applied system innovation

MDPI

Article

Efficient and Explainable Human Activity Recognition Using Deep Residual Network with Squeeze-and-Excitation Mechanism

Sakorn Mekruksavanich [1] **and Anuchit Jitpattanakul** [2,3,*]

[1] Department of Computer Engineering, School of Information and Communication Technology, University of Phayao, Phayao 56000, Thailand; sakorn.me@up.ac.th
[2] Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand
[3] Intelligent and Nonlinear Dynamic Innovations Research Center, Science and Technology Research Institute, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand
* Correspondence: anuchit.j@sci.kmutnb.ac.th

Abstract: Wearable sensors for human activity recognition (HAR) have gained significant attention across multiple domains, such as personal health monitoring and intelligent home systems. Despite notable advancements in deep learning for HAR, understanding the decision-making process of complex models remains challenging. This study introduces an advanced deep residual network integrated with a squeeze-and-excitation (SE) mechanism to improve recognition accuracy and model interpretability. The proposed model, ConvResBiGRU-SE, was tested using the UCI-HAR and WISDM datasets. It achieved remarkable accuracies of 99.18% and 98.78%, respectively, surpassing existing state-of-the-art methods. The SE mechanism enhanced the model's ability to focus on essential features, while gradient-weighted class activation mapping (Grad-CAM) increased interpretability by highlighting essential sensory data influencing predictions. Additionally, ablation experiments validated the contribution of each component to the model's overall performance. This research advances HAR technology by offering a more transparent and efficient recognition system. The enhanced transparency and predictive accuracy may increase user trust and facilitate smoother integration into real-world applications.

Keywords: human activity recognition (HAR); explainable AI (XAI); wearable sensors; squeeze-and-excitation mechanism; deep residual network

Academic Editor: Georgios Th Papadopoulos

Received: 10 February 2025
Revised: 20 March 2025
Accepted: 21 April 2025
Published: 24 April 2025

Citation: Mekruksavanich, S.; Jitpattanakul, A. Efficient and Explainable Human Activity Recognition Using Deep Residual Network with Squeeze-and-Excitation Mechanism. *Appl. Syst. Innov.* **2025**, *8*, 57. https://doi.org/10.3390/asi8030057

1. Introduction

Human activity recognition (HAR) has become an essential domain in the age of ubiquitous computing and intelligent appliances. The capacity to autonomously identify and categorize individuals' behaviors employing data from wearable sensors has significant ramifications across several sectors, including medical services, intelligent residences, fitness monitoring, and individualized solutions [1–3]. With the increasing popularity of electronic devices, including advanced sensors, the possibility for more accurate and contextually aware HAR systems also rises.

Machine learning techniques have transformed HAR, providing robust instruments for deriving significant trends from sensor data. Traditional machine learning techniques, such as support vector machines and random forests, are often employed because they are easy to use and interpret [4]. In recent years, deep learning methodologies, including convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have attained the highest possible efficiency by autonomously acquiring hierarchical representations

of features [5]. These models effectively incorporate temporal and spatial correlations in sensor data, enhancing accuracy across different tasks.

The increasing complexity of deep learning architectures introduces significant challenges, particularly in interpretability and explainability. While these models often achieve high accuracy, their internal decision-making processes remain primarily opaque, leading to the well-known black-box problem [6]. This lack of transparency may erode user trust, hinder regulatory adherence, and impede the successful deployment of HAR systems in critical applications. Recent improvements in attention mechanisms, such as squeeze-and-excitation (SE) networks [7], have shown promise in improving model efficacy by allowing focus on the most important elements [8], yet their application in HAR, particularly with explainability approaches, requires further investigation.

This study tackles these challenges by proposing an efficient deep residual neural network model incorporating the SE mechanism. Our approach utilizes data from wearable smart device sensors to improve movement tracking accuracy and provide us with a way to comprehend how the model makes decisions. The main contributions and novelties of this study are as follows:

1. This study introduces an efficient deep residual neural network model, ConvResBiGRU-SE (Convolutional Residual Bidirectional Gated Recurrent Unit with SE), a novel integration of CNNs, residual connections, bidirectional gated recurrent units (GRUs), and SE mechanisms specifically optimized for HAR tasks.
2. This research develops an explainable framework for our HAR model, allowing a transparent interpretation of the model's predictions. Unlike many deep learning HAR systems that function as black boxes, our approach incorporates gradient-weighted class activation mapping (Grad-CAM) visualizations specifically adapted for time-series sensor data. This addresses a critical gap in the literature by making complex deep-learning models for HAR more transparent and interpretable.
3. This investigation uniquely evaluates performance across multiple datasets with different sensor placements (waist-mounted in UCI-HAR vs. front pocket in WISDM), demonstrating the architecture's robustness to sensor positioning variations—a practical challenge rarely addressed in previous studies.
4. Our systematic ablation experiments quantify the specific contribution of each architectural component. This methodical approach reveals that the SE mechanism improves model stability (reducing standard deviation by over 50%) while enhancing accuracy—insights not previously established in the HAR literature.

The remainder of this paper is structured to the following: Section 2 examines relevant literature in HAR, deep learning techniques, and attention mechanisms. Section 3 delineates the suggested deep learning framework and its constituents. Section 4 delineates the experimental configuration and exhibits the findings. Section 5 analyzes the findings. Section 6 ultimately concludes the paper and delineates prospective research directions.

2. Related Work

This section reviews pertinent research in HAR using wearable sensors, deep learning techniques, and attention mechanisms. It places our work within a more comprehensive scientific context.

2.1. HAR Using Wearable Sensors

HAR using wearable sensors has been a significant research area for over a decade due to its accessibility, affordability, and portability [9]. Wearable sensor signals are typically favored over video camera signals for different explanations [10]: (1) Wearable sensors circumvent the environmental and static constraints of cameras, which are immo-

bile; (2) Considerable sensors affixed to the body enhance signal accuracy and efficiency; (3) Wearable sensors collect signals for designated objectives, in contrast to cameras that may inadvertently record non-target individuals; (4) Wearable sensors provide enhanced privacy, as video cameras perpetually capture full-body footage during daily activities; (5) Subjects must remain within the fixed field of view of a camera system, limiting mobility and practicality; and (6) Video processing is intricate and expensive.

Conventional machine learning techniques, such as support vector machine and random forest, have been extensively utilized in HAR tasks [11]. These approaches typically depend on hand-crafted features extracted from time and frequency domains. Although such methods adequately recognize simple physical activities, they often exhibit limitations in capturing complex or subtle movement patterns. Deep learning approaches have significantly advanced HAR. CNNs excel at extracting spatial features from sensor data [12]. Recurrent neural networks (RNNs), especially LSTM networks, effectively capture temporal dependencies in activity sequences [13]. Integrating models that incorporate CNNs and LSTMs utilizes both spatial and temporal data [14].

Hybrid models integrating one-dimensional CNNs and RNNs have been thoroughly investigated. Our prior research [15] demonstrated that hybrid CNN with LSTM models surpass independent CNN or LSTM models in performance. Challa et al. [16] employed a hybrid model combining CNN and bidirectional LSTM (BiLSTM) networks. Their multi-branch CNN with BiLSTM network can autonomously extract features from raw sensor data without extensive pre-processing. Canizo et al. [17] introduced a multi-headed CNN–RNN architecture, assigning a specific CNN head to each sensor. The feature maps created by the CNN are merged and transmitted to the RNN module to detect temporal patterns. Dua et al. [18] introduced a neural network model that employs CNNs and GRUs for automated feature extraction and movement identification.

2.2. Attention Mechanisms in Deep Learning

Attention mechanisms are crucial for improving models developed using deep learning by allowing them to concentrate on the most pertinent aspects of the input data [19].

SE networks, proposed by Hu et al. [7], have markedly enhanced image classification challenges through the flexible recalibration of channel-wise feature reactions. The SE block employs global data that selectively emphasize significant features while attenuating less relevant features.

Zhongkai et al. [20] investigated the feasibility of SE blocks in HAR studies by assessing the HAR capabilities of several cutting-edge CNN models (such as VGG16, Inception, ResNet18, and PyramidNet18) that were trained with them. In order to recognize transitional activities, Mekruksavanich et al. [21] suggested a deep learning model called SEResNet-BiGRU. This model integrates SE, residual, and bidirectional GRUs (BiGRUs) blocks. A multi-branch deep learning design was presented by Khan et al. [22]. In this structure, each branch extracts and re-weights feature maps using a CNN-based model with a SE unit.

3. Methodology

This section outlines a detailed methodology for recognizing human actions using our proposed ConvResBiGRU-SE model. As shown in Figure 1, the methodology includes three main components: data acquisition, data pre-processing, and model development.

In the data acquisition phase, we collect sensor data from wearable devices, focusing on the UCI-HAR and WISDM datasets, which offer diverse real-world scenarios for activity recognition. During the data pre-processing process, we perform essential phases such as

noise reduction, data normalization, and segmentation to prepare the raw sensor data for deep learning.

Finally, the model development phase describes our innovative architecture integrating CNNs, residual connections, BiGRU, and SE mechanisms to achieve robust activity recognition. Each component is meticulously designed to address specific challenges in sensor-based HAR while ensuring computational efficiency and model interpretability.

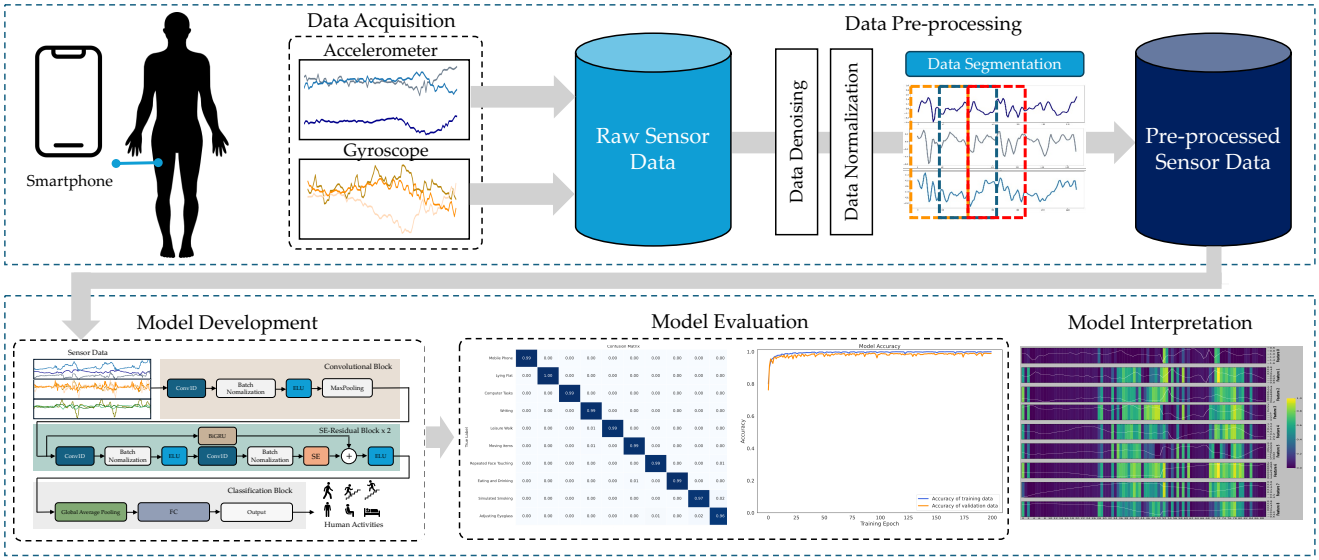


Figure 1. The HAR workflow used in this work.

