Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2024.046752

ARTICLE

An Approach for Human Posture Recognition Based on the Fusion PSE-CNN-BiGRU Model

Xianghong Cao, Xinyu Wang, Xin Geng*, Donghui Wu and Houru An

School of Building and Environment Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450000, China

*Corresponding Author: Xin Geng. Email: gengxin@email.zzuli.edu.cn

Received: 13 October 2023 Accepted: 04 January 2024 Published: 16 April 2024

ABSTRACT

This study proposes a pose estimation-convolutional neural network-bidirectional gated recurrent unit (PSE-CNN-BiGRU) fusion model for human posture recognition to address low accuracy issues in abnormal posture recognition due to the loss of some feature information and the deterioration of comprehensive performance in model detection in complex home environments. Firstly, the deep convolutional network is integrated with the Mediapipe framework to extract high-precision, multi-dimensional information from the key points of the human skeleton, thereby obtaining a human posture feature set. Thereafter, a double-layer BiGRU algorithm is utilized to extract multi-layer, bidirectional temporal features from the human posture feature set, and a CNN network with an exponential linear unit (ELU) activation function is adopted to perform deep convolution of the feature map to extract the spatial feature of the human posture. Furthermore, a squeeze and excitation networks (SENet) module is introduced to adaptively learn the importance weights of each channel, enhancing the network's focus on important features. Finally, comparative experiments are performed on available datasets, including the public human activity recognition using smartphone dataset (UCIHAR), the public human activity recognition 70 plus dataset (HAR70PLUS), and the independently developed home abnormal behavior recognition dataset (HABRD) created by the authors' team. The results show that the average accuracy of the proposed PSE-CNN-BiGRU fusion model for human posture recognition is 99.56%, 89.42%, and 98.90%, respectively, which are 5.24%, 5.83%, and 3.19% higher than the average accuracy of the five models proposed in the comparative literature, including CNN, GRU, and others. The F1-score for abnormal posture recognition reaches 98.84% (heartache), 97.18% (fall), 99.6% (bellyache), and 98.27% (climbing) on the self-built HABRD dataset, thus verifying the effectiveness, generalization, and robustness of the proposed model in enhancing human posture recognition.

KEYWORDS

Posture recognition; mediapipe; BiGRU; CNN; ELU; attention

1 Introduction

Human posture recognition aims to identify human postures through the analysis of pictures, videos, or wearable sensor data of human movements. This technology is widely utilized in the domains of smart homes, assisted rehabilitation, and automatic driving [1]. Identifying abnormal postures or states, such as falls, limping, fatigue, or even illness, can provide timely warnings or

interventions. In recent years, the increasingly severe aging population and the rising number of people living alone, including the elderly, have heightened the importance of research on real-time recognition of abnormal postures in daily living environments. Researchers are increasingly focusing on providing early warnings and timely treatment for conditions like falls and illnesses. However, due to the complexity of daily living environments and human movements, tracking, detecting, classifying, or identifying human posture in real-time activities is extremely challenging.

Human posture recognition is achieved by a machine that extracts abstract information by simulating the animal eyeball and learns relevant features by simulating human learning ability. Methods for human posture recognition can be divided into static and dynamic posture recognition. Static posture recognition is based on feature extraction from non-continuous frame sequences, where movements cannot be effectively recognized during changes. Good recognition performance is achieved when the movement is static and relatively standard compared to the training set. For example, Ashraf et al. [2] proposed a deep convolution model based on stationary posture to determine the standardization of five yoga movements. Due to the constant change of human posture in the daily home environment, similarities in different actions and occlusion of key body parts make dynamic posture-based recognition a challenging research area.

Dynamic posture recognition is realized based on feature extraction of continuous frame sequences, and dynamic postures such as surveillance video are favored by the majority of researchers due to their ease of acquisition and good real-time performance. In the early stage, the feature extraction of dynamic posture was determined by researchers based on personal experience and subjective will, so it had a high dependence on the amount of knowledge, experience, and number of experiments, which also meant that researchers had to spend a lot of time and energy resources to determine a relatively excellent feature. Later, the deep learning algorithm convolutional neural network (CNN) came into being, which is a method for computers to determine features through self-learning, and is sought after by many researchers due to its objective efficiency.

There are two main stages in the development of CNN-based human posture recognition. The first stage comprises the localization plus recognition detection algorithm, with R-CNN as a representative [3]. In order to meet real-time user requirements, some research teams have improved R-CNN, resulting in the fast R-CNN algorithm [4]; however, it still falls short of high real-time demand. The second stage involves the location-recognition fusion detection algorithm, exemplified by Single Shot MultiBox Detector (SSD) [5], Visual Geometry Group (VGG) [6], and You Only Look Once (YOLO) [7] series algorithms. In order to accurately classify yoga postures, Garg et al. [8] proposed a new Mediapipe-CNN model, which initially employs Mediapipe for human posture skeletonization and then classifies the skeletonized yoga postures. Experiments demonstrate this model's 99.62% accuracy in standard yoga posture classification, a significant improvement over VGG16, InceptionResNetV2, NASNetMobile, InceptionV3, and others. However, due to its binary classification, standard or not, the algorithm suits long-term posture stabilization exercises like yoga in home environments. Abdallah et al. [9] extracted hand models using Mediapipe and employed a CNN network for real-time gesture detection, achieving high real-time performance but less than 89% accuracy across multiple datasets. Hong et al. [10] proposed a Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU) hybrid model. Comparative experiments reveal that single CNN or GRU models suffer from low accuracy and incomplete mining of target time series features, and the Convolutional Neural Network-Long Short Term Memory (CNN-LSTM) hybrid model has lengthy running times with room for accuracy improvement. The accuracy and efficiency of the CNN-GRU hybrid model have been significantly improved, but it cannot meet people's real-time needs. Niu et al. [11] introduced a CNN-BiGRU model, incorporating BiGRU into the convolutional layer for complex high-dimensional

feature extraction and historical and time series correlation. This model's detection accuracy improved by 63.39% over traditional LSTM models, showing high performance. However, the structure design of the algorithm is only suitable for prediction, such as the changing trend of energy use, and the generalization performance is low. Xu et al. [12] proposed an electromyography (EMG) signal-based gesture recognition architecture for Squeeze-and-Excite-Convolutional Neural Network (SE-CNN), conducting experiments on multi-gesture datasets to demonstrate the convolutional network with SE's robustness, although accuracy remains low on datasets like Ninapro DB4.

In order to design a more suitable model for human posture recognition, researchers have continuously absorbed the advantages of various classification models' advantages, achieving significant results in recent years. In 2022, Samaan et al. [13] introduced a Mediapipe-GRU method, currently limited to pose recognition but showing good real-time performance for video sequence-based recognition. Abdallah et al. [9] proposed a stable but low-accuracy Mediapie-CNN-GRU pose recognition technology in 2022. In 2023, Rijayanti et al. [14] developed a multi-class classification system for Mediapipe-mask R-CNN with strong generalization ability, but accuracy still requires enhancement.

In the context of increasing aging and the growing solitary living population, accurately identifying abnormal indoor human postures is highly meaningful. Current algorithms for indoor abnormal posture recognition primarily focus on fall detection. For instance, Wang et al. [15] introduced a wearable fall detection system using an improved CNN algorithm to ascertain if elderly individuals living alone indoors have fallen. Inturi et al. [16] developed a vision-based fall detection scheme utilizing LSTM architecture for indoor fall identification. Alves et al. [17] applied classification algorithms like Multi-Layer Perception (MLP) and K Nearest Neighbors (KNN) to identify indoor falling postures, evaluating the fall detection algorithm's pros and cons. Beyond falls, abnormal behaviors should include broader meanings, such as heartache or bellyache (indicating potential acute medical conditions) and climbing behavior (suggesting possible risky activities), to effectively identify and respond to potentially dangerous actions.

The above analysis demonstrates that in recent years, many posture recognition studies have utilized the skeletonization method of Mediapipe. However, challenges such as the scarcity of diverse abnormal postures and the need for enhanced accuracy in abnormal posture recognition persist. Consequently, this study aims to address how to augment the variety of abnormal postures, extract high-dimensional and multi-class posture feature information comprehensively, identify the intrinsic links between key and other elements, and ultimately improve recognition accuracy.

This study focuses on three main points in the model construction. Firstly, the frame sequence is skeletonized by the Mediapipe model. Secondly, the exponential linear unit (ELU) and BiGRU networks are introduced to enhance the generalization and spatiotemporal feature extraction ability of the model on the foundation of convolutional neural networks. Thirdly, the SENet is incorporated into the model to strengthen attention to crucial features, reduce attention to unimportant feature information, and increase posture recognition accuracy. Two public datasets and one dataset independently developed by the author's team have been used to verify the effectiveness of the proposed algorithm. The main contributions of this study are as follows:

- A fusion model, pose estimation-convolutional neural network-bidirectional gated recurrent unit (PSE-CNN-BiGRU), is proposed based on three models (including Mediapipe, BiGRU, and CNN). In order to reduce the interference of complex environments on feature extraction, the Mediapipe model is used to skeletonize the human posture in the frame sequence. In order to fully exploit the spatiotemporal feature extraction capability of convolutional networks

and improve the accuracy of the model for posture recognition, a two-layer BiGRU temporal feature extraction network and SENet are introduced based on the spatial feature extraction of the CNN network. In order to accelerate model convergence, provide a smoother activation response, and prevent gradient disappearance, the ELU function is added to the CNN model. This hybrid strategy not only strengthens the ability of spatial feature extraction but also strengthens the ability of temporal feature extraction. To the best of the authors' knowledge, this strategy has been applied for the first time in the field of indoor human posture recognition.

