Multimodal Fall Detection Using Spatial-Temporal Attention and Bi-LSTM Based Feature Fusion

Jungpil Shin *, Abu Saleh Musa Miah *, Rei Egawa, Najmul Hassan, Koki Hirooka, Yoichi Tomioka

Posted Date: 31 March 2025

doi: 10.20944/preprints202503.2247.v1

Keywords: Human Fall Detection; Graph Convolutional Network (GCN); Multimodel; Body Pose Detection; Alpha Pose; Channel Attention; Ageing People

 doi:10.20944/preprints202503.2247.v1

Article

Multimodal Fall Detection Using Spatial-Temporal Attention and Bi-LSTM Based Feature Fusion

Jungpil Shin *, Abu Saleh Musa Miah *, Rei Egawa, Najmul Hassan, Koki Hirooka and Yoichi Tomioka

School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Fukushima 965-8580, Japan
* Correspondence: jpshin@u-aizu.ac.jp (J.S.); abusalehcse.ru@gmail.com (A.S.M.M.)

Abstract: Human fall detection is a significant healthcare concern, particularly among the elderly, due to its links to muscle weakness, cardiovascular issues, and locomotive syndrome. Accurate fall detection is crucial for timely intervention and injury prevention, which has led many researchers to work on developing effective detection systems. However, existing unimodal systems that rely solely on skeleton or sensor data face challenges such as poor robustness, computational inefficiency, and sensitivity to environmental conditions. While some multimodal approaches have been proposed, they often struggle to capture long-range dependencies effectively. In order to address these challenges, we propose a multimodal fall detection framework that integrates skeleton and sensor data. The system uses a Graph-based Spatial-Temporal Convolutional and Attention Neural Network (GSTCAN) to capture spatial and temporal relationships from skeleton and motion data information in stream-1, while a Bi-LSTM with Channel Attention (CA) processes sensor data in stream-2, extracting both spatial and temporal features. The GSTCAN model uses AlphaPose for skeleton extraction, calculates motion between consecutive frames, and applies a graph convolutional network (GCN) with a CA mechanism to focus on relevant features while suppressing noise. In parallel, the Bi-LSTM with CA processes inertial signals, with Bi-LSTM capturing long-range temporal dependencies and CA refining feature representations. The features from both branches are fused and passed through a fully connected layer for classification, providing a comprehensive understanding of human motion. The proposed system was evaluated on the Fall Up dataset, achieving a classification accuracy of 99.09% , surpassing existing methods. This robust and efficient system demonstrates strong potential for accurate fall detection and continuous healthcare monitoring.

Keywords: ageing people; alpha pose; body pose detection; channel attention; graph convolutional network (GCN); human fall detection; multimodal

1. Introduction

Response file https://docs.google.com/document/d/18aZC_snL_33gO0austOT73Nd0_Ya10k_/edit A human fall can be caused by slipping on an aisle, floor, or other surfaces or stepping off a step, projection, or floor [1]. Falls are one of the most common risks that older adults are faced with in their daily living, and living on their own may raise the possibility of death following a fall. Regarding how many older people fall, it has been reported that approximately 20% of people aged 65 and over who live at home and more than 30% of those who live in facilities fall in a year. By gender, women have a higher incidence of falls than men, and the rate increases with age [2]. Additionally, individuals with visual, neurological, orthopedic, and gait disorders are more susceptible to falls due to reduced reflexes and impaired gait [3]. Human falls are considered the second most common cause of unintentional injuries and death worldwide, as reported by the World Health Organization (WHO) [4]. The world's elderly population is expected to double (to about 2.1 billion people) over the next 30 years [5]. In Japan, where the population is ageing rapidly, there are calls for measures to prevent falls by the elderly. Data shows that if appropriate medical treatment is received immediately after a fall, the risk of death is

 doi:10.20944/preprints202503.2247.v1

reduced by 80%, and 26% reduces the length of hospital stay [6]. Systems that automatically detect falls are becoming increasingly important to protect the lives of elderly people. There are two main types of fall detection systems: sensor-based and camera-based. Sensor-based systems use devices like accelerometers, gyroscopes, and magnetometers to collect movement data [7]. These devices have become more affordable and lightweight due to advancements in computing chip technology. However, sensor data has the problem of being unable to handle complex operations due to signal and background noise, leading to false detections [8]. Camera-based systems have the advantage of being able to collect much more information than sensor-based systems. Human behavior is detected by extracting features from video frames [9,10]. More recently, Ha et al. developed a machine learning and CNN-based multimodal fall detection system by combining sensor and camera video-based dataset modalities [11]. However, camera-based detection systems suffer from privacy leakage and limitations due to lighting conditions. In addition, skeleton-based systems also face challenges in achieving satisfactory performance due to interclass similarity and lack of effective features [12]. To overcome the problem recently, researchers have focused on combining the features from the multi-modal dataset to increase the performance accuracy of the fall detection system [13–17]. However, their system still faces challenges to achieving satisfactory performance accuracy and efficiency for deployment as an accurate fall detection system due to the lack of effective long-range and short-range relationships among the data pattern features. To overcome the challenges in traditional fall detection methods, we propose a multimodal fall recognition system that combines skeleton and sensor data. This approach leverages accelerometer and skeletal data to address key issues such as robustness, computational efficiency, and recall to environmental factors. The key contributions of this work are as follows:

- **Multimodal Data Integration:** We propose a novel multimodal fall detection framework that integrates both skeleton and sensor data. This approach combines the strengths of both data modalities, addressing the limitations of unimodal systems and improving robustness, computational efficiency, and adaptability to different environments.
- **Dual-Stream Architecture:** The framework uses a Graph-based Spatial-Temporal Convolutional and Attention Neural Network (GSTCAN) to capture spatial and temporal relationships from skeleton and motion data. For sensor data, the system employs a Bi-LSTM integrated with CA. The Bi-LSTM captures long-range temporal dependencies, while the CA mechanism refines feature representations. This integration enhances feature extraction by capturing both spatial and temporal information and improving the model's sensitivity to important features.
- **Feature Fusion for Improved Classification:** The features extracted from the GSTCAN for skeleton and motion data, as well as Bi-LSTM-CA branches for sensor data, are fused and passed through a fully connected layer for classification. This fusion allows the system to leverage complementary information from both streams, improving the overall understanding of human motion and increasing fall detection accuracy.
- **State-of-the-Art Performance:** The proposed system was rigorously evaluated on the Fall Up dataset, achieving a classification accuracy of 99.09%, significantly outperforming existing methods. This demonstrates the system's robust performance and its potential for real-time fall detection and continuous healthcare monitoring.

The remainder of the paper is divided into sections. Section 2 describes the literature review, Section 3 describes the datasets used, Section 4 details the proposed system, and Section 5 details experimental results and discussion. Finally, the paper concludes with a conclusion 6 and future directions.

2. Related Work

In recent years, there has been significant progress in developing human fall detection systems. The previous systems developed can be categorized into inertial sensors and vision-based systems. We will provide details on each research.

 doi:10.20944/preprints202503.2247.v1

2.1. Inertial Sensor-Based Fall Detection Systems

Sensor-based systems use wearable sensors such as accelerometers, gyroscopes, EMG, and EEG to record features such as angles and distances [18]. Among them, Angela et al. developed a non-linear classification and Kalman filter-based human fall detection system where they reported 99.4% accuracy with the Sisfall dataset [19,20]. Desai et al. also developed a fall detection system with high accuracy within short times 0.25 seconds [21]. Faisal et al. tested k-nearest neighbors (KNN), support vector machine (SVM), and random forest (RF) on the Sisfall dataset, achieving accuracies of 99.8%, 99.78%, and 99.56%, respectively [22]. Another author combined federated learning with an extreme learning machine to develop a human fall recognition system where they showed 99.07% accuracy in experiments with only elderly subjects. In the same way, Vanilson et al. [23] developed a human fall detection system aiming to handle a variety of real-world scenarios where they evaluated the model with a combination of three datasets: UMA fall [24], WEDA fall [25], and UP fall [26]. They achieved a recall rate of 90.57% by combining three datasets. Nader et al. developed a smartwatch-based real-time fall detection system [27]. The system achieved an F-score of 85% using an LSTM model and transfer learning.

