electronics

MDPI

Article

Dynamic Fall Detection Using Graph-Based Spatial Temporal Convolution and Attention Network

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

School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Fukushima, Japan
* Correspondence: jpshin@u-aizu.ac.jp

Abstract: The prevention of falls has become crucial in the modern healthcare domain and in society for improving ageing and supporting the daily activities of older people. Falling is mainly related to age and health problems such as muscle, cardiovascular, and locomotive syndrome weakness, etc. Among elderly people, the number of falls is increasing every year, and they can become life-threatening if detected too late. Most of the time, ageing people consume prescription medication after a fall and, in the Japanese community, the prevention of suicide attempts due to taking an overdose is urgent. Many researchers have been working to develop fall detection systems to observe and notify about falls in real-time using handcrafted features and machine learning approaches. Existing methods may face difficulties in achieving a satisfactory performance, such as limited robustness and generality, high computational complexity, light illuminations, data orientation, and camera view issues. We proposed a graph-based spatial-temporal convolutional and attention neural network (GSTCAN) with an attention model to overcome the current challenges and develop an advanced medical technology system. The spatial-temporal convolutional system has recently proven the power of its efficiency and effectiveness in various fields such as human activity recognition and text recognition tasks. In the procedure, we first calculated the motion along the consecutive frame, then constructed a graph and applied a graph-based spatial and temporal convolutional neural network to extract spatial and temporal contextual relationships among the joints. Then, an attention module selected channel-wise effective features. In the same procedure, we repeat it six times as a GSTCAN and then fed the spatial-temporal features to the network. Finally, we applied a softmax function as a classifier and achieved high accuracies of 99.93%, 99.74%, and 99.12% for ImViA, UR-Fall, and FDD datasets, respectively. The high-performance accuracy with three datasets proved the proposed system's superiority, efficiency, and generality.

Keywords: fall detection (FD); graph convolutional network (GCN); human activity recognition (HAR); computer vision; body pose detection; AlphaPose; channel attention; ageing people

Citation: 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*, 3234. https://doi.org/10.3390/electronics12153234

Academic Editor: George A. Tsihrintzis

Received: 21 June 2023
Revised: 20 July 2023
Accepted: 24 July 2023
Published: 26 July 2023

1. Introduction

The ageing of the population has become a global phenomenon, and the number of elderly people in the world is projected to more than double over the next 30 years. Approximately 16.0% of the population is expected to be elderly by 2050 [1]. According to the World Health Organization (WHO) [2], falls are the second leading cause of unintentional death after traffic accidents, and adults over the age of 60 suffer the most fatal falls. Japan is one of the most ageing countries in the world, with a total of 29% of their people projected to be over 65 years old in the future; there is the possibility of increasing this ratio. When a person is unable to respond to stimuli and unable to maintain awareness of his surroundings, he becomes unconscious. As a result, they seem to be asleep and fall asleep. Falling is a significant issue for senior citizens in Japan and causes injuries and death [3]. Urgent treatment and intervention are crucial on losing consciousness, otherwise, there is a high risk to patients. It is very important to develop an automatic fall detection method to

protect life in this situation. Some of these falls are serious enough to require medical attention. It has been shown that medical attention immediately after a fall effectively reduces the likelihood of death by 80% and the need for long-term hospitalization by 26% [4]. The response time to rescue a seriously injured person from a fall is critical to the survival of the elderly. Therefore, it is very important to develop an automatic fall detection method to prevent the risk of serious injury and death because of falls. Our main goal was to develop an indoor fall detection system that will be subject- and environment-independent. There are two types of existing fall detection methods—wearable sensor-based and vision-based methods [5]. The wearable sensor-based methods have the elderly person wear a sensor, which detects the sudden acceleration changes caused by a fall [6]. However, this is inconvenient because many elderly people are often forgetful and unwilling. In addition, the method is susceptible to noise, and everyday activities such as lying down or sitting up can lead to false detection [7]. Recently, cameras have become popular in many public and private spaces, such as train stations, bus stops, and office buildings. Also, the rapid development of computer vision under the influence of deep learning [8] has led to the development of vision-based methods [9–14]. Although vision-based methods eliminate the inconvenience of wearing a device, they may be subject to false detection due to lighting or complex backgrounds. Most previous vision-based fall detection systems were developed using threshold-based methods by comparing the settings reference with input data [15]. The main problem of the threshold-based system is that there is the possibility of missing a prediction of a fall event because of the high or low threshold. In addition, it may create some sensitive issues, such as sudden changes in the human body positions, like picking up anything from the floor, or Muslim prayer, which could be considered as a fall. Currently, researchers use various machine learning algorithms for fall detection, such as random forest, support vector machine (SVM), and k-nearest neighbors (KNN) [16]. Some researchers have demonstrated the comparison between machine learning and threshold-based systems to prove the effectiveness of the machine learning algorithms [17,18].

The author collected data from various sources such as a gyroscope, an accelerometer, and magneto meters located on the subject's wrist. They compared the threshold-based and machine learning methods and reported that machine learning performs well. The main problem of the machine learning-based system is the handcrafted feature because this feature must be closely related to and connected with human activities and similar actions. The main drawback of the handcrafted feature is that it is not guaranteed to find a good description. In addition, the system's robustness is not good because it needs strong domain background knowledge to select the handcrafted feature. The choice of features is not straightforward, and finding the best feature is very important to reflect the essence of a fall [19]. The main problem occurs when multitask classification is needed, when features are closely related. Researchers have also used depth sensors, infrared sensors, optical sensors, and RGB cameras [10–12,14,20,21]. To solve the handcrafted feature problem, researchers employed a deep learning-based method to explore data and extract effective features for the specific classification task using RGB images [20,22]. RGB image data-based deep learning for fall detection still faces problems in achieving a high performance because of the redundant background light illuminations and computational complexity. To solve the problems, many researchers have employed deep learning on the redundant backgrounds of images, such as geometric multi-grid (GMG), fuzzy methods, Gaussian mixture models (GMM), and RPCA methods [23–26] Also, many researchers have used deep learning-based background removal methods such as ANN [27–29], Faster R-CNN, and Yolov3 [30,31] but these have some computational issues because of the two-time deep learning for background reduction and class action. Recently, the skeleton data points of the human body, instead of the RGB image, has been used by many researchers to solve efficiency and accuracy-related issues. Chen et al. proposed a skeleton data point-based fall detection method where they calculated different geometrical and static features

from the skeleton data. Then, using the machine learning method, they achieved 97.00% accuracy [32,33].

The main problem with these features is that the skeleton data are different from images and videos because they form a graph instead of 2D or 3D grids. Consequently, conventional feature extraction methods are not able to extract exact information and handle this data structure in its native form, which yields the preprocessing steps. The many existing approaches merge the joint points in a digestible type of data structure, such as metrics vectors. This transformation can lead to the loss of relevant, effective information and especially different joint relationships. To solve the problem, Yan et al. applied a new deep learning method, the spatial-temporal graph convolutional network (ST-GCN) [34]. It mainly extracts the various node relationships, specifically the spatial and temporal contextual relationships among the joints [32]. They constructed a graph instead of 2D grids and achieved a satisfactory performance in hand gesture and activity recognition. Keskes et al. applied the ST-GCN fall detection method to solve various challenges in the domain [35]. The main drawback of their method is that they used ten units of ST-GCN sequentially, which increases the high computational complexity. In addition, they did not consider the role of non-connected skeleton points in fall events during the spatiotemporal feature extraction. However, the positional relationship between some non-real connected points is very helpful for partially identifying events. To overcome the problems, we proposed a graph-based spatial-temporal convolution and attention network (GSTCAN) model to overcome the current challenges and developed an advanced medical technology system. The major contributions of this work are detailed below:

- We proposed a graph-based GSTCAN model in which we first calculated the motion among the consecutive frame. Then, we constructed a graph and applied a graph-based spatial-temporal convolutional neural network to extract intra-frame joints and an inter-frame joints relationship by considering the spatial and temporal domains.
- Secondly, we fed the spatial-temporal convolution feature into an attention module to select channel-wise effective features. The main purpose of the attention model is to improve the role of non-connected skeleton points in certain events during the spatial-temporal feature; we applied the attention model to GSTCAN, aiming to extract global and local features bound to impact model optimization. In the same procedure, we sequentially applied GSTCAN six times in a series, producing effective features that carry the skeleton joint's internal relationship structure.
- Finally, we applied a softmax function as a classifier and achieved high accuracies of 99.93%, 99.74%, and 99.12% for ImViA, UR-Fall, and FDD datasets, respectively. The high-performance accuracy with three datasets proved the superiority and efficiency of the proposed system.