- In order to enrich abnormal postures, ensure the diversity of abnormal postures, and meet the need to recognize human postures in actual home scenes, the home abnormal behavior recognition dataset (HABRD) is independently developed by the author's team. In this regard, five normal behaviors, including standing, sitting, lying, walking, and sweeping, and four abnormal behaviors, including heartache, falling, bellyache, and climbing, are used. There is currently no public dataset that contains so many different abnormal behaviors at the same time.

The remainder of this essay is organized as follows: Part 2 details the hybrid model, Part 3 elaborates on the experimental environment and data processing, Part 4 mainly presents the results of multi-model comparative experiments, and Part 5 concludes the study.

2 Methodology

This article presents an improved fusion model based on three modules: Mediapipe, BiGRU, and CNN (PSE-CNN-BiGRU), as shown in Fig. 1. Multi-layer high-latitude feature extraction is a benefit of this model. This model demonstrates stronger generalization ability than other comparable models and offers a higher recognition rate for frame sequence-based human posture identification. A frame sequence, formed by arranging consecutive frames in chronological order, represents an instant state, such as a video or coordinate matrix. This sequence can be captured by sensors or video. By studying this frame sequence, features of human movement, including speed, acceleration, and angular velocity, can be determined. The model is divided into four modules: Module 1, Mediapipe key point extraction; Module 2, time feature extraction; Module 3, space feature extraction; and Module 4, output.

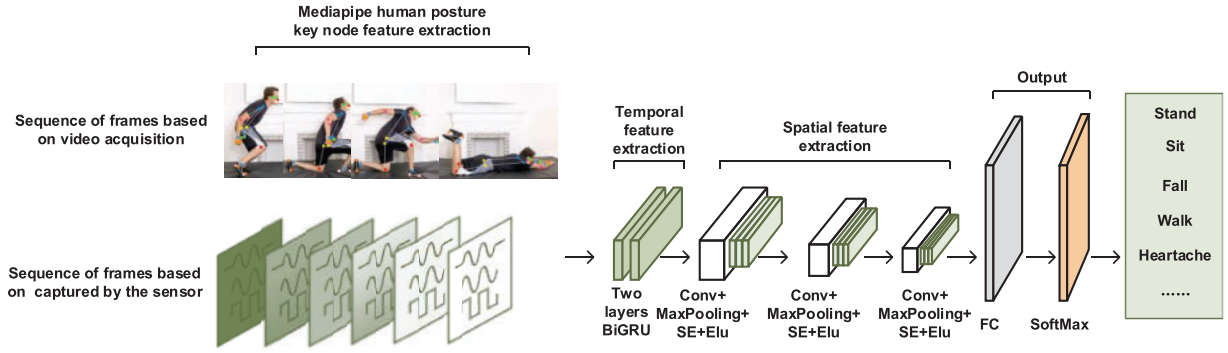


Figure 1: PSE-CNN-BiGRU model

2.1 Module 1: Mediapipe-Based Key Point Extraction

Existing video series-based posture recognition solutions, such as Openpose and Alphapose, face high requirements on algorithm performance, equipment performance, and long processing times. Previously, several scholars have proposed the Mediapipe posture recognition method to address

this issue. The Mediapipe posture identification system analyzes the real-time data stream and extracts human body posture information from the video [18]. It excels in precision, scalability, and effectiveness, offering crucial technical support for fields like human-computer interaction, virtual reality, and posture identification. The Mediapipe-based key point extraction module uses computer vision to identify critical human body parts in a video frame sequence, create a model of the human skeleton, and then input the model into Module 2. The sensor-based posture recognition mainly collects electrical signals through various wearable sensors and inputs the collected data into Module 1. Since human posture in image format cannot be identified, the module inputs the frame sequence into Module 2 before directly storing the data as a 2D information matrix.

The Mediapipe posture extraction network starts to regress after extracting the essential point information. In this process, the human head coordinate information $(H_x, H_y, H_w, H_h)$and hip coordinate information(B_x, B_y) are obtained. Combined with the Vitruvius principle [19], the coordinate information(P_x, P_y, P_w, P_h) of the human posture model and the rotation angle θ of the human posture can be calculated, as shown in Eq. (1).

$$(P_x, P_y, P_w, P_h)\begin{cases} P_x = H_x + 0.5 \times H_x - 4 \times H_h \\ P_y = H_y \\ P_w = 8 \times H_w \\ P_h = 8 \times H_h \\ \theta = \arctan \dfrac{B_x - (H_x + 0.5 \times H_w)}{B_y - (H_y + 0.5 \times H_h)} \end{cases} \tag{1}$$

The model uses a 2D video sequence as input. The data are preprocessed after obtaining enough coordinates of important human posture-related sites. The 99-dimensional information of 33 x, y, w horizontal and vertical depth of each coordinate is selected to establish a multi-type posture dataset and input the data in the form of a 3 × 33 dimensional matrix into Module 2 to extract the time feature.

2.2 Module 2: BiGRU-Based Temporal Feature Extraction

The BiGRU model, based on RNN development, emerged due to RNN's inability to solve the long-term dependency problem. In order to address this, three control gates were added, leading to the creation of LSTM. Later, to capture the relationship between the forward and backward information of input sequences and comprehensively capture contextual information in the sequences, BiLSTM, composed of two LSTMs, was developed. However, BiLSTM brought new issues, such as complex structure, a large number of parameters, and low computational efficiency. To simplify the BiLSTM, some researchers designed BiGRU [20] with only two control gates, faster convergence speed, and stronger representation capability.

In Module 2, the BiGRU network's two layers were used for time feature extraction. Fig. 2 displays the time feature extraction framework created in this study.

The BiGRU structure designed in this study is shown in Eqs. (2) and (3). First, $i = \begin{pmatrix} x_{(1,33)} \\ x_{(2,33)} \\ x_{(3,33)} \end{pmatrix}$ represents an action of human posture, and X actions are collected after the frame sequence is recognized, and at this time, the action set is represented by matrix I: $I = i_1, i_2, i_3, \ldots, i_X$, and the

long sequence matrix is input into the two-layer BiGRU structure set up.

$$S(x) = W \times \left[\begin{array}{l} \vec{g}_x = \begin{pmatrix} \sigma(W_Z I_t + U_Z h_{t-1}) \\ \sigma(W_r I_t + U_r h_{t-1}) \\ \tanh\left(W I_t + r_t \otimes U h_{t-1}\right) \\ (1 - Z_t) \otimes h_{t-1} + Z_t \otimes \tilde{h}_t \end{pmatrix} \left(\vec{g}_{x-1}, i_x\right); \\ \overleftarrow{g}_x = \begin{pmatrix} \sigma(W_Z I_t + U_Z h_{t-1}) \\ \sigma(W_r I_t + U_r h_{t-1}) \\ \tanh\left(W I_t + r_t \otimes U h_{t-1}\right) \\ (1 - Z_t) \otimes h_{t-1} + Z_t \otimes \tilde{h}_t \end{pmatrix} \left(\overleftarrow{g}_{x+1}, i_x\right) \end{array} \right] + b \tag{2}$$

$$f(O_x) = Soft\max(S(x)); Output = f(f(O_x)) \tag{3}$$

where $\begin{pmatrix} \sigma(W_Z I_t + U_Z h_{t-1}) \\ \sigma(W_r I_t + U_r h_{t-1}) \\ \tanh\left(W I_t + r_t \otimes U h_{t-1}\right) \\ (1 - Z_t) \otimes h_{t-1} + Z_t \otimes \tilde{h}_t \end{pmatrix}$ is the most basic unit of the BiGRU; t is the time; I_t is the sequence unit of the input; W_Z is the weight of I_t; h_{t-1} is the data unit that stores the past $t-1$ moments; U_z represents h_{t-1} weights; σ represents the sigmoid activation function; r_t is the reset gate; $\tilde{h}_t$ represents the introduction of new content as a new address, and tanh is the nonlinear activation function; $f(x)$ represents the SoftMax activation processing of $S(x)$; $f(f(x))$ represents the two-level setup of the BiGRU; the processed long sequence $I' = i'_1, i'_2, i'_3, \ldots, i'_X$ is passed to Module 3, the CNN-SE spatial feature extraction module.

