This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

MSS-Former: Multi-Scale Skeletal Transformer for Intelligent Fall Risk Prediction in Older Adults

Qizheng Zhao, Xiaomao Fan, Manting Chen, Yutian Xiao, Xuan Wang
Eric Hiu Kwong Yeung, Kwok Leung Tsui and Yang Zhao*

***Abstract*—Fall, a leading cause of accidental death and injury in older adults aged 65 and above, has become a rapidly growing health concern in aging populations worldwide. Data-driven methods integrating depth imaging technology have received growing attention in automated fall risk assessment owing to their non-invasiveness and less dependence on healthcare professionals. However, most existing depth image data-based models neglect the inherent physiological and potential functional connections and lack sufficient real-world data validation. To fill the research gap, we developed a novel approach named Multi-scale Skeletal Transformer (MSS-Former), leveraging depth image technology and deep learning models for effective fall risk prediction. Our contributions mainly consist of four parts. First, we introduced a multi-model output feature fusion transformer in fall risk prediction, enabling output merging and weighting from multiple model streams dynamically. Second, we developed an innovative scheme to construct inter-joint skeletal topology, systematically focusing on joints' intrinsic physiological and potential functional connections. Third, we constructed a ResNet-FPN, greatly enhancing multi-scale feature extraction capabilities. Fourth, we conducted a field study in a local hospital and performed a comprehensive validation of our developed approach. The comparison results show that our approach achieved outstanding predictive performance, surpassing state-of-the-art methods on the real-world dataset, with accuracy, precision, recall, and F1 scores of 97.84%, 97.33%, 96.97%, and 96.92%, respectively. In practice, the proposed approach would be of great value in the timely identification for individuals at high fall risk, and facilitate decision-making to take appropriate interventions.**

***Index Terms*—Older adults; Depth images; Fall risk prediction; Graph convolutional networks; Skeleton attention**

I. Introduction

THE rapid aging of the global population is intensifying healthcare challenges. As reported by The World Health Organization (WHO), it is projected that the proportion of adults aged 65 and above is projected to rise from 10% to 16% by 2050 [1]. This is expected to result in a more severe disease burden, particularly falls, which are the primary cause of injury-related deaths among the elderly population [2]. Effective fall risk prediction methods are an urgent necessity, enabling the timely identification of high-fall-risk individuals, guiding intervention strategies, and mitigating the impact of older adults' falls.

In recent years, sensor-based assessment methods have emerged as a prominent approach among various technologies for predicting fall risk. Notably, methods based on depth imaging sensors have significantly captured attention. The general modeling process for fall risk prediction based on sensor data is typically data-driven. These models usually work by first extracting features that quantify gait dynamics and balance capability, and then employing statistical or machine learning methods for predictive model development [3]. Currently, proven track records of deep learning-based models have been seen in many applications of motion recognition. Similar ideas can be developed for fall risk prediction in older adults given their common reliance on skeletal joint data generated from depth images. From Convolutional Neural Networks (CNNs) to 3D-CNNs integrated Long Short-Term Memory networks (LSTMs), deep learning methods have been widely adopted to model the temporal dynamics of skeletal movement. To better capture the spatiotemporal characteristics of skeletal motion, models such as Spatio-Temporal Graph Convolutional Networks (STGCN) and its variants, including Dynamic Group Spatio-Temporal GCN (DG-STGCN) [4], an STGCN architecture for skeleton-based human action recognition (STAR) [5], and multi-stream, multi-scale dilated spatial-temporal graph convolutional network (2M-STGCN) [6], have been proposed to enhance comprehensive skeletal feature capture over the STGCN model.

Despite the successes of deep learning in skeletal data-based posture recognition and prediction, current models commonly concentrate solely on the physiological structural topology of the skeleton, neglecting the functional interconnections between joints. In the domain of fall risk, correlations between joints reflect the stability of gait [7]. For example, fall risk classifications based on the Berg Balance Scale (BBS) reveal distinct patterns: individuals at high fall risk (BBS $\leq$ 40) exhibit stronger lower limb correlations, while those at lower fall risk (BBS $>$ 40) show more upper limb correlations (Figure 1). These insights underscore the importance of analyzing inter-joint correlations. However, adaptive graph models still face challenges. These models, as exemplified in Yang et al.'s study

This research was funded by the the National Key Research and Development Program of China, grant number 2023YFC2307305, the Shenzhen Medical Research Fund, grant number A2301041, the Shen Zhen–Hong Kong–Macao Science and Technology Project Fund, grant number SGDX20210823103403028. (*Corresponding author: Yang Zhao)
Qizheng Zhao, Manting Chen, Yutian Xiao, Xuan Wang, and Yang Zhao* are with the Intelligent Sensing and Proactive Health Research Center, School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen 518000, China (e-mail: zhaoqzh@mail2.sysu.edu.cn; chenmt33@mail2.sysu.edu.cn; xiaoyt29@mail2.sysu.edu.cn; wangx789@mail2.sysu.edu.cn; zhaoy393@mail.sysu.edu.cn).
Xiaomao Fan is with the College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China (e-mail: astrofan2008@gmail.com).
Eric Hiu Kwong Yeung is with the Department of Physiotherapy, The University of Hong Kong-Shenzhen Hospital, Shenzhen 518000, China (e-mail: yangxg@hku-szh.org).
Kwok Leung Tsui is with the Grado Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, Virginia, USA (e-mail: kltsui@vt.edu).

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

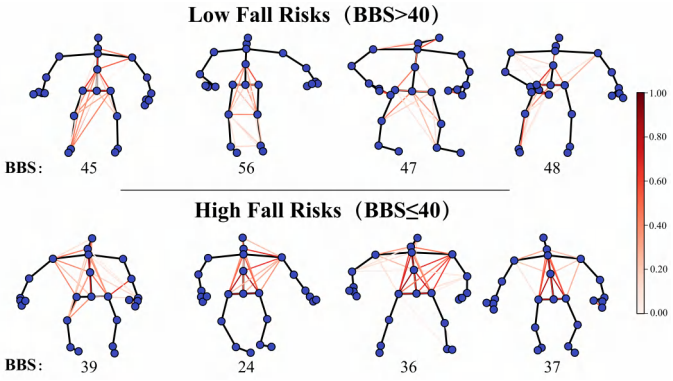

Fig. 1: Pearson's correlation between skeletal joints of participants with different BBS scores. The points represent skeletal joints, the black lines represent the connections connecting the joints, and the red lines indicate the strength of Pearson's correlation between joints, with darker colors representing stronger correlations.

on pseudo graph convolutional networks [8], attempt to address the overlooked functional interconnections between joints, still face challenges in effectively capturing these complex dynamics for accurate fall risk prediction.

Furthermore, existing deep learning models based on skeletal data still face three challenges that have not been fully addressed: i) *How to identify pixel-level spatial details in skeletal data?* While STGCN models excel at mapping the spatial structure of skeletal data beyond the capabilities of traditional CNNs, the spatial information they capture is of a superpixel-level nature represented as non-Euclidean feature. Therefore, STGCN struggles to accurately extract pixel-level information as they generally neglect the Euclidean distances between data nodes [6], [9], [10]. This limitation in detail recognition may overlook subtle yet critical differences in gait or movement associated with fall risks. ii) *How to achieve appropriate multi-model feature fusion?* To recognize more comprehensive skeletal features, cascading multiple models for feature extraction has been explored. However, the fusion techniques employed in existing studies to combine these models' outputs are often rudimentary, typically defaulting to basic or predetermined weighted fusion methods [11], [12], thereby neglecting the importance of different model outputs. iii) How to perform physiologically meaningful interpretability analysis of the model's results: Analyzing human skeletal data essentially involves understanding the physiological functions and patterns of human movement [13]. Therefore, the interpretability of models based on skeletal data is crucial. However, the inner workings of state-of-the-art (SOTA) learning methods based on skeletal data still remain mostly black-box [14]. iv) *How to validate model efficacy and clinical relevance in real-world scenarios?* Most existing models are trained and validated on public datasets, leaving their real-world efficacy and clinical significance unknown.

In this study, we develop a novel approach, a multi-scale skeletal transformer (MSS-Former) grounded in a cascaded multi-scale feature fusion network, to overcome these challenges. MSS-Former enables integrating pixel-level spatial features discerned by CNNs with a self-attention mechanism, thereby redefining the fusion of multi-model output features. Our approach begins by examining the interplay of skeletal joints in STGCN, forming a correlation-based topology for allocating attention weights. Concurrently, we leverage the multi-scale feature extraction capabilities of skeletal data via a pre-trained ResNet-FPN module. Together, these strategies enable our model to unveil a holistic spectrum of physiological-functional-multiscale features. We also conducted a pilot study in the rehabilitation department of a local hospital, performing clinical gait trials and subsequent skeletal data collection and dataset construction. On our collected dataset, we validate that our MSS-Former significantly outperforms the other models on all metrics, with an accuracy of 97.84%, Cohen's Kappa of 0.953, and MCC of 0.955. Upon validation against real-world human skeletal datasets, the MSS-Former highlights the pivotal role of Transformer-based feature fusion in augmenting the precision of fall risk prediction models.