The remainder of this paper is organized as follows: Section 2 summarizes the existing research work and related problems. Section 3 describes the three fall detection benchmark datasets, and Section 4 describes the architecture of the proposed system. Section 5 details the evaluation performed, including a comparison with a state-of-the-art approach. In Section 6, our conclusions and directions for future work are discussed.

2. Related Work

Many researchers have been working to develop fall detection systems with various feature extraction and classification approaches [18,36–40]. All the algorithms used in this domain can be divided into the following categories: (I) sensor-based systems for monitoring the person [41,42]; (II) radio frequency (RF) sensor-based systems, and (III) camera-based vision-related systems. Many researchers record various signals with various sensors such as gyroscopes, accelerometers, EMGs, and EEGs to collect information from many people, not just the elderly [43–52]. Then they extract various kinds of features, including angle, distance, the sum of X and Y with various directions and their derivatives, and geometrical, statistical, and mathematical formulas [39]. Wang et al. collected data on fall events using an accelerometer sensor and calculated the SVMA for the patients [53].

They first assigned a threshold value as an assumption and, if the SVMA value surpassed the threshold, they then calculated some features of the trunk angle and pressor pulse sensors. Moreover, if the two values were higher than the normal value, it can produce an emergency alarm and achieve 97.5% accuracy. Desai et al. used multiple sensors in combination, including an accelerometer, a gyroscope, a GSM module microcontroller, a battery, and an IMU sensor [54]. They used a logistic regression classifier and, if a fall event happened, GSM produced an emergency alarm for the helpline number. The drawback is that they only used the human activity dataset, not including any specific dataset for the fall events. Xu et al. reviewed the wearable accelerometers-based work, proving some advantages of the wearable sensor, such as low cost, portability, and efficiency at detecting falls with high-performance accuracy [55].

The main drawbacks of this type of work is that the patient needs to wear a sensor all day, as well as there being high noise, which leads to difficulty for ageing people, which badly affects their daily lives. To solve the wearable sensor problems, the second category of radar technologies and Wi-Fi was proposed to solve the mentioned problems. Tian et al. collected the RF reactions from the environment using frequency-modulated continuous-wave radio (FMCW) equipment [56]. They generated two heat maps from the reflection and applied a deep learning model for the classification, which achieved 92.00% precision and 94.00% sensitivity. RF is a non-intrusive sensor-based method that achieves good performance accuracy. It can solve the noise problem, but collecting data from each cell based on an antenna with interference is challenging. Researchers proposed a camera-based data collection system to solve the portability and high-cost of data collection problems. In recent years, camera-based fall detection approaches have been acceptable to researchers and consumers because of their low cost and portability properties.

Zerrouki et al. extracted curvelet transforms and area ratios to identify human posture in images, used SVM to identify posture, and used a hidden Markov model (HMM) [57] for activity recognition [58]. Chua et al. proposed an RGB image-based fall events detection method using human shape variation. After extracting the foreground information, they calculated three points with which to calculate the fall event-related features, reporting 90.5% accuracy [59]. Cai et al. applied the hourglass convolutional auto-encoder (HCAE) approach by combining with hourglass residual units (HRU) to extract the intermediate features from the RGB video dataset [60]. They extracted the features for the fall classification and then reconstructed the image to enhance the representation of the intermediate fall event-related features. After evaluating their model with the UR fall detection dataset, they achieved 96.20% accuracy. Chen et al. applied mask R-CNN to extract a feature, aiming to detect a fall event from the RGB image based on the CNN model [20]. Later, they applied bi-directional LSTM for the classification and achieved 96.7% accuracy for the UR fall detection dataset. Harrou et al. proposed a multi-step procedure for detecting fall events, including data processing and segmentation, splitting the foreground image into five regions based on the relevant features, extracting features from each region, and then calculating the generalized likelihood ratio (GLR) [61]. Finally, they evaluated their model with the FDD and URFD datasets, which produced 96.84% and 96.66% accuracies, respectively. Han et al. applied the MobileVGG network, which extracts the motion features from the RGB video to detect fall events, and they achieved 98.25% accuracy with their dataset [62]. Standard camera and image-based systems are sometimes not robust and their performance may be limited because of the complexity of distinguishing between foreground and background.

In addition, they still face problems of light illumination, partial occlusion, and redundant background complexity problems. To overcome the problems, many computer vision researchers have used skeleton datasets to detect fall events and human activity instead of the RGB pixel-based image to solve the mentioned problems. The skeleton-based dataset's main advantage is its robust scene variation, light illumination, and partial occlusion [34,63–65]. Yao et al. extracted the features from the skeleton joint, then applied the SVM, and achieved 93.56% accuracy with the TSTv2 dataset [63]. The main concept behind their task is that they divide the skeleton data into five parts based on the organs such as

the head, neck, spine base, and spine centre. Tsai et al. extracted features from the selected potential joints of the skeleton dataset and then applied a 1DCNN for the classification [64]. After evaluating the NTU-RGBD dataset, they achieved high-performance accuracy compared to the previous system. The main drawback of the dataset is that the NTU-RGBD dataset does not include all types of fall events. Tran et al. proposed a handcrafted feature-based fall detection method where they first calculated the plane based on the floor of the room [66]. After that, they calculated the velocity distance of the head and the spine associated with the floor. After applying the SVM method, they achieved better accuracy than the previous method. Most of the existing work on fall event detection was developed with hand-crafted features, which faces difficulties in handling large datasets. In addition, effective feature extraction and potential feature selection approaches still face many challenges. The deep learning-based approach is the most powerful classification approach, and can extract the effective features and outperforms hand-crafted features because it can obtain many more features during training; however, it needs a large dataset. In this study, we proposed a skeleton-based GSTCAN model to recognize fall events through the skeleton data provided by the AlphaPose. Our main goal was to develop a robust fall detection system with high-performance accuracy, efficiency, and generality. We tested the proposed model with three datasets to prove its high-performance accuracy, efficiency, and generality according to the standard generality system.

3. Datasets

There are few dynamic fall detection benchmark datasets available online. For this study, we selected three benchmark dynamic fall detection datasets, namely: the UR Fall Dataset [36], ImViA Datasets(le2i) [38], and FDD [67]. Table 1 provides a summary of those datasets and their specifications, including features, people, and actions, etc.

Table 1. Summary of the datasets used in this study.

Dataset Name	Type	Classes	Sample
UR Fall Detection [36]	This consists of raw video and does not contain bounding box information.	2 Class Fall/ Non-Fall	3 K images in total
ImViA Datasets (le2i) [38]	This consists of raw video and does not contain bounding box information.	2 Class Fall/ Non-Fall	40 K images in total
FDD [67]	This consists of the image	Five classes	22 K images in total

3.1. UR Fall Detection Datasets

The videos in the UR Fall Detection Dataset [68] are short and correspond to fall and non-fall sequences. This dataset contains videos of 30 falls. We also used the UR Fall detection dataset [36]. This dataset is a fall detection dataset provided by the University of Rochester's Rehabilitation Medicine Research Group. The dataset includes video data captured from multiple cameras and corresponding annotation data for fall events. The video data were captured by RGB and Depth cameras with a resolution of 640 × 480. The annotation data include information such as the time of the fall event and the posture of the person before and after the fall. It has been a useful resource for fall detection research and has been widely used in various studies. It is also used as a benchmark for fall detection tasks. Table 1 demonstrates the two most usable fall detection datasets. The videos in the UR Fall detection dataset were recorded by two different cameras, with 70 activated cameras and 3000 images. Among them, 30 activities are considered to be falls and 40 activities are normal daily living activities.

