Multimodal Technologies and Interaction

MDPI

Article

Multimodal Smart-Skin for Real-Time Sitting Posture Recognition with Cross-Session Validation

Giva Andriana Mutiara [1,*], Muhammad Rizqy Alfarisi [1], Paramita Mayadewi [2], Lisda Meisaroh [1] and Periyadi [1]

[1] Technology Computer Program Study, Department of Applied Sciences, Telkom University, Bandung 40257, Indonesia; mrizkyalfarisi@telkomuniversity.ac.id (M.R.A.); lisdameisaroh@telkomuniversity.ac.id (L.M.); periyadi@telkomuniversity.ac.id (P.)

[2] Information System Program Study, Department of Applied Sciences, Telkom University, Bandung 40257, Indonesia; paramita@telkomuniversity.ac.id

* Correspondence: givamz@telkomuniversity.ac.id

Abstract

Prolonged sitting with poor posture is associated with musculoskeletal disorders, reduced productivity, and long-term health risks. Many existing posture monitoring systems predominantly rely on single-modality sensing, such as pressure or vision-based approaches, limiting their ability to capture both static alignment and dynamic micromovements. This study proposes a multimodal smart-skin system integrating pressure, temperature, and vibration sensors for sitting posture recognition. A total of 42 sensors distributed across 14 anatomical locations were deployed, generating 15,037 samples collected over three independent sessions to evaluate cross-session temporal generalization across nine posture classes under controlled experimental conditions. Two deep learning architectures—Temporal Convolutional Networks with Attention (TCN + Attn) and Convolutional Neural Network–Long Short-Term Memory (CNN − LSTM)—were compared under Leave-One-Session-Out (LOSO) cross-validation. TCN + Attn achieved 85.23% LOSO accuracy, outperforming CNN − LSTM by 2.56 percentage points while reducing training time by 36.7% and inference latency by 33.9%. Ablation analysis revealed that temperature sensing was the most discriminative unimodal modality (71.5% accuracy), and full multimodal fusion improved LOSO accuracy by 22.93% compared to pressure-only configurations. These results demonstrate the feasibility of multimodal smart-skin sensing combined with temporal convolutional modeling for cross-session posture recognition and indicate potential for efficient real-time, privacy-preserving ergonomic monitoring. This study should be interpreted as a controlled, single-subject proof-of-concept, and further validation in multi-subject and real-world environments is required to establish broader generalizability.

Keywords: sitting posture classification; multimodal sensing; smart-skin systems; sensor fusion; cross-session validation; temporal convolutional networks; human–computer interaction

Check for updates

Academic Editor: Stephan Schlögl

Received: 18 February 2026
Revised: 2 April 2026
Accepted: 5 April 2026
Published: 9 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Sedentary behavior has become a defining characteristic of modern work environments, with office workers spending an average of 8–12 h per day in seated positions [1]. Prolonged sitting, particularly with poor posture, is associated with increased risks of musculoskeletal disorders, cardiovascular disease, metabolic syndrome, and reduced cognitive

performance [2]. The economic burden of work-related musculoskeletal disorders exceeds USD 50 billion annually in the United States alone, with lower back pain accounting for the majority of cases [3]. Despite growing awareness of these health risks, real-time posture monitoring technologies remain largely absent from everyday work environments due to technical and practical limitations in existing systems.

Traditional posture monitoring approaches predominantly rely on vision-based systems using RGB or depth cameras [4]. Although these methods can achieve high classification accuracy in controlled laboratory settings, they raise significant privacy concerns due to continuous video capture and are sensitive to lighting variations, occlusions, and environmental constraints [5]. Moreover, camera-based systems cannot directly capture pressure distributions or thermal contact patterns that reflect sustained body–chair interaction and emerging discomfort. To address these limitations, pressure-based sensing systems embedded into chair surfaces have been proposed as unobtrusive and privacy-preserving alternatives [6]. Such systems eliminate environmental dependency and enable continuous monitoring without external imaging devices.

However, most chair-based posture classification systems rely on single-modality sensing, typically employing pressure sensor arrays with 16–256 sensing points [7,8]. While pressure mapping effectively captures static posture configurations, it provides limited information regarding dynamic behaviors such as fidgeting, micro-movements, and periodic weight shifting—key indicators of discomfort and fatigue [9]. Furthermore, pressure-only systems often struggle to distinguish between postures that differ primarily in contact duration or subtle thermal signatures. Recent multimodal approaches have combined pressure with temperature sensing [10] or inertial data [11], demonstrating improved classification performance. Nevertheless, vibration sensing—capable of capturing micro-movements and dynamic postural adjustments—remains relatively underexplored. Although some studies have investigated pressure–vibration fusion, comprehensive integration of pressure, temperature, and vibration sensing for posture recognition remains limited in the literature [12–14].

Based on an algorithmic perspective, many existing systems employ traditional machine learning methods, such as Support Vector Machines, Random Forests, and k-Nearest Neighbors, which treat each time step independently and ignore temporal dependencies in posture sequences [15,16]. While recent deep learning approaches using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have improved classification accuracy, they may struggle to efficiently capture long-range temporal dependencies and often incur higher computational complexity, limiting suitability for embedded deployment [17,18]. Temporal Convolutional Networks (TCNs) provide an alternative sequence modeling paradigm through dilated causal convolutions, enabling large receptive fields and parallelizable training [19]. However, their application to multimodal posture classification remains limited.

A critical methodological limitation in the literature concerns evaluation protocols. The majority of posture classification studies report within-session or random cross-validation results, which may overestimate performance and fail to reflect real-world deployment scenarios where models must generalize across temporal and environmental variations [20,21]. Only a limited number of studies have implemented Leave-One-Session-Out (LOSO) validation, and few have systematically compared TCN and CNN − LSTM architectures under this rigorous cross-session protocol [22]. In addition, comprehensive ablation analyses quantifying the contribution of individual sensing modalities and their combinations are relatively scarce, limiting understanding of optimal sensor configurations for practical deployment [23,24].

To address these challenges, this study proposes a multimodal smart-skin system integrating pressure, temperature, and vibration sensors for real-time sitting posture recognition. In this study, smart-skin refers to a distributed body-attached multimodal sensing layer composed of pressure, temperature, and vibration sensors designed to capture human–surface interaction dynamics. The system deploys 42 sensors across 14 anatomical locations to capture complementary spatial, thermal, and dynamic interaction patterns between the human body and the chair surface. The proposed framework systematically compares a Temporal Convolutional Network with Attention (TCN + Attn) and a CNN − LSTM architecture under rigorous Leave-One-Session-Out validation across three independent sessions comprising 15,037 samples. Furthermore, this study conducts comprehensive ablation experiments across seven modality configurations to quantify the individual and combined contributions of pressure, temperature, and vibration sensing. Beyond classification accuracy, the proposed system evaluates computational efficiency in terms of training time, inference latency, memory footprint, and suitability for embedded real-time deployment.

Accordingly, the principal contributions of this research can be summarized as follows. First, it establishes a unified multimodal sensing framework integrating pressure, temperature, and vibration modalities within a smart-skin architecture for posture recognition. Second, it provides cross-session evaluation of temporal modeling strategies, highlighting the generalization capability and efficiency advantages of TCN-based approaches over recurrent architectures. Third, it offers modality-level insights through systematic ablation analysis, clarifying the relative and complementary roles of each sensing component. Finally, it demonstrates practical feasibility for real-time, privacy-preserving ergonomic monitoring through detailed computational profiling and deployment-oriented evaluation.

The remainder of this paper is organized as follows. Section 2 reviews related work in multimodal posture recognition and temporal modeling. Section 3 describes the proposed smart-skin architecture, data acquisition protocol, and learning framework. Section 4 presents experimental results and ablation findings. Section 5 discusses implications, deployment considerations, and limitations. Section 6 concludes the paper and outlines directions for future research.

2. Related Work

Sitting posture recognition has attracted increasing attention in human–computer interaction, ergonomics, and health monitoring research due to the growing prevalence of sedentary lifestyles [1]. Early systems primarily relied on vision-based approaches using RGB or depth cameras [2]. Although capable of achieving high recognition accuracy under controlled laboratory conditions, these methods raise significant privacy concerns and are sensitive to lighting variations, occlusion, and environmental constraints [5]. Consequently, research has progressively shifted toward unobtrusive embedded sensing solutions that directly measure human–chair interaction without relying on external imaging devices.

2.1. Pressure-Based and Multimodal Posture Recognition

Pressure mapping systems represent the most widely adopted approach for chair-based posture classification. High-resolution pressure arrays with up to 256 sensing points have achieved classification accuracies exceeding 90% using traditional machine learning algorithms such as Support Vector Machines and Random Forests [3,4]. Commercial pressure mats have also demonstrated reliable real-time feedback in clinical and ergonomic applications [5], while lower-resolution arrays indicate that acceptable accuracy can be maintained with reduced hardware complexity [6]. However, most pressure-based systems primarily capture static posture configurations and often lack explicit temporal modeling, limiting

their ability to characterize dynamic behaviors such as fidgeting, micro-adjustments, or periodic weight shifting.

To enhance recognition performance, researchers have increasingly explored multimodal sensing strategies. Integrating pressure with temperature sensors has been shown to improve detection of prolonged contact patterns and posture transitions [7,9]. Similarly, combining pressure with inertial measurements or accelerometers provides additional information about body movement and posture dynamics [8,11]. Hybrid systems incorporating pressure, temperature, and capacitive sensing have reported improved robustness, albeit often at the expense of increased hardware complexity [16,19]. Despite these advances, vibration sensing remains comparatively underutilized in posture classification. While vibration sensors have demonstrated effectiveness in capturing micro-movements and restless sitting behaviors [11,12], systematic integration of vibration with pressure and temperature within a unified multimodal framework remains limited.