Our study provides a novel solution for the early detection of high-fall-risk individuals. We aim to deliver real-time risk predictions, particularly for older adults, by employing affordable sensors and a multifaceted model. This approach facilitates prompt preventive actions and informs public health strategies. The study's key contributions are as follows:

- Our research introduces a novel approach for the correlation topology between skeletal joints, with a comprehensive focus on joints' inherent physiological and potential functional connections.
- We incorporate a pre-trained ResNet-FPN module to enhance multi-scale feature capture, thus improving the spatiotemporal modeling capabilities of STGCN.
- We apply a multi-model output feature fusion transformer in fall risk prediction for the first time, dynamically merging and adaptively weighting outputs from multiple model streams, with superior performance compared to the conventional feature fusion techniques.
- We conduct clinical gait trials and data collection in a hospital clinical setting, laying a foundation of data innovation for empirical research in fall risk prediction.
- Finally, we validate our MSS-Former model using a real-world human skeletal dataset, demonstrating its practicality, reliability, and accuracy in real-world scenarios.

The rest of the paper is organized as follows: Section II comprehensively reviews relevant deep learning-based research in fall risk prediction and action recognition. Section III outlines the details of the proposed MSS-Former. In Section IV, we provide a detailed account of our pilot study and the results of experiments for our model. The visualization results are shown in Section V. Section VI provides a detailed discussion. Finally, in Section VII, we describe the conclusions of our research.

II. Related work

Our methodology predominantly draws from deep learning techniques in action recognition utilizing skeletal data and adapts these principles for transfer to fall risk prediction. Accordingly, we categorize our investigation into two distinct segments: Part A focuses on the array of deep learning

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

methodologies prevalent in fall risk prediction studies, while Part B encompasses the current research on deep learning methods within the realm of action recognition.

A. Deep learning-based fall risk prediction

A variety of deep learning methods are employed for predicting the fall risk in older adults using skeletal data, primarily involving CNNs, Recurrent Neural Networks (RNNs), and LSTMs [15]. Xu et al. [16] developed a fall risk prediction method using a CNN and human skeletal graphs, applying OPENPOSE to RGB videos to convert human postures into skeletal RGB images and training with the CNN Inception-ResNet model. Their method was tested across multiple datasets, achieving an accuracy of 91.7%. However, relying on processed RGB videos instead of direct skeletal data might limit its broad applicability due to environmental and image clarity issues. Tao et al. [17] addressed this issue using skeletal data collected from Kinect devices by proposing a fall risk prediction method based on RNN and LSTM networks, which reached an accuracy of 91.7% after validation with a 264-video dataset [18]. Jun et al. [19] introduced a real-time system for older adults, encoding skeletal motion sequences into RGB images. Their method transforms motion sequences into color skeletal pseudo images, achieving a recall rate of 93.9% and an accuracy of 93.75% on the test set. By converting dynamic skeletal data into static images processable by CNNs, their method effectively leverages deep learning to identify complex motion patterns. Inspired by these studies, Section Three of our research will discuss novel spatiotemporal pseudo-image representations of dynamic skeletal data for training convolutional networks. These deep learning-based methods have realized commendable efficacy in fall risk prediction. Notably, current research predominantly relies on CNN methods, often overlooking the physiological and functional relationships inherent in skeletal structure.

B. Deep learning-based human action recognition

Both depth images-based action recognition and fall risk prediction face the same skeletal data structure. Deep learning models in both tasks require feature extraction and effective classification and prediction. This similarity allows advancements and techniques in action recognition to be extended to fall risk prediction. Therefore, this section reviews the progress and latest research in deep learning models within the action recognition domain. The skeletal structure is a natural topological architecture, and Graph Convolutional Networks (GCNs) based on this inherent physiological structure are highly effective in human action recognition [20], [21]. The STGCN model effectively captures the spatial and temporal structures inherent in skeletal data, as demonstrated by the integration of these dimensions in Yan et al.'s research work on skeletal data modeling via STGCN. Testing on large datasets such as Kinetics and NTU-RGBD demonstrates STGCN's superior performance compared to mainstream methods. Shi et al. developed an extended version of STGCN, named 2s-AGCN, by adding adaptive graph convolutional layers and second-order information for action recognition [22]. However, gait features possess unique symmetries, often neglected in previous GCN models [23], [24], impacting the performance of STGCNs in gait recognition. Liu et al. [25] introduced a Symmetry-Driven Hyper-Feature Graph Convolutional Network (SDHF-GCN), enhancing expressiveness by exploiting symmetry and multilevel feature complementarity in skeletal gait recognition.

As research in action recognition deepens, several limitations of STGCN have emerged [26], such as overlooking semantic connections, model inefficiency, and notably, the failure to utilize cross-temporal co-occurrence relationships in skeletal data effectively. Current evidence suggests that fusion strategies are a potentially effective solution to these drawbacks of STGCN [26]. Meanwhile, the development of hybrid models that combine the strengths of different networks has emerged as a popular design paradigm [27]. Two-stream adaptive graph convolutional network (2s-AGCN) introduced a dual-stream framework to model both first and second-order skeletal information [24]; SV-GCN fuses two streams of information based on skeletal and video data for action recognition [28]; Structure-feature fusion adaptive graph convolutional networks (SFAGCN) combines MLP and GCN models to adaptively fuse skeletal structure and joint features [29]. Furthermore, to address the lack of capturing pixel-level spatial features in STGCN compared to traditional CNN models, some research attempts to fuse STGCN with CNNs [30]–[33], enhancing the ability to learn spatial co-occurrence features [26].

However, in terms of merging the outputs of these multi-stream models, most existing methods are limited to basic direct fusion or predefined weighted fusion. The dynamic fusion of features based on the model's contribution to results remains an unexplored domain [11], [12]. Moreover, current action recognition prediction models mainly focus on the skeletal physiological structure, overlooking the kinematic connections in functional structures. Integrating both joint physiological and potential functional relations, comprehensively capturing pixel-level spatial features, and adaptively fusing these captured features for recognizing dynamic skeletal traits are crucial. The tasks of fall risk prediction and action recognition are similar in their approach to data input, as both necessitate analyzing the temporal and spatial characteristics of skeletal movements through human posture. Consequently, drawing inspiration from cutting-edge research in action recognition, we aim to adapt and advance these technical methodologies within the realm of fall risk prediction for enhanced performance and insight.

Notation

- J: Set of joints representing 25 skeletal joints extracted from the human skeletal data, $J = \{j_1, j_2, \ldots, j_{25}\}$.
- E: Set of edges representing connections between adjacent joints in the skeleton, defined as $E \subseteq \{(j_m, j_n) | j_m, j_n \in J, m \neq n\}$, where each edge connects a pair of adjacent joints j_m and j_n.
- A: Adjacency matrix for both physiological and functional topologies, denoted as A_{phys} and A_{func}, respectively. Each element within A is represented by a.
- $r_{mn}(t)$: Pearson's correlation coefficient between joints j_m and j_n at time t, where $m, n \in \{1, 2, \ldots, 25\}$.
- θ: Threshold for significant interactions in adjacency matrix A.
- $S_x(t), S_y(t), S_z(t)$: Temporal sequences of x, y, and z coordinates respectively, representing the positions of all 25 joints along each axis at time t.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