3.2. ImViA Datasets(le2i)

The ImViA dataset [38] is a dataset including videos from a single camera in a realistic video surveillance setting. It includes daily activities such as going from a chair to a sofa, exercising, and falling. Only one person is displayed at a time, the frame rate is 25 frames/s, and the resolution is 320 × 240 pixels. The background of the video is fixed and simple, while the texture of the images is complex.

3.3. FallDetection DATASET (FDD)

This dataset was recorded with a single uncalibrated Kinect sensor and resized at 320 × 240—the original size was 640 × 480. They collected 21,499 images in total and divided them into training and testing. The total number of images in the training dataset is 16,794, the validation dataset includes 3299 images, and the testing dataset includes 2543 images. The dataset was recorded in five different rooms and from eight different angles. They collected the dataset from five different participants, among them two male participants aged between 32 and 50, and three females aged between 10 and 40. The dataset includes five other classes: sitting, standing, bending, crawling, and lying [67].

4. Proposed Methodology

In this study, we proposed a graph-based spatial-temporal convolution and attention network (GSTCAN), mainly inspired by [34,35,65]. The main goal was to capture the pattern in the spatial domain from the motion version of the skeleton dataset. Our method was mainly a neural network designed with graphs and their structural information. We did not need to use the dimension reduction approach for the proposed model, but it works with the native form. It will overcome the limitation by extracting the complex internal pattern using contextual temporal information. Moreover, we needed a good benchmark dataset with sufficient samples representative of the action's diverse variability and a camera view to prove the power of the model. Many skeleton-based fall detection datasets still need to contain an efficient number of samples, yielding a lack of training. Many existing deep learning-based methods can adapt to the fall detection model [69,70]. These methods are not practical because of the limited size of the publicly available fall and human activity action-related datasets. This data inefficiency problem can be solved by transfer learning, a pre-trained model with a related dataset that can be used as an initial trained model for the novel task. It is highly effective at solving the existing data inefficiency problems. This method was mainly developed from the data reuse concept of learning new things. The main problem is that deep learning mainly works for specific data and domains. There is a need to newly train the model from scratch when needing to apply it to a new task or new domain. In this study, we proposed to develop an attention-based ST-GCN model to recognize fall detection by extracting complex spatial and temporal internal patterns. The working flow architecture of the proposed model is demonstrated in Figure 1, and the pseudocode of the proposed method is described in Algorithm 1.

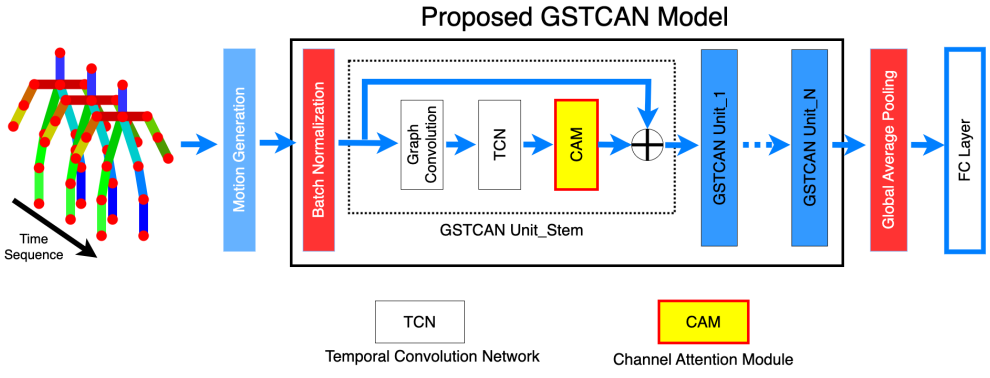

Figure 1. Proposed working flow diagram.

Algorithm 1 Pseudocode of the proposed system.

Input: Set of Input Dataset $P_i \in P(n)$
Number of Samples: N, 70% for Training and 30% for Test
Output: Set of vector s_i
define GSTCAN Model(input=InputLayer, outputs=ClassificationLayer):
$Motion_{Feature} \leftarrow MotionModule(D)$
$Normalize_{Feature} \leftarrow BatchNormalization_{Layer}(Motion_{Feature})$
$GCN_{input} \leftarrow Normalize_{Feature}$
while $i \neq 6$ **do**
$GCN \leftarrow GraphConvolutionalNetwork(GCN_{input})$
$TCN \leftarrow TemporalConvlutionalNetwork(GCN)$
$CAM \leftarrow ChannelAttention(TCN)$
$Features \leftarrow Concatenation(CAM, GCN_{input})$
$GCN_{input} \leftarrow Features$
$GAP \leftarrow GlobalAveragePooling_{Layer}(Features)$
$PredictedClass \leftarrow Classification_{Layer}(GAP)$
$returnPredictedClass$
while $i \neq NumEpochs$ **do**
// For Training
while $Batch \neq NumberBatchTraining$ **do**
$PredictedClass \leftarrow Model(Batch)$
$Loss \leftarrow Criterion(PredictedClass, Train_Class)$
$Updatetheloss \leftarrow Loss.backward(), Optimizer.Step()$
// For Testing
while $Batch \neq NumberBatchTesting$ **do**
$PredictedClass \leftarrow Model(Batch)$
$Output \leftarrow CPerformanceMatrix(PredictedClass, TestClass)$

4.1. AlphaPose Estimation

We used AlphaPose to extract skeleton joints from the fall detection dataset, an open-source library for visual image processing tools. It was developed with a deep learning-based pre-trained model which can be perceptible in real-time for various applications such as face detection, object detection, pose estimation, computer vision, and hand gesture recognition. It was built with custom overflow integration with various Python libraries, such as OpenCV and tensor flow, and we used it here to extract body landmarks for fall detection. There are two main methods for joint point detection in posture estimation: bottom-up and top-down. The bottom-up method estimates all joint points in an image and summarizes the joint points that constitute each person. The bottom-up method is vulnerable because it estimates the pose from local areas. AlphaPose [71,72] uses a top-down framework to detect human bounding boxes and then individually estimates the posture within each box. The detection of each person's joint points is accurate [73]. The Algorithm of AlphaPose we used ResNet101-based Faster R-CNN . Table 2 shows the number of frames for which AlphaPose was able to obtain skeletons for each dataset. In the study, we used those frames' skeleton points which were successfully extracted by AlphaPose and discarded the rest of the frames.

Table 2. Number of frames for which AlphaPose was able to obtain skeletons for each dataset.

Dataset Name	Number of All Frames	Number of Frames for Which the Skeleton Could Be Acquired	Number of Frames for Which the Skeleton Could Not Be Acquired
ImViA	42,066	40,631	1435
UR fall dataset	2995	2142	853
FDD fall dataset	26,911	18,704	8207

This system can read the real-time camera video or recorded video and produce the corresponding skeleton points. In the study, we provided video from the UR fall dataset and generated the skeleton points for consecutive frames. It mainly generates 18 points for each frame, including nose, mouth, ear, shoulder, elbow, wrist, finger index, hip, knee, ankle, and foot, and these points are for both the left and right sides; details of the media pipe skeleton are visualized Figure 2 and Table 3. Although this system collects 18 key points, we selected 13 by excluding eyes and ears.