2.2. Smart-Skin and Embedded Sensing Architectures

The concept of smart-skin refers to distributed sensing networks embedded within surfaces to monitor human–object interaction [13]. In seated environments, such systems enable continuous, non-intrusive posture monitoring without requiring wearable devices. Flexible and stretchable sensor arrays based on conductive fabrics or liquid metal materials have enabled conformal integration into chair surfaces [14,15]. Hybrid sensor arrays combining piezoresistive pressure sensors and thermistors have further demonstrated improved posture discrimination and emphasized the importance of optimized sensor placement [17].

Distributed multimodal architectures have also been explored to balance sensing resolution and computational efficiency. Systems incorporating pressure and temperature arrays with decision-tree classifiers have achieved moderate classification accuracy, although they may incur increased computational overhead as sensor density grows [18]. Modular smart-skin systems with reconfigurable sensor layouts have been proposed to reduce hardware complexity while maintaining acceptable performance [19]. Nevertheless, many existing designs prioritize spatial sensing density over temporal modeling capability and rarely provide systematic analysis of the relative contribution of each sensing modality.

2.3. Deep Learning for Temporal Posture Modeling

Traditional machine learning approaches treat posture recognition as a static classification problem, processing each time step independently and neglecting temporal dependencies [15,16]. Deep learning methods have addressed this limitation by enabling temporal modeling of posture transitions and dynamic behaviors.

Convolutional Neural Networks (CNNs) have been applied to pressure maps by interpreting them as spatial images, achieving competitive accuracy but often requiring high-resolution inputs [25]. One-dimensional CNN variants reduce computational cost while maintaining reasonable performance [26], and multi-scale CNN architectures have demonstrated advantages in multimodal fusion scenarios [22]. However, CNN-based approaches primarily capture short-range temporal dependencies unless explicitly combined with sequential modeling mechanisms.

Recurrent architectures, particularly Long Short-Term Memory (LSTM) networks, have been widely adopted for temporal sensor data in human activity recognition [23]. Bidirectional and stacked LSTM models have been applied to posture classification, enabling sequence modeling and improved detection of posture transitions [23,24]. Hybrid $\mathrm{CNN}-\mathrm{LSTM}$ frameworks combine spatial feature extraction with temporal sequence modeling and have demonstrated improved performance compared to standalone net-

works [24,27–29]. Despite these benefits, LSTM-based models may incur higher computational cost and can face challenges in efficiently modeling long-range dependencies, especially under resource-constrained deployment scenarios.

Temporal Convolutional Networks (TCNs) offer an alternative sequence modeling architecture based on dilated causal convolutions [30,31]. TCNs have demonstrated competitive or superior performance compared to recurrent networks in activity recognition tasks while enabling parallelizable training and improved computational efficiency [32,33]. However, applications of TCNs to posture recognition remain relatively limited, and prior studies typically focus on pressure-only inputs without systematically exploring multimodal integration [34].

2.4. Validation Protocols and Generalization Challenges

Beyond sensing and modeling strategies, evaluation methodology plays a crucial role in determining real-world applicability. Many posture recognition studies report within-session or random cross-validation results, which may overestimate generalization performance under stable experimental conditions [35]. In contrast, cross-subject validation using Leave-One-Subject-Out (LOSO) protocols often reveals performance degradation due to anthropometric variability and inter-individual differences [36–39]. Similarly, Leave-One-Session-Out validation can expose sensitivity to temporal and contextual variations that may not be captured under within-session evaluation [40–42]. Despite its importance for realistic deployment assessment, session-wise validation remains comparatively underreported in posture classification research. Moreover, systematic comparative evaluation of temporal modeling architectures under rigorous cross-session protocols remains limited. Table 1 shows the representative comparison of sitting posture recognitions system.

Table 1. Representative comparison of sitting posture recognition systems.

Study	Sensing Modalities	Temporal Architecture	Evaluation Protocol	Modality-Level Ablation
Pressure-based ML systems [3,4,6]	Pressure	SVM/Random Forest (static modeling)	Random or Within-session CV	No
Pressure–temperature systems [7,9]	Pressure + Temperature	CNN/Traditional ML	Within-session	Limited
Pressure–inertial systems [8,11]	Pressure + IMU	CNN/LSTM	Random CV	No
Hybrid multimodal smart-skin systems [16,19]	Pressure + Temperature + Capacitive	ML/CNN	Within-session	No
CNN–LSTM posture models [24,27–29]	Predominantly Pressure	CNN–LSTM	Random or Within-session CV	Partial
TCN-based activity models [32–34]	Pressure/ HAR datasets	TCN	Random or Cross-subject	No (modality-focused)

2.5. Research Gaps and Study Positioning

Although prior research has advanced pressure-based and multimodal posture recognition, several limitations remain. First, vibration sensing remains insufficiently integrated into multimodal chair-based systems, and comprehensive pressure–temperature–vibration frameworks have not been systematically examined within a unified architecture. Second, temporal modeling approaches for dynamic postures frequently rely on CNN − LSTM hybrids, while TCN-based architectures for multimodal posture recognition remain relatively underexplored. Third, many studies primarily adopt within-session validation strategies, limiting insight into cross-session robustness and temporal generalization. Finally, compre-

hensive ablation analyses that quantify the relative and complementary contributions of individual sensing modalities are still relatively scarce.

This study addresses these gaps by integrating pressure, temperature, and vibration sensing within a unified smart-skin architecture, systematically comparing TCN + Attn and CNN − LSTM models under rigorous Leave-One-Session-Out validation, and conducting detailed modality-aware ablation analysis. By combining multimodal sensing, temporal convolutional modeling, and cross-session evaluation, this work enhances both methodological rigor and practical deployment considerations for privacy-preserving ergonomic monitoring.

3. Materials and Methods

This section describes the design and implementation of the proposed multimodal smart-skin framework, including sensor configuration, data acquisition protocol, preprocessing pipeline, and deep learning architectures, as illustrated in Figure 1. The methodology is structured to ensure reproducibility and rigorous evaluation of temporal generalization. First, the anatomical sensor placement and hardware integration strategy are detailed. Next, the data collection protocol and posture class definitions are presented, followed by signal preprocessing and temporal windowing procedures. Finally, the comparative deep learning architectures and validation strategies employed to assess classification performance and cross-session robustness are described.

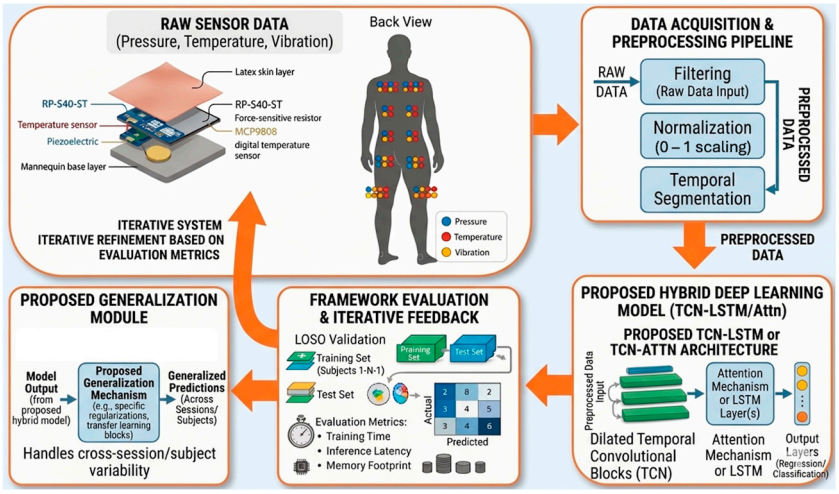


Figure 1. The architecture of the proposed multimodal smart-skin framework.

3.1. Multimodal Smart-Skin System

This section describes the sensor configuration, anatomical placement, and specifications of the proposed multimodal smart-skin system. The system consists of 42 sensing units distributed across 14 anatomically significant body locations, with each location instrumented using three sensing modalities: pressure, temperature, and vibration. In this study, the term "smart-skin" refers to a distributed, body-attached multimodal sensing layer designed to capture localized biomechanical and thermal interactions during seated postures. Unlike chair-embedded systems, the sensors are directly attached to the human body to capture localized biomechanical responses during seated postures. The selected anatomical locations correspond to primary load-bearing and contact regions identified in the ergonomic and posture analysis literature [17,43]. Figure 1 illustrates the architecture of the proposed multimodal smart-skin framework.

The 14 anatomical locations were organized into two main regions. The posterior trunk region consisted of eight locations (24 sensors in total), including the upper thoracic

region (left, center, right), mid-thoracic region (left, center, right), and lumbar region (left, right). These regions capture variations in spinal alignment, lateral leaning, and back support engagement. The pelvic and upper-leg region consisted of six locations (18 sensors in total), including the gluteal region (left, center, right) and posterior thigh region (left, center, right). These locations measure weight distribution, asymmetrical loading, forward leaning behavior, and dynamic lower-body movements.

In this study, each anatomical sensing unit integrated three co-located sensing elements: a pressure sensor (RP-S40-ST, OYMotion, Shenzhen, China), a digital temperature sensor (MCP9808, Microchip Technology Inc., Chandler, AZ, USA), and a piezoelectric vibration sensor embedded within a flexible smart-skin layer. Pressure measurements were obtained using RP-S40-ST force-sensitive resistive sensors with a nominal sensing range of 0–100 N and an approximate sensing diameter of 18 mm, enabling localized force detection at each anatomical location. The sensors were sampled at 50 Hz, and raw resistance values were converted into force units (N) using calibration curves established during system setup. For model training, pressure values were normalized using min–max scaling computed exclusively from the training data within each LOSO fold to prevent data leakage.