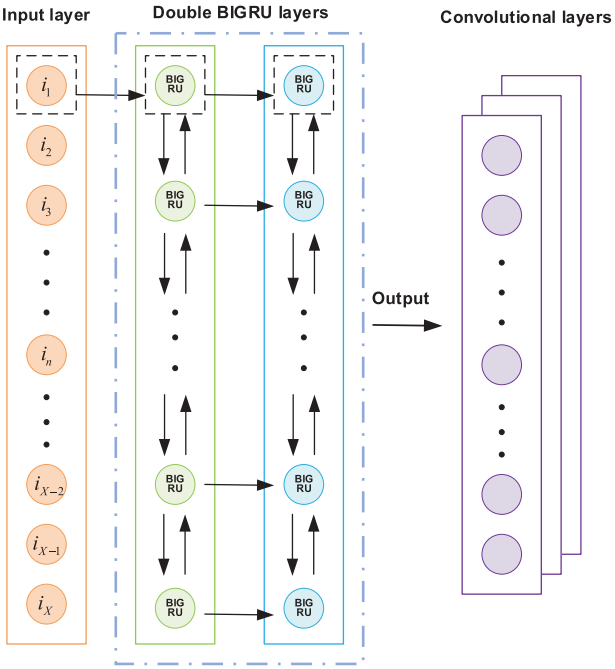


Figure 2: Temporal feature extraction framework

2.3 Module 3: CNN-SE-Based Spatial Feature Extraction

The current mainstream attention mechanisms include Selective Kernal Networks (SKNet), Convolutional Block Attention Module (CBAM), and SENet [21]. Although the first two have the advantage of considering spatial and channel interaction information, their structures are complex and have a large number of parameters, making them less suitable for deployment and application in real scenarios. SENet has a simpler structure, is easy to deploy, and can achieve good results while ensuring a lightweight design.

Module 3: Spatial Feature Extraction. This module consists of three progressive layers, each containing convolution, pooling, SENet, and ELU. Fig. 3 displays the spatial feature extraction structure proposed in this paper.

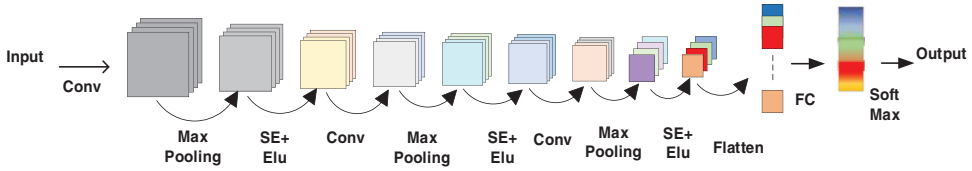


Figure 3: CNN-SE spatial feature extraction structure

Take the long sequence $I' = i'_1, i'_2, i'_3, \ldots, i'_X$ as input $I'_{input} \in \mathbb{R}^{H\times W\times C}$, the convolution kernel filter = [3, 1], the convolution kernel sliding step stride = 1, perform convolution $V = [v_1, v_2, \cdots, v_C]$ operation, and express output $I'_{output} \in \mathbb{Z}^{H\times W\times C}$ with $Z = [z_1, z_2, \cdots, z_C]$. SENet [22] are added to the dimensionality pooling P (Z) operation on top of the convolution result, and the CNN-SE binding expression described is shown in Eqs. (4) to (6).

$$P(Z)\begin{cases} H_{new} = (H-3) \div 1 + 1 \\ W_{new} = (W-1) \div 1 + 1 \end{cases} \tag{4}$$

$$\bar{Z}_C = \left(\sigma \left(W_2 \delta \left(W_1 \left(\begin{array}{l} \frac{1}{H_{new} \times W_{new}} \sum_{i=1}^{H_{new}} \sum_{j=1}^{W_{new}} z_1(i,j) \\ \frac{1}{H_{new} \times W_{new}} \sum_{i=1}^{H_{new}} \sum_{j=1}^{W_{new}} z_2(i,j) \\ \vdots \\ \frac{1}{H_{new} \times W_{new}} \sum_{i=1}^{H_{new}} \sum_{j=1}^{W_{new}} z_C(i,j) \end{array} \right) \right) \right) \right)_C \odot \left(\sum_{s=1}^{C'} v_C^S \times I'^{\,S}_{input} \right) \tag{5}$$

$$ELU(z) = \begin{cases} z, z > 0 \\ \alpha (e^z - 1), z \le 0 \end{cases} \tag{6}$$

where $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ is the fully connected layer parameter, which plays the role of the compression function $W_2 \in \mathbb{R}^{C\times \frac{C}{r}}$; σ is the Sigmoid function; δ is the ELU function; r is the compression parameter. The Rectified Linear Unit (ReLU) function in the CNN was improved by the ELU activation function [23] because the ELU function has several advantages over the ReLU function, including the elimination of the dead neuron problem, improved smoothness in negative input, a faster convergence rate, and increased resistance to noise. The structure of CNN-SE spatial feature extraction is depicted in Fig. 3. The PSE-CNN-BiGRU model's full connection and SOFTMAX output portion come next.

2.4 Module 4: PSE-CN-BiGRU-Based Output

Module 4: Output Part. This module includes the full connection layer and SOFTMAX for each layer. SoftMax probabilizes each input feature of multiple categories once features are substantially purified in the fully linked layer. After obtaining the SoftMax results, the maximum likelihood loss function is calculated [24], and the calculation process is shown in Eqs. (7) and (8).

$$\sigma\left(Z\right)_j = y_j = \frac{e^{Zj}}{\sum_{k=1}^{K} e^{Zj}}, j = 0, 1, 2, 3, \ldots, (K-1) \tag{7}$$

$$Loss_j = -\sum_{j=1}^{N} y_j \log \widehat{y}_j \tag{8}$$

where K represents the problem solved as a K classification problem; j represents class j; Z represents the input function of the input vector after feature extraction, which will then be used to calculate the probability of each class; N is the total sample size; y_j represents the label value classified into class j; $\widehat{y}_j$ represents the predicted value. In this way, the model can fully extract, recognize, and classify the spatiotemporal features of human posture.

3 Experimental Configurations

3.1 Data Sets

In order to demonstrate the effectiveness of PSE-CNN-BiGRU, the authors conducted experimental validation using two publicly available datasets, UCIHAR and HAR70PLUS, as well as a dataset developed by their team, HABRD.

The public dataset UCIHAR [25] includes triaxial accelerometer data and data gathered by the smartphone's gyroscope. The authors meticulously gathered 10,299 sets of experimental data, which included 561 features and 6 action categories. Thirty volunteers, aged between 19 and 48, participated in the study.

The authors used the integration of the camera and accelerometer in the Axivity AX3 to obtain the public dataset HAR70PLUS [26]. They collected 52,063 groups of experimental data, which included 6 features and 7 action categories. The setting for the trial was semi-structured free home living. The study involved 18 volunteers aged 70 to 95, with physical characteristics ranging from healthy to weak.

Most existing indoor abnormal posture recognition algorithms focus on the recognition of fall actions. For instance, Wang et al. [15] proposed a fall detection system. Beyond falls, abnormal behaviors should encompass a broader range of meanings, such as heartache or bellyache (indicating the possibility of an acute medical condition) and climbing behavior (indicating potential engagement in risky activities). To enrich the abnormal posture dataset for people living alone, the authors' team independently constructed the HABRD dataset. They set up three sets of monitoring devices in a fixed indoor environment to capture front, back, and side views of the subjects. The video files, recorded at 60 frames per second with a resolution of 1920 × 1080 pixels, were saved in MP4 format. Six volunteers, including three men and three women, aged between 20 and 50, were selected to perform various movements in the room for more than 5 min each. The nine types of movement included normal behaviors (standing, sitting, lying, walking, and sweeping) and abnormal behaviors (heartache, falling, bellyache, and climbing). The obtained MP4 files were input into the skeletonization module of Mediapipe, mentioned in Section 2.1, and the dataset was extracted. After filtering and cleaning out incomplete feature classes, a total of 55,082 sets of data containing 99 feature values were retained

and saved in. xlsx format. The authors adjusted the data volume and proportion of different actions to ensure that the data volume of various categories was relatively balanced.

In this study, the training set (80%) and testing set (20%) were randomly chosen from the UCIHAR, HAR70PLUS, and HABRD datasets, respectively. Relevant studies have demonstrated that random dataset splitting is beneficial for assessing the model's effectiveness, confirming its impact, and adjusting its parameters [27]. The experimental specifications of each dataset are shown in Table 1.

Table 1: Experimental specifications of the dataset

	UCIHAR		HAR70PLUS		HABRD	
Type of action	Amount of data	Percentage	Amount of data	Percentage	Amount of data	Percentage
Walking	1722	16.7%	86560	16.6%	4783	8.7%
Standing	1906	18.5%	91822	17.6%	5263	9.6%
Lying	1944	19.9%	62344	11.9%	7570	13.7%
Sitting	1777	17.3%	81506	15.7%	9210	16.7%
Walking upstairs	1544	13.9%	80120	15.4%	\	\
Walking downstairs	1406	13.7%	63840	12.3%	\	\
Shuffling	\	\	54440	10.5%	\	\
Sweeping	\	\	\	\	5921	10.8%
Heartache	\	\	\	\	5210	9.5%
Falling	\	\	\	\	6351	11.5%
Bellyache	\	\	\	\	4978	9.0%
Climbing	\	\	\	\	5796	10.5%

3.2 Assessment Indicators

The performance of the human posture recognition model put out in this research is quantitatively assessed using Accuracy, Precision, Recall, and F1-score. Table 2 displays the assessment metrics.

Table 2: Evaluation indicators

	The prediction is true	The prediction is false
The label is positive	TP (true positive)	FN (false negative)
The label is negative	FP (false positive)	TN (true negative)

According to Eq. (9), Accuracy represents the ratio of true samples to all samples.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

According to Eq. (10), Precision represents the ratio of samples with a positive label and a true prediction to those with a positive prediction.

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

According to Eq. (11), Recall represents the likelihood that the sample's true sample will be correctly anticipated.