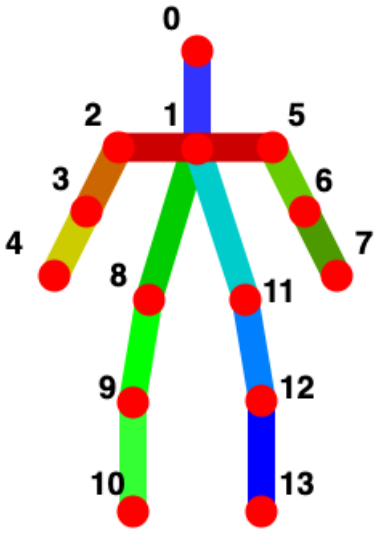

Figure 2. Body skeleton joint visualization generated with AlphaPose.

Table 3. AlphaPose landmarks name with index.

No.	Pose Name	No.	Pose Name	No.	Pose Name
0	Nose	7	Right-wrist	12	Right Knee
1	Neck	8	Left hip	14	Left eye
2	Left shoulder	11	Right hip	15	Left ear
3	Left elbow	9	Left Knee	16	Right eye
4	Left-wrist	12	Right Knee	17	Left ear
5	Right shoulder	10	Left ankle		
6	Right elbow	13	Right ankle		

4.2. Motion Calculation and Graph Construction

We mainly considered the dynamic fall detection dataset in this study; motion is one of the most effective features for the dynamic fall detection approach in terms of movement, alignment, and overall data structure effectiveness. This also directly affects the movement of the fall data. We calculated the motion using all the landmarks for X and Y as a two-dimensional vector. We mainly generated the difference between consecutive frame joint positions to calculate the motion. We calculated the motion for a specific joint by subtracting the consecutive frame joints, which are visualized in Figure 3. To calculate the motion M for a joint j, we used the formula shown in Equation (1).

$$M(j) = \begin{cases} ActivityMotion_X = X_t - X_{t-1} \\ ActivityMotion_Y = Y_t - Y_{t-1}. \end{cases} \quad (1)$$

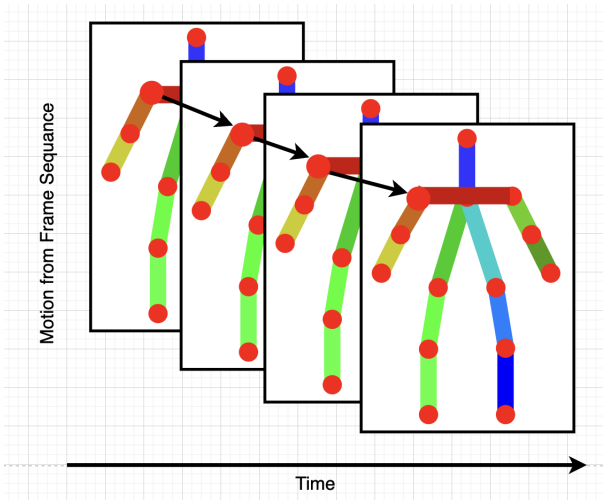

Figure 3. Example visualization of the motion calculation procedure.

Motion skeleton information represents the 2D coordinates of the human joint. In addition, full-body fall and non-fall events use multiple frames based on the sequence of relative structure and samples. The graph was mainly constructed based on the spatial and temporal domains by considering natural bone or connections among the joints. The underreacted graph was constructed using the following Equation (2).

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

Here, V and E denoted the set of nodes and edges where the graph node can be defined as $V = v_(i,t) \mid i = 1, \ldots, N, t = 1, \ldots, T$, which is mainly composed of the whole-body skeleton. After that, we constructed an adjacent matrix based on the graph using the following formulas in Equation (3):

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

4.3. Graph Convolutional Network

The study extracted the potential embedded with the whole-body skeleton based on the spatial-temporal graph convolution network. We construed the graph using the below formulas [21,34]:

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

where D, I, and A represent the diagonal, identity matrix or self-connection, and inter-body connection, respectively. Where the diagonal degree can be expressed as $(A + I)$, the weight matrix is denoted by W. For implementing the graph-based convolution, we focused on the 2D convolution, and for the spatial graph convolution, we multiplied it with $D^{-(1/2)}(A + I)D^{-(1/2)}$. In the same way for the graph-based temporal convolution, we multiplied it with a $k_t \times 1$ kernel size.

4.4. GSTCAN Algorithm

The graph-based spatial temporal convolutional and attention network (GSTCAN) model was proposed here to enhance the work of [34,35,65], and is a GCN [74]-based motion recognition method that automatically learns spatial and temporal patterns from skeleton data. The main advantage of the proposed system is that the data can be treated in its original form [34]. The data based on convolutional networks (CNN) and the data based on the skeleton were obtained in Step 2. The sequence of body joints in two-dimensional form constructs a spatiotemporal graph, with the joints as the nodes of the graph and the natural connections between the structure and time of the human body as the edges of

the graph. The inputs to GSTCAN are the joint coordinate vectors on the graph nodes, and a multi-layered spatiotemporal graph convolution operation is applied to generate higher-order feature maps. Then, whether it is falling or not is classified by the Softmax classifier. Figure 1 shows the overall flow of GSTCAN. Our proposed approach is mainly composed of a series of GSTCN + channel attention module [75] units, a pooling layer, and a fully connected layer. Each unit of the GSTCAN included the spatial and temporal convolutional neural network. Figure 2 demonstrates the node and joints where skeleton joints are considered the graph's node. As we considered dynamic fall detection, there is a sequence of frames that creates the intra-body and inter-body relationships. The intra-body connection comes from the natural connection of the human body joints, and the inter-body connection comes from the relationship between the consecutive frames established by the temporal convolution. We considered the input dimension of the tensor as (N,2, T,33, S). The batch is represented by N, the 2D joint coordinates are represented by 2 (x, y) and can be denoted as channel C, the number of frames are represented by T, the number of the skeletons from the media pipe is 33 and can be denoted by vertex V, and the total number of videos comes from the subject represented by S. After that, we modified S $\times$ N, C, T, V. After calculating the motion of the raw skeleton, we fed it into a spatial convolutional layer, aiming to extract the spatial information for each joint. This process is a little bit different from that of image convolution. Around the specific pixel location, the weight coefficient is multiplied in a spatial order for the image convolution, whereas the labeling process is followed in the GSTCAN with joint location and spatial configuration partitioning approaches. The labels considered here include root node, centripetal nodes near central gravity compared to the root node, and centrifugal nodes. After extracting the spatial features, we fed them into the temporal convolutional network (TCN) to extract temporal contextual information. The main concept of TCN is to calculate the relationship among the same joints in consecutive frames. We repeated the same process 6 times (except stem) consecutively and then applied the pooling layer to enhance the features. Finally, the output layer of the proposed model produced a vector p, which has the same size as the classes. This mainly represents the probability that is the same as the specified corresponding class. The motion of the skeletal points in the graph-based GSTCAN produced a better representation of the fall activity based on the exploitation of the spatial and temporal relationships between intra- and inter-body frame joints.

4.5. Attention Module

According to the skeleton landmark concept, there are some border skeleton points or leaf points, known as the non-connected skeleton. We can solve the problems with Graph CNN because in the graph, all key points are connected with each other through the undirected graph. In addition, we calculated the motion before feeding it into the spatial-temporal architecture. The joint motion and bone motion were calculated between the consecutive frames for each non-connected skeleton point. These motion vectors represent the motion of the non-connected points over time and can capture temporal dynamics. Moreover, our spatiotemporal and attention model can calculate and learn hierarchical representations with temporal dependencies in the long term or short term based on the sequence of non-connected skeleton points of both spatial and temporal features directly from the non-connected skeleton point sequences. We applied attention mechanisms here after the spatial, temporal feature, which can be beneficial for capturing both global and local features in a spatiotemporal model for the non-connected skeleton point, and they can significantly impact model optimization. Our study also included a channel attention model to handle the role of non-connected skeleton issues.