Surface temperature was measured using MCP9808 digital temperature sensors, featuring a measurement range from -40 °C to 125 °C and a resolution of 0.0625 °C. In this study, the effective operating range was limited to 20–40 °C, corresponding to typical human skin contact conditions. Temperature signals were sampled at 10 Hz and temporally aligned to 50 Hz using linear interpolation for multimodal synchronization. To ensure fair evaluation, temperature normalization parameters (mean and standard deviation) were computed from the training data within each LOSO fold and applied to the corresponding test data.

Dynamic micro-movements were captured using miniature piezoelectric vibration sensors embedded beneath the smart-skin layer. These sensors detect mechanical oscillations generated by fidgeting, periodic shifting, and subtle postural adjustments. The vibration signal was sampled at 50 Hz and represented as voltage amplitude proportional to acceleration magnitude. After band-pass filtering to suppress environmental noise, vibration signals were normalized to the [0, 1] range using scaling parameters derived from training data within each LOSO fold.

All sensing units were integrated into a custom data acquisition module based on an Arduino Mega 2560 (Arduino, Monza, Italy) platform equipped with high-resolution analog-to-digital conversion. The system provided synchronized sampling across all 42 channels at 50 Hz. Acquired data were transmitted to a host computer via USB serial interface for real-time logging, preprocessing, and subsequent deep learning analysis.

3.2. Data Collection Protocol

This section describes the definition of posture classes and the experimental procedure used to collect multimodal sensor data. The protocol was designed to capture both static and dynamic sitting behaviors under realistic office conditions while introducing temporal variability across sessions. All data were collected from a single participant in a controlled experimental setting; therefore, the protocol was intended to evaluate cross-session consistency rather than population-level generalization.

3.2.1. Posture Classes

Nine distinct sitting posture classes were defined based on the ergonomic literature and preliminary pilot observations of common seated behavior [1,5]. These classes represent

both recommended ergonomic postures and naturally occurring dynamic movements associated with discomfort or fatigue.

The static posture classes included Neutral Sitting, characterized by upright trunk alignment with full back support, feet flat on the floor, and thighs parallel to the ground; Full Recline, involving complete back contact with a reclined angle of approximately 15–20°, typically associated with resting behavior; and Forward Lean, where the torso inclines anteriorly with reduced back contact and increased load distribution toward the anterior pelvic region. Lateral asymmetrical postures included Right Lean and Left Lean, defined by a shift of body weight toward the corresponding side, resulting in asymmetric load distribution and altered trunk alignment. These postures commonly occur during object reaching or unilateral task engagement.

Dynamic posture classes captured movement-related behaviors. Periodic Shifting represents continuous micro-adjustments and alternating weight transfer, characterized by temporal fluctuations in pressure signals and elevated vibration activity. The three vibration intensity classes—Light Vibration, Medium Vibration, and High Vibration—correspond to increasing levels of fidgeting or postural adjustment. These classes were defined based on predefined amplitude thresholds and temporal characteristics of the vibration signal, rather than being directly inferred from the model input, to minimize potential label leakage. These classes were differentiated based on vibration signal amplitude ranges and associated pressure variability, reflecting escalating levels of restlessness during prolonged sitting. Together, these nine classes enable comprehensive modeling of both static alignment and dynamic micro-movement behaviors.

3.2.2. Data Collection Sessions

Data acquisition was conducted across three independent sessions on different days to incorporate temporal variability and environmental changes. This design enables evaluation of cross-session generalization and robustness to contextual differences such as ambient temperature and time-of-day variations.

Session 1 was conducted in the morning (09:00–12:00) under ambient temperatures ranging from 22–24 °C and yielded 5882 labeled samples. Session 2 was conducted in the afternoon (14:00–17:00) with ambient temperatures in the range of 24–26 °C, resulting in 5373 samples. Session 3 took place in the evening (18:00–20:00) under slightly cooler conditions (21–23 °C) and produced 3782 samples. In total, the dataset comprises 15,037 samples spanning nine posture classes and 42 synchronized sensor channels.

Each session followed a standardized experimental protocol. Participants began with a two-minute neutral sitting period for baseline calibration. Subsequently, posture classes were prompted verbally in randomized order to reduce sequential bias. Each posture was maintained for five minutes while participants engaged in typical office-related activities such as typing, reading, or mouse interaction. A 10-second transition interval between postures was excluded from the dataset to prevent contamination of labeled segments. Dynamic postures, including periodic shifting and vibration-related behaviors, were performed naturally to preserve ecological validity rather than through rigid scripted movement. All sessions were conducted using the same sensor configuration and placement to ensure consistency across recordings.

The experiment involved a healthy adult male participant (21 years old, 172 cm, 65 kg) with no reported musculoskeletal disorders. As the dataset was collected from a single participant, the results should be interpreted as demonstrating intra-subject cross-session consistency rather than inter-subject generalization. The study protocol received institutional review board approval, and informed consent was obtained prior to data collection.

3.3. Data Preprocessing

This section describes the preprocessing pipeline applied to the raw multimodal sensor signals prior to model training. The objective is to reduce measurement noise, mitigate session-specific variability, and transform continuous sensor streams into structured temporal inputs suitable for deep learning architectures.

3.3.1. Signal Filtering and Normalization

Raw signals from pressure, temperature, and vibration sensors were first subjected to frequency-domain filtering to remove noise while preserving posture-relevant dynamics. Pressure and temperature signals were processed using a fifth-order Butterworth low-pass filter with a cutoff frequency of 5 Hz. This configuration attenuates high-frequency measurement noise while retaining slow biomechanical variations associated with posture transitions and weight redistribution. Vibration signals, which capture dynamic micro-movements, were filtered using a band-pass filter within the range of 0.5–25 Hz. This frequency band corresponds to typical human motion frequencies and excludes low-frequency drift as well as high-frequency environmental interference.

To address potential anomalies, samples exceeding three standard deviations from the session-specific mean were identified as outliers. These instances, representing less than 0.5% of the dataset, were corrected via linear interpolation using adjacent temporal samples to preserve signal continuity.

Following filtering and outlier correction, normalization was applied to ensure cross-session comparability. Pressure and vibration channels were scaled to the [0, 1] range using min–max normalization, while temperature channels were standardized using z-score normalization to compensate for ambient temperature variations. To prevent data leakage, normalization parameters were computed using only the training data within each LOSO fold and then applied to the corresponding test fold. This ensures that no information from the test session is used during model training. Furthermore, all preprocessing steps were performed separately within each training fold. No data from the test session were used during preprocessing, model training, or hyperparameter tuning. Therefore, the LOSO evaluation reflects true cross-session generalization without information leakage.

3.3.2. Temporal Windowing

After preprocessing, continuous sensor streams were segmented into fixed-length temporal windows to construct structured inputs for sequence modeling. A sliding window approach was adopted with a window length of 120 samples, corresponding to 2.4 s at a sampling frequency of 50 Hz. A 50% overlap (60 samples) was applied between consecutive windows to increase sample diversity and preserve temporal continuity.

It should be noted that overlapping windows may introduce temporal correlation between adjacent samples. Therefore, random holdout evaluation may yield optimistic estimates of performance. To address this, Leave-One-Session-Out (LOSO) validation was adopted as the primary evaluation strategy to ensure strict temporal separation between training and test data.

The selected window duration balances two competing requirements: capturing short-term dynamic posture variations while maintaining sufficient temporal context for detecting transitions and periodic shifting behaviors. Each window was represented as a two-dimensional tensor of shape (120, 42), where 120 denotes the temporal dimension and 42 corresponds to the multimodal sensor channels (14 anatomical locations $\times$ 3 modalities).

3.3.3. Data Augmentation

To enhance model robustness and reduce overfitting, time-series data augmentation was applied during training. Three augmentation strategies were employed. First, jittering introduced small Gaussian noise ($\sigma = 0.05$) to simulate sensor measurement variability. Second, amplitude scaling multiplied sensor values by a random factor within the range of 0.9–1.1 to emulate inter-session intensity variations. Third, time warping applied controlled temporal stretching or compression ($\pm 10\%$) to simulate minor differences in movement speed.

Augmentation was applied probabilistically (50% likelihood per batch) during training only, ensuring that validation and test performance reflect true generalization without artificial perturbations. No augmentation was applied to validation or test data to maintain unbiased evaluation under LOSO conditions.

3.4. Deep Learning Architecture

To model temporal dependencies in multimodal sensor sequences, this study evaluated two representative deep learning architectures: a Temporal Convolutional Network with Attention (TCN + Attn) and a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN − LSTM) model. These architectures were selected to represent two dominant paradigms in temporal modeling: convolution-based sequence modeling (TCN) and recurrent sequence modeling (LSTM-based), enabling a fair comparative analysis under identical conditions. Both architectures receive identical preprocessed input tensors of shape (T,C) = (120,42), where T denotes the temporal length (2.4 s window at 50 Hz) and C represents the number of multimodal sensor channels.

Figure 2 illustrates an overview of both architectures, highlighting the convolution-based temporal modeling strategy of the TCN and the hierarchical convolution–recurrent structure of the CNN − LSTM model.

