weights, demonstrated superior performance, achieving accuracies of

96.95% and 99.17%, respectively. Table 6 demonstrates that our multi-modal approach using early fusion outperforms the late fusion method. These results unequivocally establish that our proposed single-modals surpass those in [16], and our multi-modals using early fusion methods show state-of-the-art results in the pathological gait classification for the GIST dataset. The proposed approach can be applied in real-world scenarios with a simple camera setup. As individuals walk through, the system captures data and classifies specific gait patterns based on their types.

Table 6. Comparison and performance of the proposed model with the baseline.

Baseline	Accuracy (%)
Jun et al. [16] (Foot pressure data)	68.82%
Our (Foot pressure data)	78.2%
Jun et al. [16] (Skeleton data)	93.40%
Our (Skeleton data)	99.04%
Jun et al. [16] (Multi-modal)	97.60%
Our (Multi-modal using early fusion)	99.86%
Our (Multi-modal using late fusion without weights)	96.95%
Our (Multi-modal using late fusion with different weights)	99.17%

7. Conclusions and Future Directions

In this study, transformer-based single-modals were introduced as the first step to categorize one normal and five abnormal gaits (antalgic, lurch, steppage, stiff-legged, and Trendelenburg), following the application of 30-degree rotation augmentations to foot pressure data. Subsequently, another single-modal approach utilizing the skeleton dataset was presented, referred to as ST-GCN. Next, a multi-modal method employing early fusion was proposed, achieved by concatenating the features from both modalities. In the final step, a multi-modal approach employing late fusion, with and without different weights, was introduced to amalgamate the outputs from both modalities. Our early fusion multi-modal approach exhibited superior performance in comparison to late fusion. Our proposed models underwent a thorough comparison with the baseline, and the results undeniably demonstrate that our work surpassed both single-modal and multi-modal approaches. The findings of this study hold the potential to enhance existing gait analysis programs through the implementation of multimodality, which captures the most important information from all the modalities, thereby offering doctors and physicians more precise results in gait categorization.

In the future, for further study, collaboration with orthopedic, otolaryngology, and rehabilitation medical facilities to collect our datasets will be useful. By utilizing actual patient datasets, the suitability of the proposed hybrid model for real-world applications will be validated.

There are several other potential applications of gait classification. First, it can be used in gait monitoring for abnormal gait detection, the recognition of human activities, fall detection, and sports performance. Second, it can be used as gait-based biometrics with applications in person identification, authentication, and re-identification, as well as gender and race recognition. Third, it can be used as a smart gait device and in environments ranging from smart socks, shoes, and other wearables to smart homes and smart retail stores that incorporate continuous monitoring and control systems.