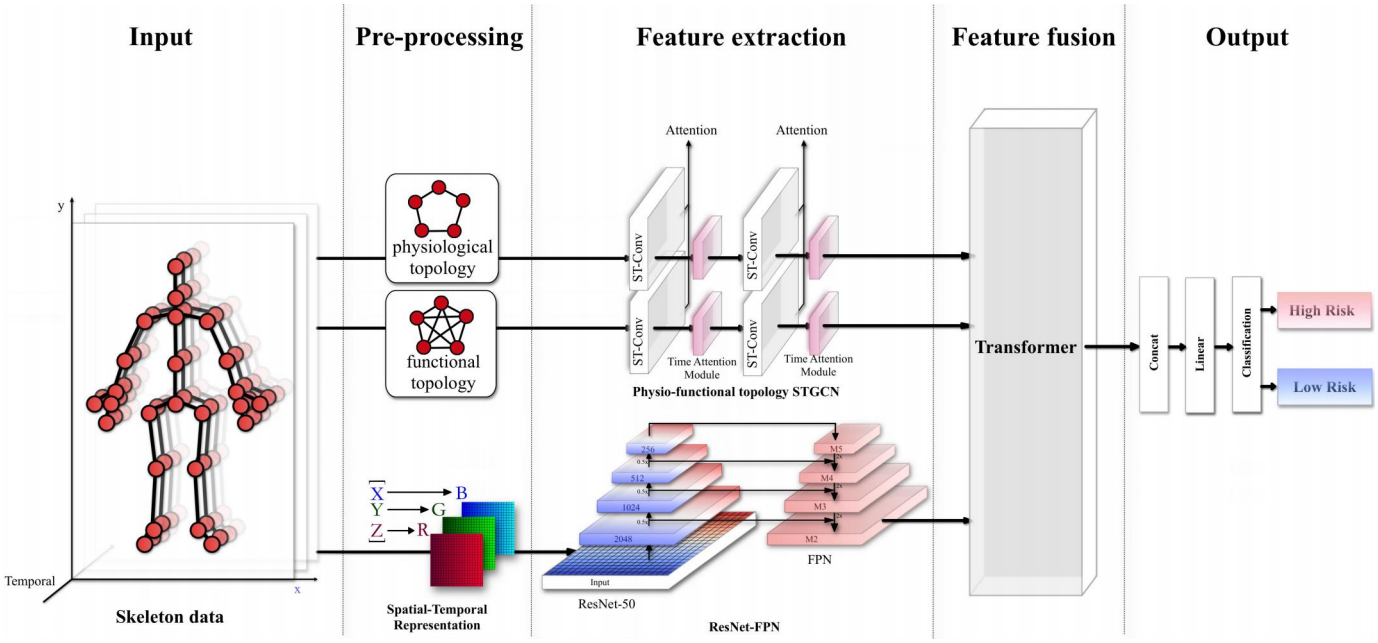

Fig. 2: Overview pipeline of the proposed MSS-Former architecture

- $M(t)$: Spatial-temporal matrix at time t, constructed from $S_x(t), S_y(t), S_z(t)$ as rows, encapsulating 3D positions of all joints at time t.
- O_i: Feature map outputs of various model streams, where $i \in \{1, 2, 3\}$. Specifically, O_1 corresponds to the physical stream, O_2 to the functional stream, and O_3 to the ResNet-FPN model's initial feature maps.
- O_i' : Attention-transformed feature maps, reflecting the refined outputs for each stream.
- d_k: Dimensionality of keys in the Transformer model, with k indicating the vector size in the attention mechanism.
- U: Unified feature representation from the model's outputs, composed of concatenated and transformed data streams.

III. Proposed Method

We propose a novel method for fall risk prediction based on skeletal data. Our MSS-Former primarily reveals the physiological-functional-multiscale feature spectrum of skeletal data through its three core components (Figure 2), thereby achieving high-accuracy early fall risk prediction. The proposed model includes three main stages. The first part focuses on data pre-processing, which involves the construction of topology and the representation of spatiotemporal dynamics in skeletal data features. The second part deals with feature extraction, specifically employing a multiscale approach within a multi-stream architecture for skeletal data. The final component pertains to feature fusion, utilizing a transformer-driven method for the adaptive integration of multistream features. In the following, an exhaustive elucidation of these three pivotal components is provided.

A. Feature Representation

skeletal data contain rich, multi-scale spatiotemporal semantic information. In the dynamic continuum of skeletal motion, nuanced shifts in posture and movement can significantly alter the precision of fall risk evaluations. Accurately capturing skeletal information in both spatial and temporal domains and dissecting skeletal data across multiple dimensions is crucial for the accuracy of fall risk assessment. In this section, we will introduce the construction of a dual physiological-functional topology structure and the pixel-level spatio-temporal representation of skeletal data.

1) Physiological-functional dual topology construction: MSS-Former integrates a dual topology framework by effectively combining physiological topology and functional topology. The physiological topology reflects the skeleton's inherent anatomy and joint interrelations. The functional topology, informed by skeletal joint interrelations, provides insights into bodily coordination and limb stability.

Leveraging the human skeleton's inherent anatomical relationships, we have developed a graph topology that encapsulates physiological structures. This approach treats the skeleton as a natural graph, capturing the spatiotemporal interactions among joints comprehensively.

In our graph topology, joints within the human skeleton are designated as nodes, characterized by their physiological and potential functional connections. The connections linking these joints are considered edges, connecting adjacent joints in the graph. The 25 joints of the human skeletal data are represented as $J = \{j_1, j_2, \ldots, j_{25}\}$, and the edges formed by these joints are denoted as $E \subseteq \{(j_m, j_n) | j_m, j_n \in J, m \neq n\}$. For any two physically adjacent joints j_m and j_n, an edge $(j_m, j_n) \in E$ exists. Thus, we construct a physically topological adjacency matrix, a 25×25 matrix A_{phys}, based on the inherent physical structure of the skeleton. For each element a_{mn}, if joints j_m and j_n are adjacent, a_{mn} is set to 1, while non-adjacent elements are represented as 0.

We developed a functional graph topology to more accurately depict the connections among skeletal joints. The current understanding of gait behavior based solely on spatial joint

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

connections is limited. The potential dependencies among non-adjacent joints reflect the intrinsic functional-structural relationships in human movement, which are crucial for a comprehensive understanding of human kinematics.

To assess the functional interactions between skeletal joints over time, we start by normalizing the coordinates (X, Y, Z) of the 25 nodes for each sample and then averaging over these three dimensions. Following this preprocessing step, we calculate Pearson's correlation coefficients r_{mn} for each pair of joints j_m and j_n across all time points. These coefficients form a 25 $\times$ 25 matrix that reflects the functional connectivity of specific sample. After calculating these matrices for all samples, we average them to obtain the final functional topology matrix A_{func}. The matrix captures the comprehensive degree of functional interaction among joint pairs, succinctly summarizing the dynamic connectivity patterns identified throughout the analysis period. Importantly, within this process, a threshold ($\theta = 0.92$) is applied to each matrix element to spotlight the most significant interactions. This threshold was validated empirically with silhouette score and the percentage of edges retained to ensure optimal performance, reflecting a careful balance between sensitivity and specificity in capturing robust functional relationships. The value of elements in A_{func} that meet or exceed this threshold are indicative of robust functional relationships, thereby offering a nuanced and dynamic depiction of the skeletal data's functionality.

By constructing a dual-topology graph structure, we have seamlessly integrated both physiological and potential functional joint connections. This holistic approach significantly enriches our understanding and augments our capability of fall risk prediction.

2) Pixel-level spatiotemporal representation: The spatial features captured within the dual topology structure are at a superpixel level. In the domain of fall risk prediction, overlooking pixel-level spatial details might miss critical differences in gait or motion patterns associated with varying fall risks. However, graph neural networks face challenges in capturing pixel-level spatial features. Innovatively building on graph convolutional networks, we represent skeletal data along spatial and temporal dimensions at the pixel level. Such pixel-level representation of skeletal data enhances the model's ability to effectively learn spatial co-occurrence features through subsequent convolutional neural network applications.

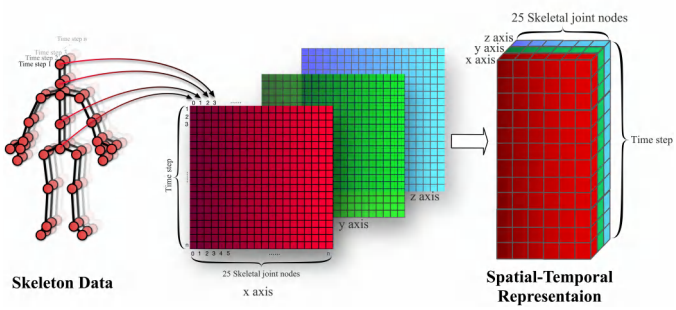

Fig. 3: Extraction and construction of spatiotemporal representations for skeletal data

We construct image-like matrices by mapping skeletal coordinates onto dimensions that resemble the RGB three-channel system. The skeletal data captured by depth images inherently forms a time series containing XYZ axis information. For each sample, we normalize the XYZ coordinates of the skeletal data to a [0,1] range, and then reconstruct an image-like matrix at each of the three channels through data transformation. The detail of this process is elaborated in Figure 3. The image-like matrices enable representing the spatial arrangement and pixel-level detail of the skeletal data, and thus can be suitably processed using CNNs in subsequent training step.

