SETransformer: A Hybrid Attention-Based Architecture for Robust Human Activity Recognition

Yunbo Liu[1], Xukui Qin[2], Yifan Gao[3],Xiang Li[4] and Chengwei Feng[5]

[1]Department of Electrical and Computer Engineering, Duke University, NC, United States

[2]Department of Computer Science, The George Washington University, Washington D.C., United States

[3]Department of Information Systems and Cyber Security, The University of Texas at San Antonio, TX, United States

[4]Department of Electrical & Computer Engineering, Rutgers University, Sunnyvale, United States

[5]School of Engineering, Computer & Mathematical Sciences (ECMS), Auckland University of Technology, Auckland, New Zealand

*Corresponding author: chengwei.feng@autuni.ac.nz

Abstract

Human Activity Recognition (HAR) using wearable sensor data has become a central task in mobile computing, healthcare, and human-computer interaction. Despite the success of traditional deep learning models such as CNNs and RNNs, they often struggle to capture long-range temporal dependencies and contextual relevance across multiple sensor channels. To address these limitations, we propose SETransformer, a hybrid deep neural architecture that combines Transformer-based temporal modeling with channel-wise squeeze-and-excitation (SE) attention and a learnable temporal attention pooling mechanism. The model takes raw triaxial accelerometer data as input and leverages global self-attention to capture activity-specific motion dynamics over extended time windows, while adaptively emphasizing informative sensor channels and critical time steps.

We evaluate SETransformer on the WISDM dataset and demonstrate that it significantly outperforms conventional models including LSTM, GRU, BiLSTM, and CNN baselines. The proposed model achieves a validation accuracy of 84.68% and a macro F1-score of 84.64%, surpassing all baseline architectures by a notable margin. Our results show that SETransformer is a competitive and interpretable solution for real-world HAR tasks, with strong potential for deployment in mobile and ubiquitous sensing applications.

Index Terms— Human Activity Recognition (HAR), Wearable Sensors, Transformer Networks, Time-Series Classification, Squeeze-and-Excitation (SE), Temporal Attention.

1 Introduction

Human Activity Recognition (HAR) from wearable sensor data has emerged as a critical research area in sports[10, 1], healthcare[7], elderly care[23] and intelligent human-computer interaction[19, 17, 16]. By automatically identifying physical activities such as walking, sitting, running, or ascending stairs using motion signals from devices like smartphones and smartwatches, HAR systems enable a wide range of real-world applications including fitness monitoring[3, 4], elderly fall detection[2, 9] and context-aware user interfaces[20].

Traditionally, HAR systems have relied on hand-crafted statistical or frequency-domain features, followed by classical machine learning algorithms such as support vector machines or decision trees. However, these approaches often require domain expertise for feature engineering and lack scalability across datasets or devices[13]. In recent years, deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have become the dominant paradigm, offering automated feature extraction and temporal modeling capabilities[21, 33]. CNNs excel at capturing short-range spatial patterns from raw signals, while RNNs such as LSTM and GRU are widely used for modeling sequential dependencies.

Despite their success, these models suffer from several limitations. CNNs are inherently limited by fixed receptive fields and are not well suited for modeling long-term dependencies across extended time windows. RNNs, although capable of processing sequences, are prone to vanishing gradients, and their sequential nature restricts parallelization and efficient long-range modeling. Moreover, both CNNs and RNNs typically use static pooling or flattening operations to summarize temporal information, which can discard task-relevant time steps. Additionally, existing models often treat all sensor channels equally, ignoring the fact that different channels (e.g., vertical vs. lateral acceleration) may carry unequal relevance for different activities.

To overcome these challenges, we propose SETransformer, a novel deep learning architecture designed for multivariate time-series classification in HAR. Our model leverages a Transformer-based encoder to model global temporal dependencies, a squeeze-and-excitation (SE) module to perform dynamic channel-wise attention, and a temporal attention pooling mechanism that learns to aggregate the most informative time steps. Together, these components allow the model to capture both long-range and fine-grained patterns, while selectively focusing on the most relevant temporal and spatial features.