3.1. Evaluation Datasets

This study utilized two publicly available HAR datasets, UCI-HAR and WISDM, to assess the proposed deep learning model. These datasets were selected because they encompass diverse real-world scenarios in HAR employing wearable sensors. They feature different sensor placements, sampling rates, and types of activities. Both datasets include measurements from inertial sensors (accelerometer and gyroscope) collected via smartphones, making them ideal for assessing HAR systems based on wearable devices.

3.1.1. UCI-HAR Dataset

This research employs the UCI human behavior recognition dataset obtained through handheld electronic devices [23] to observe events in the community. The UCI-HAR dataset comprises action data from thirty individuals, with ages, nationalities, heights, and weights varying between 18 and 48 years. Individuals carried a Samsung Galaxy S-II smartphone (Suwon, Republic of Korea) at waist height while engaging in routine activities. Each participant performed six distinct activities: walking, ascending stairs, descending stairs, sitting, standing, and lying down. Sensor data was captured utilizing integrated tri-axial measurements from the smartphone's accelerometer and gyroscope. Sensor data were recorded at a sampling rate of 50 Hz, capturing linear acceleration and angular velocity throughout the six predefined activities.

3.1.2. WISDM Dataset

The WISDM dataset [24] is an essential human activity recognition dataset from the Wireless Sensor Data Mining Laboratory. This openly accessible dataset was generated by the WISDM team and collected using a smartphone positioned in the front leg pocket of the subjects' trousers. The dataset comprises 1,098,207 samples and documents the behavior trends of 36 participants. The activities included walking, jogging, ascending,

descending, sitting, and standing. Individuals gathered information by placing an Android smartphone in their front leg pocket, employing the device's integrated accelerometer sensor at a sampling frequency of 20 Hz.

3.2. Data Pre-Processing

Raw sensor data from wearable instruments usually includes noise and variations that can negatively impact HAR model interpretation. This study implemented several pre-processing efforts to improve data quality and enhance model recognition. The pre-processing channel has three prominent phases: data denoising to eliminate undesirable noise from sensor signals, data normalization to regularize input values, and data segmentation employing a sliding window method to formulate sequential data for the deep learning model. These pre-processing phases are essential for providing the HAR system's robustness and reliability, as they support reducing the effects of sensor noise, instrument variations, and temporal inconsistencies in unprocessed data.

3.2.1. Data Denoising

In contexts involving smart wearable devices, including domestic environments, outdoor athletics, and routine employment, a smartwatch's integrated inertial measurement unit (IMU) could create noise. This distortion may arise from surrounding magnetic fields, external interference, and equipment exactness, impacting the efficacy of the individual's behavior identification technique. Employing data from various wearable inertial measurement units can mitigate noise effects in multi-sensor systems. Nonetheless, noise can substantially impact the identification method in routine scenarios where a single IMU is employed for data acquisition. Consequently, data filtration in single-sensor networks is imperative.

Diverse algorithms, such as median, low-pass, and Kalman filtering, are available. Studies demonstrate that routine interactions between individuals generally oscillate between 1 and 15 Hz, a low-frequency spectrum amenable to filtration via a low-pass filter on IMU signals [25]. The Butterworth filter is a conventional, accessible filter utilized as a low-pass filter to attenuate high-frequency noise. This filter exhibits a maximum flat frequency response curve in the passband and gradually increases attenuation to zero in the stopband. The amplitude attenuation ratio in the stopband is directly correlated to the filter's order. The Butterworth filter is extensively utilized in signal processing because of its superior response to frequency properties.

The Butterworth filter can be represented by Equation (1):

$$|H(\omega)|^2 = \frac{1}{1+(\frac{\omega}{\omega_c})^{2N}} = \frac{1}{1+\varepsilon^2(\frac{\omega}{\omega_p})^{2N}} \quad (1)$$

where N represents the filter hierarchy, ω_c is the cut-off frequency, and ω_p is the boundary frequency of the passband.

This study establishes the wearable device's IMU sampling rate at 100 Hz to ensure high-precision behavioral interpretation. The filter's cut-off frequency is established at 20 Hz with an order of 4 to preserve movement data below 15 Hz. Following the segmentation of sensor data from public datasets, the accelerometer and gyroscope data frames from a selected sensor are subjected to filtering. Figure 2 illustrates the signal waveforms prior to and subsequent to filtering.

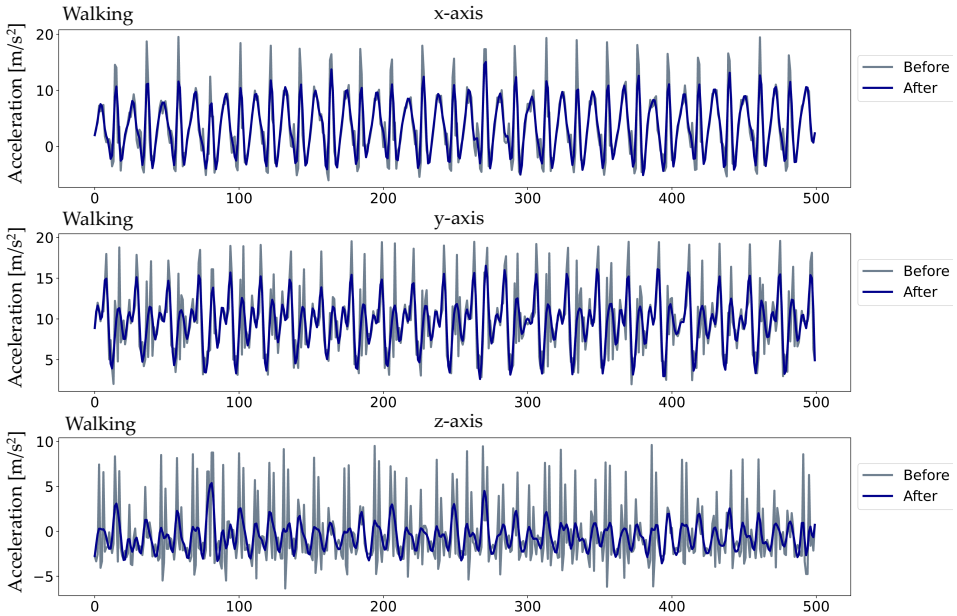


Figure 2. Sample accelerometer signal frames of walking activity, shown before and after applying filtering techniques.

3.2.2. Data Normalization

The unprocessed sensor data is subsequently normalized to a range of 0 to 1. This normalization procedure resolves the model learning concern by guaranteeing that all data values fall within a comparable scope. As a result, gradient descent algorithms can converge more efficiently [26].

$$X_i^{norm} = \frac{X_i - x_i^{min}}{x_i^{max} - x_i^{min}}, \quad i = 1, 2, \ldots, n \tag{2}$$

X_i^{norm} denotes the normalized data, n signifies the number of channels, while x_i^{max} and x_i^{min} denote the highest and lowest points of the i-th channel, respectively.

3.2.3. Data Segmentation

Due to the extensive volume of signal data gathered by wearable sensors, it is unfeasible to enter all possible information into the HAR model concurrently. Consequently, sliding window segmentation is executed before inputting the data into the model. The sliding window technique is one of the most widely used data segmentation approaches in HAR. It is effective in capturing both dynamic activities (e.g., jogging, walking) and static activities (e.g., standing, sitting, lying down) [27]. The unprocessed sensor signals are segmented into fixed-length intervals, with an overlapping section between consecutive intervals. This overlap enhances the quantity of training data samples and is beneficial in preventing the omission of changeover between operations. The windowing procedure is depicted in Figure 3.

The segmented sample data, utilizing a sliding window of length N, has dimensions of $K \times N$. The sample W_t is represented as:

$$W_t = [a_t^1 a_t^2 \ldots a_t^K] \in \mathbb{R}^{K \times N} \tag{3}$$

where the column vector $a_t^k = (a_{t_1}^k a_{t_2}^k \ldots a_{t_N}^k)^{\mathrm{T}}$ represents the signal data of sensor k at window time t. Additionally, T denotes the transpose operator, K is the number of sensors, and N is the length of the sliding window. To utilize the correlations among windows and facilitate the training process, the window data is divided into sequences of windows.

$$S = \{(W_1, y_2), (W_1, y_2), \ldots, (W_T, y_T)\} \quad (4)$$

where T denotes the duration of the window sequence and y_t signifies the associated activity label for window W. If a window comprises several activity classes, the predominant activity sample will be selected as the label for that window.

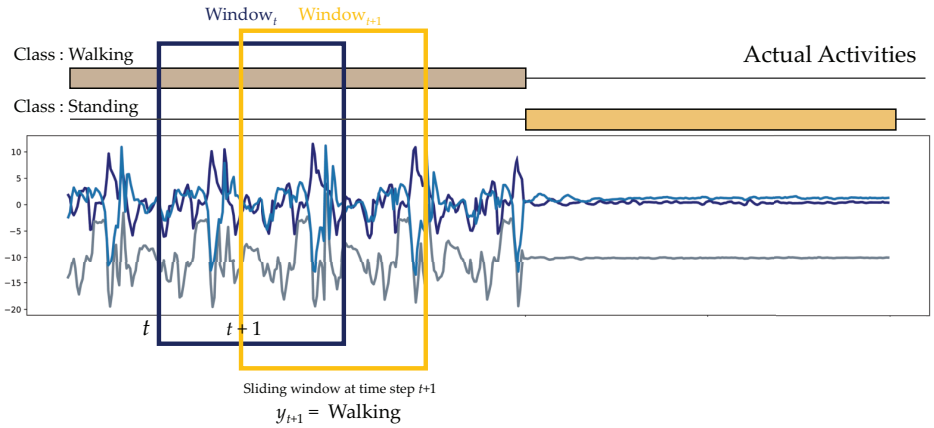


Figure 3. Fixed-length sliding window method employed in this study.

3.3. The Proposed ConvResBiGRU-SE Model

This section presents the ConvResBiGRU-SE model, an innovative deep-learning structure designed for efficient HAR utilizing data from wearable sensors. The model incorporates the benefits of CNNs, residual connections, BiGRUs, and SE mechanisms. This amalgamation ensures substantial feature acquisition and accurate behavior categorization. As shown in Figure 4, the architecture includes several essential components: a convolutional block for initial feature extraction, multiple SE-Residual blocks with bidirectional GRU units for hierarchical feature learning, and a classification module for final action recognition.