Author Contributions: Conceptualization, C.-S.L.; methodology and software, H.S. and N.-H.K.; data analysis, H.S., N.-H.K. and M.T.N.; writing—original draft preparation, M.T.N.; writing—review and editing, M.T.N. and C.-S.L.; visualization, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1A6A1A03040177).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://learn.microsoft.com/en-us/azure/kinect-dk/body-joints.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Myklebust, J.; Myklebust, B.M.; Prieto, T.; Kreis, D. Changes in motor function in the elderly: Gait, balance and joint compliance. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Orlando, FL, USA, 31 October–3 November 1991; Volume 13, pp. 863–864.
2. Ren, P.; Zhao, W.; Zhao, Z.; Bringas-Vega, M.L.; Valdes-Sosa, P.A.; Kendrick, K.M. Analysis of gait rhythm fluctuations for neurodegenerative diseases by phase synchronization and conditional entropy. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2015**, *24*, 291–299. [CrossRef]
3. Masood, H.; Farooq, H. A proposed framework for vision based gait biometric system against spoofing attacks. In Proceedings of the 2017 International Conference on communication, computing and digital systems (C-CODE), Islamabad, Pakistan, 8–9 March 2017; pp. 357–362.
4. Caldas, R.; Mundt, M.; Potthast, W.; de Lima Neto, F.B.; Markert, B. A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait Posture* **2017**, *57*, 204–210. [CrossRef] [PubMed]
5. Deschamps, K.; Matricali, G.A.; Roosen, P.; Desloovere, K.; Bruyninckx, H.; Spaepen, P.; Nobels, F.; Tits, J.; Flour, M.; Staes, F. Classification of forefoot plantar pressure distribution in persons with diabetes: A novel perspective for the mechanical management of diabetic foot? *PLoS ONE* **2013**, *8*, e79924. [CrossRef] [PubMed]
6. Seifallahi, M.; Soltanizadeh, H.; Mehraban, A.H.; Khamseh, F. Alzheimer's disease detection using skeleton data recorded with Kinect camera. *Clust. Comput.* **2020**, *23*, 1469–1481. [CrossRef]
7. Nguyen, T.N.; Huynh, H.H.; Meunier, J. Skeleton-based abnormal gait detection. *Sensors* **2016**, *16*, 1792. [CrossRef]
8. Mannini, A.; Trojaniello, D.; Cereatti, A.; Sabatini, A.M. A machine learning framework for gait classification using inertial sensors: Application to elderly, post-stroke and huntington's disease patients. *Sensors* **2016**, *16*, 134. [CrossRef] [PubMed]
9. Hsu, W.-C.; Sugiarto, T.; Lin, Y.-J.; Yang, F.-C.; Lin, Z.-Y.; Sun, C.-T.; Hsu, C.-L.; Chou, K.-N. Multiple-wearable-sensor-based gait classification and analysis in patients with neurological disorders. *Sensors* **2018**, *18*, 3397. [CrossRef] [PubMed]
10. Waldecker, U. Pedographic classification and ulcer detection in the diabetic foot. *Foot Ankle Surg.* **2012**, *18*, 42–49. [CrossRef]
11. Alharthi, A.S.; Casson, A.J.; Ozanyan, K.B. Gait spatiotemporal signal analysis for Parkinson's disease detection and severity rating. *IEEE Sens. J.* **2020**, *21*, 1838–1848. [CrossRef]
12. Li, Q.; Wang, Y.; Sharf, A.; Cao, Y.; Tu, C.; Chen, B.; Yu, S. Classification of gait anomalies from Kinect. *Vis. Comput.* **2018**, *34*, 229–241. [CrossRef]
13. Rana, S.P.; Dey, M.; Ghavami, M.; Dudley, S. Markerless gait classification employing 3D IR-UWB physiological motion sensing. *IEEE Sens. J.* **2022**, *22*, 6931–6941. [CrossRef]
14. Wang, C.; Ning, X.; Li, W.; Bai, X.; Gao, X. 3D Person Re-identification Based on Global Semantic Guidance and Local Feature Aggregation. *IEEE Trans. Circuits Syst. Video Technol.* **2023**. [CrossRef]
15. Wang, C.; Wang, H.; Ming, X.; Tian, S.; Li, W. 3D Point Cloud Classification Method Based on Dynamic Coverage of Local Area. *J. Softw.* **2022**, *34*, 1962–1976.
16. Jun, K.; Lee, S.; Lee, D.-W.; Kim, M.S. Deep learning-based multimodal abnormal gait classification using a 3D skeleton and plantar foot pressure. *IEEE Access* **2021**, *9*, 161576–161589. [CrossRef]
17. Kim, J.K.; Bae, M.N.; Lee, K.; Kim, J.C.; Hong, S.G. Explainable Artificial Intelligence and Wearable Sensor-Based Gait Analysis to Identify Patients with Osteopenia and Sarcopenia in Daily Life. *Biosensors* **2022**, *12*, 167. [CrossRef]
18. Youn, I.-H.; Won, K.; Youn, J.-H.; Scheffler, J. Wearable sensor-based biometric gait classification algorithm using WEKA. *J. Inf. Commun. Converg. Eng.* **2016**, *14*, 45–50. [CrossRef]
19. Kim, Y.K.; Visscher, R.M.; Viehweger, E.; Singh, N.B.; Taylor, W.R.; Vogl, F. A deep-learning approach for automatically detecting gait-events based on foot- marker kinematics in children with cerebral palsy—Which markers work best for which gait patterns? *PLoS ONE* **2022**, *17*, e0275878. [CrossRef]
20. Conte Alcaraz, J.; Moghaddamnia, S.; Peissig, J. Efficiency of deep neural networks for joint angle modeling in digital gait assessment. *EURASIP J. Adv. Signal Process.* **2021**, *2021*, 10. [CrossRef]
21. Hassan, S.; Wang, X.; Ishtiaq, S. Human Gait Classification Based on Convolutional Neural Network using Interferometric Radar. In Proceedings of the 2021 International Conference on Control, Automation and Information Sciences (ICCAIS), Xi'an, China, 14–17 October 2021; pp. 450–456.
22. Lee, S.-S.; Choi, S.T.; Choi, S.-I. Classification of gait type based on deep learning using various sensors with smart insole. *Sensors* **2019**, *19*, 1757. [CrossRef]