We allocate the XYZ coordinates of skeletal data across distinct layers to form a 2D spatiotemporal matrix. This matrix reflects the temporal evolution of each joint across the x, y, and z axes for all joints. Specifically, for each time t, we define three vectors: $S_x(t)$ for all x-coordinates, $S_y(t)$ for all y-coordinates, and $S_z(t)$ for all z-coordinates of the joints. These vectors represent the positions of all 25 joints along their respective axis at time t.

Consequently, the spatial-temporal matrix $M(t)$, capturing the coordinated movement of all joints at time t, is constructed as:

$$M(t) = \begin{bmatrix} S_x(t) \\ S_y(t) \\ S_z(t) \end{bmatrix} = \begin{bmatrix} x_{0,t} & \cdots & x_{24,t} \\ y_{0,t} & \cdots & y_{24,t} \\ z_{0,t} & \cdots & z_{24,t} \end{bmatrix},$$

Each row of $M(t)$ corresponds to one of the spatial dimensions (x, y, or z), encapsulating the coordinates of all joints along that axis at time t, thereby providing a complete spatiotemporal snapshot of skeletal positions.

B. Feature Extraction

1) Dual-Stream STGCN with Integrated CBAM Module: We developed a dual-stream STGCN integrated with the Convolutional Block Attention Module (CBAM) based on the previously established physiological-functional dual topology structure. For human skeletal movement dynamics data characterized by spatiotemporal complexity, we utilize the dual-stream STGCN to extract features. Meanwhile, we incorporate the CBAM attention mechanism after the convolutional network layers to focus on specific skeletal regions, forming a layer of the ST-conv module. In each workflow, we implemented two layers of ST-conv modules, with each layer followed by an internal temporal attention mechanism to focus on crucial gait movement time points (Figure 4).

The spatial-temporal convolution within the ST-conv framework blends graph-structured data G with the input feature matrix X. This process, modulated by the trainable weights Θ, yields a feature representation $Y = \Theta * (G \cdot X)$ that embodies both spatial and temporal characteristics. To mitigate overfitting, we integrate dropout, batch normalization, and ReLU activation functions. Subsequently, CBAM refines the feature maps, employing focused attention through channel and spatial mechanisms. Specifically, the channel attention map M_c is determined by a sigmoid-activated sum of multi-layer perceptron-processed average and maximum pooled features: $M_c = \sigma(MLP(AvgPool(X)) + MLP(MaxPool(X)))$. Similarly, the

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

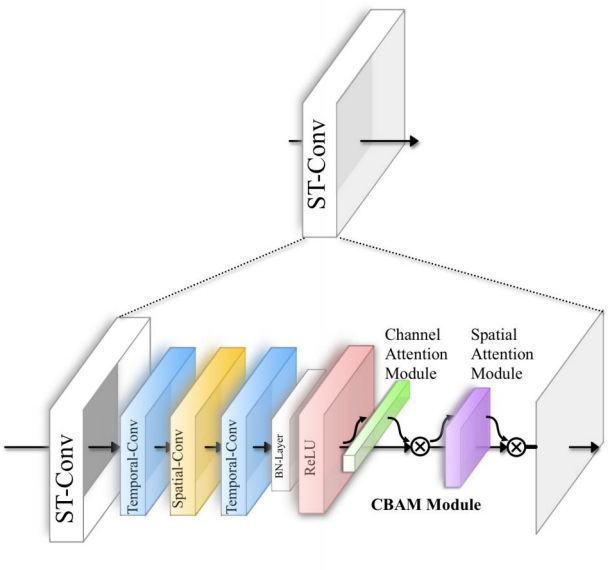

Fig. 4: Building blocks for each layer of ST-Conv modules

spatial attention map M_s emerges from a sigmoid-activated convolution of concatenated average and maximum pooled channel-wise features: $M_s = \sigma(f([AvgPool(X_c); MaxPool(X_c)]))$. The resultant refined output is given by $X' = (M_c \otimes X) \odot M_s$, where $\odot$ denotes the element-wise product. This equation sequentially integrates the attention maps, thus sharpening the focus on features.

In order to assess the performance of the skeletal topology at various time steps during movement, we incorporate a temporal attention mechanism module subsequent to the ST-conv module. This module allocates learned weights to the temporal features, directing the model's focus to critical moments in the time series. By employing dual-stream STGCN to extract features from both physiological and functional topologies, our methodology provides a detailed representation of the varied physiological and functional attributes present in human skeletal motion.

2) Enhanced ResNet with Integrated Feature Pyramid Network (FPN): Pixel-level feature extraction of skeletal data was achieved using the ResNet-FPN module, integrating the robust capabilities of ResNet architecture with FPN. This method significantly augmented and enriched pixel-level skeletal features, addressing the limitations of dual-stream STGCN in capturing the complete spectrum of such details. Using a pre-trained ResNet50 as the initial feature extraction base layer, we focused on the pixel-level spatiotemporal representation of skeletal data. To expand the receptive field of pixel-level skeletal information and comprehensively capture spatiotemporal features, we incorporated the FPN module. With the integration of the ResNet-FPN module, we were able to present the multi-scale, pixel-level features of the skeletal data in a comprehensive manner.

Integrating ResNet50 with FPN for skeletal data analysis employs a descending and ascending cascade of feature resolutions that capture multi-scale information. Following the initial input, the ResNet50 architecture methodically refines the feature channels through a sequence of layers, reducing dimensions from 2048 to 256. This streamlined approach effectively distills intricate skeletal movement details into a concise, multi-resolution feature representation.

Features undergo an inverse scaling process during their transition into the FPN. Starting with a 256-resolution output, the FPN systematically upscales it through successive stages, culminating in a 2048-resolution layer, effectively enhancing the detail and scope of feature representation. Such a top-down approach within the FPN effectively recombines high-resolution semantic details with spatial information from the initial layers (Figure 5).

The bidirectional flow of information within the network is mathematically captured as follows: At each level l in the FPN, the feature map, $\mathbf{P}_l$, results from upsampling the subsequent level $\mathbf{P}_{l+1}$ via the operation $\mathcal{U}$. This upscaled feature map then merges with the processed features from the corresponding ResNet layer $\mathbf{F}_l$ through convolution C, leading to the relationship $\mathbf{P}_l = \mathcal{U}(\mathbf{P}_{l+1}) \oplus C(\mathbf{F}_l)$. Consequently, the pyramid's apex retains the highest resolution of 2048, forming a feature-rich layer that comprehensively encapsulates the skeletal data.

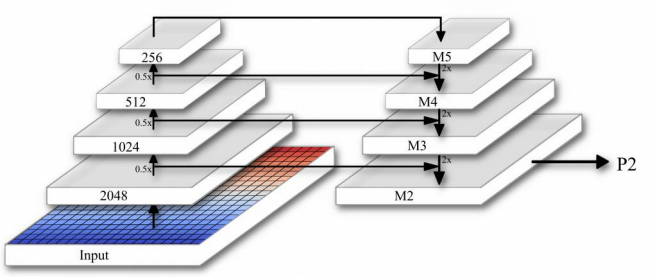

Fig. 5: ResNet-FPN architecture

C. Feature Fusion and Output

We innovatively employ the transformer-driven adaptive integration of multistream features approach for the adaptive fusion of dual-stream STGCN outputs, encompassing both the physical stream O_1 and the functional stream O_2, along with the convolutional features from the ResNet-FPN model O_3. During the feature extraction phase, the MSS-Former model processes multiple data streams in parallel, yielding a series of parallel feature outputs. This strategy mitigates potential bias in feature expression often encountered in traditional feature fusion methods, which typically merge features in a fixed and overly simplistic ratio.

In our feature fusion framework, a crucial operation is the projection of each output stream into a query (Q), key (K), and value (V) spaces. This process utilizes learnable matrices W_i^Q, W_i^K, and W_i^V within the Transformer model to compute the self-attention for each output stream. The dimension of the keys is denoted by d_k. Mathematically, the projections are represented as $Q_i = O_i W_i^Q$, $K_i = O_i W_i^K$ and $V_i = O_i W_i^V$.

The attention scores between queries and keys are computed, normalized using a softmax function, and then applied to weight the values, yielding the attention-driven output for each data stream: $O_i' = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) \cdot V_i$. This process ensures that the

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

model differentially attends to various aspects of the features based on their interrelationships and relevance.

The outputs from each data stream are then concatenated and linearly transformed within the Transformer to form a unified feature representation U:

$$U = \text{Concat}(O'_1, O'_2, O'_3) \cdot W^O$$

This unified representation U merges the multistream features from the physical, functional, and convolutional feature spaces. After processing through the output layer, this enables us to achieve precise fall risk prediction using skeletal data.