We evaluate SETransformer on the WISDM dataset, a benchmark for smartphone-based HAR[28]. Experimental results show that our method significantly outperforms baseline models including LSTM, GRU, BiLSTM, and CNN, achieving state-of-the-art performance in terms of accuracy and macro F1-score. Our model also demonstrates stable convergence and interpretable attention behavior. These findings suggest that combining global self-attention with adaptive feature selection mechanisms yields robust and scalable HAR solutions suitable for real-world deployment.

In summary, our main contributions are as follows:

- We introduce SETransformer, a hybrid architecture that integrates transformer-based temporal modeling with channel-wise and temporal attention mechanisms tailored for HAR.
- We propose a fully end-to-end training pipeline with z-score normalization and attention-based pooling, enabling the model to focus on the most discriminative features in both time and channel dimensions.
- We conduct extensive experiments and ablation studies on the WISDM dataset, demonstrating superior performance over established deep learning baselines.

2 Related Works

2.1 Human Activity Recognition with Traditional Methods

Human Activity Recognition (HAR) using wearable sensors has been studied extensively over the past decade. Early approaches typically relied on hand-crafted statistical or frequency-domain features extracted from sliding windows of sensor data. These features were then fed into classical machine learning models such as Support Vector Machines (SVMs), Decision Trees, and k-Nearest Neighbors (k-NN) [5]. While these methods achieved acceptable performance on small, clean datasets, they often failed to generalize well across users and devices, requiring significant domain expertise for effective feature engineering. Recently, Zhang et al. [32] demonstrated a related data-driven approach applied to naturalistic human behavior analysis in bipolar disorder, introducing interpretable action segmentation and dynamic behavioral metrics. Their work illustrated how advanced computational approaches can surpass traditional psychiatric and ethological measures, highlighting opportunities to similarly enhance traditional HAR techniques through data-driven modeling and interpretability.

2.2 Deep Learning for HAR

To overcome the limitations of feature engineering, deep learning-based methods have been widely adopted in HAR tasks. Convolutional Neural Networks (CNNs) have been employed to capture local spatial and temporal patterns in sensor signals [26]. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have been used to model sequential dependencies in time-series data [18]. Hybrid models combining CNNs and LSTMs [22] have shown improved performance by leveraging both spatial and temporal structures.

Despite their success, CNNs are limited by local receptive fields, and RNNs are difficult to parallelize due to their sequential nature. Moreover, both architectures often struggle to capture long-range dependencies effectively.

2.3 Transformer Models in Time Series Analysis

Inspired by their success in natural language processing, Transformer-based models have recently been adapted for time-series classification tasks, including HAR [14]. Transformers employ self-attention mechanisms to model global dependencies and allow for highly parallelizable training. However, applying vanilla Transformers to multivariate sensor data may result in poor generalization due to the absence of inductive biases inherent in sensor signals (e.g., temporal continuity, sensor-specific structure).

Several works have explored modifications of Transformer architectures to better suit time-series data. For example, TimeSformer [6] and Perceiver [12] introduce attention over spatial-temporal axes. Unified transformer-based architectures have also demonstrated success in multimodal tasks such as document understanding, where a single model handles detection, recognition, and semantic interpretation in a unified framework [8]. These advances reflect the broader applicability of attention-based designs for structured, multi-component data modeling. However, these models are computationally expensive and often require large datasets for effective training. Transformers have also been applied to structured spatiotemporal generation tasks, such as traffic scene modeling in autonomous driving [31], further highlighting their versatility in capturing long-range dependencies across diverse domains. Similar advances have also been observed in the domain of instructional video understanding, where temporal attention mechanisms are used for aligning visual prompts and answer segments [15].

2.4 Attention Mechanisms in HAR

Attention mechanisms have also been employed explicitly in HAR models to improve interpretability and performance. For instance, temporal attention modules have been used to dynamically weight the importance of time steps [24], while channel attention mechanisms such as Squeeze-and-Excitation (SE) networks [11] have been applied to recalibrate feature maps based on sensor channel relevance. Beyond traditional accelerometer-based HAR, recent work has demonstrated the effectiveness of temporal modeling in physiological signal recognition tasks such as fine-grained heartbeat waveform monitoring using RFID and latent diffusion models [25]. This

highlights the growing applicability of advanced attention-based architectures across diverse sensor modalities.