We added an attention mechanism at the end of each GSTCN unit. The added attention mechanism is shown in Figure 4. The layer was used here sequentially and we can define it as (1) GlobalAveragePooling, (2) Dense (N/4), (3) BatchNorm, (4) Dense (N), and (5) Sigmoid. A value between 0 and 1 was output for each channel using the Sigmoid function. Important features had an output of 1 or a value close to 1, and unimportant features

output 0 or a value close to 0. Then, a strong feature graph could be created by multiplying the previously learned feature graph because important features remained.

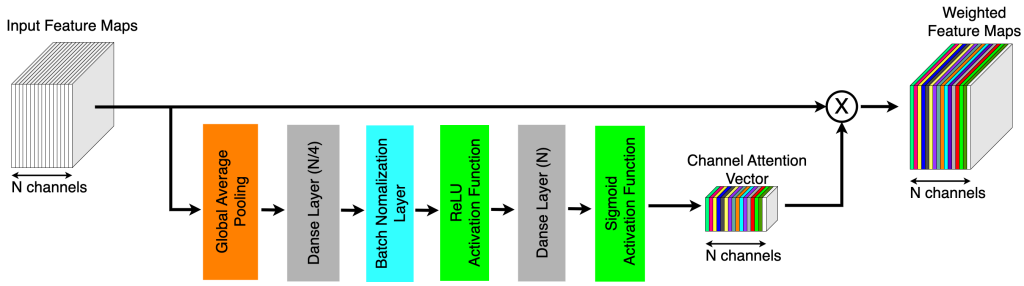

Figure 4. Channel attention mechanism.

4.6. Fully-Coupled Layer

Finally, we considered the Softmax function or Softmax activation layer as a classification or outputs layer to predict the value for each label. The loss function for classification tasks uses a cross-entropy loss function.

4.7. Network Architecture

Figure 1 demonstrates the proposed method, showing that we first calculated the motion and then fed it into a series of N. GSTCAN units in our study, $N = 6$, which leads to reducing the computational complexity. There are 64 output layers for the first two layers, 128 channels for two layers, and 256 output channels for the last two layers. The kernel size for each layer was set as 9, a residual or skip connection, and a dropout rate of 0.5 was used here to overcome the overfitting issues. We refined the feature with an attention module. Finally, we employed the Softmax activation function as a classifier. We employed an RMSprop [76] optimizer to learn the model, with 0.001 as the learning rate value.

5. Experimental Evaluation

To prove the system's superiority and effectiveness, we conducted various experiments with three benchmark datasets. We first demonstrated the training setting and evaluation matrix, then the performance of the proposed model with multiple datasets and, finally, we visualized the state-of-the-art comparison table.

5.1. Training Setting

To divide the training and testing, we followed the three-fold cross-validation approaches. In the training process, we used a learning rate of 0.001 and a batch size of 32. To implement the system, we used a GPU machine that has CUDA version 11.7, NVIDIA driver version 515, and GPU Geforce RTX 3090 24GB, with RAM 32 GB. Models were run for 100 epochs with the optimizer RMSprop [76] with the RTX3090. We also used Pytorch (version–1.13.1) [77], which has a low computational cost for deep learning, attention, transformer, OpenCV (version–4.7.0.72), pickle, and csv packages for the initial processing [78,79].

5.2. Evaluation Matrices

We used three benchmark datasets to evaluate the proposed model, mainly seen as a binary class classification problem. The evaluation metrics which we used here are included below [10]:

- Time: Perform a *t*-test from the mean and variance of processing speeds;
- Accuracy: most researchers used this, which denotes the percentage of total items classified correctly = (TP + TN)/(TP + TN + FP + FN);
- Recall/sensitivity: mainly denotes the true positive rate TP/(TP + FN);
- F1-score: denotes the harmonic mean between precision and recall = TP/(TP + 1/2 (FP + FN));

where TP comes from the true positives in our cases–the activity labeled is fall, and the system predicted it as a fall. FP denotes the false positives—the actual class is non-fall, but the system predicted a fall. TN denotes the true negatives—the actual class label is non-fall, and the system predicted non-fall. FN denotes the false negatives; the actual class label is fall, but the system predicted non-fall.

5.3. Evaluation Matrices

The processing speed of the ST-GCN model and the proposed system were compared. We evaluated the proposed system with three datasets, and the tables below visualize the proposed model's performance accuracy. Using the ImViA dataset, our proposed model achieved 99.57%, 99.68%, 99.63% and 99.93% for precision, sensitivity, F1-score, and accuracy, respectively. The UR fall detection dataset achieved 99.87%, 97.36%, 98.56%, and 99.75% for precision, sensitivity, f-score, and accuracy, respectively. In the same way, for the FDD dataset, our model achieved 97.98%, 97.21%, 97.55%, and 99.12% with precision, sensitivity, F-score, and accuracy, respectively.

5.3.1. Processing Speed Comparison

The mean and variance of the processing speeds of the ST-GCN and the proposed model are shown in Table 4. *T*-test results reject the hypothesis that the processing speeds are equal, indicating that the proposed model is more efficient.

Table 4. The mean and variance of the processing speeds.

Algorithm	Mean [sec]	Variance
ST-GCN	7.511	0.0001876
Proposed GSTCAN Model	6.787	0.0003385

5.3.2. Performance Result and State-of-the-Art Comparison for the UR Fall Dataset

The class-wise evaluation matrix table of the proposed model with the UR fall detection dataset is demonstrated in Table 5. We can see that the fall class reported 100%, 94.72%, and 97.25% for precision, sensitivity, and F1-score, respectively. In the same way for the non-fall class label, it achieved 99.73%, 100%, and 99.86% scores for precision, sensitivity, and F1-score, respectively. It also showed the performance accuracy for the all-class label simultaneously, which is 99.74%, 99.86%,97.36%, and 98.55% scores for accuracy precision, sensitivity, and F1-score, respectively.

Table 5. Class wise precision, sensitivity, and F1-score for UR fall dataset.

Label Name	Accuracy [%]	Precision [%]	Sensitivity [%]	F1-Score [%]
Fall	–	100	94.72	97.25
NonFall	–	99.73	100	99.86
Average	99.74	99.86	97.36	98.55

The state-of-the-art comparison for the proposed model is shown in Table 6 with the UR fall detection system. In this state-of-the-art comparisons table, we included the accuracy, precision, sensitivity, specificity, and F-score for a fair comparison with the previous model. We included the performances of seven previous state-of-the-art methods for the UR fall dataset. The authors of [36,37] extracted the hand-crafted features from the skeleton and depth information and, using the SVM method, they achieved 94.28% and 96.55% accuracies, respectively. The author of [60] employed a CNN-based encoder and decoder system and reported 90.50% accuracy. The author of [20] used mask-RCNN to

segment and extract the features from the fall event video dataset and applied bi-directional LSTM and achieved 96.70% accuracy with the UR fall dataset. Zheng et al. extracted the skeleton points using AlphaPose, then employed ST-GCN and achieved 97.28%, 97.15%, 97.43%, 97.30%, and 97.29% scores for accuracy [65], precision, sensitivity, specificity, and F1-score, respectively. Wang et al. [80] extracted OpenPose key points and then applied MLP (multilayer perceptron) and random forest for the classification and achieved a high-performance accuracy of 97.33%. In the same way, the author of [61] applied the GLR scheme to design the system and achieved 96.66% accuracy.

Table 6. State -of-the-art comparison for UR fall dataset.