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

According to Eq. (12), the F1-score represents the precision and recall harmonic mean in the [0, 1] range.

$$F1 - \text{score} = \frac{2PR}{P + R} \tag{12}$$

3.3 Parameter Settings

Configuration of the Model Training Environment: NVIDIA RTX 4090, Windows 10 operating system, Intel I5-10500 CPU, 16 GB of RAM. The Python platform is used for application deployment.

Training Details: Small changes in model parameters significantly affect the performance of the whole model, which is crucial for determining the optimal parameter combination of PSE-CNN-BiGRU. In selecting a model optimizer, the authors chose the Adam optimizer [28], which usually converges faster to a local optimal solution, especially in the case of large-scale data and parameters. This is advantageous for training large deep-learning models. Through numerous experiments, it was found that the accuracy rate did not improve further after 150–200 epochs, remaining stable, prompting an early stop at 200 epochs. The setting of Batch Size is critical to the model. Only by selecting the appropriate size can one balance efficiency and effectiveness and achieve better training outcomes.

As selecting a Batch Size that is a power of 2 can improve training efficiency [29], experiments on the model using Batch Sizes of 64, 128, 256, 512, and 1024 were conducted, with a fixed learning rate of 0.0001.

Fig. 4 shows the impact of different Batch Sizes on accuracy and training time. The experiment indicates that a Batch Size of 256 yields the best training performance in terms of accuracy. Regarding training time, there is little difference among different Batch Sizes, and when the Batch Size is set to 256, the training time is relatively shorter. The parameter value for the Batch Size was set to 256 through a comparative analysis.

The Learning Rate is a crucial parameter in the training process. A reasonable Learning Rate ensures the model converges to the minimum point rather than a local optimum or saddle point. Based on a Batch Size of 256, the training effects of different learning rates were compared. Fig. 5 illustrates the impact of various Learning Rates on model accuracy and training time. The experiment indicates that the model maintains high accuracy with relatively less training time when the Learning Rate is set to 0.001. Consequently, the parameter value for the learning rate is established at 0.001.

The Sliding Window's length is critical for feature extraction. A window that is too small can lead to insufficient feature extraction, whereas a window that is too large might contain multiple activities, interfering with experimental results. Similarly, the Number of Neurons significantly affects the model's performance. Experiments were conducted on the publicly available UCIHAR dataset [25] to evaluate the impact of varying Sliding Window Lengths and Numbers of Neurons on the model's

performance. The Length of The Sliding Window was set from 10 to 100, and Neuron Numbers were set to 32, 64, 128, 256, 512, and 1024. As depicted in Fig. 6, the model performs optimally with a window length of 60 and 256 neurons. Thus, the length of the sliding window is fixed at 60, and the number of neurons is at 256.

To sum up, the model's parameters are described in Table 3.

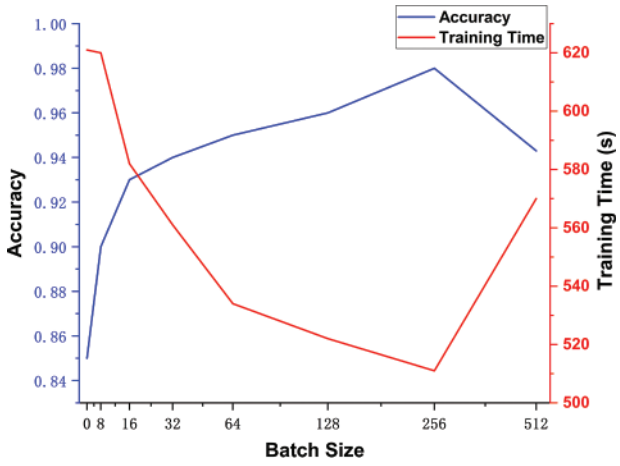


Figure 4: Effect of batch size on model training

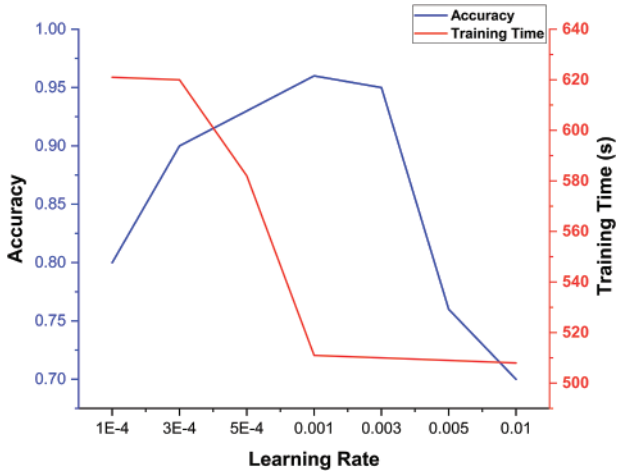


Figure 5: Effect of learning rate on model training

4 Experiments and Analysis

This study plans two experiments to test the PSE-CNN-BiGRU model's effectiveness and generalization. The first, intra-session validation, involved training and verifying the PSE-CNN-BiGRU model using the UCIHAR [25], HAR70PLUS [25] (data collection based on sensors), HAR70PLUS [26] (data collection based on sensors and video sequences), and HABRD (data collection based on video sequences). This allows for an objective assessment of the proposed human posture recognition

algorithm compared to similar models. The second experiment entails deploying an application terminal model for subjective evaluation of the PSE-CNN-BiGRU model.

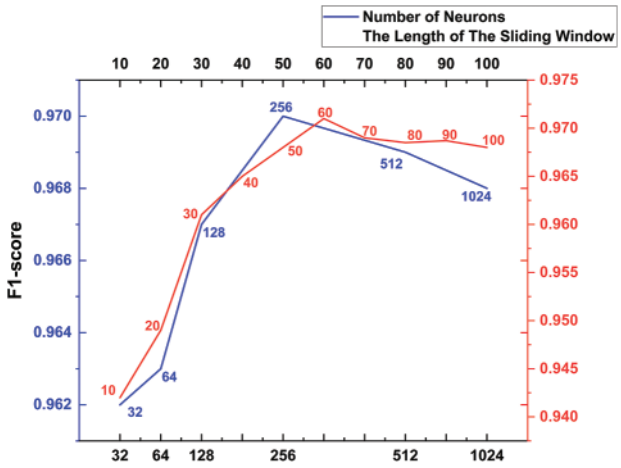


Figure 6: Effects of different sliding window lengths and number of neurons on F1-scores of PSE-CNN-BiGRU

Table 3: Model parameter

Parameters	UCIHAR	HAR70PLUS	HABRD
Enter the window length n_w	60	60	60
Number of hidden neurons n_r	256	256	256
Convolution kernel size n_f	8	16	16
The number of convolution kernels d	64	64	64
Pool kernel size n_p	4	4	4
Number of fully connected (FC) layer neurons n_c	1000	1000	1000

4.1 Intra-Session Validation

The goal of intra-session validation is to assess the evaluation indicators produced by the experiment, compare PSE-CNN-BiGRU with the prior series of models, and confirm the suitability and efficacy of the PSE-CNN-BiGRU model. Accuracy, F1-score, and confusion matrix are among the evaluation indices created for this investigation. The analysis of precision, recall, and other variables is included in the confusion matrix.

4.1.1 Accuracy

The experimental specifications for the UCIHAR, HAR70PLUS, and HABRD datasets are listed in Table 1 of Section 3.1. In order to prove the advantages of the PSE-CNN-BiGRU model compared to other alternatives, five models and the PSE-CNN-BiGRU model were selected to conduct comparative tests on three different human posture datasets. The five models are representative of the CNN model [30], BiGRU model [20], Garg et al. [8] model improved based on CNN [30],

Samaan et al. [13] model improved based on GRU [31], and Li et al. [32] model improved based on CNN-BiGRU [33]. The Accuracy of PSE-CNN-BiGRU and similar models on different datasets is shown in Table 4.

Table 4: Accuracy of PSE-CNN-BiGRU and similar models on different datasets

	UCIHAR	HAR70PLUS	HABRD
	Accuracy	Accuracy	Accuracy
CNN	90.13%	85.04%	95.35%
BiGRU	96.11%	83.23%	96.29%
Garg et al. [8]	92.33%	84.77%	95.28%
Samaan et al. [13]	94.37%	77.14%	94.62%
Li et al. [32]	98.64%	87.78%	97.02%
PSE-CNN-BiGRU	99.56%	89.42%	98.90%

On the UCIHAR dataset, the model suggested in this study was compared to other alternatives. Table 4 shows that the accuracy rates for CNN, BiGRU, Garg et al. [8], Samaan et al. [13], Li et al. [32], and the PSE-CNN-BiGRU model are 90.13%, 96.11%, 92.33%, 94.37%, 98.64%, and 99.56%, respectively. Compared to other models, PSE-CNN-BiGRU has an average improvement of 5.24%. Fig. 7 displays the accuracy curve of posture categorization. It shows that the accuracy of traditional CNN and BiGRU models fluctuates greatly with epoch, and on this dataset, the best performance of BiGRU exceeds that of CNN. This is related to feature extraction and recognition. Some datasets make it easy to recognize spatial features, and some make it easy to recognize temporal features. It can be seen that the Li et al. [32] and PSE-CNN-BiGRU models, which combine the advantages of the two, are stable and have higher accuracy. PSE-CNN-BiGRU and Li et al. [32] models are more accurate than the Garg et al. [8] and Samaan et al. [13] models. It has been demonstrated that the addition of the time feature extraction module considerably increases the capability of capturing motion data, and overfitting or underfitting issues are successfully avoided. With further training iterations, the accuracy rate reported by Li et al. [32] progressively climbed until it reached 98.64%. This shows that Li et al.'s [32] feature mining capabilities and accuracy are very good. The performance of the PSE-CNN-BiGRU model is similar to Li et al. [32] in the later stages of training. However, it converges quickly in the beginning and gets closer to accuracy faster than Li et al. [32]. This demonstrates that the addition of the Mediapipe and SE modules can enhance the PSE-CNN-BiGRU model's capacity to recognize and categorize action elements.