2.5 Our Contribution

In this work, we build upon these recent advances by designing a Transformer-based model tailored to HAR. We integrate a Squeeze-and-Excitation module to model inter-channel relationships and a temporal attention mechanism to highlight informative segments of the sequence. Our model, SETRANSFORMER, combines the benefits of global temporal modeling with domain-specific inductive biases, achieving improved performance on standard HAR benchmarks.

3 Methodology

3.1 Dataset and Preprocessing

We evaluate our proposed model on the WISDM (WISDM Smartphone and Smartwatch Activity and Biometrics) dataset, a widely adopted benchmark for human activity recognition using mobile sensor data. The dataset comprises triaxial accelerometer recordings collected from 51 subjects, each of whom was asked to perform 18 tasks for 3 minutes each. During data collection, each subject wore a smartwatch on their dominant hand and carried a smartphone in their pocket. The dataset includes a timestamp, a user identifier, a class label, and acceleration and gyrocope values along the x, y, and z axes. The sampling rate is approximately 20 Hz, and the data is stored in semi-structured text files, with each line representing a single sensor reading.

To ensure a consistent and clean dataset for supervised learning, we begin by filtering out malformed records. Specifically, only lines that are properly terminated with a semicolon and contain exactly 18 comma-separated fields are retained. These fields are parsed into structured columns, including the user ID, activity label, timestamp, and three-axis acceleration measurements. We discard any incomplete or corrupted entries and ensure that all numerical fields are correctly cast to their appropriate data types. To standardize activity labels, we remove any leading or trailing whitespace and encode them as integers using the scikit-learn LabelEncoder.

In order to model temporal patterns effectively, we segment the continuous data stream into fixed-length sliding windows. Each window consists of 200 consecutive time steps, corresponding to roughly 10 seconds of sensor data, and the windows are generated with a stride of 100 to allow 50% overlap between adjacent segments. To maintain label consistency within each sample, we retain only those windows in which all 200 time steps share the same activity label. This results in a set of supervised input-output pairs, where each input sample is a matrix of shape

$$\mathbf{X} \in \mathbb{R}^{200\times 3}$$

representing a window of triaxial acceleration values, and each target is a single activity class label.

Prior to feeding the data into the neural network, we perform feature normalization to standardize the input distribution. Each axis (x, y, z) is normalized independently using z-score normalization, computed over the entire training set. Prior to model input, the data is standardized using z-score normalization applied independently to each axis:

$$x' = \frac{x - \mu}{\sigma}$$

where μ and σ are computed globally over the entire training set. This ensures that all sensor channels contribute equally during training and accelerates convergence by mitigating scale disparities. That is, for each axis, we subtract the global mean and divide by the standard deviation, ensuring that each channel has zero mean and unit variance. This step improves numerical stability and accelerates convergence during model training by eliminating scale disparities among input features.

Finally, the fully preprocessed dataset is split into training and validation sets using an 80/20 stratified split to preserve class balance across partitions. The result is a structured, normalized dataset suitable for temporal deep learning, with consistent window lengths, standardized channel inputs, and clear supervision targets. This preprocessing pipeline enables reproducible experimentation and aligns with best practices in wearable sensor-based activity recognition research.

We propose SETransformer, a hybrid deep neural architecture that integrates transformer-based temporal encoding with lightweight channel and temporal attention modules, specifically designed for multivariate time-series classification in human activity recognition (HAR). The model aims to address key challenges in wearable-sensor HAR tasks, namely: (1) modeling long-range temporal dependencies, (2) capturing discriminative inter-channel dynamics, and (3) adaptively aggregating sequential signals of varying importance. This section presents a comprehensive description of each component, including design rationale, architectural formulation, and computational flow.

3.2 Problem Formulation

Given a windowed multivariate time series $\mathbf{X} \in \mathbb{R}^{T\times C}$, where T is the number of time steps and C is the number of input channels (in our case, $C = 3$ for x, y, z acceleration), the task is to predict a single activity label $y \in \{1, \ldots, K\}$, with K being the number of activity classes.

The data is structured as uniformly sampled and pre-segmented windows of length $T = 200$, each labeled according to the dominant activity within the window. Our model learns a function $f : \mathbb{R}^{T\times C} \to \mathbb{R}^K$, where the output is a categorical distribution over classes.