2.2. Video-Based Fall Detection System

Typically, human silhouettes are segmented from videos captured by RGB cameras or Kinect, and features are extracted for fall detection. Common approaches are based on traditional ML, such as SVM and CNN [28–30] models. Proposed fall detection using the ratio of the height and base of the triangle formed by the head and two feet [31]. Qing et al. proposed a two-stream approach using MobileVGG [32]. This model adopts human motion features. Although such ML algorithms can achieve good accuracy in some cases, it is difficult to achieve it efficiently on large-scale and dynamic video datasets. The recent development of DL techniques can also be used in fall detection. Xiaogang et al. directly applied a convolutional neural network (CNN) to each frame image in the video to perform fall detection [33]. Na et al. proposed a fall detection method based on a three-dimensional (3-D) CNN-based fall detection method [34]. 3-D CNN can extract motion features of time sequences that are important for falling detection and combine 3-D CNN with LSTM. Proposed fall detection employing a multi-stream CNN based on a multi-level image fusion approach [35]. It determines the difference in motion within 16 frames of the input video. It achieved an accuracy of 99.03% on the Le2i dataset. Traditional image and video systems have difficulty in effectively separating the foreground from the unwanted background, which can reduce the effectiveness of the entire system. In addition, the enormous amount of computation required a lot of energy and memory. Recently, more and more researchers have adopted skeleton-based systems for this task. In skeleton-based approaches, a person's pose is estimated from an RGB image, and features are extracted from the skeletal joints [36–40]. Sheldon et al. proposed a transformer model that estimates the pose of a person on a video image and employs transfer learning [40]. Leiyue et al. adopted the torso angle as a new feature for recognizing falls and adopted SVM to determine whether the action was a fall [37]. Features are calculated by extracting the human skeleton from the depth image, and this method reported 93.56% accuracy with the TSTV2 dataset. Another researcher has worked to develop a real-time fall detection system using a short-time Fourier transform (STFT) and a 1D-CNN, achieving accuracies of 91% on MCFD, 99% on UR Fall, and 98% on NTU RGB+D datasets [40]. Another researcher identified 15 key points from the Alphapose network and divided it into five segments [41]. A new feature descriptor can be generated by extracting the segmented parts' distance, angle, and tilt angle. They achieved an accuracy of 98.32% using the UPfall dataset. Proposed a fall detection model that combines an SVM and a KNN model based on important spatial features of the foreground image [42]. They select keyframes from the height-to-width ratio displacement and the center of gravity displacement in the horizontal and vertical directions. The proposed fall detection system obtained a maximum peak accuracy of 98.6% and a recall of 100% in detecting falls. Many approaches that employ skeleton data convert joint points into metric data such as distance and angle. During this conversion process,

 doi:10.20944/preprints202503.2247.v1

the original joint position data may be lost. This loss of data creates the risk of missing important information in motion analysis and recognition models. To effectively tackle the challenge above and simultaneously enhance the precision of the outcomes, Yan and his colleagues undertook the development of an innovative DL methodology, which they have designated as the spatio-temporal Graph Convolutional Network (ST-GCN), a significant contribution to the field of artificial intelligence and ML, as referenced in their research [36]. ST-GCN treats skeleton data (joint point positions) as a graph structure. It also comprehensively models the changes in motion over time rather than just looking at the joint positions. Proposed an anomaly identification framework for fall detection [43]. They incorporate ST-GCN as a feature trainer and use the reconstruction error of the encoder to recognize falls as abnormal. Sania et al. treated the skeleton stream and the motion stream separately, extracted features using CNN for each, concatenated the extracted features into a single feature vector and used GCN and ST-GCN for fall detection [44]. And achieved an average accuracy of over 94% using the URFD and UPFD datasets. To improve the accuracy, Egawa et al. adopted a GCN-based spatiotemporal feature extraction and classification model for fall detection and evaluated the model using ImViA, UR-fall, and FDD datasets, achieving 99.00% accuracy [45].

2.3. Using Multimodal Features Fall Detection System

Multimodal fall detection systems combine different data modality-based features to address individual limitations. Dina et al. utilized CNN for spatial information from images and LSTM for temporal information from signals, achieving 98.31% accuracy on the UP-Fall detection dataset, with a 10% improvement in other metrics [46]. Milon et al. employed a CNN for visual patterns from images and a ConvLSTM for time-series sensor data, achieving 97.61% accuracy on the UP-Fall detection dataset for classifying 11 activities [47]. Hafeez et al. proposed a system that combined a hybrid descriptor to recognize human activities [48] by removing noise from sensor signals and video frames, extracting movement and skeletal model features, and using logistic regression for classification. Their system achieved 91.51%, 92.98%, and 90.23% success rates on the UP fall, UR, and Sisfall datasets, respectively. This paper proposes a multimodal fall detection system that integrates the GSTCAN model for skeleton and motion data with the Bi-LSTM-CA model for sensor data to enhance detection accuracy. The system is evaluated on a single dataset.

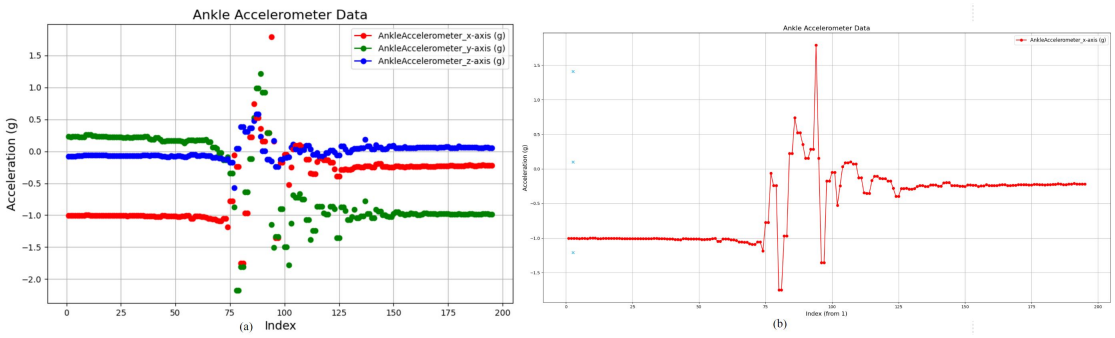

Figure 1. Visualization of the hand-gesture sensor-based signal.

3. Datasets

In the study, our goal is to develop a multimodal dataset-based fall detection system where we find a multimodal-based fall detection system, namely the SiS Fall, UP-Fall [26] and UR Fall Dataset [49]. However, in the study we used UP-Fall and UR Fall.

3.1. Fall UP Dataset

The UP-Fall dataset [26] is utilized in this study. This dataset represents a substantial resource for fall detection and activity recognition. It comprises a total of 11 distinct activities, each performed three times. The data were collected from a group of 17 healthy young adults (9 males and 8 females) aged 18 to 24 years, with an average height of 1.66 meters and an average weight of 66.8 kg. The dataset

 doi:10.20944/preprints202503.2247.v1

includes 11 activities, covering six daily movements (walking, standing, sitting, picking up an object, jumping, and lying down) and five types of falls (falling forward with hands, falling forward on knees, falling backward, falling sideways, and falling while sitting on an empty chair), as shown in the Table 1 . These specific activities were selected based on existing literature [50,51]. Numerous activities that do not involve falling were included due to their prevalence in daily routines, with picking up objects being particularly prone to misinterpretation as a fall. Jumping was executed at 30-second intervals, picking up objects every 10 seconds, and all other daily activities were conducted every 60 seconds. Each falling scenario was enacted for 10 seconds.

Table 1. Activity ID and Description on UP-Fall dataset.

Class No	Class Description	Class No	Class Description
1	Falling forward using knees	7	Standing
2	Falling forward using hands	8	Sitting
3	Falling backward	9	Picking up an object
4	Falling sideward	10	Jumping
5	Falling sitting in an empty chair	11	Laying
6	Walking		

Because this is a multimodal data set, the wearable sensor, context-aware sensor, and camera are used simultaneously to collect data. The wearable sensor collects data from a 3-axis accelerometer signal visualization shows in Figure 1, 3-axis gyroscope, and ambient light sensor. The wearable sensor is positioned in five different locations: on the left wrist, under the neck, inside the right pants pocket, at the center of the waist (on the belt), and on the left ankle, as shown in Figure 2. The reason for attaching it to the left wrist and pocket is to simulate how people normally wear a smartwatch or carry a cell phone. The subjects also wore an electroencephalograph (EEG) headset to acquire EEG signals. Context-aware sensors and cameras were installed, as shown in the figure. Two cameras were installed at a height of 1.82 m from the floor, one for the side view and the other for the front view. In our study, the only sensor data used was acceleration. This data set does not have data for the right pockets of Subjects 5 and 9, so our study experimented with data from Subjects 1 17, excluding Subjects 5 and 9. Figure 2 below visualizes the acceleration sensor position of the data used in this study. This figure shows the wrist acceleration data of Subject1 Activity1 Trial1 camera1.

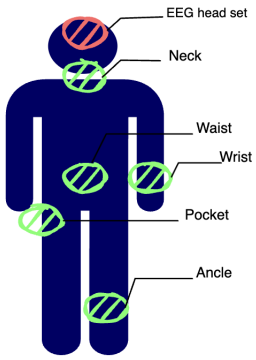

Figure 2. Wearable sensor located at human body.

3.2. UR-Fall Dataset

UR-Fall is a multimodal dataset of RGB, depth, and sensor data modalities [49]. The RGB data includes 70 sequences of 30 fall events and 40 activities of daily living (ADL). Fall events are captured using two Microsoft Kinect cameras along with accelerometric data, while ADL events are recorded with only one camera (camera 0) and an accelerometer. Sensor data is collected using PS Move (60Hz)

 doi:10.20944/preprints202503.2247.v1

and x-IMU (256Hz) devices. The dataset is organized such that each row contains a sequence of depth and RGB images from both cameras (camera 0 is parallel to the floor, and camera 1 is ceiling-mounted), synchronization data, and raw accelerometer data.

4. Proposed Methodology

Many researchers have developed systems for detecting falls using ML and DL. However, there is limited work on using multimodal datasets. This study proposes a novel multimodal fall detection system combining skeleton and sensor data. The skeleton data is processed using a GSTCAN [45], while the sensor data is processed using an integration of Bi-LSTM with the CA network. By combining both types of data, the system improves the accuracy and reliability of fall detection. The structure of the proposed system is illustrated in Figure 3.

The GSTCAN model works directly on the raw skeleton and motion data without the need for dimensionality reduction. This allows the system to capture both the spatial and temporal patterns in the data [45]. This approach, inspired by recent work in human activity recognition and skeleton-based fall detection [52–54], helps the system learn detailed motion patterns directly from the raw data. The novelty of the skeleton and stream data lies in its ability to avoid dimensionality reduction and learn complex movement patterns from the raw data, which makes it more efficient in capturing detailed and meaningful motion patterns. For sensor data, we used the Bi-LSTM to extract sensor features, followed by CA [55] to refine them. This integration enhances the system's ability to understand time-varying human motion characteristics. The novelty of our sensor stream module lies in combining the Bi-LSTM's ability to capture long-range and temporal dependencies with the feature refinement of CA. This approach improves feature extraction and strengthens the model's performance in dynamic and noisy environments, making it particularly effective for detecting events such as falls, where both temporal context and sensor channel importance are crucial.