On the HAR70PLUS dataset, the PSE-CNN-BiGRU was compared to related models. Table 4 shows that the accuracy rates for CNN, BiGRU, Garg et al. [8], Samaan et al. [13], Li et al. [32], and the PSE-CNN-BiGRU model are 85.04%, 83.23%, 84.77%, 77.14%, 87.78%, and 89.42%, respectively. Compared to other models, PSE-CNN-BiGRU has an average improvement of 5.83%. Fig. 8 displays the accuracy curve of posture categorization. The classification accuracy of all models for the posture of the elderly has decreased, as shown in Fig. 8. This shows that each model has a high error rate when analyzing elderly features and that elderly movement features are harder to extract and categorize than those of middle-aged and young persons. There is a rising market demand for technologies that can precisely detect and categorize the postures of persons who live alone, especially the elderly. The accuracy of traditional CNN and BiGRU models fluctuates greatly with epoch, and on this dataset, the best performance of CNN exceeds that of BiGRU. This indicates that the spatial features of this

dataset are easier to identify. It can be seen that the improved Garg et al. [8], Li et al. [32], and PSE-CNN-BiGRU models are stable and have higher accuracy. The PSE-CNN-BiGRU model suggested in this study is nearly equal to other models in terms of initial convergence speed, but it achieves the best accuracy of 89.4%. This shows that the PSE-CNN-BiGRU model has a reasonable structure design, integrates the advantages of many other models, and has strong robustness and high accuracy in the field of posture recognition.

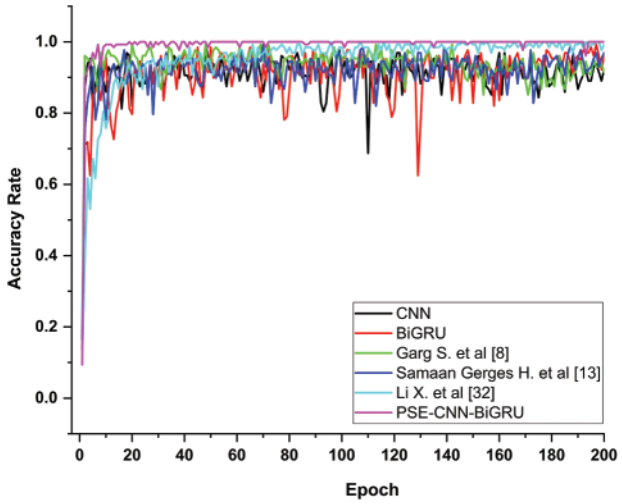


Figure 7: The posture classification accuracy curve between PSE-CNN-BiGRU and similar models on the UCIHAR dataset

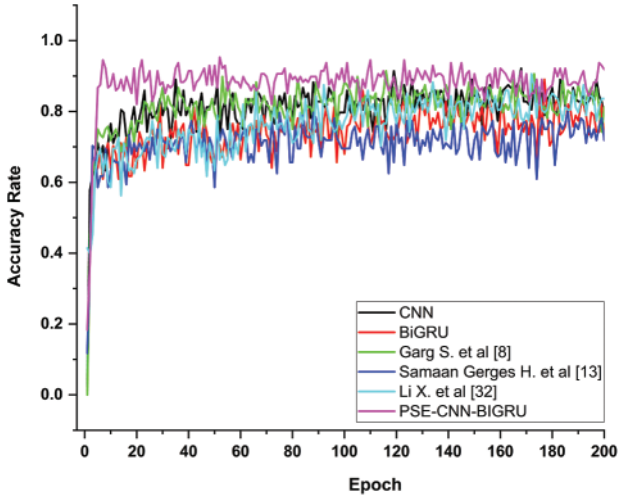


Figure 8: The posture classification accuracy curve between PSE-CNN-BiGRU and similar models on the HAR70PLUS dataset

In the HABRD dataset, PSE-CNN-BiGRU was compared with related models. Table 4 shows that the accuracy rates for CNN, BiGRU, Garg et al. [8], Samaan et al. [13], Li et al. [32], and PSE-CNN-BiGRU are 95.35%, 96.29%, 95.28%, 94.62%, 97.02%, and 98.90%, respectively. Compared to other models, PSE-CNN-BiGRU demonstrated an average improvement of 3.19%. Fig. 9 illustrates

the posture classification accuracy curve. The accuracy of the traditional CNN and BiGRU models fluctuated greatly with each epoch. In this dataset, the best performance of BiGRU exceeded that of CNN, suggesting that the temporal characteristics of this dataset are more easily identifiable. It is evident that the enhanced Garg et al. [8], Li et al. [32], and PSE-CNN-BiGRU models, based on the two, are stable and more accurate. As Fig. 9 shows, all models' convergence rates were quick and similar in the early stages, with some epoch accuracy values of CNN, Garg et al. [8] and Li et al. [32] being greater than those of the PSE-CNN-BiGRU model. This indicates that, in some circumstances, the performance of this model is comparable to that of similar models. However, the PSE-CNN-BiGRU model maintained its high accuracy in later stages when the experiment stabilized, indicating a higher level of accuracy and applicability.

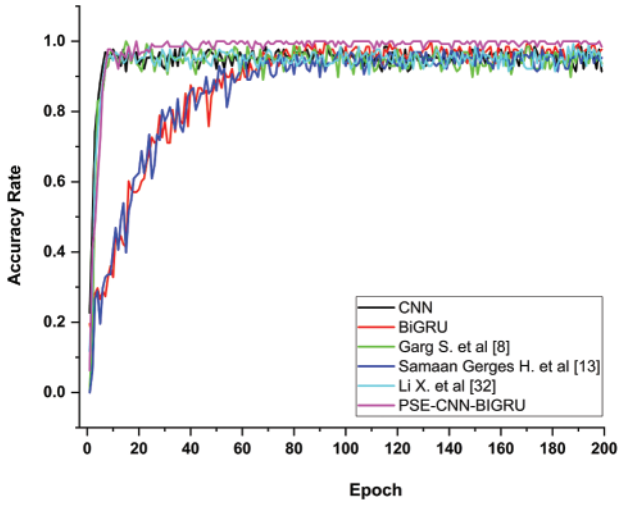


Figure 9: Posture classification accuracy curve between PSE-CNN-BiGRU and similar models on the HABRD dataset

4.1.2 F1-Score

The most direct and efficient way to assess a model is by its accuracy, but imbalanced datasets can lead to false findings. Evaluating accuracy in unbalanced and noise-sensitive datasets has inherent flaws. The experiment compared and examined the F1-score values of each model on the datasets, as shown in Table 5, to thoroughly consider the benefits and drawbacks of each indicator and ensure the efficacy of model evaluation when selecting evaluation indicators.

Table 5: F1-score for different models on various datasets

	UCIHAR	HAR70PLUS	HABRD
	F1-score	F1-score	F1-score
CNN	89.71%	65.81%	95.26%
BiGRU	96.24%	57.93%	96.11%
Garg et al. [8]	78.85%	62.83%	95.01%

(Continued)