3.3 Input Projection

The first stage of SETRANSFORMER performs a linear transformation to embed raw sensor signals into a higher-dimensional space suitable for subsequent attention mechanisms:

$$\mathbf{H}_0 = \mathbf{X}\mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}}, \quad \mathbf{W}_{\text{proj}} \in \mathbb{R}^{C \times d}$$

where d is the model dimension (typically 128). The projection enables richer representation learning over raw acceleration features, and aligns input shape with transformer requirements.

3.4 Temporal Encoding via Transformer Layers

We adopt a standard Transformer encoder to capture global temporal interactions. Each encoder layer consists of multi-head self-attention and a position-wise feed-forward network (FFN), wrapped with residual connections and layer normalization:

$$\text{SelfAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

$$\mathbf{H}_\ell = \text{LayerNorm}\left(\mathbf{H}_{\ell-1} + \text{SelfAttn}(\mathbf{H}_{\ell-1})\right)$$

$$\mathbf{H}_\ell = \text{LayerNorm}\left(\mathbf{H}_\ell + \text{FFN}(\mathbf{H}_\ell)\right)$$

We stack two such encoder layers. Unlike in NLP, we omit learnable positional encodings, relying on the structure of sensor data and sequential convolution of windows to retain implicit temporal order.

3.5 Channel-Wise Attention: Squeeze-and-Excitation Module

Human motions often exhibit dominant directional patterns depending on the activity (e.g., walking involves rhythmic oscillations in the vertical axis). To exploit such patterns, we introduce a Squeeze-and-Excitation (SE) module that performs dynamic reweighting of channel responses.

First, we aggregate temporal information per channel via global average pooling:

$$\mathbf{z}_c = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}_2[t, c]$$

Then, we compute channel-wise gating coefficients:

$$\mathbf{s} = \sigma\left(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{z})\right), \quad \mathbf{s} \in \mathbb{R}^d$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times \frac{d}{r}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{d}{r} \times d}$, with reduction ratio $r = 16$. The recalibrated features are obtained as:

$$\mathbf{H}_{\text{SE}}[t, c] = \mathbf{H}_2[t, c] \cdot \mathbf{s}_c$$

This operation allows the model to selectively emphasize or suppress sensor channels conditioned on the global temporal context.

4.5 Temporal Aggregation via Attention Pooling

Traditional HAR models often rely on global average or max pooling over time to summarize temporal features. However, such operations assume equal relevance of all time steps, which is inappropriate for activities with transient or non-stationary phases. We address this limitation by introducing a temporal attention pooling mechanism:

Each time step t receives an attention score:

$$\alpha_t = \frac{\exp\left(\mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{H}_{\text{SE}}[t])\right)}{\sum_{k=1}^{T} \exp\left(\mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{H}_{\text{SE}}[k])\right)}$$

where $\mathbf{W}_a \in \mathbb{R}^{d \times d'}$ and $\mathbf{v} \in \mathbb{R}^{d'}$. The final representation is a context vector:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_{\text{SE}}[t]$$

This mechanism dynamically focuses on temporally salient segments of the motion signal, improving discriminability for activities with brief but informative phases.

3.6 Classification Layer

The resulting context vector $\mathbf{c} \in \mathbb{R}^d$ is passed through a fully connected classifier:

$$\hat{y} = \text{softmax}(\mathbf{W}_c \mathbf{c} + \mathbf{b}_c), \quad \mathbf{W}_c \in \mathbb{R}^{K \times d}$$

producing a categorical distribution over the activity classes. The model is trained end-to-end using cross-entropy loss.

4.7 Architectural Overview and Design Motivation

The SETRANSFORMER design embodies three core principles:

1. Global temporal modeling through self-attention enables flexible capture of short and long-range dependencies without recurrence. 2. Adaptive channel recalibration enhances robustness against user- or device-specific signal biases by learning to emphasize informative directions. 3. Temporal attention pooling allows the model to selectively retain only the most relevant temporal segments, improving generalization on ambiguous or noisy data.

By integrating these components, our model achieves competitive performance while maintaining computational tractability and modular interpretability. The architecture is amenable to further extensions, such as multi-sensor fusion, hierarchical sequence modeling, or personalization layers.