The features from both data streams (skeleton, motion, and sensor) are then fused, combining complementary information from each modality. This fusion of features provides a more comprehensive and accurate understanding of human motion. After fusion, the combined features are passed through a fully connected layer for final classification. The fusion approach helps overcome the limitations of unimodal systems by providing richer, more reliable feature representations. In this study, we used video data from UP fall incidents to extract skeletal points for each frame, producing 18 points, including the nose, mouth, shoulders, elbows, wrists, hips, knees, and ankles (as shown in Figure 5). However, we selected 14 points by omitting the eyes and ears from both sides.

 doi:10.20944/preprints202503.2247.v1

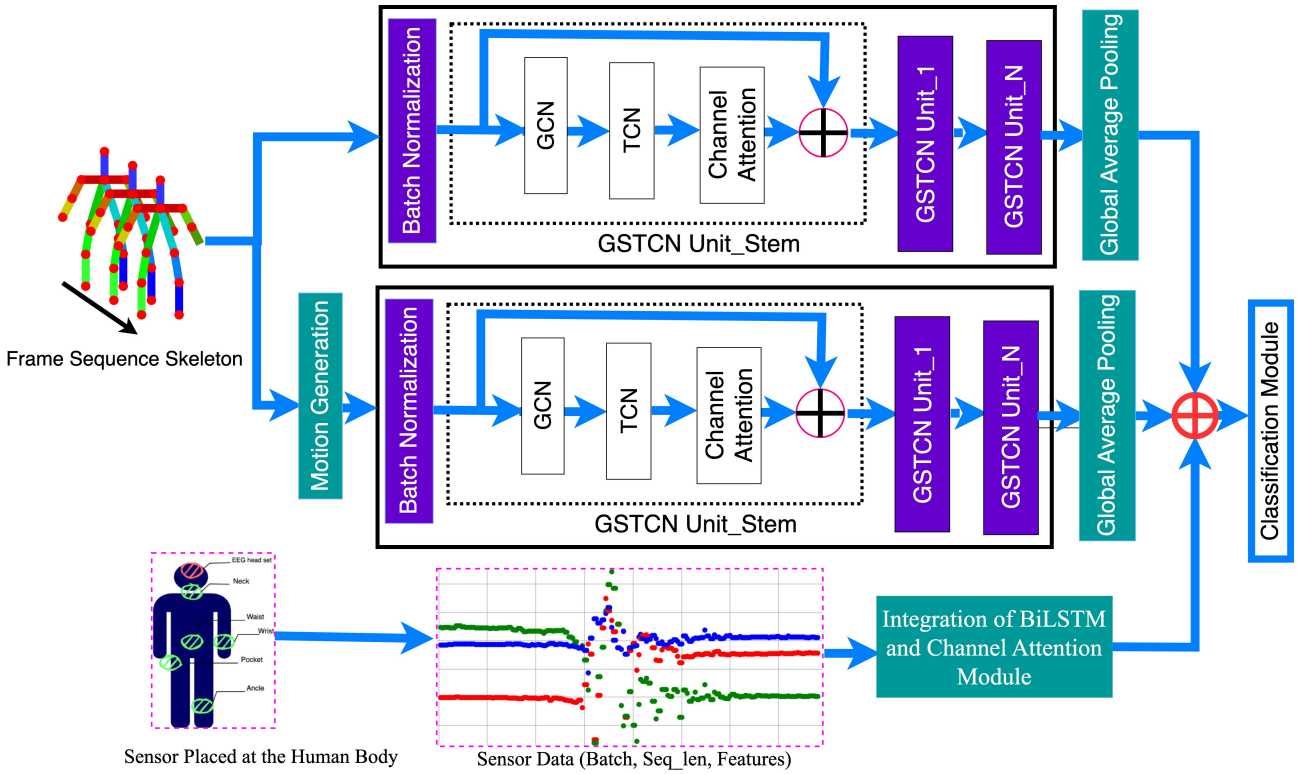

Figure 3. Proposed Working Flow Diagram. [45]

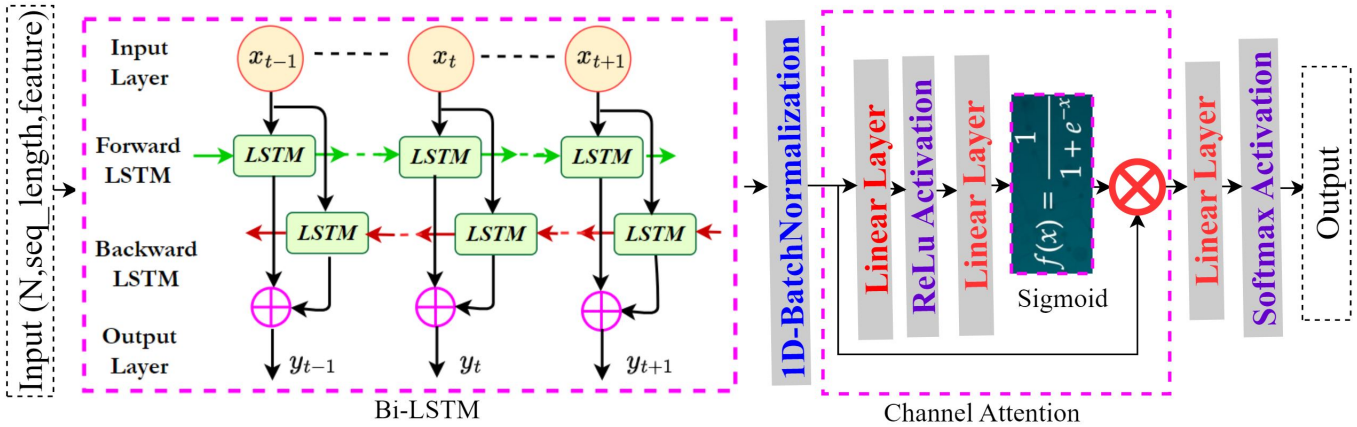

Figure 4. Integration of the CNN with Bi-LSTM model for sensor data modality feature extraction.

4.1. Stream-1 Skeleton-Based GCN

This stream employs GSTCAN, a convolutional approach grounded in graph theory that integrates spatiotemporal characteristics alongside an attention mechanism to extract features from the pose information. This framework accepts skeletal data as input, concurrently analyzes both temporal and spatial dimensions, and integrates an attention mechanism [56] to facilitate the dynamic extraction of salient features for ST-GCN, which conceptualizes data within a graph-based architecture.

4.2. Data Extraction Using Alphapose

AlphaPose [57] was used to extract skeletal data from the video here. AlphaPose identifies the human bounding box and evaluates the individual's posture encapsulated within that box. The accompanying figure illustrates the number of frames for AlphaPose, successfully acquired skeletal information. In the present study, frames from which skeletal data could not be retrieved were omitted; we elected to gather 14 key points utilizing AlphaPose, including the nose, shoulders, elbows, hips, buttocks, knees, and ankles, as depicted in Figure 5.

 doi:10.20944/preprints202503.2247.v1

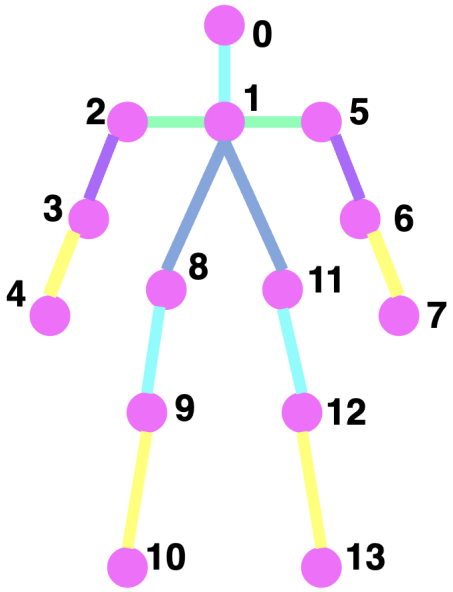

Figure 5. Human joint positions obtained with Alpha Pose.