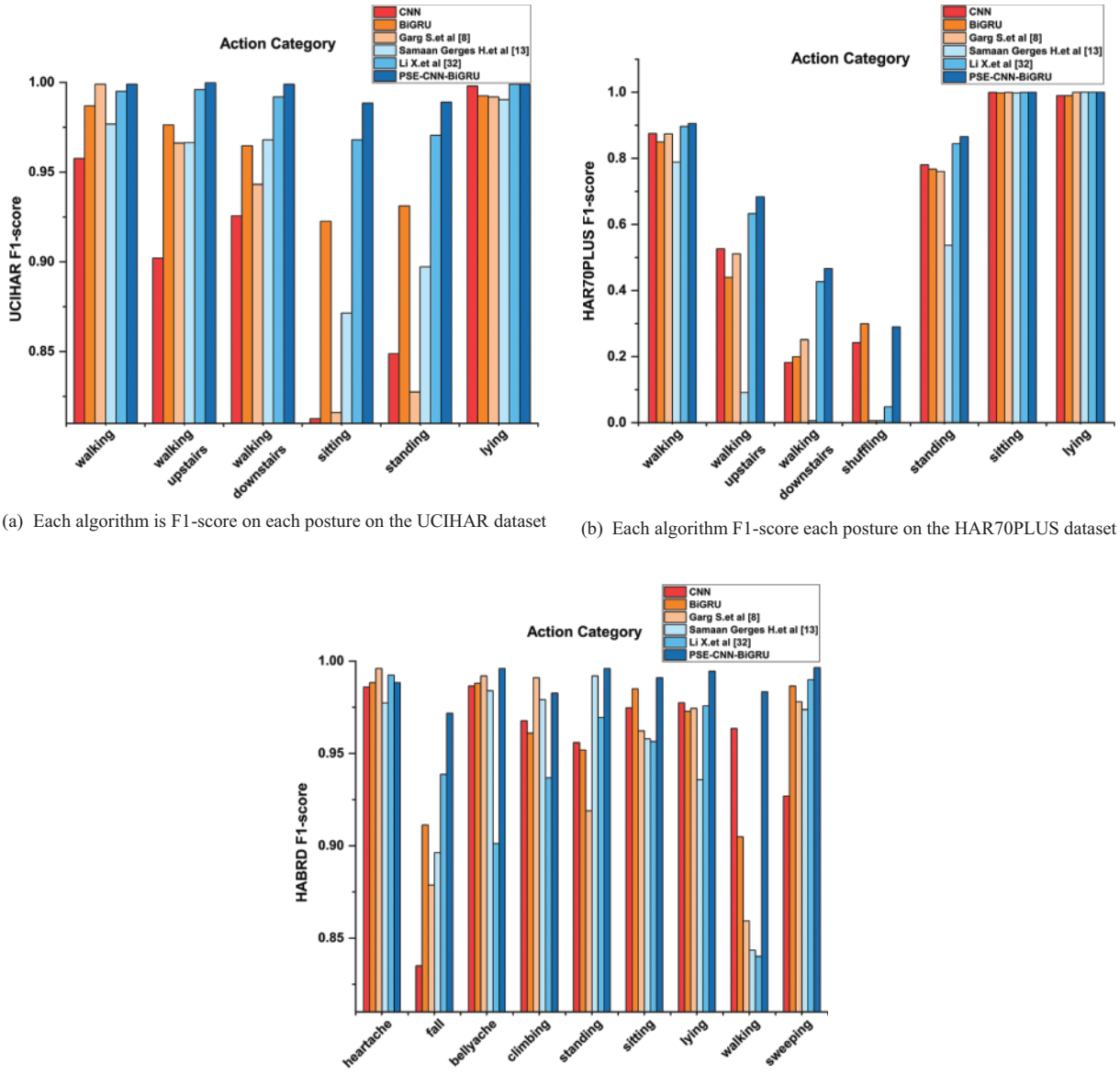
Table 5 (continued)

	UCIHAR	HAR70PLUS	HABRD
	F1-score	F1-score	F1-score
Samaan et al. [13]	94.51%	48.79%	94.89%
Li et al. [32]	98.69%	69.26%	94.45%
PSE-CNN-BiGRU	99.59%	74.44%	98.89%

Table 5 reveals that the ability of CNN and BiGRU to extract features and recognize postures cannot be judged solely on these grounds, as it is closely related to various external factors such as dataset details and device deployment. It is apparent that Li et al. [32] and PSE-CNN-BiGRU, which combine the strengths of models such as CNN and BiGRU, have significantly improved the recognition effect. Whether in public datasets like UCIHAR and HAR70PLUS or the HABRD dataset created by the authors' team, the PSE-CNN-BiGRU model maintained a relatively high F1-score. This demonstrates that PSE-CNN-BiGRU is more reasonable in structural design and more comprehensive in extracting features of different dimensions, exhibiting strong robustness and generalization.

Fig. 10 compares PSE-CNN-BiGRU and related models on various datasets in terms of F1-scores for all human postures. The analysis revealed that both CNN and BiGRU models have their advantages in feature extraction and recognition across different dimensions. As can be seen from Fig. 10, both the CNN and BiGRU models have generally good results for the recognition of each posture, whether on the published dataset or the dataset created by the author's team. At the same time, it can be clearly seen that the PSE-CNN-BiGRU model maintains a leading level of accurate posture recognition and has better stability because it integrates the advantages of multiple models. In the UCIHAR dataset, as shown in Fig. 10a, the performance of the PSE-CNN-BiGRU and Li et al. [32] models are comparable. Both achieved F1-scores above 99% in walking, walking upstairs, walking downstairs, and lying, indicating the PSE-CNN-BiGRU model is more accurate in recognizing sitting and standing motions. Fig. 10a also reveals that models developed by CNN, BiGRU, Garg et al. [8], Samaan et al. [13], and Li et al. [32] frequently misinterpret sitting and standing movements, often confusing them with walking, walking upstairs, or walking downstairs. Sitting movements can also be mistaken for walking, standing, and other motions. This also explains why sitting and standing movements have lower F1-scores. Fig. 10a shows that walking upstairs, walking downstairs, and shuffling are frequently mistaken for walking or standing, according to an analysis of senior people's daily posture. Realistic factors such as slow daily behavior and the small movement range of the elderly jointly account for why the F1-score of each model is generally low during the test. The F1-score of the PSE-CNN-BiGRU model is 68.3%, which is 0.05% greater than that of Li et al. [32] model, 60% higher than that of Samaan et al. [13] model and 17% higher than that of Garg et al. [8] model. The F1-score of the PSE-CNN-BiGRU model for downstairs is 47.2%, which is 8.05% higher than that of the Li et al. [32] model, 40% higher than that of the Samaan et al. [13] model, and 27% higher than that of the Garg et al. [8] model. The F1-score of the PSE-CNN-BiGRU model for shuffling is 30.3%, It is 28.05% higher than Li et al. [32] model, 27% higher than Samaan et al. [13] model, and 28% higher than Garg et al. [8] model. This shows that PSE-CNN-BiGRU has a reasonable structure design, integrates the advantages of multiple models for feature extraction and recognition in different dimensions, and has higher generalization ability and robustness. PSE-CNN-BiGRU has obvious benefits in the identification and classification of abnormal and non-abnormal human postures at

home in Fig. 10c, with its F1-score of diverse postures being no less than 98%. Particularly for the recognition of walking motion, the error rate is below 0.007%. This demonstrates better accuracy in the convolutional network with a spatiotemporal feature extraction module and SENet.



(a) Each algorithm is F1-score on each posture on the UCIHAR dataset

(b) Each algorithm F1-score each posture on the HAR70PLUS dataset

(c) Each algorithm F1-score each posture on the HABRD dataset

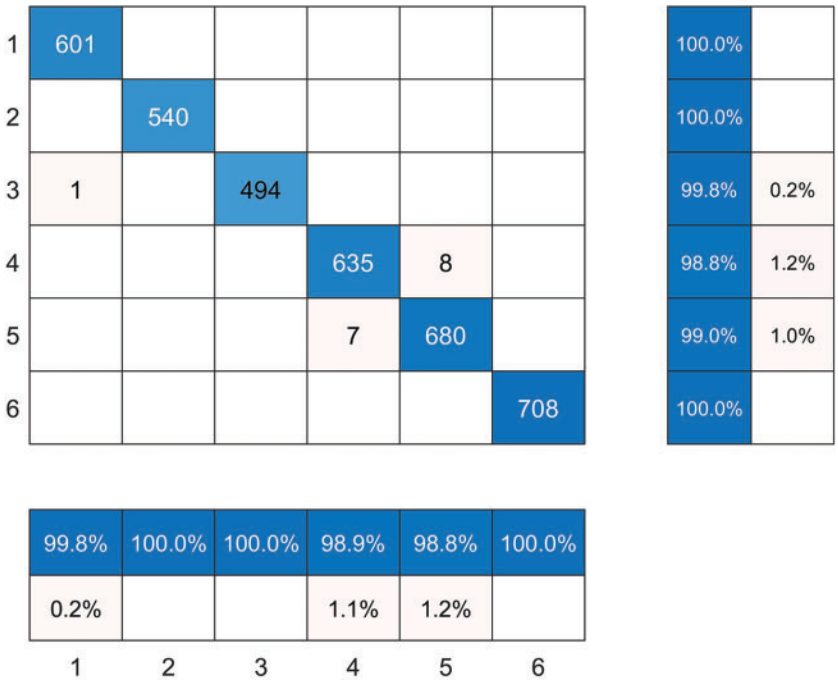
Figure 10: F1-score performance of the PSE-CN-BiGRU model and CNN, BiGRU, Garg et al. [8], Samaan et al. [13], Li et al. [32] model on different datasets: (a) F1-score for each posture of each model on the UCIHAR dataset; (b) F1-score for each model on the HAR70PLUS dataset F1-score for each posture on the dataset; (c) F1-score for each posture of each model on the HABRD dataset