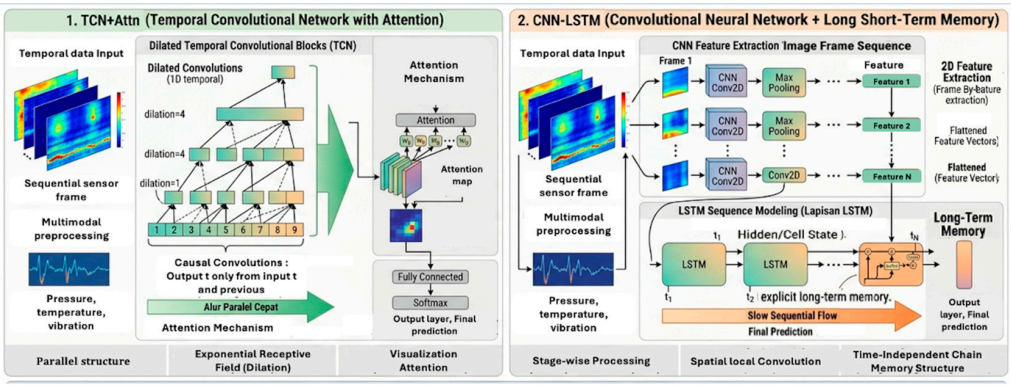


Figure 2. Neural network architectures comparison for temporal sequence modeling. Arrows indicate the data flow between processing stages, while different colors and blocks represent distinct components of the architectures (e.g., convolutional layers, attention modules, and LSTM units). The visual patterns illustrate feature transformations across stages from input sequences to final prediction.

3.4.1. Temporal Convolutional Network with Attention (McNatt)

The TCN + Attn architecture was designed to capture long-range temporal dependencies using dilated causal convolutions while maintaining parallelizable computation. The TCN formulation follows Bai et al. [30], where dilated causal convolutions serve as an alternative to recurrent sequence modeling. The network begins with an input projection layer consisting of a one-dimensional convolution that maps 42 input channels into a 64-dimensional feature space, followed by batch normalization and ReLU activation.

Temporal modeling was performed using four residual dilated convolutional blocks with exponentially increasing dilation factors $d \in \{1, 2, 4, 8\}$. For a given input sequence $x(t)$, the dilated convolution operation is defined as Equation (1), where y_t is the output at time step t, $w(k)$ denotes the convolution kernel of size K, d is the dilation factor, and $x(t - d \cdot k)$ represents the input sampled at dilated intervals [30,31].

$$y(t) = \sum_{k=0}^{K-1} w(k) \cdot x(t - d \cdot k) \quad (1)$$

Dilated convolutions expand the receptive field without increasing network depth or sacrificing temporal resolution. With four stacked residual blocks and a kernel size of 3, the effective receptive field reaches approximately 90-time steps (≈1.8 s), enabling multi-scale temporal modeling of posture transitions and micro-movements. This receptive field is sufficient to capture short-term temporal dynamics within each window; however, longer-term dependencies across windows are implicitly handled through the sliding-window segmentation strategy rather than explicit sequence continuity modeling.

To further enhance discriminative capability, a self-attention mechanism was applied after the temporal convolutional blocks. Attention weights were computed using scaled dot-product attention in Equation (2), where Q, K, and V are learned query, key, and value projections of the temporal feature matrix, and d_k denotes the dimensionality of the key vectors. This mechanism enables adaptive weighting of informative time steps within each window, allowing the model to emphasize segments associated with posture transitions and micro-movements [33,34].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

The final feature representation was obtained via global average pooling across the temporal dimension, followed by a fully connected classification head with softmax activation for nine-class prediction. The complete TCN + Attn model contains 245,128 trainable parameters, with an inference memory footprint of approximately 180 MB. These efficiency characteristics indicate potential suitability for real-time deployment scenarios, subject to hardware constraints and system-level validation.

3.4.2. CNN − LSTM Architecture

For comparative evaluation, a CNN − LSTM hybrid architecture was implemented following widely adopted paradigms in sensor-based human activity recognition [27–29]. This architecture represents hybrid spatial–temporal modeling approaches commonly applied in multimodal sensor classification tasks.

The CNN stage performs hierarchical feature extraction using stacked one-dimensional convolutions along the temporal axis. The first convolution maps 42 input channels to 64 feature maps, followed by batch normalization, ReLU activation, and max pooling. A second convolution expands the representation to 128 feature maps, again followed by max pooling for temporal downsampling.

Temporal dependencies are subsequently modeled using a two-layer bidirectional Long Short-Term Memory (BiLSTM) network. For a given sequence x_t, the LSTM updates are governed by Equations (3)–(7), where f_t, i_t, and o_t denote the forget, input, and output gates, respectively. The bidirectional configuration processes the sequence in both forward and backward directions to enhance contextual representation.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (4)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

Both architectures were trained and evaluated under identical preprocessing, windowing, and LOSO validation settings to ensure a fair and unbiased comparison of their temporal modeling capabilities.

In this implementation, two stacked bidirectional LSTM layers were employed to capture long-range contextual dependencies from both temporal directions, consistent with prior posture recognition frameworks [27,28]. The final hidden representation was passed to a fully connected classification head with softmax activation for nine-class prediction. The CNN − LSTM model contained 312,448 trainable parameters and requires approximately 235 MB of memory during inference. Compared to the TCN-based architecture, the recurrent structure introduced sequential computation, increasing computational cost and limiting parallelization efficiency. This characteristic may impact scalability for real-time or embedded deployment scenarios.

3.5. Training Protocol

This section describes the optimization strategy, regularization techniques, and validation procedures employed to ensure fair and rigorous comparison between the TCN + Attn and CNN − LSTM architectures. Both models were trained under identical hyperparameter configurations to isolate architectural effects on performance and generalization.

3.5.1. Loss Function and Optimization Strategy

Both architectures were optimized using the categorical cross-entropy loss function for multi-class classification as defined in Equation (8), where N denotes the number of training samples, $C = 9$ represents the number of posture classes, $y_{i,c}$ is the one-hot encoded ground-truth label, and $\hat{y}_{i,c}$ is the predicted probability for class c.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \quad (8)$$

Model parameters were updated using the Adam optimizer with an initial learning rate of 0.001 and momentum coefficients $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. To stabilize convergence and prevent overfitting, a ReduceLROnPlateau learning rate scheduler was applied. The learning rate was reduced by a factor of 0.5 if validation loss does not improve for five consecutive epochs, with a minimum threshold of 10^{-6}.

Training was conducted with a batch size of 64 for a maximum of 50 epochs. Early stopping with a patience of 10 epochs was implemented based on validation loss to prevent overtraining. Convolutional layers were initialized using He initialization, while fully connected layers employ Xavier initialization to promote stable gradient propagation.

Regularization strategies include L2 weight decay with coefficient $\lambda = 10^{-4}$, dropout layers with probabilities ranging from 0.2 to 0.3 depending on architectural depth, and batch normalization to mitigate internal covariate shift. These mechanisms collectively improve generalization performance and training stability.

3.5.2. Validation Strategies

To evaluate both within-session performance and cross-session generalization, two complementary validation strategies were employed. The first strategy used stratified holdout validation, in which 80% of the combined dataset (12,030 samples) was used for training and 20% (3007 samples) was reserved for testing. The split preserved class distribution across the nine posture categories. This evaluation provides an estimate of model capacity under random sampling and reflects upper-bound performance when training and testing distributions are similar. However, due to the use of overlapping sliding windows, this evaluation may introduce temporal correlation between training and test samples and should therefore be interpreted as an optimistic estimate of performance rather than a true generalization measure.

The second strategy adopted a Leave-One-Session-Out (LOSO) cross-validation protocol to assess cross-session generalization. In this scheme, data from two sessions were used for training while the remaining session was used exclusively for testing. Specifically, Fold 1 trained on Sessions 2 and 3 and evaluated on Session 1; Fold 2 trained on Sessions 1 and 3 and evaluated on Session 2; and Fold 3 trained on Sessions 1 and 2 and evaluated on Session 3. The final LOSO performance was reported as the average across the three folds.

To prevent data leakage, all preprocessing steps, including normalization and data augmentation, were performed separately within each fold using only the training data. No data from the test session were used during preprocessing, model training, or hyperparameter tuning. This ensures strict separation between training and testing data and prevents information leakage across sessions.

LOSO validation enforces strict session-wise separation between training and testing data, requiring models to generalize to unseen temporal and contextual variations [40–42]. Therefore, LOSO results are considered the primary evaluation metric in this study, as they more accurately reflect within-subject cross-session generalization under controlled experimental conditions.

3.6. Evaluation Metrics

In this study, to comprehensively assess classification performance and deployment feasibility, multiple complementary evaluation metrics were employed. These metrics quantify predictive accuracy, class-wise discrimination capability, and computational efficiency.

3.6.1. Classification Performance Metrics

Overall classification accuracy was used to measure the proportion of correctly predicted samples across all classes. It is defined as Equation (9).

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{9}$$

Although accuracy provides a global performance estimate, it may not fully reflect performance under class imbalance. Therefore, class-wise precision, recall, and F1-score were additionally computed. For each class c, precision measures the proportion of correctly predicted samples among all samples predicted as class c as seen in Equation (10), where TP_c and FP_c denote true positives and false positives for class c, respectively.

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

Recall quantifies the proportion of correctly identified samples among all true instances of class c as denote in Equation (11), where FN_c represents false negatives.