IV. Experiments

In this section, we comprehensively evaluated the necessity and strength of each component of our method and its performance in the context of real-world fall risk prediction using the skeletal datasets collected through multiple gait experiments in a local hospital's Rehabilitation Department. Then, we conducted detailed ablation studies on the proposed model using our collected dataset. In addition, we compared it with other SOTA models to ascertain the superiority and uniqueness of our approach. All our experiments were conducted using the PyTorch deep learning framework, implemented in Python 3.10.9, utilizing the computational power of one RTX A6000 GPU to ensure the accuracy and efficiency of the analysis. The preset hyperparameters during the model training process are as shown in the Table I.

TABLE I: Hyper-Parameters Of MSS-Former Models

Hyper-parameter	Value
Number of Epochs	100
Batch Size	128
Optimizer	Adam
Learning Rate	0.001
Learning rate strategy	Cosine Annealing
Warmup epochs	10
Learning Rate Scheduler Gamma	0.1
Hidden Layer Dimensions	[128, 64]
Output Dimension	10
Dropout ratio	0.2
L2 Weight decay	0.001

A. Pilot study

To validate our model, we conducted a pilot study in the rehabilitation department of a local hospital. The study recruited 61 participants for one BBS test, and multiple 3-Metres Timed Up and Go (3MTUG) tests, yielding a total of 183 data entries for 3MTUG. During the 3MTUG tests, we collected skeletal data of the participants using Microsoft Kinect v2 Kinect, constructing a dataset in a clinical context to verify our model.

We recruited participants who met the following inclusion criteria: equal or greater than 60 years old (equal or greater than 40 years old for stroke patients), engaging in a rehabilitation program, having the ability to walk independently or with a walking aid, having a normal (or corrected-to-normal) vision, and having the ability to provide informed consent. We excluded older adults who had abnormal vision, disability of walking, and/or life-threatening illnesses, as they would likely be unable to complete the gait and balance assessment. All participants were required to fill out the written informed consent before the initiation of the study. After a group of successful participant recruitment, we visited the hospital to collect their written informed consent and clinical data, including medical history, investigation outcomes, therapeutic interventions, and disease progression. During the visit, each participant was required to complete the TUG test. Next, a professional physiotherapist was invited to assess the participants' BBS performance.

This study adheres to the ethical standards delineated in the 1964 Helsinki Declaration and its subsequent amendments. It received ethical approval from the Ethics Committee of the affiliated institutes (No. 2022-023 and No. [2022]003), and was registered in the Chinese Clinical Trial Registry under ChiCTR2200061834.

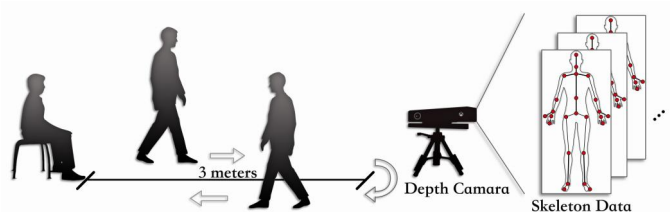

Fig. 6: Side view of the 3MTUG test procedure

The experimental phase, conducted from July 20 to October 26, 2022, involved two primary assessments. The first, the BBS consists of 14 tasks evaluating balance, each rated from 0 (unable to perform) to 4 (performed according to criteria), accumulating a total score of 56. The BBS is recognized for its objective balance control evaluation during functional tasks. A cutoff score of 40 on the BBS was used to categorize fall risk [34], with scores above and below this threshold indicating low and high fall risk, respectively, encompassing 35 and 26 participants.

The second assessment, the 3MTUG test, was captured via a Kinect, offering a nuanced analysis of each participant's gait and balance. In the 3MTUG test, the participants were asked to stand up from a chair, walk for 3 meters, turn, walk back to the chair, and sit down. In this study, a Kinect was positioned 4.16 m from the participant's chair to extract movement at all 25 points. Participants, totaling 61, underwent two assessment rounds, spaced a week apart, to avoid scheduling conflicts. Out of these, 31 completed both assessments (Figure 6).

In our clinical pilot study, each participant underwent either one or two trials, depending on their length of stay at the hospital. Specifically, 54 out of the total 61 participants completed TUG tests twice during each trial, while the remainder completed only one TUG test due to poor physical condition. As a result, our study's dataset consists of 183 entries with 66 indicating high fall-risk individuals (BBS ≤ 40) and 117 indicating low fall-risk individuals (BBS > 40).

B. Models Comparison

We evaluated our MSS-Former model alongside leading models in skeleton-based action recognition and classical models in image recognition and temporal analysis. Prior to

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

TABLE II: Comparative Performance Evaluation of the MSS-Former and Seven SOTA Models (Mean ± Standard Deviation)

Module	Accuracy	Cohen's Kappa	MCC	F1 Score	Precision	Recall
MD-STGC [35]	0.8730 (±0.0124)	0.7071 (±0.0308)	0.7176 (±0.0277)	0.7978 (±0.0232)	0.9027 (±0.0042)	0.7154 (±0.0353)
HiSTGNN [36]	0.8973 (±0.0265)	0.7718 (±0.0665)	0.7815 (±0.0556)	0.8485 (±0.0525)	0.8726 (±0.0710)	0.8462 (±0.1240)
STGCN [37]	0.8514 (±0.0607)	0.6566 (±0.1697)	0.6721 (±0.1535)	0.7577 (±0.1521)	0.8228 (±0.0603)	0.7385 (±0.2181)
GAT [38]	0.9135 (±0.0265)	0.8057 (±0.0632)	0.8091 (±0.0596)	0.8701 (±0.0455)	0.9087 (±0.0187)	0.8385 (±0.0803)
CNN-LSTM [39]	0.7655 (±0.0190)	0.5457 (±0.0313)	0.5894 (±0.0248)	0.7459 (±0.0145)	0.6165 (±0.0213)	0.9450 (±0.0150))
ResNet 18 [40]	0.8919 (±0.0270)	0.7679 (±0.0560)	0.7792 (±0.0504)	0.8522 (±0.0371)	0.8309 (±0.0749)	0.8923 (±0.1043)
ResNet 50 [40]	0.8973 (±0.0202)	0.7823 (±0.0379)	0.7896 (±0.0309)	0.8640 (±0.0212)	0.8165 (±0.0517)	0.9231 (±0.0487)
MSS-Former	**0.9784 (±0.0202)**	**0.9529 (±0.0432)**	**0.9548 (±0.0412)**	**0.9697 (±0.0273)**	**0.9733 (±0.0533)**	**0.9692 (±0.0377)**

TABLE III: Results of Ablation Experiments Evaluated Across Six Different Model Metrics and Nine Combinations (A-I)

Component	Combinations								
	A	B	C	D	E	F	G	H	I
ResNet 18		✓							
ResNet 50			✓	✓	✓	✓	✓	✓	✓
FPN				✓		✓	✓	✓	✓
CCM*							✓	✓	✓
CBAM					✓	✓		✓	✓
Transformer									✓
Accuracy	0.8649 (±0.0342)	0.8892 (±0.0189)	0.9027 (±0.0216)	0.9389 (±0.0324)	0.9222 (±0.0167)	0.9583 (±0.0139)	0.9556 (±0.0333)	0.9568 (±0.0216)	**0.9784 (±0.0202)**
Kappa	0.7033 (±0.0873)	0.7554 (±0.0424)	0.7838 (±0.0540)	0.8638 (±0.0693)	0.8287 (±0.0391)	0.9073 (±0.0295)	0.8999 (±0.0743)	0.9051 (±0.0490)	**0.9529 (±0.0432)**
MCC	0.7117 (±0.0881)	0.7577 (±0.0426)	0.7882 (±0.0497)	0.8735 (±0.0615)	0.8339 (±0.0412)	0.9116 (±0.0272)	0.9033 (±0.0712)	0.9064 (±0.0492)	**0.9548 (±0.0412)**
F1 Score	0.8063 (±0.0670)	0.8400 (±0.0288)	0.8571 (±0.0409)	0.9094 (±0.0455)	0.8879 (±0.0274)	0.9387 (±0.0189)	0.9329 (±0.0496)	0.9384 (±0.0330)	**0.9697 (±0.0273)**
Precision	0.7974 (±0.0314)	0.8537 (±0.0431)	0.8771 (±0.0357)	0.9203 (±0.0968)	0.8515 (±0.0352)	0.9352 (±0.0675)	0.9476 (±0.0726)	0.9251 (±0.0047)	**0.9733 (±0.0533)**
Recall	0.8308 (±0.1410)	0.8308 (±0.0576)	0.8462 (±0.0910)	0.9167 (±0.0986)	0.9333 (±0.0726)	0.9500 (±0.0553)	0.9250 (±0.0692)	0.9538 (±0.0615)	**0.9692 (±0.0377)**