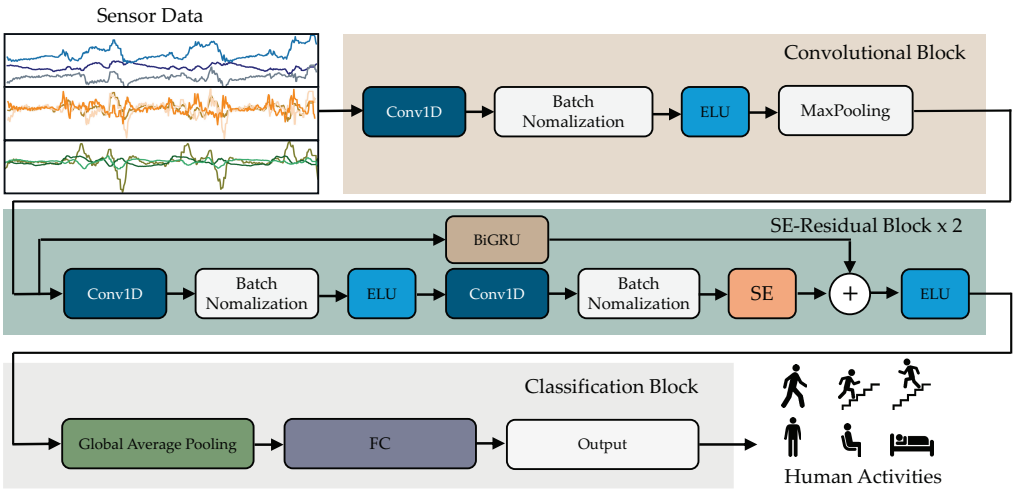


Figure 4. Comprehensive and expanded architecture of the proposed hybrid deep residual model.

3.3.1. Convolutional Block

When employing CNNs, a particular array of components is generally utilized. CNNs are frequently employed in learning under supervision, wherein each neuron is interconnected with every neuron in the following layer. The value entered by the neurons is transformed into an output value via the activation function of the neural network. The

efficiency of this function is dictated by two principal factors: sparsity and its capacity to regulate the diminished gradient flow to the network's layers below.

CNNs often utilize pooling for dimensionality reduction, implementing both maximum and average pooling functions, referred to as max-pooling and average-pooling, respectively.

This study employs a convolutional block (ConvB) to derive low-level features from unprocessed sensor data. The ConvB consists of four layers: one-dimensional convolutional (Conv1D), batch normalization (BN), exponential linear unit (ELU), and max-pooling (MP) layers, as illustrated in Figure 4. The Conv1D layer employs several trainable convolutional kernels to extract diverse features, with each kernel generating a distinct feature map. The BN layer enhances stability and accelerates the training operation. The ELU activation function enhances the model's representational capacity by introducing non-linearity while avoiding vanishing gradient issues. The MP layer reduces the spatial dimensions of the feature maps by retaining the most salient features within each pooling region.

3.3.2. Structure of GRU

The GRU was created as a novel architecture derived from RNNs to mitigate the challenges of expanding or vanishing gradients. Nonetheless, the memory cells within its architecture lead to increased memory consumption [28]. The GRU is a streamlined variant of the LSTM model, lacking a distinct memory cell in its architecture [29,30].

A GRU contains update and reset gates that regulate the revised magnitude of each hidden state. These gates dictate the data that should be transmitted to the subsequent state and withheld, as depicted in Figure 5a. The concealed state h_t at time t is determined by the output of the update gate z_t, the reset gate r_t, the current input x_t, and the preceding hidden state h_{t-1}. The mathmatical formulas of the gates in GRU are shown as follows:

$$z_t = \sigma(W_z x_t \oplus U_z h_{t-1} + b_z), \tag{5}$$

$$r_t = \sigma(W_r x_t \oplus U_r h_{t-1} + b_r), \tag{6}$$

$$g_t = \tanh(W_g x_t \oplus U_g (r_t \otimes h_{t-1}) + b_g), \tag{7}$$

$$h_t = ((1 - z_t) \otimes h_{t-1}) \oplus (z_t \otimes g_t) \tag{8}$$

where W represents the weight matrices for the current input, U represents the weight matrices for the previous hidden state, and b are bias vectors. The function σ represents a sigmoid function. The symbol $\oplus$ denotes a basic addition operation, while $\otimes$ signifies a fundamental multiplication operation.

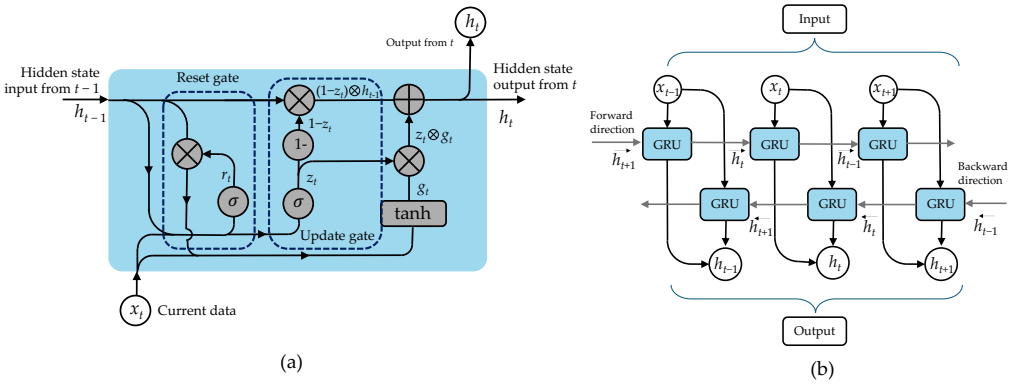


Figure 5. Structure of BiGRU: (**a**) GRU cell and (**b**) unroll BiGRU.

In 1997, Schuster and Paliwal designed the bidirectional RNN (BiRNN) to overcome the constraints of the traditional unidirectional RNN [31]. In contrast to standard RNNs, the output at any specific time step in a BiRNN incorporates information from preceding and subsequent inputs. This is achieved by simultaneously training two separate RNN layers: one processes the sequence in the forward direction, while the other processes it in reverse.

In a BiRNN, the neurons of a conventional RNN are partitioned into two segments: one for interpreting data in the forward direction and the other for the backward direction. Conversely, the outputs of positive neurons are not linked to negative neurons. This yields the overall structure illustrated in Figure 5b. The subsequent equations delineate the mathematical operations involved:

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}) \quad (9)$$

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t+1}}) \quad (10)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (11)$$

where $GRU()$ represents the GRU function that processes the current input x_t along with the previous hidden state in either the forward ($\overrightarrow{h_{t-1}}$) or backward ($\overleftarrow{h_{t+1}}$) direction.

3.3.3. SE-Residual Block

Fundamental deep learning architectures, such as LeNet, AlexNet, and VGGNet, typically consist of convolutional layers followed by fully connected layers for classification or regression tasks. These models do not incorporate residual or skip connections. Instead, each layer feeds its output directly into the next, forming what are known as sequential networks.

As the layers in a sequential network grow, concerns with vanishing or exploding gradients could arise. ResNet incorporates residual blocks that facilitate skip connections among convolutional layers. This enhances gradient propagation and facilitates the training of significantly deeper CNNs with no facing gradient vanishing issues. A residual layer could be depicted in the following format:

$$\mathrm{ELU}(x) = \begin{cases} x & \mathrm{x} \geq 0 \\ \alpha * (e^x - 1) & \mathrm{x} < 0 \end{cases} \quad (12)$$

$$R(x) = \mathrm{ELU}(x + f(x)) \quad (13)$$

In this setting, $f(x)$ denotes the layer's result, whereas x signifies the input value. The function $\mathrm{ELU}(x)$ represents the exponential linear unit function, while $R(x)$ represents the result of the residual block. The residual component $f(x)$ in this block is defined by two consecutive sequences of three processes: convolution with a 3×1 filter, batch normalization, and ELU activation. The feature map from $f(x)$ is combined with the input x. These concatenated features are subjected to the ELU activation function.

This study introduces the SEResidual block, which systematically retrieves hybrid features by integrating spatiotemporal and channel-wise data [32]. This residual block comprises Conv1D, BN, ELU, SE modules, and a shortcut connection with BiGRU, as depicted in Figure 4. The SE modules augment the model's representational capacity by emphasizing channel attention.

Figure 6 illustrates the configuration of a SE module. The convolution process provides multiple feature maps. Specific maps may include superfluous information. The SE component executes feature recalibration to accentuate the salient features and diminish the less pertinent ones. This module functions in two primary steps: squeeze and excitation.

During the squeeze stage, all data on the channels is gathered. The dimensions of the feature map U for a single channel are $C \times H \times W$, where $H \times W$ denotes the feature map's dimensions. Feature maps for each channel are condensed into 1×1 feature maps utilizing a channel descriptor function, such as global average pooling (GAP) [33]. A scalar value denoting the channel's global information is generated within this period. The squeeze process is delineated in Equation (14), where $(u_c(i,j))$ signifies a feature map for channel c after the passage of X through the convolution layer. $F_{squeeze}$ denotes the channel characterization function, and GAP was employed in this research.

$$Z_c = F_{squeeze}(U_c) \tag{14}$$

$$= \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} U_c(i,j) \tag{15}$$

Subsequently, during the excitation phase, channel-wise dependencies are modeled using the descriptors obtained from the preceding squeeze phase. This is achieved via fully-connected (FC) layers and non-linear functions. Equation (16) delineates the excitation phase, wherein z denotes the value derived from the squeeze phase, W_i signifies the ith fully connected layers, σ denotes the sigmoid function, and F_{excite} indicates the excitation function. The sigmoid function guarantees that the result of the excitation phase is confined within the interval of 0 to 1, which can subsequently serve as a calibration weight. The newly derived weight s from the excitation phase is multiplied by the current feature map U. Figure 6 illustrates the functionality of the SE module employed in the present investigation, showcasing the configuration of the squeeze and excitation phases within the SE block.

$$s = F_{excite}(z, W) \tag{16}$$

$$= \sigma(g(z, W)) \tag{17}$$

$$= \sigma(W_2 \mathrm{ReLU}(W_1 z)) \tag{18}$$

The last stage entails restructuring the outcome U to implement activations in the side path network. In this context, $X = [x_1, x_2, \ldots, x_n]$ and $s_n U_n$ represent the element-wise multiplication of the scalar s_n with the feature map. This method allocates adaptive weights to the feature channels, which is the core principle of the SE block [34].

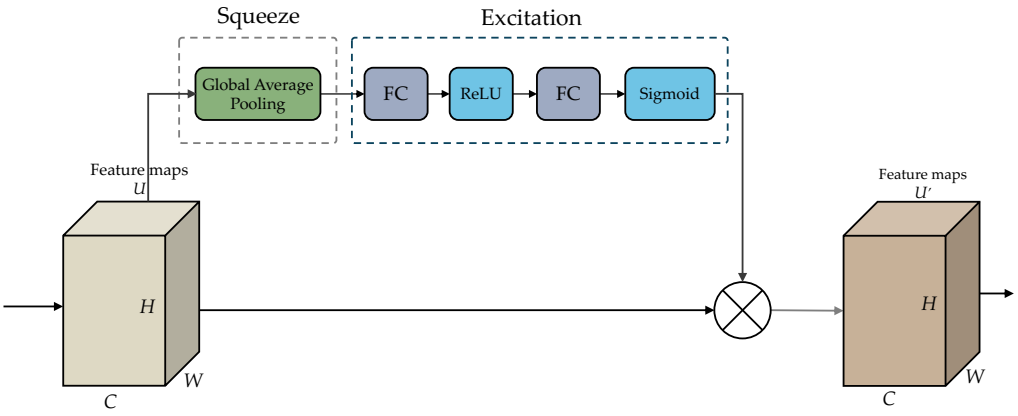


Figure 6. Structure of the SE module.