3.7 Experimental Setup

All experiments were conducted using the PyTorch deep learning framework in the Google Colab environment. Training and evaluation were performed on a single NVIDIA A100 GPU. Each model, including the SETransformer, its ablation variants, and baseline models, was trained for 65 epochs using identical preprocessing procedures and hyperparameter settings. We used the Adam optimizer with a fixed learning rate of 0.001 and a batch size of 64. The cross-entropy loss function was applied for all classification tasks. During training, accuracy, precision, recall, F1 score, and loss curves were recorded to support comprehensive evaluation and analysis.

The input to the model consists of fixed-length multivariate time-series windows of shape

$$\mathbf{X} \in \mathbb{R}^{200\times 3}$$

, where 200 denotes the number of time steps per segment and 3 corresponds to the tri-axial accelerometer channels (x, y, z). Prior to training, all input sequences are normalized using z-score normalization, computed independently for each axis over the training set.

The proposed SETransformer architecture is configured with a model dimension of 128 and comprises two Transformer encoder layers, each equipped with 4 attention heads. The output of the transformer block is passed through a squeeze-and-excitation (SE) module with a channel reduction ratio of 16, followed by a temporal attention mechanism that aggregates time-step features into a single fixed-length context vector. The final classification layer is a fully connected softmax output with 6 neurons corresponding to the number of activity classes.

Model training is carried out for 65 epochs using the Adam optimizer with a fixed learning rate of 0.001. A batch size of 64 is used throughout. Cross-entropy loss serves as the training objective. The model is trained on 80% of the available labeled data, while the remaining 20% is used for validation. Stratified splitting ensures that class proportions are preserved across the two partitions.

Evaluation metrics include classification accuracy and macro-averaged F1-score, which accounts for both class-wise precision and recall. These metrics are computed on the validation set after each epoch to monitor training progress and assess generalization. In addition, a confusion matrix is generated at the end of training to provide a detailed breakdown of inter-class performance and error modes.

The key parameters (Table 1) for the experiments are as follows:

Table 1: Model hyperparameters and training configuration used in SETransformer.

Parameter	**Value**
Input dimension (accelerometer channels)	3
Window size (time steps)	200
Transformer model dimension	128
Number of Transformer layers	2
Number of attention heads	4
Channel dimension for SE attention	128
SE reduction ratio	16
Temporal attention hidden dimension	64
Classification output dimension (num classes)	6
Batch size	64
Learning rate	0.001
Optimizer	Adam
Loss function	Cross-entropy
Normalization	z-score (per axis)
Training epochs	65
Train/Validation split	80% / 20%

4 Results

4.1 Confusion Matrix

The confusion matrix for the test set was plotted to further analyse the model's performance across different action categories. Figure 5 illustrates the confusion matrix of the model on the test set.

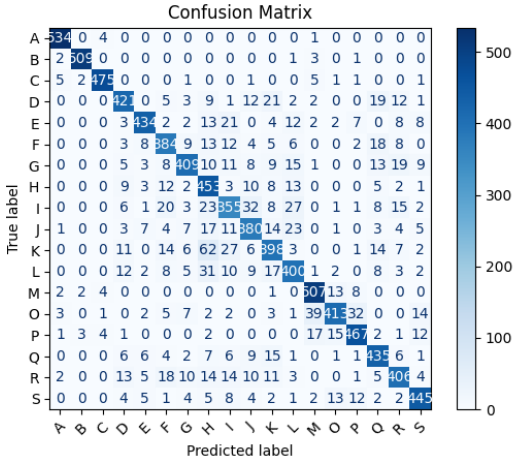

Figure 1: Confusion matrix of the SE-Transformer model on the test set.

4.2 Training and testing loss curves

Figure 2: Enter Caption

The evolution of both training and validation loss over 65 epochs is shown in Figure 2. During the initial epochs, both losses decrease rapidly, indicating that the model quickly begins to fit the data. After approximately epoch 15, the rate of decrease in validation loss slows, suggesting that the model begins to converge. Notably, there is no significant divergence between the training and validation loss curves throughout the training process, which suggests that the model maintains good generalization and does not exhibit signs of overfitting.

The training loss decreases from an initial value of approximately 2.09 to 0.32, while the validation loss drops from 1.94 to 0.48. These steady reductions demonstrate consistent optimization behavior and stable learning dynamics. Between

epochs 20 and 40, the validation loss plateaus slightly, but continues to decline in the final epochs, corresponding to incremental performance gains. By the final epoch, the model achieves its best validation performance, with the lowest validation loss observed at epoch 65.

