A Fall Detection Method Based on Channel Attention and Transformer for Wearable Sensor Data

Chen Zhao[1]

[1]Department of Informatics, University of California, Irvine, CA, United States

*Corresponding author: zerichenn@gmail.com

Abstract

Fall detection has become an important research topic in healthcare monitoring systems due to the increasing aging population. In this paper, a fall detection method based on channel attention and Transformer is proposed for multivariate sensor data. The proposed model employs a one-dimensional convolutional neural network (1D CNN) to extract local temporal features, followed by a Squeeze-and-Excitation (SE) module to enhance channel-wise feature representation. A Transformer encoder is then introduced to capture long-range temporal dependencies and model the dynamic characteristics of fall events.

Experimental results on a public dataset demonstrate that the proposed method outperforms several baseline models, including CNN, CNN+LSTM, and CNN+Transformer, in terms of accuracy, precision, recall, and F1-score. The improvements indicate that the integration of channel attention and global temporal modeling effectively enhances fall detection performance.

Furthermore, the proposed model maintains a relatively efficient structure, making it suitable for relatively practical applications in wearable devices and smart healthcare systems. The results confirm the effectiveness and practicality of the proposed approach.

Index Terms— Fall detection, time-series analysis, wearable sensing, channel attention, Transformer, deep learning

1 Introduction

Falls are one of the leading causes of injury and mortality among the elderly population, posing significant challenges to public health systems worldwide[9]. According to recent reports, a large proportion of elderly individuals experience at least one fall annually, often resulting in severe consequences such as fractures, disability, or even death. Therefore, developing reliable and relatively efficient fall detection systems has become increasingly important in healthcare monitoring and assisted living environments[6].

Existing fall detection approaches can generally be divided into vision-based and sensor-based methods[5]. Vision-based approaches utilize cameras and computer vision techniques to recognize human activities[8, 4, 1]. Although these methods can achieve high accuracy, they often suffer from privacy concerns, high computational cost, and sensitivity to environmental factors such as lighting and occlusion. In contrast, wearable sensor-based methods rely on accelerometers, gyroscopes, and other embedded sensors, offering advantages such as low cost, portability, and better privacy protection. As a result, sensor-based fall detection has gained widespread attention[3].

Traditional sensor-based methods mainly rely on threshold-based techniques or classical machine learning algorithms[2]. These methods are computationally practical but lack the ability to model complex temporal patterns in fall events[2]. With the rapid development of deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been applied to fall detection tasks[7]. CNNs are effective in extracting local temporal features, while RNN-based models such as LSTM can capture sequential dependencies. However, RNN models suffer from limitations in modeling long-range dependencies and often require high computational cost.

Recently, the Transformer architecture has demonstrated strong capability in sequence modeling tasks by leveraging self-attention mechanisms to capture global dependencies[10]. Compared with RNN-based approaches, Transformer allows parallel computation and better captures long-term temporal relationships. However, Transformer-based models are sensitive to input feature quality and may not fully utilize the importance of different sensor channels.

To address these challenges, this paper proposes a fall detection method that integrates channel attention and Transformer-based temporal modeling. Specifically, a 1D CNN is first employed to extract local temporal features from raw sensor data. Then, a Squeeze-and-Excitation (SE) module is introduced to enhance channel-wise feature representation by adaptively weighting different sensor channels. Finally, a Transformer encoder is applied to capture global temporal dependencies and model the dynamic process of falling.

The main contributions of this work are summarized as follows: The proposed method follows a progressive feature enhancement strategy that integrates channel attention and temporal modeling.

- A unified fall detection framework combining convolutional feature extraction, channel attention, and Transformer-based temporal modeling;
- An SE attention mechanism to enhance discriminative channel features in multivariate sensor data;

- A Transformer-based temporal modeling strategy to capture long-range dependencies in fall events;
- Comprehensive experimental validation demonstrating the superiority of the proposed method over baseline models.

In recent years, fall detection systems have been widely applied in smart healthcare, elderly monitoring, and assisted living environments. With the advancement of wearable technologies and Internet of Things (IoT), real-time monitoring of human activities has become increasingly feasible. However, designing an accurate and computationally low-complexity model remains a challenging task due to the complexity of human motion patterns and the variability of sensor data.

Moreover, fall events are relatively rare compared with daily activities, which leads to data imbalance issues. This further increases the difficulty of developing robust detection models. Therefore, it is necessary to design a method that can effectively extract discriminative features while maintaining generalization capability across different scenarios.