4.1.3 Confusion Matrix

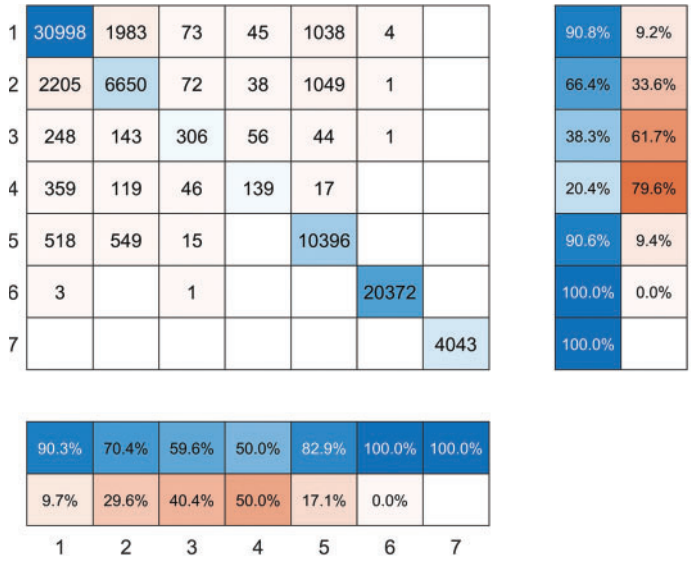
Fig. 11 displays the normalized confusion matrix produced by testing the PSE-CNN-BiGRU model on three datasets: UCIHAR, HAR70PLUS, and HABRD, as suggested in this study. Fig. 11a is a normalized confusion matrix based on the UCIHAR dataset. The six different true action types are labeled on the left, and the six different predicted action types are labeled at the bottom. The precision of each action type is shown in the blue box at the bottom, and the recall of each action type is in the blue box to the right. In the testing set, only 0.2% of actual walking movements were incorrectly predicted as downstairs movements, 1.1% of actual sitting movements as standing movements, and 1.2% of actual standing movements as sitting movements, as is intuitively obvious. The accuracy of all action types is more than 98.8%, indicating the model's high accuracy in identifying samples with positive predictions. All action types had recall rates exceeding 98.8%, demonstrating the model's ability to successfully identify more than 98.8% of the actual positive samples. The high F1-score of the PSE-CNN-BiGRU model in Fig. 10a demonstrates its validity. Fig. 11b presents a normalized confusion matrix based on the HAR70PLUS dataset. As shown, 40.4% of the movements of walking downstairs and shuffling were incorrectly predicted as movements other than lying, and 50% as walking, walking upstairs, or walking downstairs, respectively. The low accuracy of these three motions suggests the model's low accuracy in identifying positive samples predicted by these movements. Similarly, the three types of action recall are not high, with shuffling being the lowest at 20.4%, indicating the model's difficulty in accurately identifying the real positive samples of such actions. This supports the poor F1-score outcomes of the two acts in Fig. 10b and provides additional justification for the low resolution of the elderly posture dataset. It highlights the impact of genuine factors like slow daily behavior and small motion magnitude on the posture detection effect of the elderly and the strong consumer demand for elderly posture recognition at home. Figs. 8, 10b and 11 demonstrate the PSE-CNN-BiGRU model's considerable advantages in recognizing geriatric postures, showcasing its generalizability and applicability. Fig. 11c shows a normalized confusion matrix based on the HABRD dataset. Only 2.5% of the 9 postures seen in this lab's collection of 8 walking motions were mistaken for falling and standing. The model has high accuracy for the samples expected to be positive samples, as the accuracy of all types of actions reached more than 95.7%. Additionally, all types of action recall reached 98.7%, indicating the model's ability to correctly identify more than 98.7% of the true positive samples. Further evidence of the PSE-CNN-BiGRU model's significant benefits over other models is shown in Fig. 10c, which highlights the model's high F1-score for all types of actions, particularly walking movements.

Based on Tables 4, 5, and Figs. 7 to 11, it can be concluded that:

- PSE-CNN-BiGRU possesses a well-structured design, integrating spatiotemporal multi-dimensional feature extraction capabilities and enhancing the network's selective attention to important features, resulting in robustness and generalization. The PSE-CNN-BiGRU model has demonstrated high accuracy across all experiments. On the UCIHAR dataset, the testing set accuracy was 99.60%, and the training set accuracy was 99.56%. On the HAR70PLUS dataset, the training set accuracy reached 89.41%, and the testing set accuracy matched this figure. On the HABRD dataset, the training set accuracy was 99.54%, with the testing set accuracy at 98.90%.
- PSE-CNN-BiGRU performs better in accuracy and F1-score when compared to similar approaches. Additionally, practical challenges such as delayed behavior and limited motion range in the elderly make it difficult to extract and accurately identify posture features.

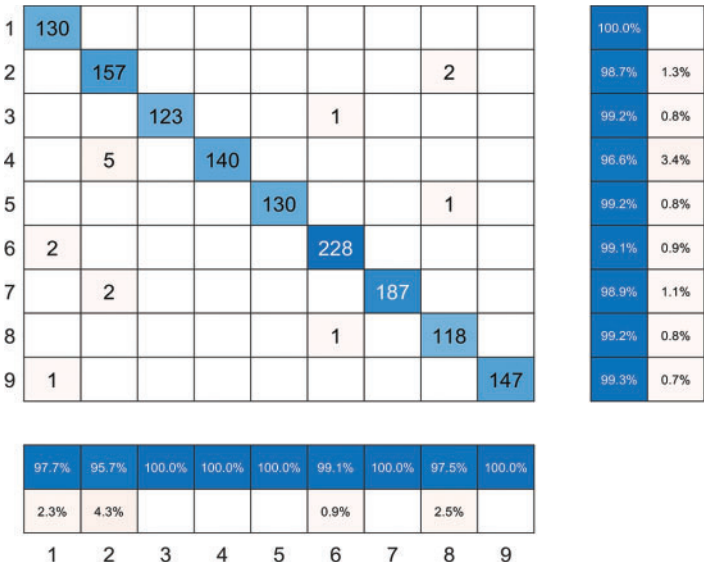


(a) Normalized confusion matrix based on the UCIHAR dataset (1:walking, 2:walking upstairs, 3:walking downstairs, 4:sitting, 5:standing, 6:lying)



(b) Normalized confusion matrix based on the HAR70PLUS dataset (1:walking, 2:walking upstairs, 3:walking downstairs, 4:shuffling, 5:standing, 6:sitting, 7:lying)

Figure 11: (Continued)



(c) Normalized confusion matrix based on the HABRD dataset (1:heartache, 2:falling, 3:bellyache, 4:climbing, 5:standing, 6:sitting, 7:lying, 8:walking, 9:sweeping)

Figure 11: Shows how the PSE-CN-BiGGRU model performed on the three datasets, UCIHAR, HAR70PLUS, and HABRD, that were suggested in this study. (a) Using the UCIHAR dataset as a basis, a normalized confusion matrix; (b) Using the HAR70PLUS dataset as a basis, a normalized confusion matrix; (c) Using the HABRD dataset as a basis, a normalized confusion matrix

4.2 Model Deployment of Application Terminal

In this project, a web-based program on the Python platform was developed to recognize human body posture in real-time through a camera and display the skeleton. The back end recognizes and categorizes human posture using the trained PSE-CNN-BiGRU model. Depending on the action type, it decides whether to send an alarm. System logs display alarm information. Fig. 12 illustrates how standing normally and falling abnormally have the same detection and timely alert effects. The PSE-CNN-BiGRU model can be demonstrated to quickly execute posture detection, continually determine whether the posture is aberrant after identifying the abnormal behavior, and continuously deliver an alarm message. This model provides alert information at the onset of abnormal actions, detects abnormal postures in daily human life, and demonstrates high accuracy in posture identification.

4.3 Discussion

A new PSE-CNN-BiGRU model for posture recognition based on Mediapipe was developed in this study. It is possible to fully extract the essential components of human posture and create a connection between many channels by adding the Mediapipe skeletalized module and the double-layer BiGRU time extraction module to the convolutional layer's input. The PSE-CNN-BiGRU model's accuracy and robustness are further increased by the addition of SENet after the convolutional layer, which improves the model's capacity to choose and learn features. The experiment results show the model's high applicability and practicability in posture recognition. PSE-CNN-BiGRU performs effectively on several datasets, including UCIHAR (training set accuracy of 99.6% and testing set accuracy of 99.56%), HAR70PLUS (training set accuracy of 89.40% and testing set accuracy of 89.41%), and HABRD (training set accuracy of 99.54% and testing set accuracy of 98.90%). It is

demonstrated that the PSE-CNN-BiGRU model performs well in terms of precision, recall, F1-score, and other evaluation indicators. It demonstrates the model's strong classification and generalization capabilities and lays the groundwork for creating this study's abnormal posture recognition system. At the same time, it is interesting to find that in the HAR70PLUS dataset for the elderly, due to realistic factors such as slow behavior and small movement amplitude of the elderly, it is more difficult to extract features that can be used for classification, some movements are easy to be confused, and the accuracy of posture recognition is generally low. For example, shuffling is incorrectly predicted as walking, walking upstairs, or standing. Future experiments, including further analysis of postural characteristics in elderly individuals with pathological or disabled states, are necessary. Additionally, optimizing the abnormal posture recognition system on the website and applying the model to real-life scenarios based on improved posture classification is essential.

(a) Normal posture: standing, no alarm generated

(b) Abnormal posture: fall, alarm generated

(c) Abnormal posture: fall, alarm continuously generated

Figure 12: PSE-CN-BiGRU is applied to the visualization of the performance of the human posture recognition model on the web end: (a) normal posture: standing, no alarm generation; (b) abnormal posture: fall, alarm generation; (c) abnormal posture: fall, alarm continuously

5 Conclusion

An innovative PSE-CNN-BiGRU deep learning network model is presented in this study for recognizing everyday postures. It comprises a Mediapipe skeletal model, a two-layer BiGRU temporal feature extraction model, a CNN model with an ELU activation function, and a SENet. This model has four primary advantages: (1) The Mediapipe skeletonization model skeletonizes the human body, drastically reducing the extraction of irrelevant features such as clothing. (2) To address the issue of most models overlooking time series feature extraction, the two-layer BiGRU model is adopted. (3) The CNN network's ReLU function is upgraded to an ELU activation function, addressing the issue of dead neurons, speeding up convergence, and increasing the model's robustness. (4) SENet is included to enhance the model's ability to learn and extract features by enabling it to autonomously identify significant feature channels. Experiments were conducted on (1) intra-session validation and (2) model deployment at the application terminal. The results demonstrate that the proposed PSE-CNN-BiGRU model improves the accuracy and robustness of the system. Through model deployment at the application terminal, PSE-CNN-BiGRU showed superior stability and applicability in addressing the classification of multiple postures. Additionally, due to its affordability, the model is well-suited for actual application and terminal promotion.