Algorithm	Dataset	Accuracy [%]	Precision [%]	Sensitivity [%]	Specificity [%]	F-Score [%]
Depth +SVM [36]	UR Fall	94.28	n/a	n/a	n/a	n/a
Skeleton +SVM [37]	UR Fall	96.55	n/a	n/a	n/a	n/a
HCAE [60]	UR Fall	90.5%	n/a	n/a	n/a	n/a
Bi-Directional LSTM [20]	UR Fall	96.70%	n/a	n/a	n/a	n/a
Hontago [65]	UR Fall	97.28	97.15	97.43	97.30	97.29
Wang [80]	UR Fall	97.33	97.78	97.78	96.67	97.78
Harrou [61]	UR Fall	96.66	94	100	94.93	96.91
Proposed GSTCAN Model	UR Fall	99.74	99.86	97.36	n/a	98.55

5.3.3. Performance Result and State-of-the-Art Comparison of the ImViA Dataset

In this section, we compared the performance of the proposed model with the state-of-the-art model. Table 7 demonstrates the state-of-the-art comparison for the ImViA dataset, where the proposed model achieved 99.93% accuracy whereas the previous model reported 96.86% accuracy, proving that our model has high effectiveness and efficiency.

Table 7. Class wise precision, sensitivity, and F1-score for ImViA fall dataset.

Label Name	Accuracy [%]	Precision [%]	Sensitivity [%]	F1-Score [%]
Fall	n/a	99.17	99.39	99.28
NonFal	n/a	99.96	99.95	99.96
Average	99.93	99.57	99.67	99.92

The state-of-the-art comparison of the proposed model using the ImViA dataset is shown in Table 8. Wang et al. [80] extracted OpenPose key points and then applied MLP (multilayer perceptron) and random forest for the classification and achieved a high-performance accuracy of 96.91%. Chalme et al. [81] demonstrated a performance of 79.31%, 79.41%, 83.47%, 73.07% and 81.39% accuracies. This accuracy proved that we could evaluate the system.

Table 8. State-of-the-art comparison for ImViA fall dataset.

Algorithm	Dataset	Accuracy [%]	Precision [%]	Sensitivity [%]	Specificity [%]	F-Score [%]
Hontago [65]	ImViA	96.86	97.01	96.71	96.81	96.77
Wang [80]	ImViA	96.91	97.65	96.51	97.37	97.08
Chamle [81]	ImViA	79.31	79.41	83.47	73.07	81.39
Proposed GSTCAN Model	ImViA	99.93	99.57	99.67	n/a	99.92

5.3.4. Performance Result and State-of-Art Comparison for the FDD Fall Dataset

In this section, we compared the performance of the proposed model with the state-of-the-art model. Table 9 demonstrates the state-of-the-art comparison performance for the ImViA dataset. The table shows that the FDD dataset has five rows including accuracy, precision, sensitivity, and F1-score.

Table 9. Class wise precision, sensitivity, and F1-score for FDD fall dataset.

Label Name	Accuracy [%]	Precision [%]	Sensitivity [%]	F1-Score [%]
Sitting	n/a	99.54	99.49	99.52
Standing	n/a	98.96	99.04	98.99
Bending	n/a	98.01	99.00	98.50
Crawling	n/a	96.64	89.92	93.09
Lying	n/a	96.75	98.60	97.63
Total	99.12	97.98	97.21	97.55

A state-of-the-art comparison for the FDD dataset is demonstrated in Table 10. The author of [61] applied the GLR scheme to design the system and achieved 96.6% accuracy for the FDD dataset, whereas our proposed method achieved 99.22% accuracy using the FDD dataset.

Table 10. State-of-the-art comparison for FDD fall dataset.

Algorithm	Dataset	Accuracy [%]	Precision [%]	Sensitivity [%]	F-Score[%]
Haroo [61]	FDD	96.84			
Proposed GSTCAN Model	FDD	99.12	97.98	97.21	97.55

5.4. Deliberation

In this study, we proposed using AlphaPose to extract and select the skeleton data points instead of the RGB image. Then, we constructed an undirected graph and applied a graph-based CNN like GSTCAN. The positional relationships between some of the non-real connected points are very helpful for partially identifying events. We proposed a graph-based spatial-temporal convolution and attention network (GSTCAN) model to overcome the current challenges and developed an advanced medical technology system. In the procedure, we first calculated the motion among the consecutive frames, then constructed a graph and applied a graph convolutional neural network (GCN). We repeated the same procedure six times as GSTCAN and then applied it to the fully connected layer. To improve the role of non-connected skeleton points in certain events during the spatial-temporal feature, we applied the attention model with GSTCAN, aiming to extract global and local

features which are bound to impact model optimization. Finally, we applied a sigmoid function as a classifier and achieved a high accuracy of 99.93%, 99.74%, and 99.12% for ImViA, UR-Fall, and FDD datasets, respectively. The high-performance accuracy with the three datasets proved the superiority and efficiency of the proposed system. According to comparison Tables 5, 7, and 9, we can say that our all datasets can be considered balanced, because our method achieved high precision, sensitivity, and F-score as well as high-performance accuracy. The state-of-the-art comparison Tables 6, 8, and 10 demonstrated high performance of the proposed model for all three fall-event datasets compared to the existing state-of-the-art systems. In addition, the existing fall detection systems achieved lower performance accuracy with various models, which sometimes need high computational complexity. Our proposed system generated better performance accuracy than the hand-crafted feature and machine learning algorithms with lower computational complexity than the state-of-the-art systems. Based on the state-of-the-art comparison table, we can see that the high performance of our method with the three datasets proves the proposed system's superiority in terms of performance and efficiency. We can differentiate our model with the following: (a) It can effectively detect the motion of the fall events; (b) It achieved a more than 5% higher performance accuracy compared to the existing work; (c) It takes less time compared to the existing work because we efficiently used fewer GSTCAN units. We can conclude that our model is suitable for discriminating fall events from human activity-based video datasets with a small cost of average classification rate.

6. Conclusions

This paper proposed a graph-based spatial-temporal convolution and attention network (GSTCAN) model to extract intra- and inter-frame joint relationships to improve performance accuracy and efficiency to confirm whether a person has fallen. To emphasize the role of a non-skeleton joint, we employed a modified channel attention model to the GSTCAN feature for selecting the channel-wise effective feature. We achieved higher accuracy than the existing models on the two datasets. Our model achieved high-performance accuracy for the three benchmark fall event datasets. The high-performance accuracy with less complexity proved the superiority and efficiency of the proposed model. In the future, we plan to increase the number of hand-crafted features with spatial-temporal features to reduce the number of parameters of the model to achieve high performance with a low computational cost and apply this model to the field of movement disorder detection. In the future, we will train the model with human action recognition datasets, aiming to make it a pre-trained model for human action recognition as well as fall detection [42,82].

Author Contributions: Conceptualization, R.E., K.H. and A.S.M.M.; methodology, R.E., K.H., A.S.M.M., Y.T. and J.S.; investigation, R.E., A.S.M.M. and J.S.; data curation, R.E., A.S.M.M., Y.T. and J.S.; writing—original draft preparation, R.E., A.S.M.M. and J.S.; writing—review and editing, A.S.M.M., Y.T. and J.S.; visualization, R.E. and A.S.M.M.; supervision, J.S.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI Grant Number JP23H03477.

Data Availability Statement: ImViA dataset is accessible at https://imvia.u-bourgogne.fr/en/database/fall-detection-dataset-2.html. UR fall dataset can be found at http://fenix.ur.edu.pl/~mkepski/ds/uf.html. FDD fall dataset is accessible at https://falldataset.com/.