these comparisons, all data underwent a uniform preprocessing protocol which included data filtering, augmentation, and balancing. For data filtering, we utilized a Butterworth low-pass filter with a sampling frequency of 100 Hz and a cutoff frequency set at 5 Hz, employing a 6th order filter. Data augmentation was achieved using a sliding window technique with a window size of 400 and a step size of 200, increasing the data amount to 3,843 sequences. Then SMOTE method was applied to balance the dataset. To ensure a robust evaluation, we divided the augmented data into training (80%, 3074 sequences) and testing sets (20%, 769 sequences), with 10% of the training set (307 sequences) used for validation to fine-tune model parameters. All models were compared on the identical dataset collected, and were appropriately fine-tuned for fall risk prediction. This comparison involved MD-STGC [35], HiSTGNN [36], STGCN [37], Graph Attention Networks (GAT) [38], CNN-LSTM [39], ResNet 18, and ResNet 50 [40], effectively showcasing our model's comparative performance (Table II). Our MSS-Former model significantly outperforms the other models on all metrics with an accuracy of 97.84%, Cohen's Kappa of 0.953, and MCC of 0.955, and its excellent performance proves the superiority of its unique architecture. In addition to the high average performance, the relatively low standard deviation indicates that the model's performance is consistent and unaffected by data variations or training epochs. Significantly, the GAT model attains an accuracy of 91.35%, a feat surpassed only by our MSS-Former model. This suggests the advantage of enabling selective focus on pertinent features for gait detection and recognition. Conversely, the CNN-LSTM model, designed to capture temporal dependencies, falls short in performance metrics compared to other models. However, it achieves a high recall rate of 94.50%, indicating its relative potential in certain aspects of the task.

C. Ablation Study

Detailed ablation studies confirm that our method's existing model components and architecture are rational and optimal compared to other combinations. In our study, we extended the STGCN model to include a preliminary integration of ResNet [40] for capturing deeper data layers not reachable by the graph neural network alone (Table III). Our MSS-Former model's infrastructure combines various elements, with Combination B and C showing improvements in metrics over Combination A. Notably, ResNet 50 in Combination C outperformed ResNet 18 in Combination B, aligning with the understanding that deeper networks more effectively discern complex data patterns [41]. Consequently, ResNet 50 was selected as the foundational structure. The integration of Feature Pyramid Networks (FPNs)

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

TABLE IV: The 25 Skeletal Joint Numbers Output by the Kinect Depth Image and Their Corresponding Joint Names

Joint num.	Joint name	Joint num.	Joint name
0	SPINEBASE	13	KNEELEFT
1	SPINEMID	14	ANKLELEFT
2	NECK	15	FOOTLEFT
3	HEAD	16	HIPRIGHT
4	SHOULDERLEFT	17	KNEERIGHT
5	ELBOWLEFT	18	ANKLERIGHT
6	WRISTLEFT	19	FOOTRIGHT
7	HANDTIPLEFT	20	SPINESHOULDER
8	SHOULDER RIGHT	21	HANDTIPLEFT
9	ELBOWRIGHT	22	THUMBLEFT
10	WRISTRIGHT	23	HANDTIPRIGHT
11	HANDTIPRIGHT	24	THUMB RIGHT
12	HIPLEFT		

[42] from Combination C to G further enhanced performance, indicating FPN's efficacy in providing multi-scale features and expanding the model's sensory field. In Combinations G, H, and I, the correlation adjacency matrix [43] was employed for neural graph construction within the STGCN, resulting in incremental enhancements over Combination F, where solely FPN and CBAM were utilized. Regarding the attention mechanism, incorporating CBAM modules [44] at each STGCN layer led to significant performance enhancements in Combinations D and E over Combination C, particularly in test accuracy and Cohen's Kappa. This improvement suggests that leveraging attention mechanisms on skeletal spatiotemporal data is advantageous for fall risk prediction in older adults. However, whether introducing Transformer-Driven [45] Adaptive Integration of Multistream Features improves performance over previous approaches has not been explored in existing research. To validate our innovative architecture, we conducted comparisons between Combination I and Combination H, demonstrating that including the Transformer architecture for adaptive feature fusion significantly enhances performance. Our thorough ablation experiments validated the necessity and superiority of each component in the proposed method.

V. Attention weight visualization

After integrating the CBAM attention mechanism into the STGCN basic framework, a rigorous iterative process refines the weights of the 25 skeletal joints by employing a channeled attention strategy (see Table IV for the correspondence of the 25 skeletal joints).

Throughout the entire cohort, the CBAM model consistently allocates significant attention weights to the lower limb joints for both high and low-fall risk categories. The visualization of the ultimate time-channel attention weights for the initial subjects across both categories is shown in (Figure 7). A recurring pattern is observed, particularly involving bilateral hip, knee, and toe joints, with a pronounced emphasis on the left ankle. This concentration on the lower limb joints corresponds with their well-documented significance in motion analysis and fall risk assessment.

We synthesized the channel and spatial attention weight distributions into a color-coded visualization of the temporal skeletal graph. Additionally, we separately displayed the attention feature

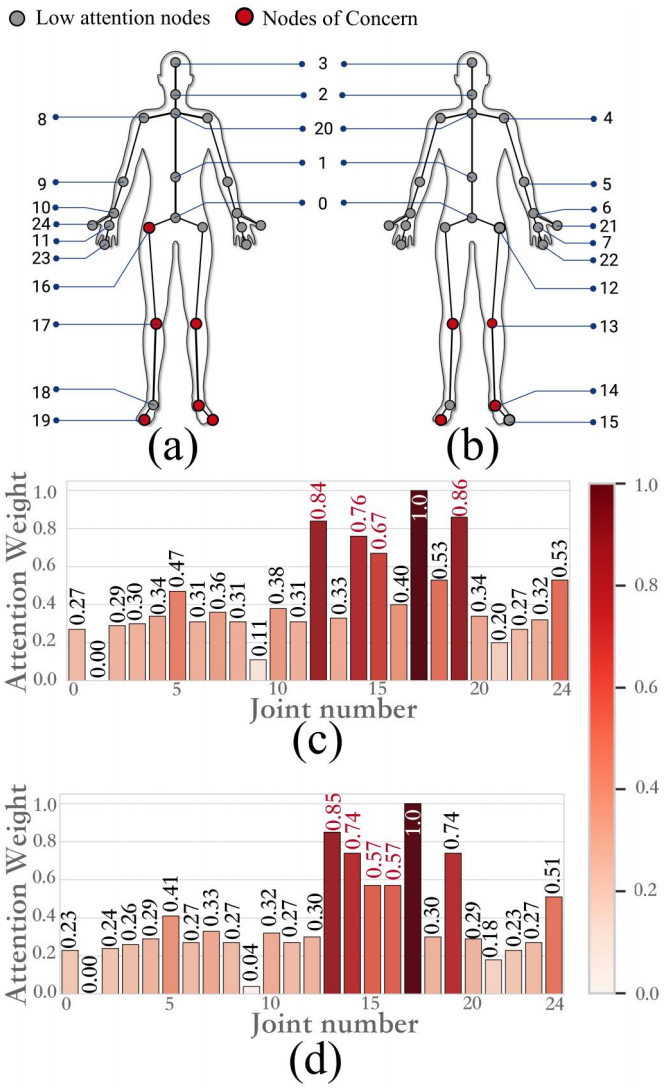

Fig. 7: Channel attention weights and skeletal joint importance distributions for high fall risk and low fall risk participants. (a) Attention weights for high fall risk. (b) Attention weights for low fall risk. (c) Joint importance for high fall risk. (d) Joint importance for low fall risk.

map for spatial attention weights in the adjacency matrix (Figure 8). The weights between distal joint pairs in high fall-risk participants are predominantly zero, suggesting impaired motor control in distal joint movement states. In contrast, the weight distribution is more uniform in the low fall-risk group, indicating preserved communication and coordination among skeletal components. There is also significant interaction sparsity between the base of the spine and the pelvis in high fall-risk participants, implying poorer trunk stability during the 3MTUG task, significantly affecting balance. The interaction between the shoulders and wrists is also diminished, suggesting insufficient upper limb support. Conversely, the weight distribution at the lower limb connections in low-fall-risk participants is denser, indicating more synchronized limb movements conducive to task stability and efficiency. Differences in foot interactions further differentiate the groups, with poorer ankle and foot