2 Methodology

2.1 Problem Definition

Fall detection is formulated as a binary classification problem based on multivariate time-series signals collected from wearable sensors. Given an input sequence:

$$X = \{x_1, x_2, ..., x_L\}, \quad x_t \in \mathbb{R}^C \tag{1}$$

where C represents the number of sensor channels and L denotes the sequence length, the objective is to learn a mapping function:

$$f : \mathbb{R}^{C \times L} \rightarrow \{0, 1\} \tag{2}$$

which classifies each input sequence into fall or non-fall categories.

In practical scenarios, fall events are characterized by sudden changes in motion patterns, often accompanied by complex temporal dynamics across multiple sensor channels. Therefore, an effective fall detection model should be capable of capturing both local variations and long-term temporal dependencies, while also emphasizing the most informative sensor channels.

2.2 Data Preprocessing

Before training, raw sensor data are normalized to reduce the influence of scale differences across channels. A sliding window technique is applied to segment the continuous data into fixed-length sequences. Each segment is labeled based on the corresponding activity type.

To improve generalization, overlapping windows are used, which increases the number of training samples and helps the model capture transitional patterns between activities.

2.3 Overall Architecture

The proposed model adopts a hierarchical feature learning framework that integrates convolutional feature extraction, channel attention enhancement, and Transformer-based temporal modeling. The overall processing pipeline can be summarized as:

$$X \rightarrow F_{cnn} \rightarrow F_{se} \rightarrow F_{trans} \rightarrow \hat{y} \tag{3}$$

Specifically, the raw sensor signal is first processed by a 1D CNN to extract local temporal features, which capture short-term motion patterns such as abrupt acceleration changes. These features are then refined by the SE module, which adaptively reweights different channels based on their importance. Finally, the Transformer encoder is applied to model global temporal dependencies, enabling the system to capture the dynamic evolution of fall events over time.

This progressive design ensures that feature representation is gradually enhanced, moving from low-level local patterns to high-level global temporal structures.

2.4 1D CNN Feature Extraction Module

To effectively extract discriminative patterns from raw sensor signals, a one-dimensional convolutional neural network is employed as the initial feature extractor. The convolution operation is defined as:

$$F^{(l)} = \sigma(W^{(l)} * F^{(l-1)} + b^{(l)}) \tag{4}$$

where $W^{(l)}$ denotes the convolution kernel, $b^{(l)}$ is the bias term, and $\sigma(\cdot)$ represents the ReLU activation function.

In order to improve model stability and generalization capability, batch normalization is applied after each convolution layer, followed by max pooling to reduce feature dimensionality:

$$F^{(l)} = \text{MaxPool}(\text{BN}(F^{(l)})) \tag{5}$$

Through stacked convolutional layers, the CNN is able to capture local temporal variations such as sudden peaks or abrupt changes in sensor signals, which are key indicators of fall events. These local features serve as the foundation for subsequent attention-based enhancement and temporal modeling.

2.5 SE Channel Attention Module

To further improve feature representation, a Squeeze-and-Excitation (SE) module is introduced to model channel-wise dependencies and enhance important sensor signals.

First, a global average pooling operation is applied along the temporal dimension to obtain channel-wise statistical information:

$$z_c = \frac{1}{L} \sum_{t=1}^{L} F_c(t) \tag{6}$$

This operation compresses the temporal information into a compact descriptor vector, allowing the model to summarize the overall contribution of each channel.

Next, the descriptor is passed through two fully connected layers to learn non-linear channel relationships:

$$s = \sigma(W_2\delta(W_1 z)) \tag{7}$$

where $\delta(\cdot)$ denotes the ReLU function and $\sigma(\cdot)$ represents the sigmoid activation.

Finally, the learned channel weights are used to recalibrate the original features:

$$F_{se,c} = s_c \cdot F_c \tag{8}$$

This mechanism enables the model to assign higher importance to informative channels (e.g., acceleration-related signals) while suppressing irrelevant or noisy features, thereby improving the overall feature quality before temporal modeling.

2.6 Transformer-Based Temporal Modeling

To capture long-range temporal dependencies and model the dynamic evolution of fall events, a Transformer encoder is employed after the SE module.