3.4. Model Training and Hyperparameters

The proposed ConvResBiGRU-SE model's performance is dependent on the availability of sufficient and diverse training data and the careful tuning of architectural design parameters, commonly known as hyperparameters. The hyperparameters comprise epoch

counts, learning rates, and batch size. To guarantee strong model performance, we utilized a conventional method that entails partitioning the data into training and validation sets. The training set facilitated hyperparameter optimization, whereas the holdout validation set served for independent comparative evaluation.

We identified the optimal hyperparameter settings that maximized the model's accuracy through trial and error. The parameters consist of a batch size of 128, a learning rate of 1×10^{-3}, and a maximum epoch limit of 200. An early stopping strategy was employed to prevent overfitting and ensure that the model learns generalizable patterns rather than merely memorizing the training data. The training was automatically halted after 10 consecutive epochs without any enhancement in validation accuracy. Furthermore, we instituted an adaptive learning rate strategy that decreases the learning rate by 25% if no enhancement is detected during the specified interval. This method facilitates more efficient model convergence and prevents entrapment in suboptimal solutions.

Table 1 presents a comprehensive overview of the hyperparameters employed in our model architecture. The convolutional block employed a kernel size of 3 and 256 filters, facilitating efficient spatial feature extraction from the raw sensor data. The SE-Residual blocks, central to our proposed architecture, were designed with BiGRU units featuring 128 hidden states to capture temporal dependencies, while the convolutional paths utilized varying kernel sizes (3 and 5) to extract multi-scale features. The classification block employed global average pooling, succeeded by fully connected layers, resulting in an output layer with neurons corresponding to the number of activity classes.

Table 1. Summary of hyperparameters of the ConvResBiGRU-SE model.

Stage	Hyperparameters		Values
Architecture	Convolution Block		
	1D Convolution	Kernel Size	3
		Stride	1
		Filters	256
	Batch Normalization		-
	Activation		ELU
	Max Pooling		2
	SE-Residual Block - 2 (Path 1)		
	BiGRU Unit		128
	(Path 2)		
	Conv1D	Kernel Size	3
		Stride	1
		Filters	256
	Batch Normalization		-
	Activation		ELU
	Conv1D	Kernel Size	5
		Stride	1
		Filters	256
	Batch Normalization		-
	SE		-
	Classification Block		
	Global Average Pooling		-
	Dense		64
	Dense		No. of activity classes
Training	Loss Function		Cross-entropy
	Optimizer		Adam
	Batch Size		128
	Number of Epochs		200

This carefully tuned hyperparameter configuration yielded a model that achieved outstanding performance while maintaining reasonable computational complexity, making it suitable for practical applications in human activity recognition using wearable sensors.

3.5. Model Evaluation

The efficacy of identifying activities in the suggested deep learning approach is evaluated through a 5-fold cross-validation technique. The subsequent equations delineate the mathematical representations for all five metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{22}$$

The following are the principal assessment requirements employed in HAR investigations. Recognition is categorized based on true positive (TP) classification for the target class and true negative (TN) identification for all other classes. Occasionally, activity sensor data from one category may be erroneously classified as belonging to another, resulting in a false positive (FP) identification. Likewise, data from a different class may be erroneously classified as the target class, leading to a false negative (FN) classification.

The efficacy of the deep learning models in the present research was evaluated utilizing a confusion matrix. This square matrix with k classes offers comprehensive findings for a multiclass classification issue. The confusion matrix provides an in-depth evaluation of accurately and inaccurately classified occurrences by supervised learning models. An element $c_{i,j}$ in the matrix denotes the frequency of instances of class C_i that have been categorized as class C_j. The confusion matrix also emphasizes classification inaccuracies.

Let C represent a confusion matrix derived from the identical testing method. In these expressions, $C_1, C_2, C_3, \ldots, C_k$ represent the k categories of activities in a HAR dataset, and $n = \sum\sum c_{i,j}$ denotes the overall amount of data elements classified within the matrix C. Diagonal elements represent concordant components that are accurately classified within the same category, whereas $c_{i,j}$ for $i \neq j$ denotes discordant components that are part of C_i but are classified as C_j. The confusion matrix C is depicted in Figure 7.

$$C = \begin{array}{c} \\ \text{True Class} \end{array} \begin{array}{c} \text{Predicted Class} \\ \begin{array}{cccc} C_1 & C_2 & C_j & C_k \end{array} \\ \begin{array}{c} C_1 \\ C_2 \\ C_i \\ C_k \end{array} \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,j} & c_{1,k} \\ c_{2,1} & c_{2,2} & c_{2,j} & c_{2,k} \\ c_{i,1} & c_{i,2} & c_{i,j} & c_{i,k} \\ c_{k,1} & c_{k,2} & c_{k,j} & c_{k,k} \end{bmatrix} \end{array}$$

Figure 7. The confusion matrix C.

3.6. Model Interpretation

Deep learning has emerged as a prominent and developed domain across multiple sectors that adopt modern innovations. The advancement of community life depends on deep learning to address intricate issues and deliver dependable alternatives. Deep learning is often regarded as having the potential to automate human-centric tasks, reducing or eliminating the need for direct human intervention. However, the inherently opaque nature of deep learning models has led many communities to hesitate to adopt them for

routine decision-making tasks. As a result, there is a growing demand for transparency and interpretability in these emerging methodologies.

Since 2018, many investigators have developed a novel discipline known as eXplainable Artificial Intelligence (XAI) [35]. XAI tackles the scientific aspects of deep learning models to guarantee a degree of interpretability while incorporating principles of confidentiality and accountability.

From a technical standpoint, the interpretability of a recently created deep learning model can improve its implementation. Primarily, constructing an interpretable model ensures equity in the decision-making process. Additionally, interpretability can detect possible adversarial changes that influence forecasting, facilitating targeted enhancements to the model's foundation. Lastly, interpretability guarantees that only significant features affect the intended output, elucidating the fundamental causality in the data and model rationale.

Grad-CAM is an interpretability method employed in deep learning, especially for CNNs, to emphasize the areas in an image that significantly influence the prediction caused by a model for a specific class. The procedure entails calculating the gradients of the target class score concerning the feature maps in the last convolutional layer. The gradients are subsequently weighted according to their significance, and the weighted sum is utilized to generate a heatmap that illustrates the discriminative areas that affected the model's conclusion [36,37].

Grad-CAM is utilized in multiple applications, such as image classification, object identification, and clinical image analysis. It offers perspectives on essential features and enhances the interpretability of CNNs by generating spatially accurate visualizations of class-specific activations.

4. Experimental Environments and Findings

This section considerably evaluates our suggested ConvResBiGRU-SE model via experimental validation on two public HAR datasets: UCI-HAR and WISDM. Initially, we outline the experimental setup, detailing the hardware configuration, software environment, and implementation specifics. Subsequently, we provide a comprehensive performance analysis, contrasting our model with various baseline methods utilizing standard assessment indicators, including accuracy, precision, recall, and F1-score. The experimental findings demonstrate the efficacy of our suggested approach in identifying human actions from wearable sensor data.

4.1. Experimental Settings

This research utilized Google Colab Pro+, equipped with a Tesla L4 GPU, to accelerate the training of deep learning models. The ConvResBiGRU-SE model, in conjunction with various standard deep learning networks (CNN, LSTM, BiLSTM, GRU, and BiGRU), was constructed utilizing a Python library with TensorFlow 2.17.1 and CUDA backends. The following Python libraries were also employed:

- Numpy 1.26.4 and Pandas 2.2.2 for data management, including retrieval, processing, and sensor data investigation.
- Matplotlib 3.10.0 and Seaborn 0.13.2 are utilized to visualize and demonstrate data exploration and model assessment outcomes.
- Scikit-learn 1.5.2 for sampling and data generation in experiments.
- TensorFlow 2.17.1 and Keras 2.5.0 for establishment and training deep learning models.

Analyses were conducted on the UCI-HAR and WISDM datasets to determine the most efficacious methodology. By employing the data segmentation technique in Section 3.2.3, the dataset derived from the two datasets utilized in our study was initially partitioned into

10,299 segments for UCI-HAR and 10,864 segments for WISDM. The quantity of segmented samples in each activity class is presented in Table 2.

Table 2. Numbers of segmented samples in each activity class of the UCI-HAR and WISDM datasets.

Dataset	Walking	Walking Upstairs	Walking Downstairs	Sitting	Standing	Laying	Jogging	Total
UCI-HAR	1722	1544	1406	1777	1906	1944	-	10,299
WISDM	4148	1236	1001	597	484	-	3366	10,864

To mitigate the reliance of findings on specific data partitioning into training and testing subsets, we employed a stratified 5-fold cross-validation methodology as illustrated in Figure 8. This validation scheme involves partitioning the dataset into five equal folds, with each fold (20% of the data) utilized sequentially for testing, while the remaining four folds (80% of the data) serve for training. This validation technique preserves the original dataset's class label proportions in each fold.

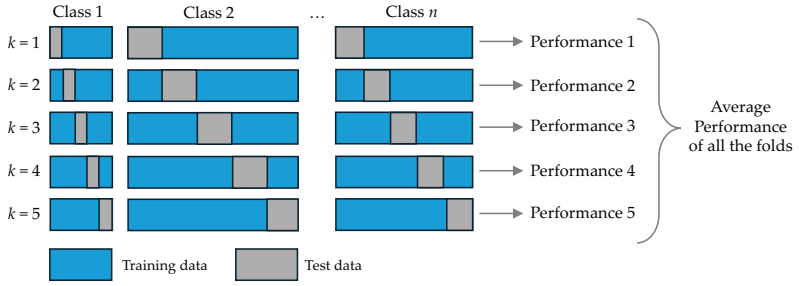


Figure 8. Schematic diagram of Stratified 5-fold cross-validation.

4.2. Experimental Findings

4.2.1. Experimental Results from UCI-HAR Dataset

The experimental findings on the UCI-HAR dataset highlight the exceptional interpretation of our proposed ConvResBiGRU-SE model compared to baseline methods. These results are presented in Table 3.

Table 3. Performance comparison of deep learning models on the UCI-HAR dataset (results are presented as mean percentage ± standard deviation over 5-fold cross-validation, with top-performing results highlighted in bold).