23. Chen, I.-M.; Yeh, P.-Y.; Chang, T.-C.; Hsieh, Y.-C.; Chin, C.-L. Sarcopenia Recognition System Combined with Electromyography and Gait Obtained by the Multiple Sensor Module and Deep Learning Algorithm. *Sens. Mater.* **2022**, *34*, 2403–2425. [CrossRef]
24. Pandey, C.; Roy, D.S.; Poonia, R.C.; Altameem, A.; Nayak, S.R.; Verma, A.; Saudagar AK, J. GaitRec-Net: A deep neural network for gait disorder detection using ground reaction force. *PPAR Res.* **2022**, *2022*, 9355015. [CrossRef]
25. Naji, A.; A Abboud, S.; A Jumaa, B.; Abdullah, M.N. Gait Classification Using Machine Learning for Foot Disseises Diagnosis. *Tech. Rom. J. Appl. Sci. Technol.* **2022**, *4*, 37–49. [CrossRef]
26. Jani, D.; Varadarajan, V.; Parmar, R.; Bohara, M.H.; Garg, D.; Ganatra, A.; Kotecha, K. An Efficient Gait Abnormality Detection Method Based on Classification. *J. Sens. Actuator Netw.* **2022**, *11*, 31. [CrossRef]
27. Kim, J.-K.; Bae, M.-N.; Lee, K.B.; Hong, S.G. Identification of patients with sarcopenia using gait parameters based on inertial sensors. *Sensors* **2021**, *21*, 1786. [CrossRef] [PubMed]
28. Chen, B.; Chen, C.; Hu, J.; Sayeed, Z.; Qi, J.; Darwiche, H.F.; Little, B.E.; Lou, S.; Darwish, M.; Foote, C.; et al. Computer Vision and Machine Learning-Based Gait Pattern Recognition for Flat Fall Prediction. *Sensors* **2022**, *22*, 7960. [CrossRef] [PubMed]
29. Wolf, T.; Babaee, M.; Rigoll, G. Multi-view gait recognition using 3D convolutional neural networks. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 4165–4169.
30. Wen, J. Gait recognition based on GF-CNN and metric learning. *J. Inf. Process. Syst.* **2020**, *16*, 1105–1112.
31. Liu, W.; Zhang, C.; Ma, H.; Li, S. Learning efficient spatial-temporal gait features with deep learning for human identification. *Neuroinformatics* **2018**, *16*, 457–471. [CrossRef] [PubMed]
32. Jun, K.; Lee, Y.; Lee, S.; Lee, D.-W.; Kim, M.S. Pathological gait classification using kinect v2 and gated recurrent neural networks. *IEEE Access* **2020**, *8*, 139881–139891. [CrossRef]
33. Guo, Y.; Deligianni, F.; Gu, X.; Yang, G.-Z. 3-D canonical pose estimation and abnormal gait recognition with a single RGB-D camera. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3617–3624. [CrossRef]
34. Jun, K.; Lee, D.-W.; Lee, K.; Lee, S.; Kim, M.S. Feature extraction using an RNN autoencoder for skeleton-based abnormal gait recognition. *IEEE Access* **2020**, *8*, 19196–19207. [CrossRef]
35. Kwon, S.B.; Han, H.S.; Lee, M.C.; Kim, H.C.; Ku, Y. Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images. *IEEE Access* **2020**, *8*, 120597–120603. [CrossRef]
36. Shah, S.A.; Tahir, A.; Ahmad, J.; Zahid, A.; Pervaiz, H.; Shah, S.Y.; Ashleibta, A.M.A.; Hasanali, A.; Khattak, S.; Abbasi, Q.H. Sensor fusion for identification of freezing of gait episodes using Wi-Fi and radar imaging. *IEEE Sens. J.* **2020**, *20*, 14410–14422. [CrossRef]
37. Yunas, S.U.; Alharthi, A.; Ozanyan, K.B. Multi-modality sensor fusion for gait classification using deep learning. In Proceedings of the 2020 IEEE Sensors Applications Symposium (SAS), Kuala Lumpur, Malaysia, 9–11 March 2020; pp. 1–6.
38. Weng, Y.; Pan, Z.; Han, M.; Chang, X.; Zhuang, B. An efficient spatio-temporal pyramid transformer for action detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 358–375.
39. Li, Y.; Wu, C.Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; Feichtenhofer, C. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4804–4814.
40. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 183–192.
41. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Trans. Image Process.* **2020**, *29*, 9532–9545. [CrossRef] [PubMed]
42. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. No. 1.
43. Kim, J.; Seo, H.; Naseem, M.T.; Lee, C.-S. Pathological-Gait Recognition Using Spatiotemporal Graph Convolutional Networks and Attention Model. *Sensors* **2022**, *22*, 4863. [CrossRef]
44. Available online: https://learn.microsoft.com/en-us/azure/kinect-dk/body-joints (accessed on 12 April 2023).
45. Srivastava, N.; Salakhutdinov, R.R. Multimodal learning with deep Boltzmann machines. *Adv. Neural Inf. Process. Syst.* **2012**, *25*.
46. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, DC, USA, 28 June–2 July 2011; pp. 689–696.