The refined feature sequence is first projected into an embedding space:

$$Z = F_{se}W_e \tag{9}$$

To preserve temporal order information, positional encoding is incorporated:

$$Z' = Z + PE \tag{10}$$

The core of the Transformer is the self-attention mechanism, defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

which allows the model to learn relationships between all time steps simultaneously.

Multi-head attention further enhances representation capacity:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, ..., head_h)W^O \tag{12}$$

Following the attention layer, a feed-forward network is applied:

$$FFN(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{13}$$

Residual connections and layer normalization are used to stabilize training and improve convergence:

$$\tilde{x} = \text{LayerNorm}(x + \text{MultiHead}(x)) \tag{14}$$

Compared with traditional sequential models such as LSTM, the Transformer enables parallel computation and more effectively captures long-range temporal dependencies. This is particularly important for fall detection, where the temporal context before and after the fall plays a critical role in accurate classification.

2.7 Classification Layer

The output features from the Transformer encoder are aggregated and fed into a fully connected layer for classification:

$$\hat{y} = \text{Softmax}(Wx + b) \tag{15}$$

The predicted output $\hat{y}$ represents the probability distribution over fall and non-fall classes.

2.8 Loss Function

The model is trained using the cross-entropy loss function:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{16}$$

This loss function measures the discrepancy between predicted probabilities and ground truth labels, guiding the model to learn more discriminative representations.

The model is trained using the Adam optimizer with an initial learning rate of 0.001. The learning rate is gradually reduced during training using a step decay strategy. Dropout is applied to prevent overfitting, and early stopping is employed based on validation loss.

All experiments are repeated three times, and the average results are reported to ensure robustness and reliability.

2.9 Model Analysis

The proposed framework integrates local feature extraction, channel-wise attention, and global temporal modeling in a unified manner. By enhancing feature quality before applying temporal modeling, the model is able to achieve more accurate predictions.

Furthermore, the introduction of the SE module brings only a small computational overhead, while the Transformer is configured with a limited number of layers to maintain efficiency. As a result, the model achieves a favorable balance between performance and computational cost, making it suitable for relatively efficient fall detection applications.

2.10 Computational Complexity Analysis

Compared with traditional deep learning models, the proposed framework maintains a relatively low computational complexity. The SE module introduces only a small number of additional parameters through two fully connected layers, while the Transformer encoder is designed with a limited number of layers and attention heads.

Assuming the input sequence length is L and the feature dimension is d, the computational complexity of the self-attention mechanism is $\mathcal{O}(L^2 d)$. In this work, the sequence length is controlled by segmentation, which ensures that the overall computational cost remains manageable.

Therefore, the proposed model achieves a balance between detection performance and computational efficiency, making it suitable for relatively low-complexity deployment in wearable systems.

3 Experiments

3.1 Dataset Description

To evaluate the effectiveness of the proposed method, experiments are conducted on a public fall detection dataset. In this study, we adopt the SisFall dataset, which contains a variety of simulated fall and daily activity (ADL) data collected from wearable sensors.

The dataset includes multiple types of fall events (e.g., forward fall, backward fall) and non-fall activities (e.g., walking, sitting, standing), providing a comprehensive benchmark for evaluation.

All sensor signals are segmented into fixed-length sequences with a sliding window approach. Each segment is labeled as either fall or non-fall. The SisFall dataset contains data collected from multiple subjects using wearable sensors. The signals include acceleration and angular velocity measurements captured from different body positions.

The dataset consists of various types of falls and daily activities, providing a realistic benchmark for evaluating fall detection methods.

To ensure fair evaluation, the dataset is split into training, validation, and testing subsets, with no overlap between subjects.

3.2 Experimental Settings

The dataset is divided into training, validation, and test sets with a ratio of 70%, 10%, and 20%, respectively.

The proposed model is implemented using PyTorch and trained on a GPU platform. The main hyperparameters are summarized as follows:

- Batch size: 64
- Learning rate: 0.001
- Optimizer: Adam
- Number of epochs: 100
- Transformer layers: 2
- Number of attention heads: 4
- Dropout rate: 0.3

Early stopping is applied to prevent overfitting.

3.3 Evaluation Metrics

To comprehensively evaluate the performance, the following metrics are used:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{20}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

3.4 Comparison with Baseline Methods

To validate the effectiveness of the proposed model, we compare it with several baseline methods:

- CNN
- CNN + LSTM
- CNN + Transformer
- CNN + SE

The comparison results are shown in Table 1.