Model	Recognition Performance (Mean% (±std.%))			
	Accuracy	Precision	Recall	F1-Score
CNN	98.02% (±0.21%)	98.11% (±0.21%)	98.13% (±0.20%)	98.12% (±0.21%)
LSTM	96.75% (±0.74%)	96.97% (±0.67%)	96.95% (±0.71%)	96.95% (±0.70%)
BiLSTM	97.86% (±0.64%)	98.05% (±0.54%)	97.98% (±0.62%)	97.99% (±0.61%)
GRU	98.27% (±0.61%)	98.37% (±0.58%)	98.37% (±0.56%)	98.36% (±0.57%)
BiGRU	98.77% (±0.26%)	98.85% (±0.22%)	98.83% (±0.24%)	98.84% (±0.24%)
ConvResBiGRU-SE	**99.18% (±0.20%)**	**99.25% (±0.17%)**	**99.23% (±0.20%)**	**99.24% (±0.19%)**

Table 3 demonstrates that the model attained an accuracy of 99.18% (±0.20%), markedly exceeding all baseline models. The model's comprehensive performance indicators reveal a precision of 99.25% (±0.17%), underscoring its effectiveness in reducing false positives. The recall rate of 99.23% (±0.20%) illustrates its efficacy in accurately determining true positives. The F1-score of 99.24% (±0.19%) signifies a well-balanced efficiency in precision and recall.

Among the baseline models, BiGRU performed second best with an accuracy of 98.77%, showcasing the benefits of bidirectional temporal information processing. GRU and CNN also performed well, with accuracies of 98.27% and 98.02% respectively. The basic LSTM

model had a relatively lower performance at 96.75%, while BiLSTM improved upon this with an accuracy of 97.86%.

The low standard deviations (around ±0.2%) across all metrics for our proposed model indicate its stability and reliability. Integrating the SE mechanism with the residual architecture enhanced the model's feature learning capability, leading to more accurate and robust activity recognition. These results validate our approach of combining attention mechanisms with deep residual learning for HAR tasks.

4.2.2. Experimental Results from WISDM Dataset

The experimental findings on the WISDM dataset confirm the efficacy of the suggested ConvResBiGRU-SE model. The outcomes are presented in Table 4.

Table 4. Performance comparison of deep learning models on the WISDM dataset (results are presented as mean percentage ± standard deviation over 5-fold cross-validation, with top-performing results highlighted in bold).

Model	Recognition Performance (Mean% (±std.%))			
	Accuracy	Precision	Recall	F1-Score
CNN	92.59% (±0.77%)	90.67% (±0.87%)	89.43% (±0.98%)	89.90% (±0.91%)
LSTM	97.40% (±0.50%)	96.13% (±0.60%)	96.08% (±1.08%)	96.09% (±0.82%)
BiLSTM	97.86% (±0.22%)	96.88% (±0.58%)	96.82% (±0.58%)	96.84% (±0.57%)
GRU	97.45% (±0.24%)	96.34% (±0.42%)	96.10% (±0.37%)	96.20% (±0.40%)
BiGRU	97.96% (±0.20%)	96.96% (±0.50%)	96.70% (±0.61%)	96.82% (±0.56%)
ConvResBiGRU-SE	**98.78% (±0.24%)**	**98.04% (±0.49%)**	**98.16% (±0.56%)**	**98.09% (±0.53%)**

As indicated in Table 4, the ConvResBiGRU-SE model attained the highest accuracy of 98.78% (±0.24%) among all the methods compared. This highlights its superior ability to recognize human activities from wearable sensor data. The model's detailed metrics include a precision of 98.04% (±0.49%), a recall of 98.16% (±0.56%), and an F1-score of 98.09% (±0.53%), demonstrating consistent performance across various evaluation criteria.

Among the baseline models, BiGRU performed well with an accuracy of 97.96% (±0.20%). BiLSTM followed closely with an accuracy of 97.86% (±0.22%), showcasing the effectiveness of bidirectional processing in capturing temporal dependencies. GRU and LSTM models achieved similar performance levels, with accuracies of 97.45% and 97.40% respectively. This suggests that unidirectional recurrent architectures can still effectively model sequential patterns in human activity data. The basic CNN model had the lowest performance, with an accuracy of 92.59%, indicating that simple convolutional architectures alone may not be sufficient to capture the complex temporal patterns present in the WISDM dataset.

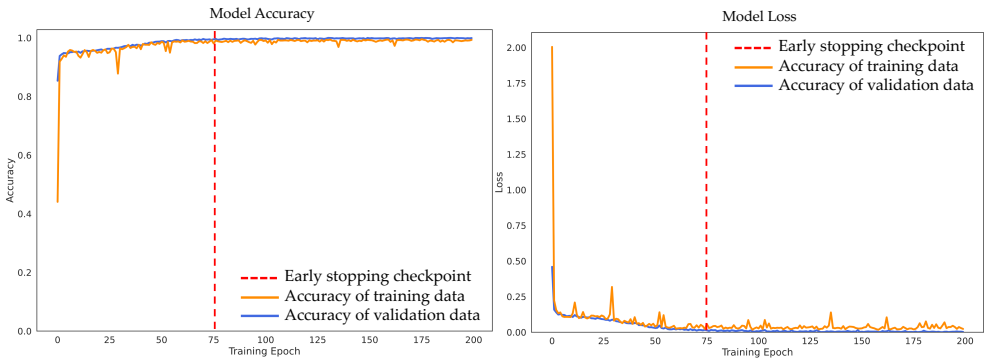
5. Discussion

This section shows an in-depth analysis of the performance and attributes of our ConvResBiGRU-SE model. We analyze the model's behavior through three main aspects: performance across various datasets, ablation studies of its architectural components, and the interpretability of its decision-making process. This analysis expects to deliver insights into the quantitative outcomes and qualitative understanding of our model's effectiveness in HAR tasks.

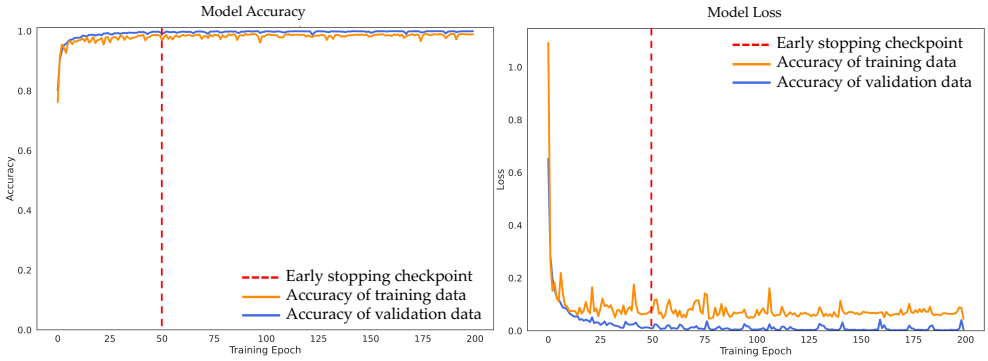
5.1. Performance Analysis

The performance of our ConvResBiGRU-SE model can be evaluated using quantitative metrics and training dynamics. Figure 9 illustrates that the model maintains stable training behavior on the UCI-HAR and WISDM datasets. For the UCI-HAR dataset (Figure 9a), the model converges quickly within the first 25 epochs. Training and validation accuracy curves

show steady improvement before leveling off at approximately 99% accuracy. The loss curves initially drop sharply and stabilize, indicating effective learning without overfitting.



(**a**) UCI-HAR dataset



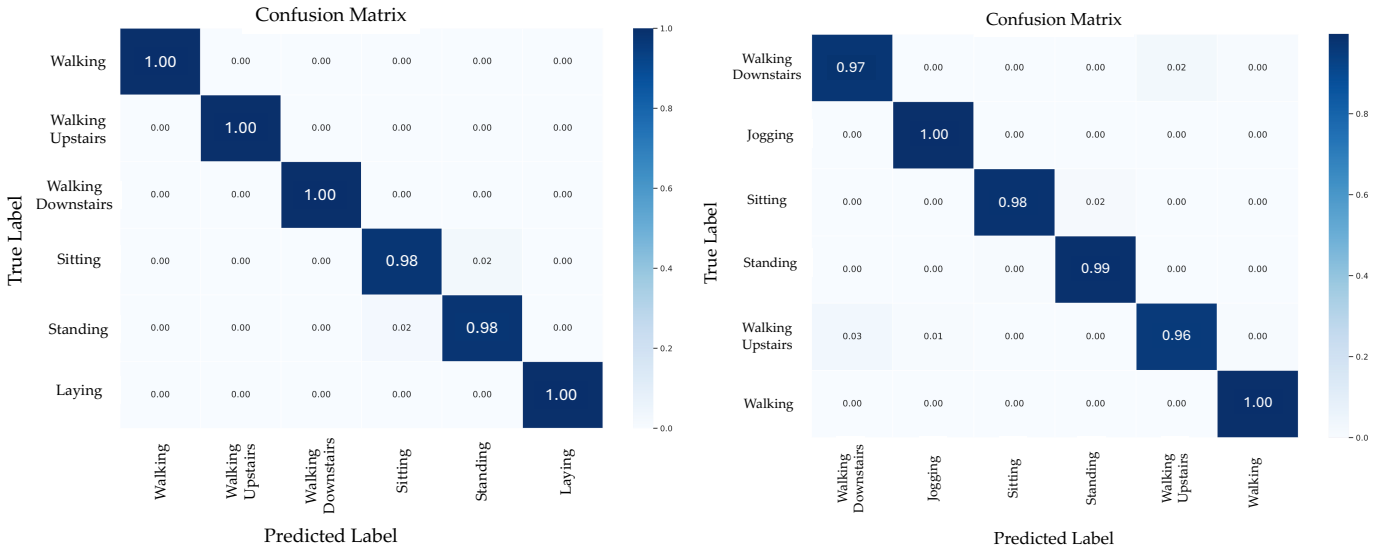
(**b**) WISDM dataset

Figure 9. Accuracy and loss curves of the proposed model using different datasets: (**a**) UCI-HAR (**b**) WISDM.

Comparable training dynamics are noted for the WISDM dataset (Figure 9b), albeit with some distinct features. The model converges around epoch 50, with the final accuracy stabilizing at about 98.7%. The slower convergence rate compared to the UCI-HAR dataset is likely due to the unique challenges of the WISDM dataset, such as its larger size and more varied activity patterns.

Our implementation of early stopping proved critical to model performance. While Figure 9 displays the entire 200-epoch trajectory for completeness, the vertical lines indicate where early stopping typically occurs (around epoch 75 for UCI-HAR and 50 for WISDM). This approach effectively prevented overfitting, as evidenced by the consistent accuracy of validation and loss curves following convergence. The difference in convergence speed between the two datasets highlights how our training strategy adapts to dataset complexity, terminating earlier for more straightforward recognition tasks while allowing more epochs for more challenging ones.

The confusion matrices in Figure 10 offer deeper insights into the model's classification performance. For the UCI-HAR dataset (Figure 10a), the model demonstrates excellent discrimination across all activity classes, with most diagonal elements exceeding 0.98. Notably, the model effectively distinguishes between similar activities like walking and walking upstairs, a common challenge in HAR systems.



(**a**) UCI-HAR dataset (**b**) WISDM dataset

Figure 10. Confusion matrices s of the proposed model using different datasets: (**a**) UCI-HAR (**b**) WISDM.

The results for the WISDM dataset (Figure 10b) show similarly strong performance, though with slightly more variation across activity classes. Static activities such as sitting and standing are recognized with near-perfect accuracy (>0.99). Dynamic activities have slightly lower but impressive recognition rates (>0.96). The minor confusion between walking and walking upstairs (around 0.02) reflects the inherent similarity in their sensor patterns.