Certainly, PSE-CNN-BiGRU still has limitations requiring further improvement. In the context of an aging population, the model's accuracy for elderly posture recognition on the HAR70PLUS dataset is 89.4%, indicating room for significant improvement. Future research will focus on elderly posture recognition, with enhancements and experiments on the model to better extract and recognize elderly movement characteristics. Additionally, efforts will be made to refine abnormal posture recognition and create a comprehensive dataset specifically for abnormal posture in elderly home settings.

Acknowledgement: The authors would like to thank the editor and reviewers for their valuable comments and suggestions. In addition, the author especially thanks Kevin Parry for visually verifying videos such as 50 Ways to Fall.mp4.

Funding Statement: This paper is a research result funded by the Henan Provincial Science and Technology Research Project (222102210086) and the Starry Sky Creative Space Innovation Space Innovation Incubation Project of Zhengzhou University of Light Industry (2023ZCKJ211).

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: Xianghong Cao, Xinyu Wang; data collection: Xinyu Wang, Houru An; analysis and interpretation of results: Xianghong Cao, Xinyu Wang, Xin Geng; draft manuscript preparation: Xianghong Cao, Xinyu Wang, Donghui Wu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: UCIHAR and HAR70PLUS datasets utilized in this study are publicly available datasets. Readers can access these datasets through the sources cited in the references section of this paper. Readers can also access the datasets by visiting https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones and http://archive.ics.uci.edu/dataset/780/har70. As the data originates from publicly accessible repositories, there are no restrictions on its availability. Therefore, all materials and data used in this study are readily accessible to interested readers.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Varona, J., Jaume-i-Capó, A., Gonzàlez, J., Perales, F. J. (2009). Toward natural interaction through visual recognition of body gestures in real-time. *Interacting with Computers, 21(1/2),* 3–10.
2. Ashraf, F. B., Islam, M. U., Kabir, M. R., Uddin, J. (2023). YoNet: A neural network for yoga pose classification. *SN Computer Science, 4,* 198.
3. Girshick, B. R., Donahue, J., Darrell, T., Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587. Columbus, OH, USA. https://doi.org/10.1109/CVPR.2014.81
4. Sri, S. M., Naik, R. B., Sankar, J. K. (2021). Object detection based on faster R-CNN. *International Journal of Engineering and Advanced Technology (IJEAT), 10(3),* 72–76.
5. Liu, W., Anguelov, D., Erhan, D. (2015). SSD: Single shot multibox detector. *ECCV 2016: Computer Vision–ECCV 2016*, pp. 21–37. Amsterdam, The Netherlands. https://doi.org/10.1007/978-3-319-46448-0_2
6. Mahanta, D., Hazarika, D., Nath, V. (2023). Automated diagnosis of COVID-19 using synthetic chest X-ray images from generative adversarial networks and blend of inception-v3 and Vgg-19 features. *SN Computer Science, 4(5),* 558. https://doi.org/10.1007/s42979-023-02002-w
7. Zhao, X., Xia, Y., Zhang, W., Zheng, C., Zhang, Z. (2023). YOLO-ViT-based method for unmanned aerial vehicle infrared vehicle target detection. *Remote Sensing, 15(15),* 3778. https://doi.org/10.1007/s42979-023-02002-w
8. Garg, S., Saxena, A., Gupta, R. (2022). Yoga pose classification: A CNN and Mediapipe inspired deep learning approach for real-world application. *Journal of Ambient Intelligence and Humanized Computing, 16(3),* 16551–16562. https://doi.org/10.1007/s12652-022-03910-0
9. Abdallah, M. S., Samaan, G. H., Wadie, A. R., Makhmudov, F., Cho, Y. I. Lightweight deep learning techniques with advanced processing for real-time hand gesture recognition. *Sensors, 31(1),* 2. https://doi.org/10.3390/s23010002
10. Wan, H. L., Pan, J. C., Zhen, R., Shi, Z. Q. (2022). CNN-GRU based ship trajectory prediction. *Journal of Guangzhou Institute of Navigation, 30(2),* 12–18.
11. Niu, D. H., Yu, M., Sun, L. J., Gao, T., Wang, K. K. (2022). Short-term multi-energy load forecasting for integrated energy systems based on CNN-BiGRU optimized by attention mechanism. *Applied Energy, 313,* 0306–2619. https://doi.org/10.1016/j.apenergy.2022.118801
12. Samaan, G. H., Wadie, A. R., Attia, A. K., Asaad, A. M., Kamel, A. E. et al. (2022). A novel SE-CNN attention architecture for sEMG-based hand gesture recognition. *Computer Modeling in Engineering & Sciences, 134(1),* 157–177. https://doi.org/10.32604/cmes.2022.020035
13. Samaan, G. H., Wadie, A. R., Attia, A. K., Asaad, A. M., Kamel, A. E. et al. (2022). Mediapipe's landmarks with RNN for dynamic sign language recognition. *Electronics, 11(19),* 3228.
14. Rijayanti, R., Hwang, M., Jin, K. (2023). Detection of anomalous behavior of manufacturing workers using deep learning-based recognition of human-object interaction. *Applied Sciences, 13(15),* 8584. https://doi.org/10.3390/app13158584
15. Wang, S. B., Wu, J. (2023). Patch-transformer network: A wearable-sensor-based fall detection method. *Sensors, 23(14),* 6360. https://doi.org/10.3390/s23146360
16. Inturi, A. R., Manikandan, V. M., Garrapally, V. (2023). A novel vision-based fall detection scheme using keypoints of human skeleton with long short-term memory network. *Arabian Journal for Science & Engineering, 48(2),* 1143–1155.
17. Alves, D. S. D., Lima, F. N. B. C., Sepúlveda, L. M. A. D., Galvão, C. B. H. M., Guilherme, D. A. B. (2024). Electromyography and dynamometry in the prediction of risk of falls in the elderly using machine learning tools. *Biomedical Signal Processing and Control, 88,* 105635. https://doi.org/10.1016/j.bspc.2023.105635
18. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E. et al. (2019). MediaPipe: A framework for building perception pipelines. https://doi.org/10.1109/CONECCT57959.2023.10234829

19. Vivienne, B., Owen, J. L. (2012). *A green vitruvius: Principles and practice of sustainable architectural design*, pp. 6–25. Taylor and Francis; CRC Press: New York, USA.
20. Schuster, M., Paliwal, KK. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45(11),* 2673–2681. https://doi.org/10.1109/78.650093
21. Wang, X., Cao, W. (2023). Bit-plane and correlation spatial attention modules for plant disease classification. *IEEE Access, 11,* 93852–93863.
22. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E. H. (2019). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8),* 2011–2023.
23. Clevert, D., Unterthiner, T., Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). https://doi.org/10.48550/arXiv.1511.07289
24. Lury, A. D. (1972). Statistical methods for research workers. *Journal of the Royal Statistical Society: Series D (The Statistician), 21(3),* 229. https://doi.org/10.2307/2986695
25. Astrid, U., Aleksej, L., Øverengen, S. T., Pernille, T., Beatrix, V. et al. (2023). Validation of an activity type recognition model classifying daily physical behavior in older adults: The HAR70 + Model. *Sensors, 23(5),* 2368. https://doi.org/10.3390/s23052368
26. Garcia-Gonzalez, D., Rivero, D., Fernandez-Blanco, E., Luaces, M. R. (2020). A public domain dataset for real-life human activity recognition using smartphone sensors. *Sensors, 20(8),* 2200. https://doi.org/10.3390/s20082200
27. Du, P. F., Wang, Z. H., Li, X. W., Zhu, Y. W. (2021). A quick way to build anomalous behavior detection datasets. *Computer Technology and Development, 31(9),* 155–160.
28. Yi, D., Ahn, J., Ji, S. (2020). An effective optimization method for machine learning based on ADAM. *Applied Sciences, 10(3),* 1073. https://doi.org/10.3390/app10031073
29. Hinton, E. S. K. A. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60(6),* 84–90.
30. Toshev, A., Szegedy, C. (2014). DeepPose: Human pose estimation via deep neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660. Columbus, OH, USA. https://doi.org/10.1109/CVPR.2014.214
31. Yoonkyu, K., Heeyong, K. (2021). Fall detection method based on pose estimation using GRU. *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing: 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2021-Winter)*, pp. 28–30. Toyama, Japan.
32. Xue, L., Zheng, H. X., Wu, H. C. (2023). Research on academic text classification based on CNN-BiGRU. *Journal of Zhengzhou Institute of Aeronautical Industry Management, 41(3),* 61–68.
33. Yadav, H., Shah, P., Gandhi, N., Vyas, T., Nair, A. et al. (2023). CNN and bidirectional GRU-based heartbeat sound classification architecture for elderly people. *Mathematics, 11(6),* 2227–7390.