Table 1: Performance comparison with baseline methods

Method	Accuracy	Precision	Recall	F1-score
CNN	91.38%	90.72%	89.95%	90.33%
CNN + LSTM	93.12%	92.45%	91.88%	92.16%
CNN + Transformer	94.08%	93.74%	92.96%	93.35%
CNN + SE	94.76%	94.21%	93.67%	93.94%
Proposed Method	**95.42%**	**95.08%**	**94.36%**	**94.72%**

It can be observed that the proposed method achieves the best performance across all metrics, demonstrating the effectiveness of combining channel attention and Transformer-based temporal modeling.

From Table 1, it can be observed that the proposed method achieves consistent improvements across all evaluation metrics. Compared with CNN-based models, the proposed approach demonstrates better ability to capture complex temporal patterns.

In particular, the improvement in F1-score indicates that the model achieves a better balance between precision and recall, which is crucial for fall detection tasks. A high recall ensures that most fall events are correctly detected, while a high precision reduces false alarms.

Furthermore, the relatively moderate improvement suggests that the performance gain comes from effective feature enhancement rather than overfitting. This indicates that the proposed method is robust and generalizable.

Although the proposed method achieves high performance, some misclassifications still occur. Most errors are observed in activities that involve sudden movements, such as sitting down quickly or jumping, which may resemble fall events in sensor patterns.

These results suggest that distinguishing between similar motion patterns remains a challenging problem. Future work

may focus on incorporating additional contextual information or multimodal data to improve discrimination capability.

In addition to accuracy, computational efficiency is an important factor for real-world deployment. The proposed model maintains a relatively low complexity by using a lightweight SE module and a shallow Transformer encoder.

Compared with CNN+LSTM models, the Transformer-based architecture allows parallel computation, which improves inference speed. Therefore, the proposed method is suitable for practical applications in wearable devices.

3.5 Ablation Study

To further analyze the contribution of each module, an ablation study is conducted. The results are presented in Table 2.

Table 2: Ablation study of different components

Model	Accuracy	Precision	Recall	F1-score
CNN	91.38%	90.72%	89.95%	90.33%
CNN + SE	94.76%	94.21%	93.67%	93.94%
CNN + Transformer	94.08%	93.74%	92.96%	93.35%
CNN + SE + Transformer	**95.42%**	**95.08%**	**94.36%**	**94.72%**

From the results, it is clear that:

- The SE module leads to consistent gains in channel feature representation;
- The Transformer enhances temporal modeling capability;
- The combination of both modules yields the best performance.

The experimental results demonstrate that the proposed model effectively integrates local feature extraction, channel attention, and global temporal modeling. Compared with traditional methods, it shows superior performance in detecting fall events.

In addition, the model maintains a relatively low-complexity structure, making it suitable for relatively efficient applications in wearable devices.

From Table 1, it can be observed that the proposed method achieves consistent improvements across all evaluation metrics. Compared with CNN-based models, the proposed approach demonstrates better ability to capture complex temporal patterns.

In particular, the improvement in F1-score indicates that the model achieves a better balance between precision and recall, which is crucial for fall detection tasks. A high recall ensures that most fall events are correctly detected, while a high precision reduces false alarms.

Furthermore, the relatively moderate improvement suggests that the performance gain comes from effective feature enhancement rather than overfitting. This indicates that the proposed method is robust and generalizable.

4 Discussion

The experimental results demonstrate that the proposed model achieves superior performance compared with baseline methods, which can be attributed to the effective integration of convolutional feature extraction, channel attention enhancement, and Transformer-based temporal modeling.

First, the introduction of the SE module leads to consistent gains in the representation of multichannel sensor data. In fall detection tasks, different sensor channels contribute unequally to the identification of fall events. For example, sudden changes in acceleration are more informative than relatively stable signals. By applying channel-wise recalibration, the SE module enables the model to automatically emphasize informative channels while suppressing less relevant ones. This mechanism reduces the impact of noise and redundancy in the input signals, leading to improved feature quality for subsequent processing.

Second, the Transformer module plays a crucial role in capturing long-range temporal dependencies. Unlike traditional methods such as LSTM, which process sequences in a step-by-step manner, the Transformer leverages self-attention to model global relationships across all time steps simultaneously. This is particularly important for fall detection, where a fall event is characterized not only by a sudden change but also by its temporal context, including pre-fall and post-fall patterns. The ability of the Transformer to model such temporal dynamics contributes significantly to the overall performance improvement.