The confusion matrices depicted in Figure 10 illustrate normalized values compiled from all five folds of our stratified cross-validation process. Each fold generates a distinct confusion matrix, and these five matrices were aggregated and normalized to create the final visualization. The aggregation and subsequent rounding of decimal values for display purposes may cause some rows to total slightly more or less than 100%. In the WISDM dataset results (Figure 10b), 99% of standing activities are accurately classified, whereas roughly 2% are erroneously classified as sitting. This apparent discrepancy (99% + 2% = 101%) is a byproduct of the rounding process applied to the normalized aggregated values from all five folds, where the actual decimal values total precisely 100%.

The consistently low standard deviations in our performance metrics (approximately ±0.2% for UCI-HAR and ±0.25% for WISDM) across multiple runs indicate the model's reliability and stability. This robust performance is due to several key architectural components: the SE mechanism's effective feature recalibration, the residual connections' optimization of gradient flow, and the bidirectional GRU's ability to capture temporal patterns.

5.2. Ablation Studies

The ablation study is now frequently employed in neural networks [38] to analyze a model's comprehension by modifying specific aspects [39]. Consequently, we examine the effects of ablation on our proposed model through three case studies. Altering various blocks and layers enables us to assess their influence on the exhibited structure [40]. Upon concluding each of the research circumstances, we can ascertain the optimal configuration of our ConvResBiGRU-SE model, attaining the highest possible identification efficacy.

5.2.1. Impact of Convolution Blocks

The first ablation study examines how the convolutional block affects the model's performance by comparing the entire ConvResBiGRU-SE architecture with the convolutional block and a version without it, as revealed in Table 5. The experimental outcomes highlight the convolutional block's significant contribution to the model's overall interpretation of both datasets.

For the UCI-HAR dataset, adding the convolutional block increased the model's accuracy from 97.29% to 99.18%, an absolute improvement of about 1.89%. Similarly, the F1-score improved from 97.34% to 99.24%. The relatively small standard deviations (±0.20% for accuracy) indicate consistent performance across different runs. This improvement suggests that the convolutional block effectively extracts relevant spatial features from the raw sensor data, providing a better foundation for subsequent layers.

In the case of the WISDM dataset, including the convolutional block led to an accuracy increase from 97.91% to 98.78% and an F1-score increase from 97.11% to 98.09%. Although the performance gain is smaller than on UCI-HAR, it remains significant and confirms the block's effectiveness across different datasets. The slightly more significant standard deviations on WISDM (±0.24% for accuracy) suggest more variability in the model's effectiveness, possibly due to the different characteristics and complexity of datasets.

Table 5. Impact of convolution blocks.

Model	Recognition Performance			
	UCI-HAR		WISDM	
	Accuracy	F1-Score	Accuracy	F1-Score
ConvResBiGRU-SE without the convolutional block	97.29% (±0.16%)	97.34% (±0.15%)	97.91% (±0.16%)	97.11% (±0.29%)
Our ConvResBiGRU-SE with the convolutional block	99.18% (±0.20%)	99.24% (±0.19%)	98.78% (±0.24%)	98.09% (±0.53%)

5.2.2. Impact of the SE-Residual Blocks

The second ablation study evaluates the effect of SE-Residual blocks on the model's effectiveness by comparing the ConvResBiGRU-SE architecture with a version that excludes these blocks, as shown in Table 6. The findings reveal that SE-Residual blocks are essential for enhancing the model's ability to determine human movements across both datasets.

For the UCI-HAR dataset, incorporating SE-Residual blocks significantly improved performance. Accuracy increased from 96.80% to 99.18%, an absolute gain of 2.39%. The F1-score also improved, rising from 96.97% to 99.24%. Additionally, the standard deviation for accuracy decreased from ±1.06% to ±0.20%, indicating that SE-Residual blocks boost performance and contribute to more stable and consistent predictions.

The impact on the WISDM dataset was similarly significant. The model with SE-Residual blocks achieved an accuracy of 98.78%, compared to 96.03% without them. The F1-score increased from 94.57% to 98.09%, showing enhanced balanced performance across all activity classes. The reduction in standard deviation from ±1.20% to ±0.24% further confirms the stabilizing effect of these blocks.

Table 6. Effect the SE-Residual blocks.

Model	Recognition Performance			
	UCI-HAR		WISDM	
	Accuracy	F1-Score	Accuracy	F1-Score
ConvResBiGRU-SE without the SE-Residual blocks	96.80% (±1.06%)	96.97% (±1.02%)	96.03% (±1.20%)	94.57% (±1.27%)
Our ConvResBiGRU-SE with the SE-Residual blocks	99.18% (±0.20%)	99.24% (±0.19%)	98.78% (±0.24%)	98.09% (±0.53%)

5.2.3. Impact of the SE Mechanism

The third ablation study evaluates the impact of the SE mechanism by comparing the full ConvResBiGRU-SE model with a version that excludes the SE mechanism. The findings in Table 7 show that the SE mechanism improves the model's capability to recalibrate features and support overall recognition performance.

For the UCI-HAR dataset, adding the SE mechanism increased the model's accuracy from 98.86% to 99.18%, an absolute improvement of 0.32%. The F1-score also rose from 98.93% to 99.24%. Although these improvements might seem modest, they are significant given the high baseline performance. More importantly, including the SE mechanism resulted in more stable predictions, as indicated by the reduction in standard deviation from ±0.54% to ±0.20% for accuracy.

The SE mechanism had a more pronounced effect on the WISDM dataset. Accuracy improved from 98.25% to 98.78%, and the F1-score increased from 97.29% to 98.09%. The substantial decrease in standard deviation from ±1.14% to ±0.24% for accuracy indicates that the SE mechanism significantly enhances the model's stability and reliability. This improvement is particularly valuable for real-world applications where consistent performance is crucial.

Table 7. Effect the SE mechanism.

Model	Recognition Performance			
	UCI-HAR		WISDM	
	Accuracy	F1-Score	Accuracy	F1-Score
ConvResBiGRU-SE without the SE mechanism	98.86% (±0.54%)	98.93% (±0.51%)	98.25% (±1.14%)	97.29% (±1.90%)
Our ConvResBiGRU-SE with the SE mechanism	99.18% (±0.20%)	99.24% (±0.19%)	98.78% (±0.24%)	98.09% (±0.53%)

5.2.4. Intrepretablility of the Proposed Model

We employed Grad-CAM [36] to enhance the interpretability and transparency of the model we developed. This approach enables us to interpret and pinpoint particular areas within the sensor data that substantially influence the forecasts made by the model. In time-series data, Grad-CAM emphasizes significant occurrences in the sequence that affect the network's classification selections. The visualization features a demonstration sequence with a colormap accentuating these significant areas [41]. In this research, we extracted and normalized the activation features produced by Grad-CAM to a scale of 0 to 1. Figures 11 and 12 illustrate the Grad-CAM activations for arbitrarily chosen portions from the UCI-HAR and WISDM datasets utilizing our ConvResBiGRU-SE model. Brighter shades denote elevated activation outcomes, indicating a more significant influence on the forecast.

As illustrated in Figure 11, Grad-CAM visualizations display distinct activation patterns for various activities for the UCI-HAR dataset. During walking movements (walking, stepping upstairs, and downstairs), the model emphasizes the periodic patterns in accelerometer and gyroscope signals, highlighted by bright regions in the activation maps. For static activities like sitting and standing, the activation patterns are more evenly distributed, indicating that the model recognizes the sustained nature of these postures. The laying activity exhibits minimal activations across the sensor channels, reflecting the low-intensity nature of this static position.

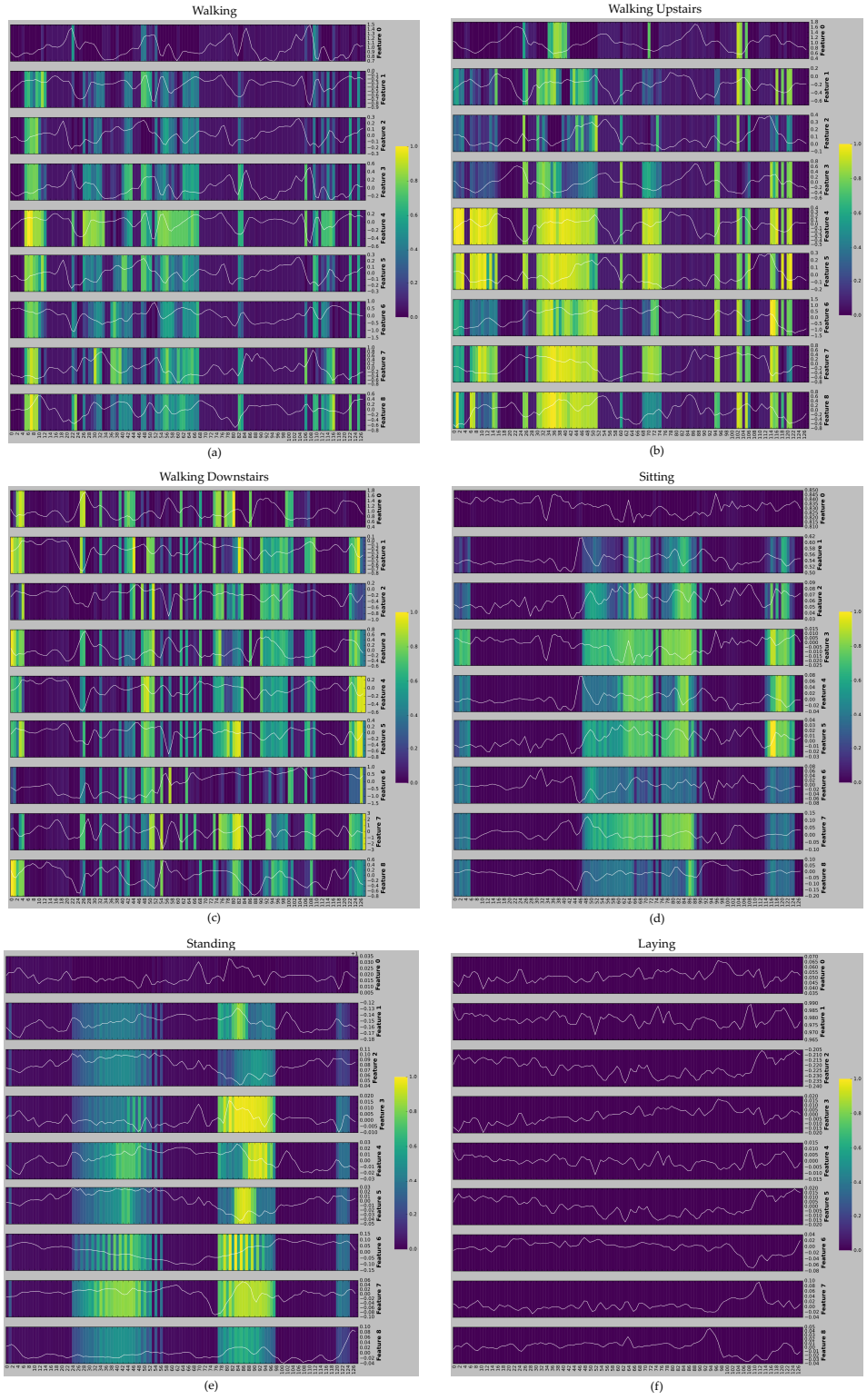


Figure 11. Visualization of the Grad-CAM method using UCI-HAR dataset: (**a**) Walking, (**b**) Walking upstairs, (**c**) Walking downstairs, (**d**) Sitting, (**e**) Standing, and (**f**) Laying. The white lines in the figures represent the original raw sensor signals from the accelerometer and gyroscope channels, allowing for direct visual comparison between the model's activation patterns (shown by the colored heatmap regions) and the underlying sensor data.