4.2.1. Motion Calculation and Graph Construction

In this study, we focus on the **dynamic fall detection dataset**, where motion is a critical characteristic for an effective fall detection framework, particularly in terms of movement, alignment, and data structure efficiency. Motion directly influences the dynamics inherent in fall-related data. To quantify motion, we utilize all landmark positions for x and y coordinates, represented as a two-dimensional vector. The motion for a given joint is computed by differentiating the joint positions between consecutive frames:

$$X_{motion} = M_j^t = J_j^t - J_j^{t-1} \tag{1}$$

where, M_j^t represents the motion of joint j at time t, $J_j^t = (x_j^t, y_j^t)$ is the position of joint j at frame t, $J_j^{t-1} = (x_j^{t-1}, y_j^{t-1})$ is the position of the same joint in the previous frame. The X_{motion} represent the motion dataset here, which serves as a structured representation of human movement, capturing the two-dimensional coordinates of various joints over time. The classification of fall and non-fall events is based on analyzing multiple frames, utilizing the sequential arrangement of relative joint positions. To model spatiotemporal relationships, we construct a **graph-based representation** that integrates both spatial and temporal domains while considering the anatomical relationships among joints. The primary graph structure, which is initially underconnected, is formulated as follows:

$$G = (V, E) \tag{2}$$

Here, V and E denote the nodes and edges of the graph. Then the node can written as $V = v_{(i,t)} \mid i = 1, \ldots, N, t = 1, \ldots, T$, that considered the whole body skeleton as a graph. Consequently, the adjacent matrix can be defined as in Equation (3).

$$f(x) = \begin{cases} 0 & \text{if they are not adjacent} \\ 1 & \text{if the nodes are adjacent.} \end{cases} \tag{3}$$

4.2.2. Graph Convolutional Network (GCN)

Our research work involved the systematic extraction and rigorous analysis of the latent capabilities inherent within the complete human skeletal structure, utilizing a robust framework established on the foundational principles of the spatial-temporal graph convolutional network. To accurately

 doi:10.20944/preprints202503.2247.v1

represent the intricate relationships within this graph structure, we systematically constructed it utilizing the following mathematical formulations as delineated in previous scholarly works [52,58]:

$$G_{out} = D^{-(1/2)}(A + I)D^{-(1/2)} \times W, \tag{4}$$

In this context, the symbols D, I, and A correspond to the diagonal matrix that encompasses the degree of nodes, the identity matrix that signifies self-connections among the vertices, and the adjacency matrix that delineates connections between different bodies, respectively. It is important to note that the diagonal degree can be represented by the expression (A+I), while the weight matrix that conveys the significance of the connections is denoted by the symbol W. To implement the graph-based convolutional operations, our scholarly focus was directed towards the utilization of two-dimensional convolutional techniques, whereby we engaged in multiplication with the matrix $D^{-(1/2)}(A + I)D^{-(1/2)}$ specifically for the spatial graph convolution. Comparably, for the execution of the graph-oriented temporal convolution, we conducted multiplicative operations employing a kernel dimension of $k_t \times 1$.

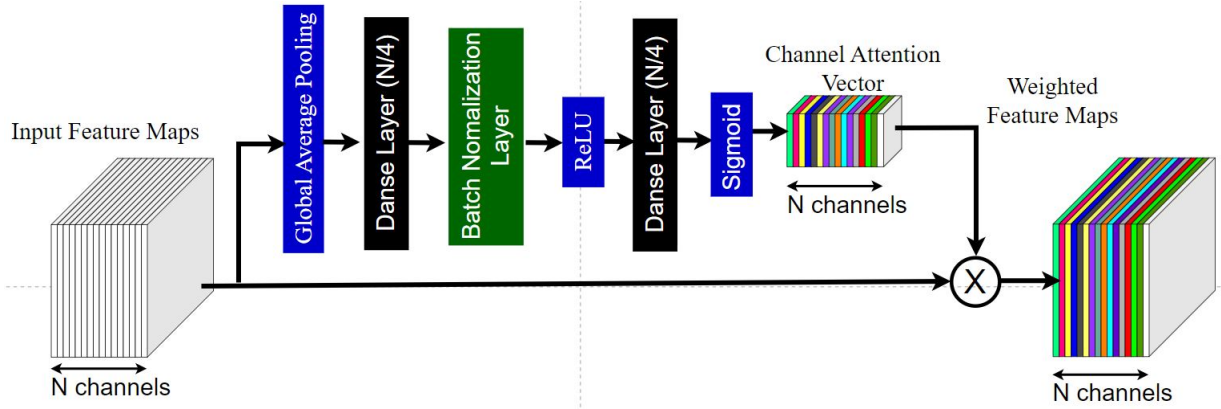

Figure 6. Visualization of Channel attention mechanism [56,59,60].

4.3. Skeleton Feature Using GSTCAN

The GSTCAN model executes action recognition grounded in a Graph Convolutional Network (GCN) framework, which autonomously identifies spatial and temporal patterns through data procured via AlphaPose as its input. The joints of the human skeleton are used as nodes of the graph, and the human body structures and their connections in time are used as edges of the graph. GSTCAN is mainly composed of a GSTCAN unit and a pooling layer. The GCTCAN unit is constructed by the convolution of the spatial and temporal graph and is constructed by the CA module. The inputs are of the form (N, C, T, V), where N is the batch size, C is the number of channels, T is the sequence length, and V is the number of nodes. Spatial convolution extracts each joint's features based on the relationship between nodes. The distance connections between nodes are classified into three categories: connections between nodes with the same distance from the center of gravity, connections between nodes in the direction closer to the center of gravity, and connections between nodes in the direction away from the center of gravity. Temporal convolution captures changes in the same joint over time. For example, it learns how the elbow position moves in the far frame. This GSTCAN unit is applied 6 times to the input data to generate higher-order feature maps on the graph. The number of channels is 64 for the first 2 layers, 128 for the next two layers, and 256 for the last two layers. A channel-attention mechanism is added at the end of each unit [58–63]. The addition of the CA mechanism distinguishes between "important" and "unimportant" learned features, allowing us to focus on more important features. It enables more accurate behaviour recognition. Figure 6 shows the CA approach's working architecture.

The CA mechanism is constructed utilizing the following components: (1) GlobalAveragePooling, (2) Dense (N/4), (3) BatchNorm, (4) ReLU, (5) Dense (N), and (6) Sigmoid. Through the final Sigmoid

 doi:10.20944/preprints202503.2247.v1

function, a scalar value ranging from 0 to 1 is produced for each respective channel. This scalar value signifies the relative importance of each channel, whereby significant features are allocated a value approaching 1 or 1, while features of lesser significance are assigned a value nearing 0 or 0. After that, to preserve important features, these values are multiplied by the previously learned feature graph to produce a stronger feature graph. Finally, the pooling layer takes this feature map as input and calculates the average value of all elements, which is converted to a single scalar value.

The final skeleton feature f_{skeleton} is computed as:

$$f_{\text{skeleton}} = \mathcal{P}\left(\sum_{i=1}^{N} \mathcal{A}(\mathcal{T}(\mathcal{G}(\mathcal{B}(X)))) \oplus X\right) \tag{5}$$

Where,X dentes input skeleton information (N, C, T, V),N is the batch size, C is the number of channels, T is the length of the sequence, V is the number of nodes (joints in the skeleton). $\mathcal{G}(X)$ denotes GCN, which extracts spatial features based on the skeleton's graph structure. $\mathcal{T}(\cdot)$ denotes TCN, which captures temporal dependencies in joint movements. $\mathcal{A}(\cdot)$ denotes Channel Attention (CA) module, which assigns importance to features by weighting different channels.$\oplus$ dented Skip connection, which concatenates or adds the refined features to the original input.$\mathcal{P}(\cdot)$ denotes Global Average Pooling (GAP), which computes the final compact representation of the skeleton features.

4.4. Motion Feature Using GSTCAN

Similarly to the skeleton feature process mentioned in the previous section, we extract the motion feature using another stream. Concretely, the skeleton data were provided by AlphaPose and fed to the Motion Generation module, which calculates the displacement of each coordinate along the temporal dimension. After that, motion data is passed to GSTCAN modules that are in another stream. The motion of joints among consecutive frames of the human skeleton is used as nodes of the graph, and the human body structures and their connections in time are used as edges of the graph. The final motion feature f_{motion} is computed as:

$$f_{\text{motion}} = \mathcal{P}\left(\sum_{i=1}^{N} \mathcal{A}(\mathcal{T}(\mathcal{G}(\mathcal{B}(X_{motion})))) \oplus X\right) \tag{6}$$

Where,X_{motion} denotes input motion feature of shape (N, C, T, V),N is the batch size, C is the number of channels, T is the length of the sequence, V is the number of nodes (joints in the skeleton). $\mathcal{G}(X)$ denotes GCN, which extracts spatial features based on the skeleton's graph structure. $\mathcal{T}(\cdot)$ denotes TCN, which captures temporal dependencies in joint movements. $\mathcal{A}(\cdot)$ denotes the Channel Attention (CA) module, which assigns importance to features by weighting different channels.$\oplus$ dented Skip connection, which concatenates or adds the refined features to the original input.$\mathcal{P}(\cdot)$ denotes Global Average Pooling (GAP), which computes the final compact representation of the motion features.

4.5. Stream-2: Sensor Stream Methodology - Bi-LSTM Integration with Channel Attention (CA) Model

We propose an integration of the Bi-LSTM with the CA approach, using sensor data as input. Figure 4 illustrates the architecture of the integrated Bi-LSTM and CA model. The motivation for adopting this model is discussed in the results section.

4.5.1. Bi-LSTM and Channel Attention Integration

The data is first processed by a Bi-LSTM layer. Bi-LSTM, a type of recurrent neural network (RNN), is designed to capture both past and future temporal dependencies in time-series data, making it more effective than unidirectional LSTMs for handling sequences with long-term dependencies. The Bi-LSTM processes the input data from both the past and future, providing a richer context for more accurate predictions at each time step [64–68].

 doi:10.20944/preprints202503.2247.v1

The output from the Bi-LSTM layer is considered Feature-1. This feature is then passed through a CA mechanism, which selectively refines the feature representations by highlighting important channels and suppressing irrelevant ones. This attention mechanism helps the model focus on the most significant parts of the data, improving the overall feature extraction process. After applying CA, the refined feature is passed through a linear layer, followed by ReLU activation and another linear layer. A sigmoid activation function is applied to this layer to produce a final scaled output. To further enhance model performance, we incorporate a residual (skip) connection by multiplying Feature-1 with the output from the sigmoid layer. This step helps retain important information from earlier layers, improving the model's ability to capture relevant patterns. The final result is passed through a linear layer followed by a Softmax activation function to output the classification probability.

4.5.2. Model Derivation

In this model, the input sensor data is represented as a sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$, where T is the total number of time steps in the sequence.