Moreover, the combination of SE and Transformer modules forms a complementary framework. The SE module focuses on enhancing feature quality at the channel level, while the Transformer captures temporal dependencies at the sequence level. This hierarchical design ensures that the input to the Transformer is already refined, allowing it to more effectively learn meaningful temporal relationships. As demonstrated in the ablation study, the integration of both modules yields better results than using either module individually.

In addition, the proposed model maintains a relatively efficient structure compared to more complex deep learning architectures. By using a single SE module and a shallow Transformer encoder, the model achieves a balance between performance and computational efficiency. This makes it suitable for relatively resource-efficient applications in wearable devices, where computational resources are often limited. The experimental results indicate that improving feature quality before temporal modeling is crucial for performance enhancement. By introducing channel attention prior to the Transformer, the model is able to focus on more informative signals, which leads to more stable and accurate temporal modeling. This demonstrates the effectiveness of the proposed hierarchical feature enhancement strategy. Despite the promising results, several limitations should be noted. First, the experiments are conducted on a single public dataset, which may not fully represent the variability of real-world scenarios. Differences in sensor placement, user behavior, and environmental conditions may affect model performance. Second, the model relies

on supervised learning, which requires labeled data that can be costly to obtain. Finally, although the model is relatively resource-efficient, further optimization may be necessary for deployment on low-power embedded systems.

Future work will focus on improving the generalization ability of the model across different datasets and real-world environments. In addition, exploring relatively efficient model compression techniques and semi-supervised learning approaches may further enhance the practicality of the proposed method.

5 Conclusion

In this paper, we proposed a fall detection method that integrates channel attention and Transformer-based temporal modeling. The model combines a 1D CNN for local feature extraction, an SE module for channel-wise feature enhancement, and a Transformer encoder for capturing long-range temporal dependencies in multivariate sensor data.

Experimental results on a public dataset demonstrate that the proposed method outperforms several baseline models, including CNN, CNN+LSTM, and CNN+Transformer. The improvements can be attributed to the effective collaboration between channel attention and global temporal modeling, which enables the model to better capture both discriminative features and temporal dynamics of fall events.

In addition, the proposed architecture maintains a relatively lightweight design, making it suitable for relatively efficient fall detection in wearable or mobile devices. This provides practical value for applications such as elderly monitoring and smart healthcare systems.

Future work will focus on improving the robustness and generalization ability of the model in real-world scenarios. Specifically, cross-dataset validation, model compression techniques, and the integration of multimodal data will be explored to further enhance performance and applicability.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest.

Informed Consent

Not applicable.

Consent to Publish

Not applicable.

References

[1] Boris Bačić, Chengwei Feng, and Weihua Li. Jy61 imu sensor external validity: A framework for advanced pedometer algorithm personalisation. *ISBS Proceedings Archive*, 42(1):60, 2024.

[2] Alan K Bourke and Gerald M Lyons. A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor. *Medical engineering & physics*, 30(1):84–90, 2008.

[3] Yueng Santiago Delahoz and Miguel Angel Labrador. Survey on fall detection and fall prevention using wearable and external sensors. *Sensors*, 14(10):19806–19842, 2014.

[4] Chengwei Feng, Boris Bačić, Weihua Li, and Hongqi Xu. Sks-transformer: multi-scale and direction-aware attention for inertial sensor-based activity recognition. *Frontiers in Sports and Active Living*, 8:1754717, 2026.

[5] Raul Igual, Carlos Medrano, and Inmaculada Plaza. Challenges, issues and trends in fall detection systems. *Biomedical engineering online*, 12(1):66, 2013.

[6] Nishat Tasnim Newaz and Eisuke Hanada. The methods of fall detection: A literature review. *Sensors*, 23(11):5212, 2023.

[7] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[8] Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Fall detection from human shape and motion history using video surveillance. In *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, volume 2, pages 875–880. IEEE, 2007.

[9] Xueyi Wang, Joshua Ellul, and George Azzopardi. Elderly fall detection systems: A literature survey. *Frontiers in Robotics and AI*, 7:71, 2020.

[10] Zongfei Zhang and Haoze Ni. A lightweight skeleton-based fall detection framework using multi-dimensional attention mechanisms. *INNO-PRESS: Journal of Emerging Applied AI*, 1(9), 2025.