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

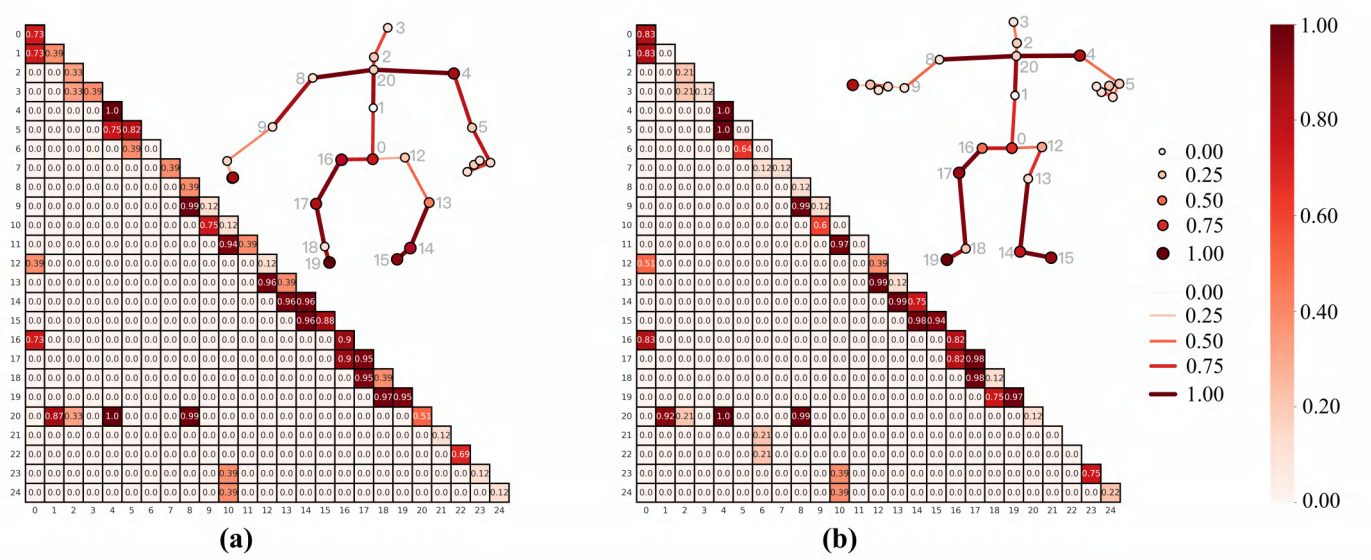

Fig. 8: Visualization of attention weights for adjacent edges and joints in the CBAM, standardized to [0, 1]. (a) High fall risk participants (b) Low fall risk participants.

coordination in the high-fall-risk group potentially exacerbating the risk of falling.

Investigating skeletal joints enriches our comprehension of skeletal movement patterns, providing insights into the intricate coordination and dynamics underlying human locomotion and posture. This also explains the efficacy of our MSS-Former model, enhancing its interpretability and further refining the stratification of different population groups regarding fall risk.

VI. Discussion

The performance characteristics of deep learning models are often closely related to the nature of the data and the specifics of particular tasks, as evidenced by our empirical studies [46], [47]. In skeletal data-based fall risk prediction, the precision in data representation and the development of models adept at thoroughly extracting these representations are fundamental to ensuring accurate fall risk prediction.

In our research, we constructed a Physiological-functional dual topology and utilized an STGCN model integrated with a CBAM module for feature extraction of the skeletal physiological and functional structures. Concurrently, we applied pixel-level spatiotemporal representation to uncover features that graph neural networks typically overlook, utilizing an Enhanced ResNet integrated with an FPN [48]. The necessity of integrating outputs from parallel multistream models led us to use a transformer-driven adaptive integration of multistream features.

Our ablation experiments showed that the standalone ResNet 50 outperformed ResNet 18, emphasizing depth's importance in deciphering complex data patterns, albeit with an increased risk of overfitting [49]. In our proposed method, the contribution of the FPN is undeniable. By providing multi-scale features, it effectively extracted both global and local information from pixel-level skeletal features [50]–[52]. Ablation experiments spanning combinations G to I revealed performance improvements following incorporating the correlation adjacency matrix. This suggests that the dual-topology structure, grounded in both physiological and functional significance, more effectively captures the inherent connections of human skeletal movements. To more effectively discern the spatial weights of different joints within the skeletal structure, we integrated the CBAM module into the feature extraction phase of the graph neural network. This enhancement not only underscored the attention mechanism's role in enabling the model to focus on pertinent features, thereby significantly boosting its discernibility [53], but it also opened up opportunities for a deeper interpretability analysis of the model. Currently, the methods for integrating outputs from multistream models have remained simple, merging or fixed weighting. Our research introduces the Transformer model's adaptive weight distribution for the first time. This adaptive integration avoids weight biases based on prior knowledge, ensuring that the final decision fully utilizes the strengths of each module. Through the model's self-driven dynamic decision-making process, our MSS-Former model not only enhances its adaptability to various datasets and tasks but also retains the flexibility to adjust the outputs of each model structure. However, progress comes with trade-offs. The increased complexity of the current model, attributed to the correlation adjacency matrix and CBAM, may heighten the risk of overfitting. Future work could explore regularization techniques or network pruning to balance this complexity [54].

Given the significance of fall risk prediction in healthcare, geriatric care, and public health prevention and control, our method has profound implications [55]. We can conduct extensive early fall risk prediction and screening based on simple, portable, and widely applicable non-invasive gait capture images, significantly reducing severe injuries or prolonged hospitalizations. Moreover, while the MSS-Former model was originally conceived for fall risk prediction, the adaptability of the skeletal data it processes opens avenues for applications in posture correction, physiotherapy, and early neurological

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

disorder detection. This versatility is notably enhanced by the model's integrated attention mechanism within the graph neural network, which adeptly assigns varying degrees of importance to different skeletal regions. This research not only introduces a novel model architecture for the comprehensive utilization of skeletal data but also imbues the model with an innovative structure for enhanced interpretability. The enhancements made in the MSS-Former model offer promising prospects for skeletal data analysis and fall risk prediction, potentially yielding significant contributions to the field of fall risk assessment.

VII. Conclusion

This article proposes an innovative deep learning model, MSS-Former, using depth image technology to predict fall risks in older adults. The model achieves significant improvement in recognizing skeletal dynamics through a Transformer-driven, multi-scale feature fusion architecture. It constructs a topology of inter-joint skeletal correlations and incorporates the ResNet-FPN for multi-scale feature extraction, revealing the physiological-functional-multiscale feature spectrum of older adults' gait. Clinical gait trials in a hospital setting provided data for model validation. In comparative tests, MSS-Former excelled, markedly surpassing other methods in all key metrics, achieving accuracy, precision, recall, and F1 scores of 97.84%, 97.33%, 96.97%, and 96.92%, respectively. The model's attention mechanism was analyzed for interpretability, proving its robustness and rationality in learning movement physiology. Moving forward, our focus will be on refining MSS-Former by broadening our dataset and continuously optimizing parameters. These efforts are directed toward enhancing the model's interpretability and boosting its practical effectiveness in real-world scenarios. Such advancements are expected to further sharpen the accuracy of fall risk predictions, contributing significantly to safer and improved living conditions for older adults.

References

[1] United Nations. World population prospects 2022: Summary of results. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf, 2022. Accessed: December 13, 2023.

[2] World Health Organization. Falls. https://www.who.int/news-room/fact-sheets/detail/falls. Accessed: Feb 2, 2022.

[3] Ahmed Abobakr, Mohammed Hossny, and Saeid Nahavandi. A Skeleton-Free Fall Detection System From Depth Images Using Random Decision Forest. *IEEE Systems Journal*, 12(3):2994–3005, 2018.

[4] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. DG-STGCN: dynamic spatial-temporal modeling for skeleton-based action recognition. *arXiv preprint arXiv:2210.05895*, 2022.

[5] Weiwei Wu, Fengbin Tu, Mengqi Niu, Zhiheng Yue, Leibo Liu, Shaojun Wei, Xiangyu Li, Yang Hu, and Shouyi Yin. STAR: An STGCN ARchitecture for Skeleton-Based Human Action Recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2023.

[6] Haiping Zhang, Xu Liu, Dongjin Yu, Liming Guan, Dongjing Wang, Conghao Ma, and Zepeng Hu. Skeleton-based action recognition with multi-stream, multi-scale dilated spatial-temporal graph convolution network. *Applied Intelligence*, pages 1–15, 2023.

[7] Saman Shahid, Anup Nandy, Soumik Mondal, Maksud Ahamad, Pavan Chakraborty, and Gora Chand Nandi. A study on human gait analysis. In *Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*, pages 358–364, 2012.

[8] Hongye Yang, Yuzhang Gu, Jianchao Zhu, Keli Hu, and Xiaolin Zhang. PGCN-TCA: Pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition. *IEEE Access*, 8:10040–10047, 2020.