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

References

1. United Nations. *World Population Ageing 2020: Highlights: Living Arrangements of Older Persons*; United Nations Department of Economic and Social Affairs: New York, NY, USA, 2021.
2. Zahedian-Nasab, N.; Jaberi, A.; Shirazi, F.; Kavousipor, S. Effect of virtual reality exercises on balance and fall in elderly people with fall risk: A randomized controlled trial. *BMC Geriatr.* **2021**, *21*, 509. [CrossRef] [PubMed]
3. Lord, S.R. Visual risk factors for falls in older people. *Age Ageing* **2006**, *35*, ii42–ii45. [CrossRef] [PubMed]

4. 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: 19th International Conference on Ad-Hoc Networks and Wireless, ADHOC-NOW 2020, Bari, Italy, 19–21 October 2020; Springer: Cham, Switzerland, 2020; pp. 125–130.
5. Gutiérrez, J.; Rodríguez, V.; Martin, S. Comprehensive review of vision-based fall detection systems. *Sensors* **2021**, *21*, 947. [CrossRef]
6. Lu, K.L.; Chu, E.T.H. An image-based fall detection system for the elderly. *Appl. Sci.* **2018**, *8*, 1995. [CrossRef]
7. 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), Boston, MA, USA, 21–24 October 2018; pp. 1–4.
8. Dong, S.; Wang, P.; Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
9. 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. [CrossRef]
10. 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. *Appl. Sci.* **2022**, *12*, 3933. [CrossRef]
11. Miah, A.S.M.; Shin, J.; Al Mehedi Hasan, M.; Rahim, M.A.; Okuyama, Y. Rotation, Translation And Scale Invariant Sign Word Recognition Using Deep Learning. *Comput. Syst. Sci. Eng.* **2023**, *44*, 2521–2536. [CrossRef]
12. Shin, J.; Musa Miah, A.S.; Hasan, M.A.M.; Hirooka, K.; Suzuki, K.; Lee, H.S.; Jang, S.W. Korean Sign Language Recognition Using Transformer-Based Deep Neural Network. *Appl. Sci.* **2023**, *13*, 3029. [CrossRef]
13. Rahim, M.A.; Miah, A.S.M.; Sayeed, A.; Shin, J. Hand gesture recognition based on optimal segmentation in human-computer interaction. In Proceedings of the 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII), Kaohsiung, Taiwan, 21–23 August 2020; pp. 163–166.
14. Miah, A.S.M.; Hasan, M.J.S.L.H.S.J. Multi-Stream General and Graph-Based Deep Neural Networks for Skeleton-Based Sign Language Recognition. *Electronics* **2023**, *12*, 2841. [CrossRef]
15. Ren, L.; Peng, Y. Research of fall detection and fall prevention technologies: A systematic review. *IEEE Access* **2019**, *7*, 77702–77722. [CrossRef]
16. Xu, T.; Zhou, Y. Elders' fall detection based on biomechanical features using depth camera. *Int. J. Wavelets Multiresolution Inf. Process.* **2018**, *16*, 1840005. [CrossRef]
17. De Quadros, T.; Lazzaretti, A.E.; Schneider, F.K. A movement decomposition and machine learning-based fall detection system using wrist wearable device. *IEEE Sensors J.* **2018**, *18*, 5082–5089. [CrossRef]
18. Kibria, K.A.; Noman, A.S.; Hossain, M.A.; Islam Bulbul, M.S.; Rashid, M.M.; Musa Miah, A.S. Creation of a Cost-Efficient and Effective Personal Assistant Robot using Arduino & Machine Learning Algorithm. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 5–7 June 2020; pp. 477–482. [CrossRef]
19. Rubenstein, L.Z. Falls in older people: Epidemiology, risk factors and strategies for prevention. *Age Ageing* **2006**, *35*, ii37–ii41. [CrossRef] [PubMed]
20. 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. [CrossRef]
21. 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**, *11*, 4703–4716. [CrossRef]
22. Gasparrini, S.; Cippitelli, E.; Gambi, E.; Spinsante, S.; Wåhslén, J.; Orhan, I.; Lindh, T. Proposal and Experimental Evaluation of Fall Detection Solution Based on Wearable and Depth Data Fusion. In *Proceedings of the ICT Innovations 2015: Emerging Technologies for Better Living;* Springer: Berlin/Heidelberg, Germany, 2016; pp. 99–108.
23. Maddalena, L.; Petrosino, A. Background subtraction for moving object detection in RGBD data: A survey. *J. Imaging* **2018**, *4*, 71. [CrossRef]
24. Kreković, M.; Čerić, P.; Dominko, T.; Ilijaš, M.; Ivančić, K.; Skolan, V.; Šarlija, J. A method for real-time detection of human fall from video. In Proceedings of the 2012 Proceedings of the 35th International Convention MIPRO, Opatija, Croatia, 21–25 May 2012; pp. 1709–1712.
25. El Baf, F.; Bouwmans, T.; Vachon, B. Type-2 Fuzzy Mixture of Gaussians Model: Application to Background Modeling. In *Proceedings of Advances in Visual Computing*; Springer: Berlin/Heidelberg, Germany, 2008; Number Part I; pp. 772–781.
26. Guo, H.; Qiu, C.; Vaswani, N. An online algorithm for separating sparse and low-dimensional signal sequences from their sum. *IEEE Trans. Signal Process.* **2014**, *62*, 4284–4297. [CrossRef]
27. Dong, S.; Li, R. Traffic identification method based on multiple probabilistic neural network model. *Neural Comput. Appl.* **2019**, *31*, 473–487. [CrossRef]
28. Miah, A.S.M.; Mamunur Rashid, M.; Redwanur Rahman, M.; Tofayel Hossain, M.; Shahidujjaman Sujon, M.; Nawal, N.; Hasan, M.; Shin, J. Alzheimer's Disease Detection Using CNN Based on Effective Dimensionality Reduction Approach. In *Proceedings of the Intelligent Computing and Optimization*; Vasant, P., Zelinka, I., Weber, G.W., Eds.; Springer: Cham, Switzerland, 2021; pp. 801–811.
29. Kafi, H.M.; Miah, A.S.M.; Shin, J.; Siddique, M.N. A Lite-Weight Clinical Features Based Chronic Kidney Disease Diagnosis System Using 1D Convolutional Neural Network. In Proceedings of the 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), Gazipur, Bangladesh, 24–26 February 2022; pp. 1–5. [CrossRef]