First, the sensor data $\mathbf{x}$ is passed through the Bi-LSTM layer, which captures temporal dependencies in both directions (past and future):

$$\mathbf{h}_{bilstm} = \text{BiLSTM}(\mathbf{h}_x) \tag{7}$$

The output of the Bi-LSTM is considered as Feature-1:

$$\mathbf{f}_1 = \mathbf{h}_{bilstm} \tag{8}$$

Next, Feature-1 is passed through a CA mechanism, which refines the feature representations by selectively focusing on important channels while suppressing irrelevant ones. The output after applying CA is given by:

$$\mathbf{f}_2 = \text{ChannelAttention}(\mathbf{f}_1) \tag{9}$$

The refined feature $\mathbf{f}_2$ is then passed through a fully connected layer, followed by ReLU activation:

$$\mathbf{f}_3 = \text{ReLU}(\mathbf{W}_1\mathbf{f}_2 + \mathbf{b}_1) \tag{10}$$

Where $\mathbf{W}_1$ and $\mathbf{b}_1$ are the weights and bias of the first linear layer.

The output of the fully connected layer is passed through a sigmoid activation function to produce a final scaled output:

$$\mathbf{y}_1 = \sigma(\mathbf{W}_2\mathbf{f}_3 + \mathbf{b}_2) \tag{11}$$

Where σ is the sigmoid activation function, and $\mathbf{W}_2$ and $\mathbf{b}_2$ are the weights and biases of the second linear layer.

Then a residual (skip) connection by multiplying Feature-1 $\mathbf{f}_1$ with the output from the sigmoid layer $\mathbf{y}_1$:

$$\mathbf{f}_{sensor} = \mathbf{f}_1 \odot \mathbf{y}_1 \tag{12}$$

Where $\odot$ denotes element-wise multiplication.

Finally, for the single modality or sensor modality output, the residual final feature vector $\mathbf{f}_sensor$ is passed through another linear layer followed by Softmax activation to obtain the final classification probabilities:

$$\mathbf{y}_{final} = \text{Softmax}(\mathbf{W}_3\mathbf{f}_{sensor} + \mathbf{b}_3) \tag{13}$$

Where $\mathbf{W}_3$ and $\mathbf{b}_3$ are the weights and bias of the final linear layer, and $\mathbf{y}_{final}$ represents the output vector containing the classification probabilities.

For the multimodal case, the concatenated skeleton and sensor features are passed through the linear layer and concatenated with the GSTCAN feature to produce the multimodal feature fusion.

 doi:10.20944/preprints202503.2247.v1

4.6. Multimodal Feature Fusion and Classification

For channel-wise feature fusion, we concatenate the skeleton feature produced by GSTCAN ($\mathbf{f}_{skeleton}$), the motion feature produced by another GSTCAN ($\mathbf{f}_{motion}$) and the sensor feature produced from the residual feature vector ($\mathbf{f}_{sensor}$). The equation for the channel-wise feature fusion is given by:

$$\mathbf{f}_{fusion} = \text{Concat}(\mathbf{f}_{skeleton}, \mathbf{f}_{motion}, \mathbf{f}_{sensor}) \tag{14}$$

Where: $\text{Concat}(\mathbf{f}_{skeleton}, \mathbf{f}_{motion}, \mathbf{f}_{sensor})$ represents the concatenation of the skeleton feature, motion feature and sensor feature along the feature (channel) dimension. $\mathbf{f}_{fusion}$ is the resulting fused feature vector. Once the features are fused, they are passed through a Softmax activation function for classification:

$$\mathbf{f}_{final} = \text{Softmax}(\mathbf{f}_{fusion}) \tag{15}$$

Where Softmax denotes the activation function that outputs the classification probabilities. Finally, the optimizer uses RMSProp with a learning rate of 0.001, and the loss function uses cross-entropy.

5. Experimental Evaluation

We use a 10-fold cross-validation approach on the UP-Fall dataset and evaluate the model performance with the highest validation in that epoch. In the experiment, 30 consecutive frames of one sample are saved as a pkl. Since we separate training and testing for each video, there are no overlapping frames between training and testing. There are 100 epochs, and the batch size is 32. The system was experimented on a GPU machine with CUDA version 11.7, NVIDIA driver version 515, GPU Geforce RTX 3090 24GB, and RAM 32GB.

5.1. Environmental Setting

Experiments will evaluate models using only sensor data, skeleton, motion data, and a combination. For the models using sensor data, CNN and Bi-LSTM will be compared, along with models combining Bi-LSTM and CA. The model achieving the highest accuracy among them will be selected. The system's performance will be evaluated by averaging accuracy, goodness-of-fit, repeatability, and F1-score for each fold in the 10-fold cross-validation.

5.2. Ablation Study

In the ablation study, we first experimented with our stream-1, which employed GSTCAN to produce the result for skeleton data. In the stream-2 or sensor data stream with CNN, CNN-BiLSTM, and Bi-LSTM-CA model; then, we experimented with our only. In both single-stream cases, we included the label-wise precision-recall, f1-score, and accuracy, and then the average in the section below.

5.2.1. Ablation Study with UP-Fall Dataset

Table 2 presents the ablation study of the proposed model using the UP-FALL multimodal dataset. The models are evaluated with both skeleton data in Stream-1 and sensor data in Stream-2. The study demonstrates the performance of three configurations: Stream-1 only, Stream-2 only, and a multi-stream combination using both skeleton and sensor features. In the Stream-1 column, "Yes" or "No" indicates whether Stream-1 is used, and if used, the number of GSTCN modules is provided in the third column. Similarly, Stream-2 follows the same strategy, with "Yes" or "No" indicating if Stream-2 is used, and the model name specified if applicable. The results for Stream-2 (sensor data only) are shown first. Three models were evaluated: CNN, CNN with Bi-LSTM, and Bi-LSTM with Channel Attention (CA). For the CNN model, the accuracy was 97.78%, with the precision of 93.79%, recall of 92.92%, and F1-score of 93.02%. The combination of CNN and Bi-LSTM achieved 99.04% accuracy, with a precision of 96.92%, recall of 97.24%, and F1-score of 96.91%. The Bi-LSTM-CA model produced the highest accuracy of 99.07%, with a precision of 96.63%, recall of 97.21%, and F1-score of 96.75%.

 doi:10.20944/preprints202503.2247.v1

For the Stream-1 skeleton dataset with various GSTCN modules, the highest accuracy achieved was 91.86% with 6 GSTCN modules. The combination of 6 GSTCN modules in Stream-1 and Bi-LSTM-CA in Stream-2 resulted in the best performance with 99.09% accuracy, 97.06% precision, 97.18% recall, and 96.99% F1-score. Additionally, results with 9 GSTCN modules in the multimodal stream are also included.

Table 2. Ablation study of the proposed model for UP-Fall multi-modal dataset.

Ablation	Stream-1 Skeleton		Stream-2 Sensor		Result with UR-FALL (10 fold mean)			
	Yes or No Stream-1	No of GSTCN Skeleton	Yes or No Stream-2	Model Name	Accuracy	Precision	Recall	F1-score
1	No	-	Yes	Only CNN	97.78	93.79	92.92	93.02
2	No	-	Yes	Bi-LSTM with CNN	99.04	96.92	97.24	96.91
3	No	-	Yes	Bi-LSTM with Channel Attention	99.07	96.63	97.21	96.75
4	Yes	3	No	-	91.57	-	-	-
5	Yes	4	No	-	91.56	-	-	-
6	Yes	6	No	-	91.86	-	-	-
7	Yes	9	No	-	91.67	-	-	-
8	Yes	3	Yes	Bi-LSTM with Channel Attention	98.53	-	-	-
8	Yes	9	Yes	Bi-LSTM with Channel Attention	98.66	-	-	-
9	Yes	6	Yes	Bi-LSTM with Channel Attention	99.09	97.06	97.18	96.99

5.2.2. Ablation Study with UR-Fall Dataset

Table 3 presents the ablation study of the proposed model using the UR-FALL multimodal dataset. The models are evaluated with both skeleton data in Stream-1 and sensor data in Stream-2. The study examines five different configurations by varying the number of GSTCN modules in Stream-1 while using a fixed BiLSTM-CNN combination in Stream-2. The results show the performance of each configuration in terms of accuracy, precision, recall, and F1-score. The configuration with 3 GSTCN modules achieves an accuracy of 99.14%, with a precision of 99.06%, recall of 99.041%, and F1-score of 99.04%. Increasing the number of GSTCN modules to 4 results in a slight improvement, with an accuracy of 99.15%, precision of 99.19%, recall of 98.81%, and F1-score of 98.99%. The performance continues to improve with 5 GSTCN modules, reaching an accuracy of 99.24%, a precision of 99.12%, recall of 99.19%, and F1-score of 99.15%. The best performance is achieved with 9 GSTCN modules, which yield an accuracy of **99.32%**, a precision of 99.23%, recall of 99.19%, and F1-score of 99.21%. These results demonstrate that increasing the number of GSTCN modules significantly enhances the model's performance, with the highest accuracy and F1-score achieved with 9 GSTCN modules.

 doi:10.20944/preprints202503.2247.v1

Table 3. Ablation study of the proposed model for UR Fall multi-modal dataset.

Ablation	Stream-1 Num of GSTCN	Stream-2 BiLSTM-CNN	Result with UR-FALL (10 fold mean)			
			Accuracy	Precision	Recall	F1-Score
1	3	1	99.14	99.06	99.041	99.04
2	4	1	99.15	99.19	98.81	98.99
3	5	1	99.24	99.12	99.19	99.15
4	6	1	99.16	99.20	98.48	98.82
5	9	1	**99.32**	99.23	99.19	99.21

5.3. Performance Result of the Proposed Model with UP-Fall Dataset

The integration of the Bi-LSTM and CA model using sensor data achieved higher accuracy than the other two models, so we conducted experiments using a combination of the Bi-LSTM-CA and GSTCAN systems in a multimodal system. The performance accuracy of each is shown in Table 4.