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

The F1-score, defined as the harmonic mean of precision and recall, balances both measures and is defined in Equation (12).

$$\text{F1}_c = 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{12}$$

Meanwhile, to evaluate performance across all nine posture classes while treating each class equally, macro-averaged metrics were computed. In particular, the Macro-F1 score is defined as Equation (13), where $C = 9$ denotes the number of posture classes. Macro averaging ensures that minority or dynamically defined classes contribute equally to the final performance measure. Under LOSO validation, Macro-F1 was computed as the average across folds and is used as the primary evaluation metric to assess cross-session generalization performance.

$$\text{Macro-F1} = \frac{1}{\text{C}} \sum_{\text{c}=1}^{\text{C}} \text{F1}_c \tag{13}$$

In addition to scalar metrics, a 9×9 confusion matrix was reported to visualize inter-class misclassification patterns. The confusion matrix provides insights into posture similarities, asymmetric prediction errors, and the discriminative capability of multimodal sensing under cross-session conditions.

3.6.2. Computational Efficiency Metrics

Beyond predictive performance, computational efficiency was evaluated to assess real-time deployment feasibility. The following system-level metrics were recorded: training time per epoch (seconds), measuring optimization cost during model development; inference latency per sample (milliseconds), quantifying real-time responsiveness; number of trainable parameters, reflecting model complexity; and memory footprint (MB) during inference mode.

These metrics are essential for determining the practicality of embedded smart-skin deployment, particularly under resource-constrained hardware conditions. By jointly analyzing classification accuracy and computational overhead, this study evaluates not only predictive robustness but also system-level efficiency.

3.7. Ablation Study Design

A structured ablation study was conducted to systematically evaluate the contribution of each sensing modality and their combinations. The objective of this analysis was to quantify the discriminative power of individual modalities and to determine whether multimodal fusion provides additive performance improvements under cross-session validation.

Seven sensor configurations were examined. These included three unimodal settings—pressure only (P), temperature only (T), and vibration only (V)—each utilizing 14 anatomical sensing locations. In addition, three bimodal combinations were evaluated: pressure–temperature (P + T), pressure–vibration (P + V), and temperature–vibration (T + V), each comprising 28 sensor channels. Finally, the complete multimodal configuration (P + T + V) incorporated all 42 sensors distributed across the 14 anatomical regions.

For all ablation experiments, the TCN + Attn architecture was employed with identical hyperparameters, preprocessing procedures, and training settings to ensure a controlled

comparison. All preprocessing steps, including normalization and data augmentation, were applied within each LOSO fold using only training data to prevent data leakage and ensure fair evaluation across configurations. The Leave-One-Session-Out (LOSO) validation protocol was consistently applied across all configurations to assess cross-session generalization rather than within-session performance.

This structured evaluation served three analytical objectives. First, it identified the most discriminative unimodal sensing modality for sitting posture recognition. Second, it quantified performance changes obtained through bimodal fusion. Third, it examined whether full multimodal integration yields performance improvements beyond pairwise combinations. This design also enables distinguishing whether performance improvements arise from additive effects of multiple modalities or from complementary interactions between sensing channels. By maintaining architectural and training consistency, observed performance differences can be attributed directly to sensing modality contributions rather than modeling variations.

4. Results

4.1. Performance Comparison Result

A summary of the performance comparison between TCN + Attn and CNN- LSTM results is provided in Table 2. Meanwhile, Figure 3 presents a comprehensive comparison architectures across both predictive performance and computational efficiency metrics.

Table 2. Performance comparison between TCN + Attn and CNN − LSTM.

Metric	TCN + Attn	CNN − LSTM	Difference	Winner
Holdout Accuracy	89.89%	91.34%	−1.45%	CNN − LSTM
LOSO Accuracy	85.23%	82.67%	+2.56%	TCN + Attn
Macro F1 (Holdout)	88.45%	89.78%	−1.33%	CNN − LSTM
Macro F1 (LOSO)	84.87%	86.12%	−1.25%	CNN − LSTM
Training Time/Epoch	245 s	387 s	−36.7%	TCN + Attn
Inference Time	14.2 ms	21.5 ms	−33.9%	TCN + Attn
Parameters	245 k	312 k	−21.5%	TCN + Attn
Memory Footprint	180 MB	235 MB	−23.4%	TCN + Attn

According to Table 2, under stratified holdout validation, the CNN-LSTM architecture achieves slightly higher accuracy (91.34%) compared to TCN + Attn (89.89%). A similar trend is observed in the Macro-F1 score under holdout evaluation.

In contrast, under the Leave-One-Session-Out (LOSO) protocol, the TCN + Attn architecture achieves an accuracy of 85.23%, outperforming CNN − LSTM by 2.56 percentage points. In contrast, under the Leave-One-Session-Out (LOSO) protocol, the TCN + Attn architecture achieves an accuracy of 85.23%, outperforming CNN − LSTM by 2.56 percentage points. However, in terms of Macro-F1 under LOSO evaluation, CNN − LSTM achieves a slightly higher score (86.12%) compared to TCN + Attn (84.87%), indicating that, while TCN + Attn provides higher overall accuracy, CNN − LSTM maintains slightly better class-balanced performance across posture categories.

The generalization gap, defined as the difference between holdout and LOSO accuracy, is 4.66% for TCN + Attn and 8.67% for CNN − LSTM. The larger performance drop observed in CNN − LSTM indicates greater sensitivity to session-level distribution shifts.

Under within-session evaluation, CNN − LSTM demonstrates marginally higher performance. However, this advantage diminishes under cross-session testing, where stricter separation between training and testing data is enforced.

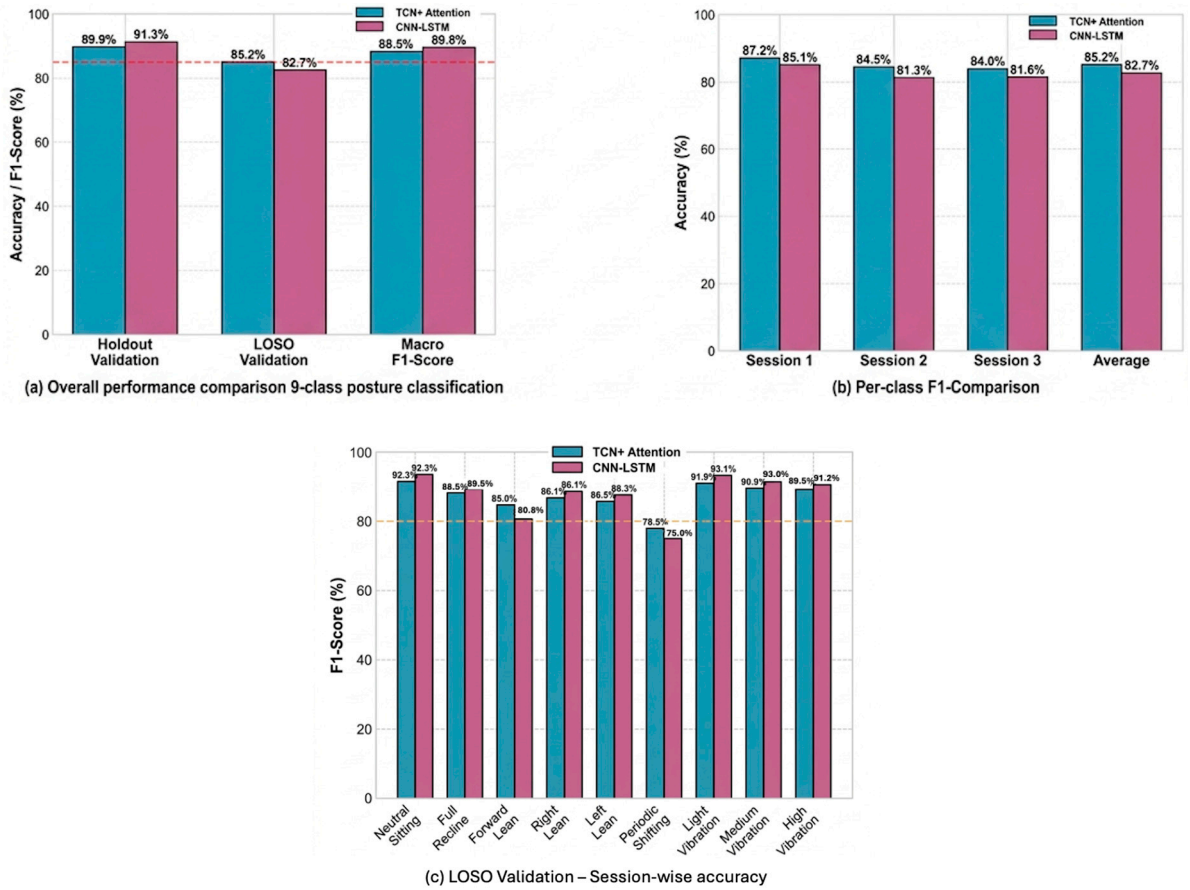


Figure 3. Performance comparison of TCN + Attn and CNN − LSTM.

From a computational perspective, TCN + Attn demonstrates reduced training time per epoch (245 s vs. 387 s) and lower inference latency (14.2 ms vs. 21.5 ms). Additionally, TCN + Attn contains fewer trainable parameters (245 k vs. 312 k) and requires less memory during inference (180 MB vs. 235 MB). These differences reflect the architectural distinction between convolutional temporal modeling and recurrent sequence processing. These results are consistent with LOSO-based evaluation, which serves as the primary metric in this study, as it provides a more realistic estimate of cross-session generalization performance compared to random holdout validation.

Overall, the results indicate that, while CNN − LSTM achieves slightly higher performance under random holdout validation, TCN + Attn demonstrates stronger robustness under cross-session evaluation and improved computational efficiency.