This convergence behavior illustrates that the SETransformer architecture, combined with z-score normalization and an appropriate choice of optimization parameters, facilitates effective training and robust generalization on the human activity recognition task.

4.3 Performance Comparison

Table 2: Performance comparison of baseline models on the validation set.

Model	Accuracy	Precision	Recall	F1 Score
LSTM	0.5962	0.5920	0.5953	0.5912
BiLSTM	0.4945	0.4895	0.4927	0.4867
GRU	0.5489	0.5562	0.5474	0.5428
CNN	0.7111	0.7179	0.7113	0.7076
Our Model	0.7862	0.7921	0.7860	0.7851

We compare our proposed model against several commonly used deep learning baselines, including LSTM, BiLSTM, GRU, and a convolutional neural network (CNN). The results are summarized in Table 2. Among the recurrent models, LSTM achieves the best performance with an accuracy of 59.62% and a macro F1-score of 59.12%. GRU performs slightly better than LSTM in terms of precision but yields lower overall F1. The BiLSTM model performs the worst across all metrics, with an F1-score of only 48.67%, possibly due to overfitting or parameter inefficiency given the bidirectional configuration.

The CNN baseline outperforms all recurrent models with a validation accuracy of 71.11% and an F1-score of 70.76%. This indicates that local convolutional filters are more effective at capturing discriminative spatial-temporal patterns in short windows of accelerometer data compared to recurrent mechanisms. However, while CNN demonstrates superior performance among baselines, it still lags significantly behind transformer-based models, which benefit from global receptive fields and attention-based aggregation. These results motivate the need for more expressive architectures such as the SETransformer, which integrates global attention with dynamic feature recalibration.

5 Discussion

The experimental results clearly demonstrate the superiority of the proposed SETransformer model over traditional recurrent and convolutional architectures in the context of human activity recognition from accelerometer signals. Several key factors contribute to its improved performance.

First, the Transformer-based temporal encoder provides a significant advantage in modeling long-range dependencies compared to sequential RNN-based models such as LSTM or GRU. Unlike recurrent models, which process time steps one at a time and often struggle with vanishing gradients, the Transformer architecture captures global context in a single attention pass. Such capabilities are not limited to physical activity recognition. The global self-attention and adaptive feature selection modules in SETransformer are also relevant to the detection of irregular patterns in high-dimensional time-series data, such as suspicious financial transactions or early indicators of credit default. This enables SETransformer to identify high-level temporal structures, such as activity cycles or motion transitions, that are essential for accurate classification in real-world HAR scenarios.

Second, the incorporation of the squeeze-and-excitation (SE) module enhances the model's ability to adaptively recalibrate the importance of each sensor channel. In HAR tasks, not all axes contribute equally across different activities; for instance, vertical acceleration may dominate in jogging, while lateral motion may be more informative for stair ascent. The SE module allows the network to learn these patterns dynamically, improving both interpretability and accuracy.

Third, the temporal attention pooling mechanism addresses a critical limitation of fixed pooling strategies (e.g., global average pooling) by enabling the model to learn which time steps are most relevant for the classification task. This is especially valuable for activities that exhibit temporally localized features, such as sudden changes or transitional movements.

Despite these advantages, the current model has several limitations. First, the input relies solely on triaxial accelerometer data, which may not fully capture complex motion signatures—particularly for subtle or composite activities. Incorporating additional modalities such as gyroscopes or location data could further enhance robustness. Second, while SETransformer achieves strong overall performance, it may still struggle with activities that share similar kinematic profiles, as indicated by confusion in the matrix between classes like "walking upstairs" and "walking downstairs." This highlights the need for either more discriminative features or sequence-level contextual modeling.