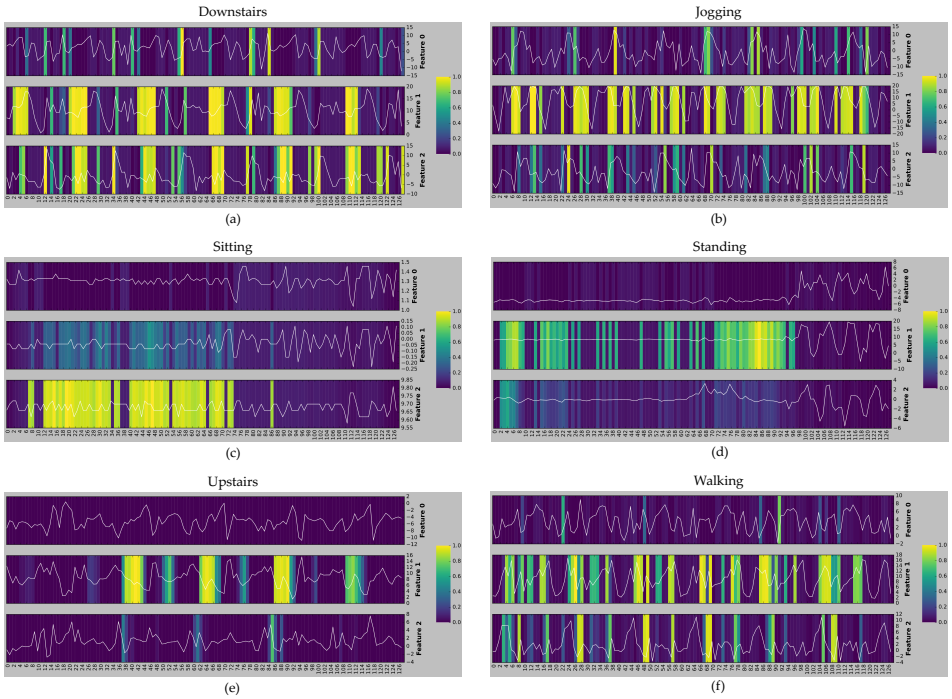


Figure 12. Visualization of the Grad-CAM method using WISDM dataset: (**a**) Downstairs, (**b**) Jogging, (**c**) Sitting, (**d**) Standing, (**e**) Upstairs, and (**f**) Walking. The white lines in the figures represent the original raw sensor signals from the accelerometer channels, allowing for direct visual comparison between the model's activation patterns (shown by the colored heatmap regions) and the underlying sensor data.

The results from the WISDM dataset (Figure 12) reveal similar interpretability patterns with some unique characteristics specific to the dataset. The jogging activity displays intense activations at regular intervals, reflecting the rhythmic nature of the movement. The activations are more concentrated in specific sensor channels for standing and sitting activities, indicating that the model distinguishes these static postures through subtle differences in sensor orientations.

5.3. Resource Utilization and Processing Speed

This study aimed to enhance recognition accuracy and model interpretability; however, computational efficiency is also vital for the practical implementation of HAR systems, especially on resource-limited wearable devices.

Our ConvResBiGRU-SE model, with its multiple specialized components, requires more computational resources than simpler architectures such as basic CNN or LSTM models. During our experiments on the Google Colab Pro+ environment with Tesla L4 GPU, the average training time per epoch was approximately 5.85 s for the UCI-HAR dataset and 6.39 s for the WISDM dataset. The inference time for a single activity sample was approximately 0.5352 ms for the UCI-HAR dataset and 0.5171 ms for the WISDM dataset, which is acceptable for most real-time applications.

The model size is approximately 1.1 MB, which might present challenges for deployment on extremely resource-limited devices without optimization. However, several approaches could mitigate these constraints, including model compression techniques (pruning, quantization), knowledge distillation to smaller models, or on-device optimization frameworks.

5.4. Comparison Results with State-of-the-Art Models

To thoroughly assess the effectiveness of our ConvResBiGRU-SE model, we perform extensive comparisons with state-of-the-art methods using the UCI-HAR and WISDM datasets. These comparisons cover a range of architectural paradigms, from traditional deep learning models to modern hybrid architectures, offering a comprehensive evaluation of our model's capabilities. The evaluation includes accuracy and detailed performance metrics to ensure a fair and thorough comparison.

We divide this analysis into two parts. First, we examine the results of the UCI-HAR dataset, a standard benchmark in the field. Then, we evaluate the performance of the WISDM dataset, which presents unique challenges due to its distinct data collection methods and activity patterns. Through these comparisons, we aim to highlight the advantages of our proposed architecture and its contributions to advancing human activity recognition.

5.4.1. Comparison Results from UCI-HAR Dataset

To confirm the effectiveness of our ConvResBiGRU-SE model, we compare its performance with several state-of-the-art methods on the UCI-HAR dataset. Table 8 shows the comparative analysis, which includes various deep learning architectures. These methods encompass different strategies for HAR, from traditional convolutional and recurrent architectures to more advanced hybrid models.

Table 8. Performance comparison of state-of-the-art model on the UCI-HAR dataset.

Reference	Year of Publication	Model	Accuracy (%)
Zhao et al. [42]	2018	Residual BiLSTM	93.60%
Xia et al. [43]	2020	LSTM-CNN	95.78%
Wang and Lie [44]	2020	Hierarchical Deep LSTM	91.65%
Cruciani et al. [45]	2020	CNN	91.98%
Ronald et al. [46]	2021	iSPLInception	95.09%
Bhattacharya et al. [47]	2022	Ens-HAR	95.05%
The proposed model	-	ConvResBiGRU-SE	99.18%

Table 8 compares the proposed ConvResBiGRU-SE model with various prominent HAR methodologies utilizing the UCI-HAR dataset. The results underscore the remarkable efficacy of our model, attaining an accuracy of 99.18%, substantially surpassing current methodologies.

Historically, earlier methods, such as Zhao et al.'s Residual BiLSTM [42] and Cruciani et al.'s CNN [45], achieved accuracies of 93.6% and 91.98%, respectively. More recent advancements, including Ronald et al.'s iSPLInception [46] and Bhattacharya et al.'s Ens-HAR [47], showed improved accuracies of 95.09% and 95.05%, respectively. This steady increase in accuracy over the years indicates continuous progress in HAR model architectures.

When evaluating different architectural strategies, simpler models like CNN and Hierarchical Deep LSTM had relatively modest accuracies of 91.98% and 91.65%, respectively. Hybrid architectures, such as Xia et al.'s LSTM-CNN model [43], performed better with an accuracy of 95.78%, suggesting the advantages of combining various neural network components. Nonetheless, our proposed ConvResBiGRU-SE model surpasses these hybrid approaches by over 3%.

5.4.2. Comparison Results from WISDM Dataset

We perform comparison research using the WISDM dataset to evaluate the generalization capability and robustness of the ConvResBiGRU-SE model. Table 9 displays this comparative analysis, which includes several deep learning architectures.

Table 9. Performance comparison of state-of-the-art model on the WISDM dataset.

Reference	Year of Publication	Model	Accuracy (%)
Zhang et al. [48]	2018	U-Net	97.00%
Quispe et al. [49]	2018	KNN	96.20%
Pienaar et al. [50]	2020	RNN-LSTM	94.00%
Bhattacharya et al [47]	2022	Ens-HAR	98.70%
The proposed model	-	ConvResBiGRU-SE	98.78%

Table 9 provides a comparative analysis of the ConvResBiGRU-SE model we propose against several leading methods for human activity recognition on the WISDM dataset. The proposed model attains an accuracy of 98.78%, showcasing competitive performance and slightly surpassing existing methods.

Examining the chronological development of HAR models, Zhang et al.'s U-Net architecture [48] set a strong baseline with 97% accuracy, while Quispe et al.'s KNN approach [49] achieved 96.20% accuracy. The evolution continued with Pienaar et al.'s RNN-LSTM model [50], which showed relatively lower performance at 94%. This indicates that not all architectural advancements necessarily lead to improved results on this specific dataset.

A notable recent development is Bhattacharya et al.'s Ens-HAR [47], which achieved an impressive accuracy of 98.70%. Our ConvResBiGRU-SE model builds upon this progress, achieving a marginal improvement with 98.78% accuracy. Although this 0.08% improvement may seem modest, it represents a significant advancement in a field where performance gains become increasingly challenging to achieve at higher accuracy levels.

6. Conclusions and Future Works

This paper introduces ConvResBiGRU-SE, an inventive deep-learning architecture for HAR employing wearable sensor data. The proposed model effectively integrates CNNS, residual connections, BiGRU, and SE mechanisms to attain state-of-the-art implementation on standard HAR benchmarks. Our experimental results highlight several significant contributions to the field: (1) ConvResBiGRU-SE model achieves superior recognition accuracy compared to existing methods, reaching 99.18% accuracy on the UCI-HAR dataset and 98.78% on the WISDM dataset. These results represent substantial advancements over previous state-of-the-art approaches, demonstrating the efficacy of our architectural design choices; (2). Our comprehensive ablation studies validate the significance of each component in the architecture. The convolutional blocks enhance feature extraction capabilities, while the SE-Residual blocks enhance feature recalibration and model stability. Integrating bidirectional GRU units captures temporal dependencies in sensor data, leading to more robust activity recognition; and (3). We address the critical challenge of model interpretability in HAR systems by integrating Grad-CAM visualization techniques. These visualizations deliver insights into the decision-making procedure of the model, revealing how different sensor channels and temporal regions contribute to activity classification. This advancement towards explainable HAR is essential for creating confidence in real-world applications.

Future research should optimize the model for real-time use on resource-constrained devices through compression and quantization. Transfer learning can enhance cross-dataset generalization, and personalization strategies can tailor the model for individual users while ensuring privacy and efficiency. Additionally, extending the framework to recognize complex, long-duration activities, incorporating more sensor modalities, and developing advanced visualization methods will improve interpretability. Energy efficiency is also crucial, balancing accuracy and power consumption for wearable devices.

Author Contributions: Conceptualization, S.M. and A.J.; methodology, S.M.; software, A.J.; validation, A.J.; formal analysis, S.M.; investigation, S.M.; resources, A.J.; data curation, A.J.; writing—original draft preparation, S.M.; writing—review and editing, A.J.; visualization, S.M.; supervision, A.J.; project administration, A.J.; funding acquisition, S.M. and A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research budget was allocated by the University of Phayao; the Thailand Science Research and Innovation Fund (Fundamental Fund 2025, Grant No. 5014/2567); National Science, Research and Innovation Fund (NSRF); and King Mongkut's University of Technology North Bangkok (Project no. KMUTNB-FF-68-B-02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: To clarify, our research utilizes a pre-existing, publicly available dataset. The dataset has been anonymized and does not contain any personally identifiable information. We have cited the source of the dataset in our manuscript and have complied with the terms of use set forth by the dataset provider.