4.2. Confusion Matrix Analysis

Figure 4 presents the 9 × 9 confusion matrices for both holdout and Leave-One-Session-Out (LOSO) validation using the TCN + Attn architecture. Detailed per-class performance metrics under LOSO evaluation are summarized in Table 3. This analysis provides class-level insight into discriminative capability and error distribution beyond aggregate accuracy metrics.

According to Table 3, which presents class-level performance characteristics under Leave-One-Session-Out (LOSO) validation, most posture classes achieve F1-scores above 88%, indicating consistent discrimination across sessions. The highest-performing classes include Medium Vibration (92.09%), Periodic Shifting (91.31%), and Forward Lean (91.38%). Notably, the Periodic Shifting class achieves 100% precision, indicating that all samples

predicted as this class were correctly classified. This suggests that its temporal dynamics exhibit distinct patterns that are well captured by the model.

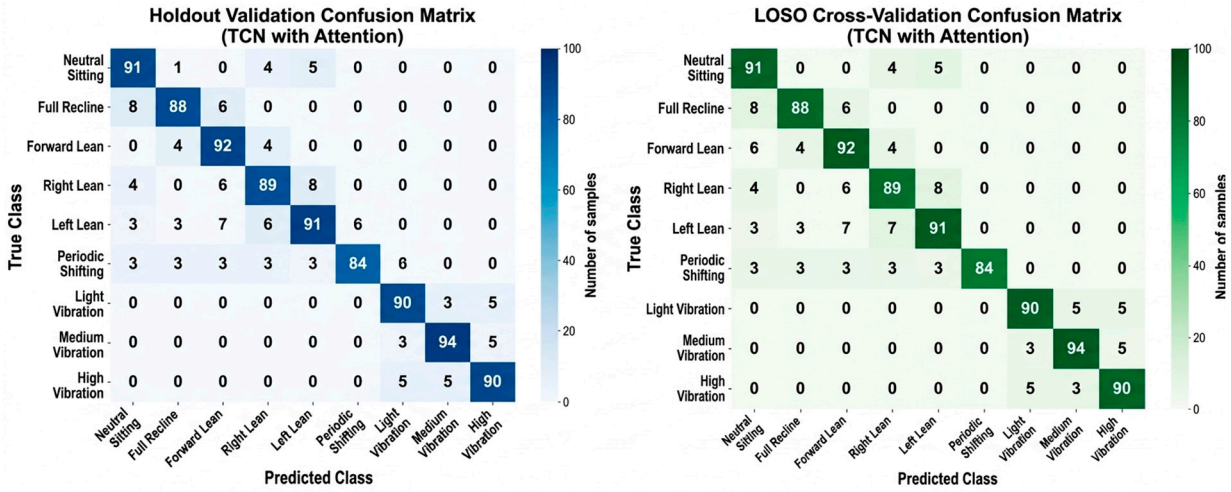


Figure 4. Posture classification confusion matrices under holdout and LOSO validation.

Table 3. Per-class performance metrics (TCN + Attn and LOSO validation).

Posture Class	Precision	Recall	F1-Score	Support
Neutral Sitting	85.51%	91.15%	88.24%	1673
Full Recline	92.44%	88.35%	90.35%	1582
Forward Lean	91.10%	91.65%	91.38%	1698
Right Lean	83.32%	88.51%	85.84%	1621
Left Lean	84.92%	90.86%	87.79%	1645
Periodic Shifting	100.00%	84.01%	91.31%	1456
Light Vibration	91.77%	90.21%	90.98%	1734
Medium Vibration	90.39%	93.84%	92.09%	1812
High Vibration	91.88%	89.92%	90.89%	1816
Macro Average	90.15%	89.83%	89.87%	15,037

The lowest F1-scores are observed for Right Lean (85.84%) and Left Lean (87.79%). Confusion primarily occurs between these two lateral lean postures, with a mutual misclassification rate of approximately 7%. This pattern indicates that lateral asymmetry is more challenging to distinguish compared to vibration-dominant or forward-lean configurations. The confusion likely arises from subtle differences in lateral pressure distribution, which may be influenced by minor anatomical asymmetry or sensor placement variability across sessions.

Additional confusion patterns include Neutral Sitting versus Full Recline (approximately 4.8% confusion), reflecting partial similarity in backrest contact distribution. Similarly, Light Vibration versus Medium Vibration exhibits approximately 3.9% confusion, suggesting that intermediate vibration intensities can overlap in signal magnitude. These results indicate that misclassifications predominantly occur between semantically or biomechanically adjacent classes rather than across unrelated posture categories.

It is important to note that the Macro-F1 value reported in Table 3 (89.87%) is computed from aggregated predictions across all LOSO folds. In contrast, the primary evaluation metric used in this study is the fold-wise averaged Macro-F1 of 84.87%, as reported in Table 4. The discrepancy arises from different aggregation methods, where fold-wise averaging provides a more conservative and realistic estimate of cross-session generalization performance.

Table 4. LOSO cross-validation results (TCN + Attn).

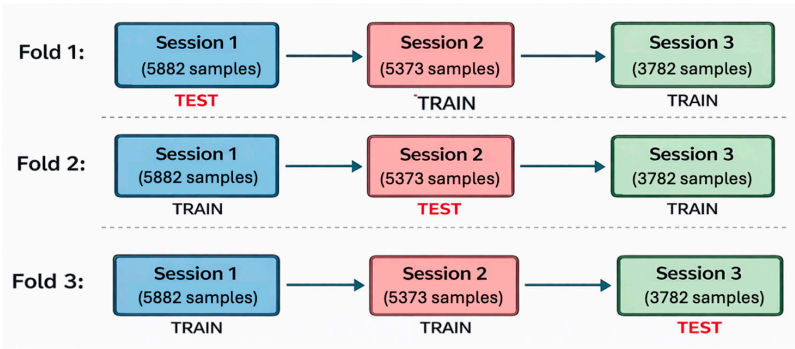
Test Session	Train Sessions	Test Samples	Accuracy	Macro-F1	Training Time
Session 1	S2 + S3	5882	87.2%	86.8%	11,450 s
Session 2	S1 + S3	5373	84.5%	84.1%	12,230 s
Session 3	S1 + S2	3782	84.0%	83.7%	13,120 s
Average	-	15,037	85.23%	84.87%	12,267 s

When comparing holdout and LOSO validation, per-class F1-scores decrease by an average of approximately 1.2%, confirming the impact of temporal distribution shifts across sessions. The largest performance reductions are observed in lateral lean postures (Right Lean, −2.8%; Left Lean, −2.4%), suggesting greater sensitivity to inter-session variability. In contrast, vibration-dominant classes demonstrate relatively stable performance across sessions, indicating that dynamic signal characteristics generalize more robustly than subtle spatial pressure asymmetries.

Overall, the confusion matrix analysis demonstrates that the TCN + Attn model maintains strong class-level discrimination under cross-session validation, with errors primarily concentrated among closely related posture categories. These findings highlight the importance of multimodal sensing and temporal modeling in improving robustness under realistic deployment conditions.

4.3. Leave-One-Session-Out (LOSO) Validation Results

Figure 5 illustrates the LOSO validation scheme and the session-wise performance distribution of the TCN + Attn model. Detailed quantitative results for each fold are summarized in Table 4.



(a) Leave-One-Session-Out (LOSO) Cross-Validation Scheme



(b) Leave-One-Session-Out (LOSO) - Validation results (3 – Folds)

Figure 5. LOSO validation scheme with three sessions (S1–S3). The dashed horizontal line represents the average LOSO accuracy, serving as a reference benchmark for comparing model performance across different folds. TCN + Attn consistently achieves higher accuracy compared to CNN − LSTM across all validation folds.

According to Table 4, based on Session-Wise Generalization Analysis, the LOSO results demonstrate consistent performance across all three temporal folds. The highest accuracy (87.2%) is achieved when Session 1 is used as the test set. This session was conducted in the morning under moderate ambient temperature (22–24 °C), potentially providing more stable thermal conditions and reduced environmental variability. When Session 2 is used for testing, accuracy decreases to 84.5%. This session was conducted during the afternoon under higher ambient temperatures (24–26 °C), which may have introduced distribution shifts in temperature channels and subtle sensor response variations. Meanwhile, Session 3 contains the smallest sample size (3782 samples), which may limit the diversity of training patterns available for generalization. However, these interpretations are provided as plausible explanations rather than definitive causal conclusions, as the study does not explicitly model environmental variables.

Despite these variations, the standard deviation of LOSO accuracy across sessions is only 1.7%, indicating relatively stable cross-session generalization under temporal and environmental variability.

Figure 6 further illustrates per-class F1-scores for each LOSO fold. Several posture categories demonstrate strong robustness across sessions. In particular, Medium Vibration, High Vibration, and Forward Lean consistently maintain F1-scores above 88%, indicating that dynamic and posture-shift classes are reliably captured by multimodal sensing. In contrast, Right Lean and Left Lean exhibit greater variability, with F1-score fluctuations ranging from 5% to 8% across sessions. These lateral postures are inherently sensitive to subtle differences in body alignment and anatomical asymmetry, making them more susceptible to distribution shifts between sessions.