30. Bouwmans, T.; Javed, S.; Sultana, M.; Jung, S.K. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Netw.* **2019**, *117*, 8–66. [CrossRef]
31. Maldonado-Bascon, S.; Iglesias-Iglesias, C.; Martín-Martín, P.; Lafuente-Arroyo, S. Fallen people detection capabilities using assistive robot. *Electronics* **2019**, *8*, 915. [CrossRef]
32. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the 30 th International Conference on Ma- chine Learning, Atlanta, GA, USA, 16–21 June 2013; Volume 30, p. 3.
33. Al Nahian, M.J.; Ghosh, T.; Al Banna, M.H.; Aseeri, M.A.; Uddin, M.N.; Ahmed, M.R.; Mahmud, M.; Kaiser, M.S. Towards an accelerometer-based elderly fall detection system using cross-disciplinary time series features. *IEEE Access* **2021**, *9*, 39413–39431. [CrossRef]
34. Yan, S.; Xiong, Y.; Lin, D. Spatial, temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
35. Keskes, O.; Noumeir, R. Vision-based fall detection using st-gcn. *IEEE Access* **2021**, *9*, 28224–28236. [CrossRef]
36. Kwolek, B.; Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Comput. Methods Programs Biomed.* **2014**, *117*, 489–501. [CrossRef] [PubMed]
37. 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. *J. Imaging* **2021**, *7*, 109. [CrossRef]
38. Charfi, I.; Miteran, J.; Dubois, J.; Atri, M.; Tourki, R. Optimised spatio-temporal descriptors for real-time fall detection: Comparison of SVM and Adaboost based classification. *J. Electron. Imaging JEI* **2013**, *22*, 17.
39. Mubashir, M.; Shao, L.; Seed, L. A survey on fall detection: Principles and approaches. *Neurocomputing* **2013**, *100*, 144–152. [CrossRef]
40. Miah, A.S.M.; Ahmed, S.R.A.; Ahmed, M.R.; Bayat, O.; Duru, A.D.; Molla, M.I. Motor-Imagery BCI Task Classification Using Riemannian Geometry and Averaging with Mean Absolute Deviation. In Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 24–26 April 2019; pp. 1–7. [CrossRef]
41. Liu, H.; Hartmann, Y.; Schultz, T. Motion Units: Generalized Sequence Modeling of Human Activities for Sensor-Based Activity Recognition. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 1506–1510. [CrossRef]
42. Liu, H.; Gamboa, H.; Schultz, T. Sensor-Based Human Activity and Behavior Research: Where Advanced Sensing and Recognition Technologies Meet. *Sensors* **2023**, *23*, 125. [CrossRef]
43. Miah, A.S.M.; Rahim, M.A.; Shin, J. Motor-imagery classification using Riemannian geometry with median absolute deviation. *Electronics* **2020**, *9*, 1584. [CrossRef]
44. Miah, A.S.M.; Islam, M.R.; Molla, M.K.I. EEG classification for MI-BCI using CSP with averaging covariance matrices: An experimental study. In Proceedings of the 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 11–12 July 2019; pp. 1–5.
45. Joy, M.M.H.; Hasan, M.; Miah, A.S.M.; Ahmed, A.; Tohfa, S.A.; Bhuaiyan, M.F.I.; Zannat, A.; Rashid, M.M. Multiclass mi-task classification using logistic regression and filter bank common spatial patterns. In Proceedings of the Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, 26–27 March 2020; Revised Selected Papers; Springer: Berlin/Heidelberg, Germany, 2020; pp. 160–170.
46. Zobaed, T.; Ahmed, S.R.A.; Miah, A.S.M.; Binta, S.M.; Ahmed, M.R.A.; Rashid, M. Real time sleep onset detection from single channel EEG signal using block sample entropy. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *928*, 032021. [CrossRef]
47. Miah, A.S.M.; Islam, M.R.; Molla, M.K.I. Motor imagery classification using subband tangent space mapping. In Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2017; pp. 1–5. [CrossRef]
48. Kabir, M.H.; Mahmood, S.; Al Shiam, A.; Musa Miah, A.S.; Shin, J.; Molla, M.K.I. Investigating Feature Selection Techniques to Enhance the Perfor-mance of EEG-Based Motor Imagery Tasks Classification. *Mathematics* **2023**, *11*, 1921. [CrossRef]
49. Miah, A.S.M.; Mouly, M.A.; Debnath, C.; Shin, J.; Sadakatul Bari, S. Event-Related Potential Classification Based on EEG Data Using xDWAN with MDM and KNN. In *Proceedings of the International Conference on Computing Science, Communication and Security*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 112–126.
50. Miah, A.S.M.; Shin, J.; Hasan, M.A.M.; Molla, M.K.I.; Okuyama, Y.; Tomioka, Y. Movie Oriented Positive Negative Emotion Classification from EEG Signal using Wavelet transformation and Machine learning Approaches. In Proceedings of the 2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Penang, Malaysia, 19–22 December 2022; pp. 26–31.
51. Miah, A.S.M.; Shin, J.; Islam, M.M.; Molla, M.K.I. Natural Human Emotion Recognition Based on Various Mixed Reality (MR) Games and Electroencephalography (EEG) Signals. In Proceedings of the 2022 IEEE 5th Eurasian Conference on Educational Innovation (ECEI), Taipei, Taiwan, 10–12 February 2022; pp. 408–411.
52. Daniela, M.; Marco, M.; Paolo, N. UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones. *Appl. Sci.* **2017**, *7*, 1101.
53. Wang, J.; Zhang, Z.; Li, B.; Lee, S.; Sherratt, R.S. An enhanced fall detection system for elderly person monitoring using consumer home networks. *IEEE Trans. Consum. Electron.* **2014**, *60*, 23–29. [CrossRef]

54. 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), Greater Noida, India, 2–4 October 2020; pp. 502–505.
55. Xu, T.; Zhou, Y.; Zhu, J. New advances and challenges of fall detection systems: A survey. *Appl. Sci.* **2018**, *8*, 418. [CrossRef]
56. Tian, Y.; Lee, G.H.; He, H.; Hsu, C.Y.; Katabi, D. RF-based fall monitoring using convolutional neural networks. *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.* **2018**, *2*, 1–24. [CrossRef]
57. Xue, T.; Liu, H. Hidden Markov Model and Its Application in Human Activity Recognition and Fall Detection: A Review. In *Proceedings of the Communications, Signal Processing, and Systems*; Liang, Q., Wang, W., Liu, X., Na, Z., Zhang, B., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; pp. 863–869.
58. Zerrouki, N.; Houacine, A. Combined curvelets and hidden Markov models for human fall detection. *Multimed. Tools Appl.* **2017**, *77*, 6405–6424. [CrossRef]
59. Chua, J.L.; Chang, Y.C.; Lim, W.K. A simple vision-based fall detection technique for indoor video surveillance. *Signal Image Video Process.* **2015**, *9*, 623–633. [CrossRef]
60. 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. [CrossRef]
61. Harrou, F.; Zerrouki, N.; Sun, Y.; Houacine, A. An integrated vision-based approach for efficient human fall detection in a home environment. *IEEE Access* **2019**, *7*, 114966–114974. [CrossRef]
62. 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. [CrossRef]
63. Yao, L.; Yang, W.; Huang, W. An improved feature-based method for fall detection. *Teh. Vjesn.* **2019**, *26*, 1363–1368.
64. 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. [CrossRef]
65. Zheng, H.; Liu, Y. Lightweight fall detection algorithm based on AlphaPose optimization model and ST-GCN. *Math. Probl. Eng.* **2022**, *2022*, 9962666. [CrossRef]
66. Tran, T.T.H.; Le, T.L.; Morel, J. An analysis on human fall detection using skeleton from Microsoft Kinect. In Proceedings of the 2014 IEEE Fifth International Conference on Communications and Electronics (ICCE), Danang, Vietnam, 30 July–1 August 2014; pp. 484–489.
67. Adhikari, K.; Bouchachia, H.; Nait-Charif, H. Activity recognition for indoor fall detection using convolutional neural network. In Proceedings of the 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017; pp. 81–84. [CrossRef]
68. Pathak, D.; Bhosale, V. Fall Detection for Elderly People in Indoor Environment using Kinect Sensor. *Nternational J. Sci. Res.* **2015**, *6*, 1956–1960.
69. Hwang, S.; Ahn, D.; Park, H.; Park, T. Maximizing accuracy of fall detection and alert systems based on 3D convolutional neural network. In Proceedings of the Second International Conference on Internet-of-Things Design and Implementation, Pittsburgh, PA, USA, 18–21 April 2017; pp. 343–344.
70. Fakhrulddin, A.H.; Fei, X.; Li, H. Convolutional neural networks (CNN) based human fall detection on body sensor networks (BSN) sensor data. In Proceedings of the 2017 4th International Conference on Systems and informatics (ICSAI), Hangzhou, China, 1–13 November 2017; pp. 1461–1465.
71. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
72. Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C. Pose Flow: Efficient online pose tracking. *arXiv* **2018**, arXiv:1802.00977.
73. 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 Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7157–7173. [CrossRef]
74. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Networks Learn. Syst.* **2020**, *32*, 4–24. [CrossRef] [PubMed]
75. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
76. Tieleman, T.; Hinton, G. Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. *COURSERA Neural Netw. Mach. Learn.* **2012**, *17*, 26–31.
77. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
78. Gollapudi, S. *Learn Computer Vision Using OPENCV*; Springer: Berlin/Heidelberg, Germany, 2019.
79. Dozat, T. Incorporating Nesterov Momentum into Adam. 2016. Available online: https://openreview.net/pdf?id=OM0jvwB8jIp57ZJjtNEZ (accessed on 21 June 2023).
80. 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. [CrossRef]

81. Chamle, M.; Gunale, K.; Warhade, K. Automated unusual event detection in video surveillance. In Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 26–27 August 2016; Volume 2, pp. 1–4.
82. Hartmann, Y.; Liu, H.; Schultz, T. Interactive and Interpretable Online Human Activity Recognition. In Proceedings of the 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), Pisa, Italy, 21–25 March 2022.