Data Availability Statement: The original data presented in the study are openly available for the UCI-HAR dataset at https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones, (accessed on 28 November 2024) and the WISDM dataset at https://www.cis.fordham.edu/wisdm/dataset.php, (accessed on 28 November 2024).

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

References

1. Sannasi Chakravarthy, S.R.; Bharanidharan, N.; Vinoth Kumar, V.; Mahesh, T.R.; Khan, S.B.; Almusharraf, A.; Albalawi, E. Intelligent Recognition of Multimodal Human Activities for Personal Healthcare. *IEEE Access* **2024**, *12*, 79776–79786. [CrossRef]
2. Mekruksavanich, S.; Jitpattanakul, A. LSTM Networks Using Smartphone Data for Sensor-Based Human Activity Recognition in Smart Homes. *Sensors* **2021**, *21*, 1636. [CrossRef] [PubMed]
3. Mekruksavanich, S.; Jitpattanakul, A. A Residual Deep Learning Method for Accurate and Efficient Recognition of Gym Exercise Activities Using Electromyography and IMU Sensors. *Appl. Syst. Innov.* **2024**, *7*, 59. [CrossRef]
4. Kulsoom, F.; Narejo, S.; Mehmood, Z.; Chaudhry, H.N.; Butt, A.; Bashir, A.K. A review of machine learning-based human activity recognition for diverse applications. *Neural Comput. Appl.* **2022**, *34*, 18289–18324. [CrossRef]
5. Kaseris, M.; Kostavelis, I.; Malassiotis, S. A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 842–876. [CrossRef]
6. Aquino, G.; Costa, M.G.F.; Filho, C.F.F.C. Explaining and Visualizing Embeddings of One-Dimensional Convolutional Models in Human Activity Recognition Tasks. *Sensors* **2023**, *23*, 4409. [CrossRef]
7. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
8. Mekruksavanich, S.; Jitpattanakul, A. Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition. *Sci. Rep.* **2023**, *13*, 12067. [CrossRef]
9. Ramanujam, E.; Perumal, T.; Padmavathi, S. Human Activity Recognition With Smartphone and Wearable Sensors Using Deep Learning Techniques: A Review. *IEEE Sens. J.* **2021**, *21*, 13029–13040. [CrossRef]
10. Serpush, F.; Menhaj, M.B.; Masoumi, B.; Karasfi, B. Wearable Sensor-Based Human Activity Recognition in the Smart Healthcare System. *Comput. Intell. Neurosci.* **2022**, *2022*, 1391906. [CrossRef]
11. Baldominos, A.; Cervantes, A.; Saez, Y.; Isasi, P. A Comparison of Machine Learning and Deep Learning Techniques for Activity Recognition using Mobile Devices. *Sensors* **2019**, *19*, 521. [CrossRef]
12. Aquino, G.; Costa, M.G.F.; Costa Filho, C.F.F. Explaining One-Dimensional Convolutional Models in Human Activity Recognition and Biometric Identification Tasks. *Sensors* **2022**, *22*, 5644. [CrossRef] [PubMed]
13. Mekruksavanich, S.; Jitpattanakul, A. Deep Convolutional Neural Network with RNNs for Complex Activity Recognition Using Wrist-Worn Wearable Sensor Data. *Electronics* **2021**, *10*, 1685. [CrossRef]
14. Ordóñez, F.J.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* **2016**, *16*, 115. [CrossRef] [PubMed]
15. Mekruksavanich, S.; Jitpattanakul, A. Biometric User Identification Based on Human Activity Recognition Using Wearable Sensors: An Experiment Using Deep Learning Models. *Electronics* **2021**, *10*, 308. [CrossRef]

16. Challa, S.K.; Kumar, A.; Semwal, V.B. A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data. *Vis. Comput.* **2022**, *38*, 4095–4109. [CrossRef]
17. Canizo, M.; Triguero, I.; Conde, A.; Onieva, E. Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. *Neurocomputing* **2019**, *363*, 246–260. [CrossRef]
18. Dua, N.; Singh, S.N.; Semwal, V.B. Multi-input CNN-GRU based human activity recognition using wearable sensors. *Computing* **2021**, *103*, 1461–1478. [CrossRef]
19. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. https://doi.orgttps://doi.org/10.1016/j.neucom.2021.03.091. [CrossRef]
20. Zhongkai, Z.; Kobayashi, S.; Kondo, K.; Hasegawa, T.; Koshino, M. A Comparative Study: Toward an Effective Convolutional Neural Network Architecture for Sensor-Based Human Activity Recognition. *IEEE Access* **2022**, *10*, 20547–20558. [CrossRef]
21. Mekruksavanich, S.; Hnoohom, N.; Jitpattanakul, A. A Hybrid Deep Residual Network for Efficient Transitional Activity Recognition Based on Wearable Sensors. *Appl. Sci.* **2022**, *12*, 4988. [CrossRef]
22. Khan, Z.N.; Ahmad, J. Attention induced multi-head convolutional neural network for human activity recognition. *Appl. Soft Comput.* **2021**, *110*, 107671. [CrossRef]
23. Anguita, D.; Ghio, A.; Oneto, L.; Parra Perez, X.; Reyes Ortiz, J.L. A public domain dataset for human activity recognition using smartphones. In Proceedings of the 21th International European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 6–8 October 2013; pp. 437–442.
24. Kwapisz, J.R.; Weiss, G.M.; Moore, S.A. Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.* **2011**, *12*, 74–82. [CrossRef]
25. Lin, L.; Wu, J.; An, R.; Ma, S.; Zhao, K.; Ding, H. LIMUNet: A Lightweight Neural Network for Human Activity Recognition Using Smartwatches. *Appl. Sci.* **2024**, *14*, 10515. [CrossRef]
26. Pires, I.M.; Hussain, F.; Garcia, M.N.; Lameski, P.; Zdravevski, E. Homogeneous Data Normalization and Deep Learning: A Case Study in Human Activity Classification. *Future Internet* **2020**, *12*, 194. [CrossRef]
27. Banos, O.; Galvez, J.M.; Damas, M.; Pomares, H.; Rojas, I. Window Size Impact in Human Activity Recognition. *Sensors* **2014**, *14*, 6474–6499. [CrossRef]
28. Cho, K.; van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In Proceedings of the SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014; Wu, D., Carpuat, M., Carreras, X., Vecchi, E.M., Eds.; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 103–111. [CrossRef]
29. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555. [CrossRef]
30. Wang, X.; Shang, J. Human Activity Recognition Based on Two-Channel Residual–GRU–ECA Module with Two Types of Sensors. *Electronics* **2023**, *12*, 1622. [CrossRef]
31. Schuster, M.; Paliwal, K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
32. Muqeet, A.; Iqbal, M.T.B.; Bae, S.H. HRAN: Hybrid Residual Attention Network for Single Image Super-Resolution. *IEEE Access* **2019**, *7*, 137020–137029. [CrossRef]
33. Lin, M.; Chen, Q.; Yan, S. Network in Network. *arXiv* **2014**. [CrossRef]
34. Rundo, L.; Han, C.; Nagano, Y.; Zhang, J.; Hataya, R.; Militello, C.; Tangherloni, A.; Nobile, M.S.; Ferretti, C.; Besozzi, D.; et al. USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing* **2019**, *365*, 31–43. [CrossRef]
35. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
36. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
37. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847. [CrossRef]
38. Montaha, S.; Azam, S.; Rafid, A.K.M.R.H.; Ghosh, P.; Hasan, M.Z.; Jonkman, M.; De Boer, F. BreastNet18: A High Accuracy Fine-Tuned VGG16 Model Evaluated Using Ablation Study for Diagnosing Breast Cancer from Enhanced Mammography Images. *Biology* **2021**, *10*, 1347. [CrossRef] [PubMed]
39. de Vente, C.; Boulogne, L.H.; Venkadesh, K.V.; Sital, C.; Lessmann, N.; Jacobs, C.; Sánchez, C.I.; van Ginneken, B. Improving Automated COVID-19 Grading with Convolutional Neural Networks in Computed Tomography Scans: An Ablation Study. *arXiv* **2020**. [CrossRef]

40. Meyes, R.; Lu, M.; de Puiseau, C.W.; Meisen, T. Ablation Studies in Artificial Neural Networks. *arXiv* **2019**. [CrossRef]
41. Choi, H.; Jung, C.; Kang, T.; Kim, H.J.; Kwak, I.Y. Explainable Time-Series Prediction Using a Residual Network and Gradient-Based Methods. *IEEE Access* **2022**, *10*, 108469–108482. [CrossRef]
42. Zhao, Y.; Yang, R.; Chevalier, G.; Xu, X.; Zhang, Z. Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors. *Math. Probl. Eng.* **2018**, *2018*, 7316954. [CrossRef]
43. Xia, K.; Huang, J.; Wang, H. LSTM-CNN Architecture for Human Activity Recognition. *IEEE Access* **2020**, *8*, 56855–56866. [CrossRef]
44. Wang, L.; Liu, R. Human Activity Recognition Based on Wearable Sensor Using Hierarchical Deep LSTM Networks. *Circuits Syst. Signal Process.* **2020**, *39*, 837–856. [CrossRef]
45. Cruciani, F.; Vafeiadis, A.; Nugent, C.; Cleland, I.; McCullagh, P.; Votis, K.; Giakoumis, D.; Tzovaras, D.; Chen, L.; Hamzaoui, R. Feature learning for Human Activity Recognition using Convolutional Neural Networks. *CCF Trans. Pervasive Comput. Interact.* **2020**, *2*, 18–32. [CrossRef]
46. Ronald, M.; Poulose, A.; Han, D.S. iSPLInception: An Inception-ResNet Deep Learning Architecture for Human Activity Recognition. *IEEE Access* **2021**, *9*, 68985–69001. [CrossRef]
47. Bhattacharya, D.; Sharma, D.; Kim, W.; Ijaz, M.F.; Singh, P.K. Ensem-HAR: An Ensemble Deep Learning Model for Smartphone Sensor-Based Human Activity Recognition for Measurement of Elderly Health Monitoring. *Biosensors* **2022**, *12*, 393. [CrossRef] [PubMed]
48. Zhang, Y.; Zhang, Y.; Zhang, Z.; Bao, J.; Song, Y. Human activity recognition based on time series analysis using U-Net. *arXiv* **2018**. [CrossRef]
49. Montero Quispe, K.G.; Sousa Lima, W.; Macêdo Batista, D.; Souto, E. MBOSS: A Symbolic Representation of Human Activity Recognition Using Mobile Sensors. *Sensors* **2018**, *18*, 4354. [CrossRef]
50. Pienaar, S.W.; Malekian, R. Human Activity Recognition using LSTM-RNN Deep Neural Network Architecture. In Proceedings of the 2019 IEEE 2nd Wireless Africa Conference (WAC), Pretoria, South Africa, 18–20 August 2019; pp. 1–5. [CrossRef]