Table 4. Cross Validation Fold wise precision, recall, and F1-score for UP-Fall multimodal dataset.

Fold	**Accuracy [%]**	**Precision [%]**	**Recall [%]**	**F1-Score [%]**
k = 1	99.35	98.65	98.29	98.45
2	99.44	98.25	98.15	98.14
3	99.45	97.48	98.28	97.84
4	99.58	98.90	98.67	98.76
5	98.2	94.62	95.16	94.86
6	98.96	95.97	97.20	96.31
7	98.75	96.46	96.25	96.15
8	98.42	95.79	95.44	95.43
9	99.38	97.69	97.27	97.38
10	99.33	96.83	97.08	96.67
Average	99.09	97.064	97.18	96.99

The results of classifying sensor data using a combination of the Bi-LSTM-Channel Attention and GSTCAN systems are shown in Table 4. Using a combination of the Bi-LSTM-Channel Attention and GSTCAN features, we obtained an average accuracy of 99.09%, precision of 97.06%, recall of 97.18%, and F1-score of 96.99%.

5.4. State of the Art Comparison for UP-FAll Dataset

Table 5 compares our proposed model's performance with several state-of-the-art fall detection models. The results in this table are taken directly from the corresponding papers, where the performance of each model on similar datasets is reported. We aim to highlight how our model compares to these existing systems using identical or comparable multimodal datasets.

 doi:10.20944/preprints202503.2247.v1

Table 5. State-of-the-Art Comparison of the Proposed Model on the UP Fall Multimodal Dataset.

Author	Data Modality	Method Name	Accuracy [%]	Precision [%]	Recall [%]
Martínez et al. [13]	Multi-Sensor	SVM (IMU)+EEG System	90.77	-	-
Ghadi et al. [14]	Multi-Sensor	MS-DLD System	88.75	-	-
Le et al. [16]	Multi-Sensor	Naive Bayes Classifier	88.61	-	-
Li et al. [15]	Skelton	JDM	88.10	-	-
Hafeez et al. [69]	Skeleton+Multi-Sensor	Logistic Regression (LR)	91.51	90.00	91.00
Our Proposed System	Sensor+Skeleton	Two-Stream DNN	**99.09**	**97.06**	**97.18**

Hafeez et al. [69] employed a multimodal approach that combined skeleton data with multi-sensor time-series signals. Their system achieved 91.51% accuracy, 90.00% precision, and 91.00% recall. The results were derived from preprocessing techniques such as noise reduction, silhouette extraction, and keypoint identification. In contrast, our proposed model outperforms Hafeez et al.'s system across all metrics, achieving 99.09% accuracy, 97.06% precision, 97.18% recall, and 96.99% F1 score. Our model integrates multimodal features using a dynamic fusion mechanism, enabling it to capture complex temporal and spatial dependencies more effectively than Hafeez et al.'s basic fusion approach. Additionally, our advanced noise reduction and augmentation techniques contribute to more accurate feature extraction, even in noisy environments. Martínez et al. [13] proposed a fall detection system using a multi-sensor approach, combining Inertial Measurement Unit (IMU) and Electroencephalogram (EEG) data, and applied a Support Vector Machine (SVM) classifier. Their system achieved 90.77% accuracy, though precision and recall metrics were not reported. This method emphasizes the combination of sensor signals, with EEG offering a unique but untested modality for fall detection. Ghadi et al. [14] utilized a multi-sensor system, the MS-DLD, integrating data from multiple inertial sensors. The system achieved 88.75% accuracy, but precision and recall were not reported. Although their approach highlights the importance of multi-sensor data, it shows lower accuracy compared to others in the comparison. Le et al. [16] used a Naive Bayes classifier to process multi-sensor data for fall detection. Their system achieved 88.61% accuracy, with precision and recall metrics not disclosed. The simplicity of the Naive Bayes classifier makes it an interesting choice, though it may not capture complex patterns as effectively as more advanced methods. The primary difference lies in how we handle multimodal data. While Hafeez et al. [69] utilized a basic fusion of visual and inertial features, our model integrates these modalities through a dynamic fusion mechanism, which optimally combines visual and inertial signal data. This method ensures a more robust representation of actions and helps the model capture complex temporal and spatial dependencies more effectively than traditional approaches. Additionally, we use advanced noise reduction and augmentation techniques that improve feature extraction, resulting in better performance even in noisy environments.

5.5. Performance Result of the Proposed Model with UR-FAll Dataset

We performed a fold-wise cross-validation and experimental evaluation of the proposed model on the UR-Fall multimodal dataset, as shown in Table 6. The model achieved an average accuracy of 99.21%, a precision of 98.98%, a recall of 99.07%, and an F1-score of 98.99%, respectively. These results demonstrate that the proposed model generalizes well to other datasets.

 doi:10.20944/preprints202503.2247.v1

Table 6. Cross Validation Fold wise precision, recall, and F1-score for UR-Fall multimodal dataset.

Fold	Accuracy [%]	Precision [%]	Recall [%]	F1-Score [%]
1	100	100	100	100
2	99.31	99.53	98.70	99.11
3	100	100	100	100
4	99.68	99.80	99.23	99.51
5	96.69	96.68	96.71	96.68
6	99.68	99.17	99.80	99.48
7	100	100	100	100
8	99.42	98.65	99.63	99.13
9	100	100	100	100
10	98.38	98.47	97.83	98.14
Average	99.32	99.23	99.19	99.21

5.6. State of the Art Comparison for the UR-FALL Multimodal Dataset

We conduct the experimental evaluation on the UR Fall multimodal dataset and compare our results with state-of-the-art methods, as shown in Table 7, including data modalities, methods, and performance metrics. Kwolek et al. [49] propose a low-cost, 24/7 embedded fall detection system that combines accelerometer data and depth maps. They use an SVM classifier to reduce false positives and report an accuracy of 94.28%. Youssfi et al. [70] introduce a fall detection method based on video and 2D body pose estimation. By tracking body movements over time and applying an SVM model, they achieve an accuracy of 96.55%. Cai et al. [71] propose a deep learning-based method using an hourglass-shaped network that also reconstructs video frames to reduce information loss. Their approach reaches an accuracy of 90.50%. Chen et al. [72] present a method designed for complex visual backgrounds. They use Mask R-CNN to extract moving objects and an attention-guided Bi-LSTM to detect falls, reporting an accuracy of 96.70%. Zheng et al. [53] combine an improved YOLOv4 with GhostNet and BiFPN for fast human detection. AlphaPose is used for pose estimation, and ST-GCN is applied for action recognition, achieving a high accuracy of 97.28% on the UR Fall dataset. Wang et al. [73] propose a visual fall detection method based on Dual-Channel Feature Integration, which separates fall events into "falling" and "fallen" states. They use YOLO and OpenPose to extract dynamic and static features, classified using MLP and Random Forest. Their combined approach achieves an accuracy of 97.33%. In comparison to these studies, our proposed model achieves a significantly higher accuracy of 99.32% using multimodal data. This demonstrates the strong performance and effectiveness of our system in leveraging multimodality for fall detection.

Table 7. State-of-the-Art Comparison of the Proposed Model on the UR Fall Multimodal Dataset.

Author	Data Modality	Method Name	Accuracy [%]	Precision [%]	Recall [%]
Kwolek [49]	Depth	SVM	94.28	-	-
Youssfi [70]	Skeleton	SVM	96.55	-	-
Cai [71]	-	HCAE	90.50	-	-
Chen et al. [72]	RGB	Bi-LsTM	96.70	-	-
Zheng [53]	Skeleton		97.28	97.15	97.43
Wang [73]	Keypoints	-	97.33	97.78	97.78
Our Proposed System	Sensor+Skeleton	Two-Stream DNN	99.32	99.23	99.19

5.7. Discussion

In this study, we proposed a multimodal fall detection system using a dataset that combines skeleton, motion, and sensor data. The performance of the multimodal system surpasses that of the only skeleton or sensor data. The accuracy of the multimodal system improved due to the

 doi:10.20944/preprints202503.2247.v1

effective integration of features from the skeleton, motion, and sensor modalities. Table 4 and Table 6 show the performance accuracy on the UP-Fall and UR-Fall datasets. Additionally, Table 5 and Table 7 compare the proposed model with state-of-the-art systems, demonstrating that our model achieves higher performance accuracy than existing methods. One key limitation identified is the dataset imbalance, where fall actions represent only 20% of the total samples in the UP-Fall dataset. The GSTCAN model, which utilizes skeleton data, showed lower accuracy for fall detection on the UP-Fall dataset, as shown in Table 2, especially for activities like walking, sitting, and standing. However, the model demonstrated high accuracy on the UR-Fall dataset, as shown in Table 3. Reviewing the fall videos, we observed that nearly half of the incidents were motionless or involved lying down. This suggests a need for more diverse fall scenarios to improve the model's recognition capability. On the other hand, the accuracy of the accelerometer for actions like standing and jumping was almost 100%, as shown in Table 2, due to the distinct and consistent changes in sensor values, which made these actions easier for the model to recognize. The sudden acceleration changes during falls also contributed to the higher accuracy in the sensor stream compared to skeleton data. However, the recognition accuracy for actions such as falling over or grabbing something using hands and knees was lower, likely due to the gentler acceleration changes during these actions. These subtler changes are harder for the model to detect, leading to misclassification. To improve the accuracy of the multimodal system on the UP-Fall dataset, it is necessary to increase the amount of skeleton data for falls. Future experiments will focus on incorporating additional features, such as the distance between joints, alongside the existing joint movement features, to capture the dynamics of fall incidents better. In addition to addressing the dataset imbalance and improving recognition performance, we plan to explore synthetic data generation techniques. Although we initially refrained from applying SMOTE (Synthetic Minority Over-sampling Technique) due to the complexity of handling video data and preserving temporal dependencies, we recognize its potential for balancing class distribution. In future work, we aim to investigate the application of SMOTE to generate additional fall samples while maintaining the integrity of the temporal relationships in video sequences. Additionally, we will explore complementary approaches, such as class-weight adjustments and further data augmentation techniques, including temporal dependency and synthetic video generation, to address the imbalance and improve the model's robustness.