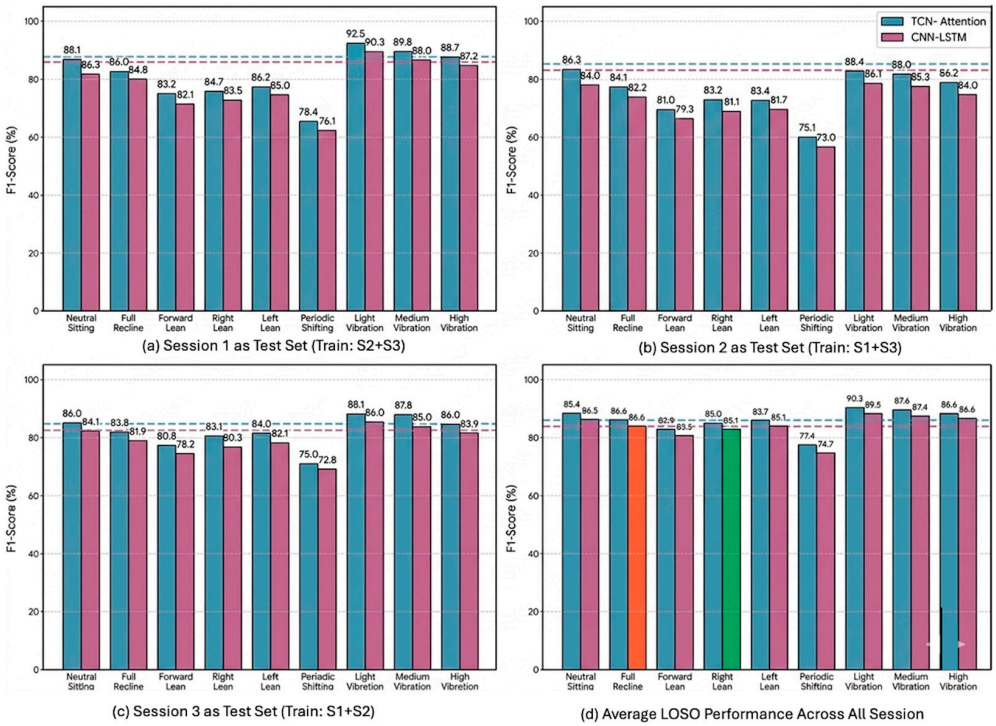


Figure 6. LOSO session-wise performance. LOSO session-wise performance. The dashed horizontal lines represent the average F1-score for each session, serving as a reference benchmark across posture classes. Different colors indicate the performance of the compared models (TCN + Attention and CNN − LSTM) across each posture category.

When Session 3 serves as the test fold, most classes experience a modest F1-score reduction of approximately 2–4%, aligning with its smaller sample size and different

temporal context. To further assess temporal robustness, Table 5 compares session-wise LOSO accuracy between TCN + Attn and CNN − LSTM architectures.

Table 5. Architecture comparison—LOSO session-wise accuracy.

Test Session	TCN + Attn	CNN − LSTM	Difference
Session 1	87.2%	85.1%	+2.1%
Session 2	84.5%	81.3%	+3.2%
Session 3	84.0%	81.6%	+2.4%
Average	85.23%	82.67%	+2.56%

Table 5 illustrates of architecture comparison across all three folds, TCN + Attn consistently outperforms CNN − LSTM. The largest improvement (+3.2%) is observed when testing on Session 2, the fold most affected by environmental variation. This consistent performance gap across folds indicates that the TCN-based architecture provides more stable cross-session generalization under the evaluated conditions. It should be noted that these results are specific to the single-subject experimental setting and should be interpreted as evidence of cross-session consistency rather than population-level generalization. Overall, the LOSO analysis demonstrates that the TCN + Attn model achieves stable performance across varying temporal conditions, with consistent advantages over CNN − LSTM in both accuracy and robustness.

4.4. Ablation Study: Sensor Modality Contribution

To quantify the relative contribution of each sensing modality and their combinations, a structured ablation study was conducted under the TCN + Attn architecture using the LOSO validation protocol. Figure 7 visualizes the comparative performance trends, while detailed quantitative results are presented in Table 6.

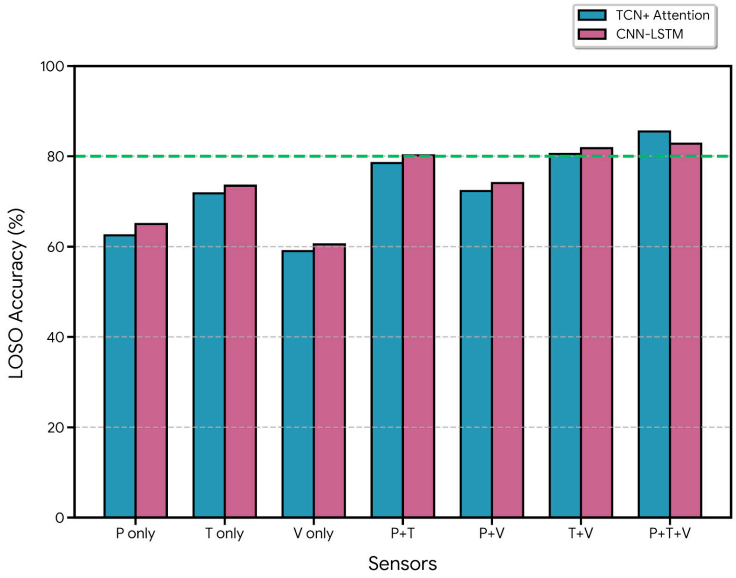


Figure 7. Ablation study—modality contribution: P = pressure, T = temperature, V = vibration. The dashed line indicates the reference performance threshold (80% LOSO accuracy), serving as a benchmark for comparing model performance across different sensor modality configurations.

For single-modality performance, among unimodal configurations, temperature sensing achieves the highest LOSO accuracy (71.5%), outperforming pressure-only (62.3%) and vibration-only (58.7%) settings by 9.2% and 12.8%, respectively. This result indicates that thermal contact patterns provide highly discriminative posture information. Several

factors may explain this behavior. First, prolonged body contact generates stable thermal signatures that reflect posture-specific contact distributions. Second, temperature gradients indirectly encode weight redistribution and sustained pressure zones. Third, thermal inertia introduces temporal smoothing, reducing instantaneous variability compared to pressure measurements.

Table 6. Ablation study results (TCN + Attn and LOSO validation).

Modality Configuration	Sensors	LOSO Accuracy	Macro-F1	Δ vs. P + T + V
Single Modality				
Temperature only (T)	14	71.5%	69.8%	−13.7%
Pressure only (P)	14	62.3%	60.1%	−22.9%
Vibration only (V)	14	58.7%	56.4%	−26.5%
Dual Modality				
Temperature + Vibration (T + V)	28	80.3%	78.9%	−4.9%
Pressure + Temperature (P + T)	28	78.4%	76.7%	−6.8%
Pressure + Vibration (P + V)	28	72.1%	70.3%	−13.1%
Multimodal				
Pressure + Temperature + Vibration (P + T + V)	42	85.23%	84.87%	-

In contrast, pressure-only performance is substantially lower than commonly reported in chair-based [3–5]. This discrepancy likely arises from the increased complexity of cross-session validation and sensitivity to body morphology, clothing thickness, and subtle posture similarities. Vibration-only sensing yields the lowest overall accuracy, as static postures generate limited dynamic signals.

However, per-class analysis reveals that vibration sensors are highly effective for dynamic posture categories. For example, vibration-only accuracy reaches 89.2% for High Vibration and 82.7% for Periodic Shifting, while averaging only 42.3% across static postures. This confirms that vibration signals are modality-specific and primarily beneficial for movement-related behaviors.

Meanwhile, in dual-modality fusion or among bimodal configurations, Temperature + Vibration (T + V) achieves the highest LOSO accuracy (80.3%), outperforming Pressure + Temperature (78.4%) and Pressure + Vibration (72.1%). This finding suggests that temperature and vibration provide complementary information: temperature captures static and sustained contact patterns, whereas vibration captures dynamic micro-movements. The relatively weaker performance of Pressure + Vibration further supports the hypothesis that pressure alone lacks sufficient discriminatory power under cross-session variability, particularly for distinguishing subtle static postures.

Furthermore, the full multimodal configuration (P + T + V) achieves 85.23% LOSO accuracy, representing a +4.9% improvement over the best dual-modality configuration (T + V), a +13.7% improvement over the best single modality (Temperature), and a +22.9% improvement over pressure-only systems. These results indicate that multimodal fusion provides additive performance benefits under cross-session evaluation. However, the observed improvements should be interpreted as empirical performance gains rather than strict causal contributions, as interactions between modalities may vary depending on sensor configuration and experimental conditions.

Specifically, the incremental gains follow a diminishing-return pattern T → T + V: +8.8% improvement, and T + V → P + T + V: +4.9% improvement. This trend suggests that temperature and vibration form the dominant discriminative modalities, while pressure contributes complementary but comparatively smaller improvements in cross-session scenarios.

Although vibration-only sensing yields the lowest overall accuracy, per-class analysis indicates strong performance for dynamic posture categories. For example, vibration-only accuracy reaches 89.2% for High Vibration and 82.7% for Periodic Shifting, while averaging significantly lower performance across static postures. This confirms that vibration sensing is particularly effective for capturing movement-related behaviors rather than static posture configurations.

From a system design perspective, the ablation results provide practical insights for deployment. A reduced Temperature + Vibration (T + V) configuration using 28 sensors achieves 80.3% LOSO accuracy, corresponding to approximately 94.2% of full-system performance while reducing sensor count by 33%. This configuration may be suitable for resource-constrained deployments where hardware complexity, power consumption, or cost must be minimized. However, the full multimodal configuration remains preferable for applications requiring maximum robustness and cross-session stability.