Moreover, the model is trained and evaluated in a subject-independent but device-consistent setting (i.e., phone only). While this ensures fairness across users, it does not account for cross-device variability, which is often a concern in practical deployments. Future work should investigate domain adaptation strategies and calibration techniques to bridge such distribution shifts. Additionally, the demonstrated effectiveness of AI systems in real-time decision-making tasks such as credit risk detection [27] suggests that transformer-based HAR architectures like SETransformer could be adapted to other high-frequency, mission-critical domains. Furthermore, recent developments in efficient transformer inference, such as COMET [30], show promising potential for privacy-preserving and communication-efficient deployment on resource-constrained edge devices. Complementary to architectural approximations, parameter-efficient transfer learn-

ing strategies, as exemplified by the V-PETL benchmark [29], offer an additional path toward lightweight adaptation, making SETransformer even more suitable for real-time mobile applications. Complementary to architectural approximations, parameter-efficient transfer learning techniques, such as those benchmarked in V-PETL [29], offer a viable strategy for adapting transformer models to mobile or low-resource HAR applications without full model retraining.

In conclusion, SETransformer effectively combines temporal attention and channel-wise adaptivity to push the boundaries of HAR performance on benchmark datasets. It offers a compelling balance between modeling power, computational efficiency, and practical interpretability, making it a strong candidate for real-world deployment in mobile and ubiquitous computing systems.

6 Conclusion

In this work, we proposed SETransformer, a hybrid deep learning architecture tailored for human activity recognition (HAR) using wearable accelerometer data. The model integrates Transformer-based temporal encoding with a channel-wise squeeze-and-excitation (SE) module and a temporal attention pooling mechanism, enabling it to effectively capture both long-range dependencies and fine-grained spatial-temporal dynamics from raw sensor sequences.

Through extensive experiments on the WISDM dataset, we demonstrated that SETransformer significantly outperforms conventional sequence models such as LSTM, GRU, and CNN, achieving a validation accuracy of 84.68% and a macro-averaged F1 score of 84.64%. The model shows stable convergence, strong generalization, and interpretable attention mechanisms that focus on discriminative time segments. Ablation results further validate the individual contributions of the SE and temporal attention modules.

The effectiveness of SETransformer suggests its strong potential for real-world mobile sensing and context-aware applications. In future work, we plan to extend the model to incorporate multi-modal sensor inputs (e.g., gyroscope, magnetometer), investigate domain adaptation across users and devices, and explore its deployment efficiency on resource-constrained embedded systems.

References

[1] Basant Adel, Asmaa Badran, Nada E Elshami, Ahmad Salah, Ahmed Fathalla, and Mahmoud Bekhit. A survey on deep learning architectures in human activities recognition application in sports science, healthcare, and security. In *The International Conference on Innovations in Computing Research*, pages 121–134. Springer, 2022.

[2] Mohammed AA Al-Qaness, Ahmed M Helmi, Abdelghani Dahou, and Mohamed Abd Elaziz. The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis. *Biosensors*, 12(10):821, 2022.

[3] 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.

[4] Boris Bačić, Claudiu Vasile, Chengwei Feng, and Marian G Ciucă. Towards nation-wide analytical healthcare infrastructures: A privacy-preserving augmented knee rehabilitation case study. *arXiv preprint arXiv:2412.20733*, 2024.

[5] Malti Bansal, Apoorva Goyal, and Apoorva Choudhary. A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, 3:100071, 2022.

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[7] Luigi Bibbò, Riccardo Carotenuto, and Francesco Della Corte. An overview of indoor localization system for human activity recognition (har) in healthcare. *Sensors*, 22(21):8119, 2022.

[8] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*, 2023.

[9] F Xavier Gaya-Morey, Cristina Manresa-Yee, and José M Buades-Rubio. Deep learning for computer vision based activity recognition and fall detection of the elderly: a systematic review. *Applied Intelligence*, 54(19):8982–9007, 2024.

[10] Alexander Hoelzemann, Julia Lee Romero, Marius Bock, Kristof Van Laerhoven, and Qin Lv. Hang-time har: a benchmark dataset for basketball activity recognition using wrist-worn inertial sensors. *Sensors*, 23(13):5879, 2023.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[13] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. Human activity recognition: A survey. *Procedia Computer Science*, 155:698–703, 2019.

[14] Clayton Souza Leite, Henry Mauranen, Aziza Zhanabatyrova, and Yu Xiao. Transformer-based approaches for sensor-based human activity recognition: Opportunities and challenges. *arXiv preprint arXiv:2410.13605*, 2024.

[15] Shutao Li, Bin Li, Bin Sun, and Yixuan Weng. Towards visual-prompt temporal answer grounding in instructional video. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):8836–8853, 2024.