6. Conclusions

In this study, we propose a multimodal fall detection system that leverages advanced DL techniques to improve accuracy and robustness. By combining the Graph-based Spatial-Temporal Convolution and Attention Network (GSTCAN) for skeleton and motion data and then the Bi-LSTM-CA integration for sensor data, we enhance the system's ability to capture both spatial and temporal features. For the sensor stream, we compared Bi-LSTM-CA and its integration, ultimately selecting Bi-LSTM-CA due to its superior performance in capturing long-range dependencies. The GSTCAN model extracts skeleton data using AlphaPose, calculates motion between consecutive frames, builds a graph, applies Graph Convolutional Networks (GCN), and integrates a CA mechanism to highlight important features. The features from both the skeleton and sensor streams are fused, enabling more accurate and comprehensive fall detection. Experiments on the UP-Fall dataset showed that the multimodal system outperformed individual models, demonstrating that combining both modalities leads to improved fall detection accuracy. Future work will focus on enhancing the skeleton data stream through data augmentation, which will further improve the multimodal system's performance. This will enable more stable and accurate fall detection, even with limited sensor data, ultimately improving safety and reducing the physical and psychological burden on the elderly.

Author Contributions: Conceptualization, R. Egawa, K. Hirooka, A.S.M. Miah; methodology, R. Egawa, K. Hirooka, A.S.M. Miah, Y. Tomioka, and J. Shin; investigation, R. Egawa, A.S.M. Miah, and J. Shin; data curation, R. Egawa, A.S.M. Miah, Y. Tomioka, and J. Shin; writing—original draft preparation, R. Egawa, A.S.M. Miah,N.

 doi:10.20944/preprints202503.2247.v1

Hassan, and J. Shin; writing—review and editing, A.S.M. Miah, Y. Tomioka, and J. Shin; visualization, R. Egawa, A.S.M. Miah, N. Hassan supervision, J. Shin; funding acquisition, J. Shin.

Data Availability Statement: UP-Fall Dataset: http://sites.google.com/up.edu.mx/har-up/; UR Fall Dataset : https://fenix.ur.edu.pl/mkepski/ds/uf.html

Conflicts of Interest: The authors declare no conflict of interest

References

1. Ministry of Health, L.; Welfare. Fall Accidents. [Accessed: 08 January 2025].
2. NCGG. What makes it easier to fall? [Accessed: 08 January 2025].
3. Semwal, V.B.; Katiyar, S.A.; Chakraborty, R.; Nandi, G.C. Biologically-inspired push recovery capable bipedal locomotion modeling through hybrid automata. *Robotics and Autonomous Systems* **2015**, *70*, 181–190.
4. WHO. Falls. [Accessed: 08 January 2025].
5. Naja, S.; Makhlouf, M.; Chehab, M.A.H. An ageing world of the 21st century: a literature review. *Int J Community Med Public Health* **2017**, *4*, 4363–9.
6. Romeo, L.; Marani, R.; Petitti, A.; Milella, A.; D'Orazio, T.; Cicirelli, G. Image-Based Mobility Assessment in Elderly People from Low-Cost Systems of Cameras: A Skeletal Dataset for Experimental Evaluations. In Proceedings of the Ad-Hoc, Mobile, and Wireless Networks; Grieco, L.A.; Boggia, G.; Piro, G.; Jararweh, Y.; Campolo, C., Eds. Springer International Publishing, 2020, pp. 125–130.
7. Azmat, U.; Jalal, A. Smartphone inertial sensors for human locomotion activity recognition based on template matching and codebook generation. In Proceedings of the 2021 International Conference on Communication Technologies (ComTech). IEEE, 2021, pp. 109–114.
8. Huang, Z.; Liu, Y.; Fang, Y.; Horn, B.K. Video-based fall detection for seniors with human pose estimation. In Proceedings of the 2018 4th international conference on Universal Village (UV). IEEE, 2018, pp. 1–4.
9. Shanmughapriya, M.; Gunasundari, S.; Bharathy, S. Loitering detection in home surveillance system. In Proceedings of the 2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22). IEEE, 2022, pp. 1–6.
10. Abdullah, F.; Jalal, A. Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system. *Arabian Journal for Science and Engineering* **2023**, *48*, 2173–2190.
11. Ha, T.V.; Nguyen, H.; Huynh, S.T.; Nguyen, T.T.; Nguyen, B.T. Fall detection using multimodal data. In Proceedings of the International Conference on Multimedia Modeling. Springer, 2022, pp. 392–403.
12. Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.L.; Grundmann, M. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* **2020**.
13. Martínez-Villaseñor, L.; Ponce, H.; Brieva, J.; Moya-Albor, E.; Núñez-Martínez, J.; Peñafort-Asturiano, C. UP-fall detection dataset: A multimodal approach. *Sensors* **2019**, *19*, 1988.
14. Ghadi, Y.; Javeed, M.; Alarfaj, M.; Shloul, T.; Alsuhibany, S.; Jalal, A.; Kamal, S.; Kim, D.S. MS-DLD: Multi-sensors based daily locomotion detection via kinematic-static energy and body-specific HMMs. *IEEE Access* **2022**, *10*, 23964–23979.
15. Li, C.; Hou, Y.; Wang, P.; Li, W. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Process. Lett.* **2017**, *24*, 624–628.
16. Le, T.M.; Tran, L.V.; Dao, S.V.T. A feature selection approach for fall detection using various machine learning classifiers. *IEEE Access* **2021**, *9*, 115895–115908.
17. Xu, Q.; Huang, G.; Yu, M.; Guo, Y. Fall prediction based on key points of human bones. *Physica A: Statistical Mechanics and its Applications* **2020**, *540*, 123205.
18. Mubashir, M.; Shao, L.; Seed, L. A survey on fall detection: Principles and approaches. *Neurocomputing* **2013**, *100*, 144–152.
19. Sucerquia, A.; López, J.D.; Vargas-Bonilla, J.F. Real-life/real-time elderly fall detection with a triaxial accelerometer. *Sensors* **2018**, *18*, 1101.
20. Sucerquia, A.; López, J.D.; Vargas-Bonilla, J.F. SisFall: A fall and movement dataset. *Sensors* **2017**, *17*, 198.
21. Desai, K.; Mane, P.; Dsilva, M.; Zare, A.; Shingala, P.; Ambawade, D. A novel machine learning based wearable belt for fall detection. In Proceedings of the 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 2020, pp. 502–505.
22. Hussain, F.; Hussain, F.; Ehatisham-ul Haq, M.; Azam, M.A. Activity-Aware Fall Detection and Recognition Based on Wearable Sensors. *IEEE Sensors Journal* **2019**, *19*, 4528–4536. https://doi.org/10.1109/JSEN.2019.2898891.