[9] Qichao Liu, Liang Xiao, Jingxiang Yang, and Zhihui Wei. CNN-Enhanced Graph Convolutional Network With Pixel- and Superpixel-Level Feature Fusion for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(10):8657–8671, 2021.

[10] Junru Yin, Xuan Liu, Ruixia Hou, Qiqiang Chen, Wei Huang, Aiguang Li, and Peng Wang. Multiscale Pixel-Level and Superpixel-Level Method for Hyperspectral Image Classification: Adaptive Attention and Parallel Multi-Hop Graph Convolution. *Remote Sensing*, 15(17):4235, 2023.

[11] Ying Zhang, Rongrong Zhang, Qunfei Ma, Yanhao Wang, Qingqing Wang, Zihao Huang, and Linyan Huang. A feature selection and multi-model fusion-based approach of predicting air quality. *ISA Transactions*, 100:210–220, 2020.

[12] Mingqiang Lin, Denggao Wu, Jinhao Meng, Ji Wu, and Haitao Wu. A multi-feature-based multi-model fusion method for state of health estimation of lithium-ion batteries. *Journal of Power Sources*, 518:230774, 2022.

[13] Yuling Xing and Jia Zhu. Deep learning-based action recognition with 3D skeleton: A survey, 2021.

[14] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017.

[15] Raoudha Nouisser, Salma Kammoun Jarraya, and Mohamed Hammami. Deep Learning and Kinect Skeleton-based Approach for Fall Prediction of Elderly Physically Disabled. In *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE, 2022.

[16] Qingzhen Xu, Guangyi Huang, Mengjing Yu, and Yanliang Guo. Fall prediction based on key points of human bones. *Physica A: Statistical Mechanics and its Applications*, 540:123205, 2020.

[17] Xu Tao and Zhou Yun. Fall prediction based on biomechanics equilibrium using Kinect. *International Journal of Distributed Sensor Networks*, 13(4):1550147717703257, 2017.

[18] Caroline Rougier, Jean Meunier, et al. Fall detection using 3d head trajectory extracted from a single camera video sequence. In *First International Workshop on Video Processing for Security (VP4S-06), June*, pages 7–9, 2006.

[19] Jun Wu, Ke Wang, Baoping Cheng, Ruifeng Li, Changfan Chen, and Tianxiang Zhou. Skeleton Based Fall Detection with Convolutional Neural Network. In *2019 Chinese Control And Decision Conference (CCDC)*, pages 5266–5271, 2019.

[20] Jeffrey M Hausdorff. Gait dynamics, fractals and falls: finding meaning in the stride-to-stride fluctuations of human walking. *Human movement science*, 26(4):555–589, 2007.

[21] Claire E Adam, Annette L Fitzpatrick, Cindy S Leary, Anjum Hajat, Sindana D Ilango, Christina Park, Elizabeth A Phelan, and Erin O Semmens. Change in gait speed and fall risk among community-dwelling older adults with and without mild cognitive impairment: a retrospective cohort analysis. *BMC Geriatrics*, 23(1):1–11, 2023.

[22] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020.

[23] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[24] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[25] Xiaokai Liu, Zhaoyang You, Yuxiang He, Sheng Bi, and Jie Wang. Symmetry-Driven hyper feature GCN for skeleton-based gait recognition. *Pattern Recognition*, 125:108520, 2022.

[26] Liqi Feng, Yaqin Zhao, Wenxuan Zhao, and Jiaxi Tang. A comparative review of graph convolutional networks for human skeleton-based action recognition. *Artificial Intelligence Review*, pages 1–31, 2022.

[27] Chuanlin Zhang, Kai Cao, Limeng Lu, and Tao Deng. A multi-scale feature extraction fusion model for human activity recognition. *Scientific Reports*, 12(1):20620, 2022.

[28] Yun Liu, Ruidi Ma, Hui Li, Chuanxu Wang, and Ye Tao. RGB-D human action recognition of deep feature enhancement and fusion using two-stream ConvNet. *Journal of Sensors*, 2021:1–10, 2021.

[29] Zhitao Zhang, Zhengyou Wang, Shanna Zhuang, and Jiahui Wang. Toward action recognition and assessment using SFAGCN and combinative regression model of spatiotemporal features. *Applied Intelligence*, 53(1):757–768, 2023.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3420789

[30] Amsaprabhaa M, Nancy Jane Y, and Khanna Nehemiah H. Multimodal spatiotemporal skeletal kinematic gait feature fusion for vision-based fall detection. *Expert Systems with Applications*, 212:118681, 2023.

[31] Huaigang Yang, Ziliang Ren, Huaqiang Yuan, Wenhong Wei, Qieshi Zhang, and Zhaolong Zhang. Multi-scale and attention enhanced graph convolution network for skeleton-based violence action recognition. *Frontiers in Neurorobotics*, 16, 2022.

[32] Wenjie Yang, Jianlin Zhang, Jingju Cai, and Zhiyong Xu. HybridNet: Integrating GCN and CNN for skeleton-based action recognition. *Applied Intelligence*, 53(1):574–585, 2023.

[33] Zhiwei Li, Anyu Zhang, Fangfang Han, Junchao Zhu, and Yawen Wang. Worker Abnormal Behavior Recognition Based on Spatio-Temporal Graph Convolution and Attention Model. *Electronics*, 12(13):2915, 2023.

[34] Natalia Miranda-Cantellops and Timothy K Tiu. Berg balance testing. 2021.

[35] Silu He, Qinyao Luo, Ronghua Du, Ling Zhao, Guangjun He, Han Fu, and Haifeng Li. STGC-GNNs: A GNN-based traffic prediction framework with a spatial–temporal Granger causality graph. *Physica A: Statistical Mechanics and its Applications*, page 128913, 2023.

[36] Minbo Ma, Peng Xie, Fei Teng, Bin Wang, Shenggong Ji, Junbo Zhang, and Tianrui Li. HiSTGNN: Hierarchical spatio-temporal graph neural network for weather forecasting. *Information Sciences*, 648:119580, 2023.

[37] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[38] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[39] Ronald Mutegeki and Dong Seog Han. A CNN-LSTM approach to human activity recognition. In *2020 international conference on artificial intelligence in information and communication (ICAIIC)*, pages 362–366. IEEE, 2020.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Elsa J. Harris, I-Hung Khoo, and Emel Demircan. A survey of human gait-based artificial intelligence applications. *Frontiers in Robotics and AI*, 8, 2022.

[42] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[43] Taotao Cai, Xiangyu Song, and Adnan Mahmood. Dynamic Correlation Adjacency-Matrix-Based Graph Neural Networks for Traffic Flow. *Sensors*, 2023.

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] Sarthak Pati, Siddhesh P Thakur, İbrahim Ethem Hamamcı, Ujjwal Baid, Bhakti Baheti, Megh Bhalerao, Orhun Güley, Sofia Mouchtaris, David Lang, Spyridon Thermos, et al. GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows. *Communications Engineering*, 2(1):23, 2023.

[47] Rohan Shad, John P Cunningham, Euan A Ashley, Curtis P Langlotz, and William Hiesinger. Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. *Nature Machine Intelligence*, 3(11):929–935, 2021.

[48] Ying Xia, Chun-Qiu Xia, Xiaoyong Pan, and Hong-Bin Shen. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9):e51–e51, 2021.

[49] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[50] Weisheng Li, Xuesong Liang, and Meilin Dong. Mdecnn: A multiscale perception dense encoding convolutional neural network for multispectral pan-sharpening. *Remote Sensing*, 13, 2021.

[51] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[52] Mark Alber, Adrian Buganza Tepole, William R Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W Lytton, Paris Perdikaris, Linda Petzold, et al. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ digital medicine*, 2(1):115, 2019.

[53] Gang Yan, Alexander Schmitz, Satoshi Funabashi, Sophon Somlor, Tito Pradhono Tomo, and Shigeki Sugano. SCT-CNN: A Spatio-Channel-Temporal Attention CNN for Grasp Stability Prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2627–2634, 2021.

[54] Zahraa Saddi Kadhim, Hasanen S. Abdullah, and Khalil I. Ghathwan. Automatically Avoiding Overfitting in Deep Neural Networks by Using Hyper-Parameters Optimization Methods. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(05):pp. 146–162, Apr. 2023.

[55] R. Jayakarthik, Aravindan Srinivasan, Sohan Goswami, Shivaranjini, and Mahaveerakannan R. Fall Detection Scheme based on Deep Learning Model for High-Quality Life. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1582–1588, 2022.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on August 10,2024 at 07:26:26 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.