[16] Wanxin Li. The impact of apple's digital design on its success: An analysis of interaction and interface design. *Academic Journal of Sociology and Management*, 2(4):14–19, 2024.

[17] Wanxin Li. Transforming logistics with innovative interaction design and digital ux solutions. *Journal of Computer Technology and Applied Mathematics*, 1(3):91–96, 2024.

[18] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565, 2024.

[19] Zhihan Lv, Fabio Poiesi, Qi Dong, Jaime Lloret, and Houbing Song. Deep learning for intelligent human–computer interaction. *Applied Sciences*, 12(22):11457, 2022.

[20] Johannes Meyer, Adrian Frank, Thomas Schlebusch, and Enkelejda Kasneci. U-har: A convolutional approach to human activity recognition combining head and eye movements for context-aware smart glasses. *Proceedings of the ACM on Human-Computer Interaction*, 6(ETRA):1–19, 2022.

[21] Nafiul Rashid, Berken Utku Demirel, and Mohammad Abdullah Al Faruque. Ahar: Adaptive cnn for energy-efficient human activity recognition in low-power edge devices. *IEEE Internet of Things Journal*, 9(15):13041–13051, 2022.

[22] Muhammad Asfandyar Rustam, Muhammad Yasir Ali Khan, Tasawar Abbas, and Bilal Khan. Distributed secondary frequency control scheme with a-symmetric time varying communication delays and switching topology. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 9:100650, 2024.

[23] Lisa Schrader, Agustín Vargas Toro, Sebastian Konietzny, Stefan Rüping, Barbara Schäpers, Martina Steinböck, Carmen Krewer, Friedemann Müller, Jörg Güttler, and Thomas Bock. Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people. *Journal of Population Ageing*, 13:139–165, 2020.

[24] Jiangrong Shen, Yulin Xie, Qi Xu, Gang Pan, Huajin Tang, and Badong Chen. Spiking neural networks with temporal attention-guided adaptive fusion for imbalanced multi-modal learning. *arXiv preprint arXiv:2505.14535*, 2025.

[25] Yiting Wang, Tianya Zhao, and Xuyu Wang. Fine-grained heartbeat waveform monitoring with rfid: A latent diffusion model. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, pages 86–91, 2025.

[26] Yucheng Wang, Yuecong Xu, Jianfei Yang, Min Wu, Xiaoli Li, Lihua Xie, and Zhenghua Chen. Fully-connected spatial-temporal graph for multivariate time-series data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 15715–15724, 2024.

[27] Zhuqi Wang, Qinghe Zhang, and Zhuopei Cheng. Application of ai in real-time credit risk detection. *Preprints*, February 2025.

[28] Gary M Weiss. Wisdm smartphone and smartwatch activity and biometrics dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, 7(133190-133202):5, 2019.

[29] Yi Xin, Siqi Luo, Xuyang Liu, Haodi Zhou, Xinyu Cheng, Christina E Lee, Junlong Du, Haozhe Wang, MingCai Chen, Ting Liu, et al. V-petl bench: A unified visual parameter-efficient transfer learning benchmark. *Advances in Neural Information Processing Systems*, 37:80522–80535, 2024.

[30] Xiangrui Xu, Qiao Zhang, Rui Ning, Chunsheng Xin, and Hongyi Wu. Comet: A communication-efficient and performant approximation for private transformer inference. *arXiv preprint arXiv:2405.17485*, 2024.

[31] Chen Yang, Yangfan He, Aaron Xuxiang Tian, Dong Chen, Jianhui Wang, Tianyu Shi, Arsalan Heydarian, and Pei Liu. Wcdt: World-centric diffusion transformer for traffic scene generation. *arXiv preprint arXiv:2404.02082*, 2024.

[32] Zhanqi Zhang, Chi K Chou, Holden Rosberg, William Perry, Jared W Young, Arpi Minassian, Gal Mishne, and Mikio Aoi. A computational ethology approach for characterizing behavioral dynamics in bipolar disorder. *medRxiv*, 2024. Preprint, not peer-reviewed.

[33] Simin Zhu, Ronny Gerhard Guendel, Alexander Yarovoy, and Francesco Fioranelli. Continuous human activity recognition with distributed radar sensor networks and cnn–rnn architectures. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.