23. Fula, V.; Moreno, P. Wrist-based fall detection: towards generalization across datasets. *Sensors* **2024**, *24*, 1679.
24. Casilari, E.; Santoyo-Ramón, J.A.; Cano-García, J.M. Umafall: A multisensor dataset for the research on automatic fall detection. *Procedia Computer Science* **2017**, *110*, 32–39.
25. Marques, J.; Moreno, P. Online Fall Detection Using Wrist Devices. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23031146.
26. Martínez-Villaseñor, L.; Ponce, H.; Brieva, J.; Moya-Albor, E.; Núñez-Martínez, J.; Peñafort-Asturiano, C. UP-Fall Detection Dataset: A Multimodal Approach. *Sensors* **2019**, *19*. https://doi.org/10.3390/s19091988.
27. Maray, N.; Ngu, A.H.; Ni, J.; Debnath, M.; Wang, L. Transfer Learning on Small Datasets for Improved Fall Detection. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23031105.
28. Miah, A.S.M.; Shin, J.; Hasan, M.A.M.; Rahim, M.A. BenSignNet: Bengali Sign Language Alphabet Recognition Using Concatenated Segmentation and Convolutional Neural Network. *Applied Sciences* **2022**, *12*, 3933.
29. Miah, A.S.M.; Hasan, M.A.M.; Jang, S.W.; Lee, H.S.; Shin, J. Multi-Stream General and Graph-Based Deep Neural Networks for Skeleton-Based Sign Language Recognition. *Electronics* **2023**, *12*.
30. Miah, Abu Saleh Musa, S.J.; Hasan, M.A.M.; Rahim, M.A.; Okuyama, Y. Rotation, Translation And Scale Invariant Sign Word Recognition Using Deep Learning. *Computer Systems Science and Engineering* **2023**, *44*, 2521–2536.
31. Juang, L.H.; Wu, M.N. Fall down detection under smart home system. *Journal of medical systems* **2015**, *39*, 1–12.
32. Han, Q.; Zhao, H.; Min, W.; Cui, H.; Zhou, X.; Zuo, K.; Liu, R. A Two-Stream Approach to Fall Detection With MobileVGG. *IEEE Access* **2020**, *8*, 17556–17566. https://doi.org/10.1109/ACCESS.2019.2962778.
33. Li, X.; Pang, T.; Liu, W.; Wang, T. Fall detection for elderly person care using convolutional neural networks. In Proceedings of the 2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI). IEEE, 2017, pp. 1–6.
34. Lu, N.; Wu, Y.; Feng, L.; Song, J. Deep Learning for Fall Detection: Three-Dimensional CNN Combined With LSTM on Video Kinematic Data. *IEEE Journal of Biomedical and Health Informatics* **2019**, *23*, 314–323. https://doi.org/10.1109/JBHI.2018.2808281.
35. Alanazi, T.; Muhammad, G. Human Fall Detection Using 3D Multi-Stream Convolutional Neural Networks with Fusion. *Diagnostics* **2022**, *12*. https://doi.org/10.3390/diagnostics12123060.
36. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* **2018**, *32*. https://doi.org/10.1609/aaai.v32i1.12328.
37. Leiyue.; Wei. An Improved Feature-Based Method for Fall Detection. *Tehnicki vjesnik - Technical Gazette* **2019**.
38. Tsai, T.H.; Hsu, C.W. Implementation of Fall Detection System Based on 3D Skeleton for Deep Learning Technique. *IEEE Access* **2019**, *7*, 153049–153059. https://doi.org/10.1109/ACCESS.2019.2947518.
39. Zheng, H.; Liu, Y. Lightweight Fall Detection Algorithm Based on AlphaPose Optimization Model and ST-GCN. *Mathematical Problems in Engineering* **2022**, *2022*, 9962666.
40. McCall, S.; Kolawole, S.S.; Naz, A.; Gong, L.; Ahmed, S.W.; Prasad, P.S.; Yu, M.; Wingate, J.; Ardakani, S.P. Computer Vision Based Transfer Learning-Aided Transformer Model for Fall Detection and Prediction. *IEEE Access* **2024**, *12*, 28798–28809. https://doi.org/10.1109/ACCESS.2024.3368065.
41. Inturi, A.R.; Manikandan, V.M.; Kumar, M.N.; Wang, S.; Zhang, Y. Synergistic Integration of Skeletal Kinematic Features for Vision-Based Fall Detection. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23146283.
42. De, A.; Saha, A.; Kumar, P.; Pal, G. Fall detection approach based on combined two-channel body activity classification for innovative indoor environment. *Journal of Ambient Intelligence and Humanized Computing* **2022**, *14*, 1–12. https://doi.org/10.1007/s12652-022-03714-2.
43. Galvão, Y.M.; Portela, L.; Ferreira, J.; Barros, P.; De Araújo Fagundes, O.A.; Fernandes, B.J.T. A Framework for Anomaly Identification Applied on Fall Detection. *IEEE Access* **2021**, *9*, 77264–77274. https://doi.org/10.1109/ACCESS.2021.3083064.
44. Zahan, S.; Hassan, G.M.; Mian, A. SDFA: Structure-Aware Discriminative Feature Aggregation for Efficient Human Fall Detection in Video. *IEEE Transactions on Industrial Informatics* **2023**, *19*, 8713–8721. https://doi.org/10.1109/TII.2022.3221208.
45. Egawa, R.; Miah, A.S.M.; Hirooka, K.; Tomioka, Y.; Shin, J. Dynamic Fall Detection Using Graph-Based Spatial Temporal Convolution and Attention Network. *Electronics* **2023**, *12*. https://doi.org/10.3390/electronics12153234.

 doi:10.20944/preprints202503.2247.v1

46. Chahyati, D.; Hawari, R. Fall Detection on Multimodal Dataset using Convolutional Neural Netwok and Long Short Term Memory. In Proceedings of the 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2020, pp. 371–376. https://doi.org/10.1109/ICACSIS51025.2020.9263201.
47. Islam, M.M.; Nooruddin, S.; Karray, F. Multimodal Human Activity Recognition for Smart Healthcare Applications. In Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2022, pp. 196–203. https://doi.org/10.1109/SMC53654.2022.9945513.
48. Ray, S.; Alshouiliy, K.; Agrawal, D.P. Dimensionality reduction for human activity recognition using google colab. *Information* **2020**, *12*, 6.
49. Kwolek, B.; Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer methods and programs in biomedicine* **2014**, *117*, 489–501.
50. Igual, R.; Medrano, C.; Plaza, I. Challenges, issues and trends in fall detection systems. *Biomedical engineering online* **2013**, *12*, 66.
51. Zhang, Z.; Conly, C.; Athitsos, V. A survey on vision-based fall detection. In Proceedings of the Proceedings of the 8th ACM international conference on PErvasive technologies related to assistive environments, 2015, pp. 1–7.
52. Yan, S.; Xiong, Y.; Lin, D. Spatial, temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.
53. Zheng, H.; Liu, Y. Lightweight fall detection algorithm based on AlphaPose optimization model and ST-GCN. *Mathematical Problems in Engineering* **2022**, *2022*.
54. Keskes, O.; Noumeir, R. Vision-based fall detection using st-gcn. *IEEE Access* **2021**, *9*, 28224–28236.
55. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
56. Hassan, N.; Miah, A.S.M.; Suzuki, T.; Shin, J. Gradual Variation-Based Dual-Stream Deep Learning for Spatial Feature Enhancement With Dimensionality Reduction in Early Alzheimer's Disease Detection. *IEEE Access* **2025**, *13*, 31701–31717. https://doi.org/10.1109/ACCESS.2025.3542458.
57. Fang, H.S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.L.; Lu, C. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**.
58. Miah, A.S.M.; Hasan, M.A.M.; Shin, J. Dynamic Hand Gesture Recognition using Multi-Branch Attention Based Graph and General Deep Learning Model. *IEEE Access* **2023**.
59. Miah, A.S.M.; Hasan, M.A.M.; Nishimura, S.; Shin, J. Sign Language Recognition Using Graph and General Deep Neural Network Based on Large Scale Dataset. *IEEE Access* **2024**, *12*, 34553–34569.
60. Shin, J.; Miah, A.S.M.; Suzuki, K.; Hirooka, K.; Hasan, M.A.M. Dynamic Korean Sign Language Recognition Using Pose Estimation Based and Attention-Based Neural Network. *IEEE Access* **2023**, *11*, 143501–143513. https://doi.org/10.1109/ACCESS.2023.3343404.
61. Hassan, N.; Miah, A.S.M.; Suzuki, K.; Okuyama, Y.; Shin, J. Stacked CNN-based multichannel attention networks for Alzheimer disease detection. *Scientific Reports* **2025**, *15*, 5815.
62. Miah, A.S.M.; Hasan, M.A.M.; Shin, J.; Okuyama, Y.; Tomioka, Y. Multistage Spatial Attention-Based Neural Network for Hand Gesture Recognition. *Computers* **2023**, *12*, 13.
63. Miah, A.S.M.; Hasan, M.A.M.; Okuyama, Y.; Tomioka, Y.; Shin, J. Spatial–temporal attention with graph and general neural network-based sign language recognition. *Pattern Analysis and Applications* **2024**, *27*, 37.
64. Gurbuz, S.Z.; Amin, M.G. Radar-Based Human-Motion Recognition With Deep Learning: Promising Applications for Indoor Monitoring. *IEEE Signal Processing Magazine* **2019**, *36*, 16–28. https://doi.org/10.1109/MSP.2018.2890128.
65. Le Kernec, J.; Fioranelli, F.; Ding, C.; Zhao, H.; Sun, L.; Hong, H.; Lorandel, J.; Romain, O. Radar Signal Processing for Sensing in Assisted Living: The Challenges Associated With Real-Time Implementation of Emerging Algorithms. *IEEE Signal Processing Magazine* **2019**, *36*, 29–41. https://doi.org/10.1109/MSP.2019.2903715.
66. Hassan, N.; Miah, A.S.M.; Shin, J. A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition. *Applied Sciences* **2024**, *14*, 603.
67. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* **2005**, *18*, 602–610. IJCNN 2005, https://doi.org/https://doi.org/10.1016/j.neunet.2005.06.042.

 doi:10.20944/preprints202503.2247.v1

68. Graves, A.; Mohamed, A.r.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing. Ieee, 2013, pp. 6645–6649.
69. Hafeez, S.; Alotaibi, S.S.; Alazeb, A.; Mudawi, N.A.; Kim, W. Multi-Sensor-Based Action Monitoring and Recognition via Hybrid Descriptors and Logistic Regression. *IEEE Access* **2023**, *11*, 48145–48157. https://doi.org/10.1109/ACCESS.2023.3275733.
70. Youssfi Alaoui, A.; Tabii, Y.; Oulad Haj Thami, R.; Daoudi, M.; Berretti, S.; Pala, P. Fall detection of elderly people using the manifold of positive semidefinite matrices. *Journal of Imaging* **2021**, *7*, 109.
71. Cai, X.; Li, S.; Liu, X.; Han, G. Vision-based fall detection with multi-task hourglass convolutional auto-encoder. *IEEE Access* **2020**, *8*, 44493–44502.
72. Chen, Y.; Li, W.; Wang, L.; Hu, J.; Ye, M. Vision-based fall event detection in complex background using attention guided bi-directional LSTM. *IEEE Access* **2020**, *8*, 161337–161348.
73. Wang, B.H.; Yu, J.; Wang, K.; Bao, X.Y.; Mao, K.M. Fall detection based on dual-channel feature integration. *IEEE Access* **2020**, *8*, 103443–103453.