4.5. Computational Efficiency Analysis

In addition to predictive performance, computational efficiency was evaluated to assess the feasibility of real-time and embedded deployment. Figure 8 illustrates training dynamics, inference latency distribution, and resource utilization, while quantitative comparisons between TCN + Attn and CNN − LSTM are summarized in Table 7.

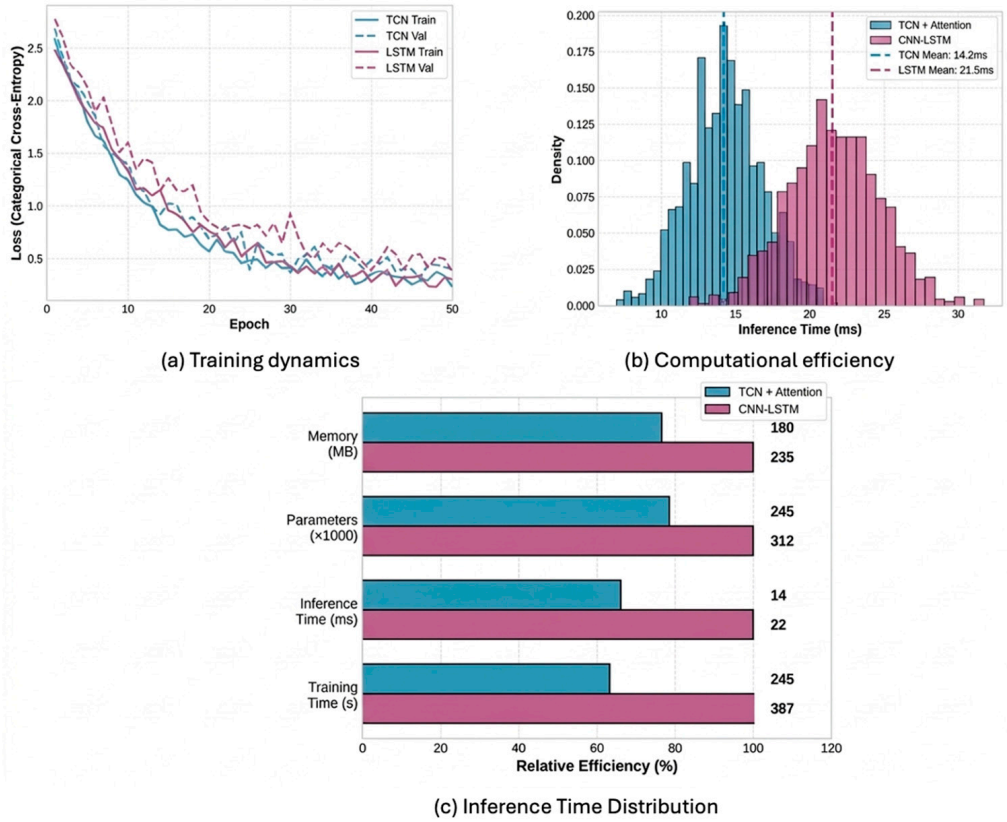


Figure 8. Training dynamics, computational efficiency, and inference time distribution.

This behavior is consistent with the parallelizable nature of convolutional operations in TCN, in contrast to the sequential processing required by recurrent architectures. The reduction in per-epoch training time (245 s vs. 387 s) results in a total 37% reduction in training duration over 50 epochs. This acceleration significantly benefits hyperparameter tuning and experimental iteration cycles. Moreover, TCN + Attn requires substantially less GPU memory during training (4.2 GB vs. 6.8 GB), enabling operation on lower-end GPUs or facilitating larger batch sizes.

Meanwhile, real-time inference performance further highlights the efficiency of the TCN architecture. With a mean inference time of 14.2 ms per sample, TCN + Attn achieves a throughput of 70.4 samples per second, exceeding the 50 Hz sensor sampling rate. This indicates that the model can operate in real time without introducing processing bottlenecks under the evaluated conditions. Latency stability is also improved, with a standard deviation of 1.8 ms for TCN + Attn compared to 3.2 ms for CNN − LSTM, suggesting more predictable inference behavior—an important consideration for real-time monitoring systems.

Table 7. Computational efficiency comparison.

Metric	TCN + Attn	CNN + LSTM	Improvement
Training Metrics			
Time per Epoch	245 s	387 s	36.7% faster
Total Training Time (50 epochs)	12,250 s	19,350 s	37.0% faster
GPU Memory (Training)	4.2 GB	6.8 GB	38.2% less
Convergence Epoch	32	38	15.8% faster
Inference Metrics			
Mean Inference Time	14.2 ms	21.5 ms	33.9% faster
Median Inference Time	13.8 ms	20.9 ms	34.0% faster
95th Percentile Latency	16.7 ms	25.3 ms	34.0% faster
Throughput (samples/s)	70.4	46.5	51.4% higher
Model Complexity			
Parameters	245,128	312,448	21.5% fewer
Memory Footprint (Inference)	180 MB	235 MB	23.4% less
FLOPs per Sample	12.3 M	18.7 M	34.2% fewer

The lower inference memory footprint (180 MB) and reduced parameter count further enhance deployability on embedded platforms such as NVIDIA Jetson Nano (4 GB RAM) or Raspberry Pi 4 (8 GB RAM), supporting real-time multimodal processing under resource constraints. However, the actual deployment performance may vary depending on hardware-specific optimization, runtime frameworks, and system integration constraints.

Energy consumption was theoretically estimated under the assumption of 250 W GPU power draw during training and 50 W during inference. Under these conditions, total training energy consumption is approximately 850 Wh for TCN + Attn compared to 1350 Wh for CNN − LSTM, representing a 58.8% increase for the recurrent model.

During inference, per-sample energy consumption is estimated at 0.197 mJ for TCN + Attn versus 0.298 mJ for CNN − LSTM. These estimates are based on theoretical calculations and should be interpreted as indicative comparisons rather than precise measurements, as actual energy consumption depends on hardware characteristics and system-level implementation.

Figure 9 presents precision–recall (PR) curves for both architectures. TCN + Attn achieves an Average Precision (AP) of 0.801 compared to 0.797 for CNN − LSTM, indicating slightly improved precision–recall trade-offs. Across most recall levels, TCN + Attn maintains comparable or slightly higher precision, suggesting improved calibration and reduced false-positive rates. This characteristic is particularly relevant in real-world monitoring scenarios, where excessive false alarms may reduce system usability.

Overall, TCN + Attn demonstrates improved computational efficiency across multiple dimensions, including training time, inference latency, model complexity, and estimated energy consumption, while maintaining stronger cross-session generalization performance. The fully convolutional structure enables parallel computation and may provide advantages for scalable and real-time deployment, subject to hardware and system-level validation.

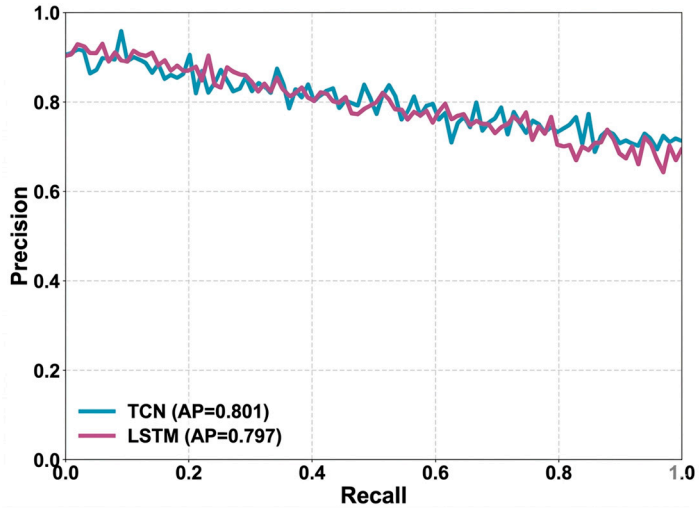


Figure 9. Macro-average precision–recall curves.

5. Discussion

5.1. Architectural Analysis: Cross-Session Robustness of TCN

Although CNN − LSTM achieved slightly higher within-session accuracy (91.34% vs. 89.89%), TCN + Attn demonstrated improved cross-session generalization under LOSO validation (85.23% vs. 82.67%). The smaller generalization gap observed for TCN (4.66% vs. 8.67%) suggests increased robustness to session-level distribution shifts. Several architectural characteristics may contribute to this behavior.

5.1.1. Multi-Scale Temporal Modeling via Dilated Convolutions

TCN employs dilated causal convolutions with dilation rates {1, 2, 4, 8}, yielding an effective receptive field of approximately 90 time steps using 245 k parameters [30]. This structure enables simultaneous modeling of short-term fluctuations, intermediate adjustments, and longer temporal patterns without relying on recurrent state propagation.

Dynamic posture categories—such as periodic shifting and vibration-related behaviors—exhibit hierarchical temporal characteristics, combining slower weight redistribution with higher-frequency oscillatory components. Dilated convolutions are well-suited to capture such multi-scale dependencies [31]. In contrast, recurrent architectures process sequences sequentially and may introduce bias toward recent temporal states, which can affect generalization under varying temporal conditions.

5.1.2. Parallel Temporal Processing and Regularization Effects

LSTM-based architectures process temporal information sequentially, limiting parallelization and increasing computational cost. In this study, CNN − LSTM required longer training time (387 s vs. 245 s per epoch) and exhibited a larger generalization gap under LOSO validation. Fully convolutional temporal processing in TCN enables parallel computation across time steps. This structural property may act as an implicit regularization mechanism, reducing reliance on sequence-specific trajectories and encouraging the learning of more stable and generalizable spatiotemporal representations.

5.1.3. Attention-Based Temporal Emphasis

The self-attention module enables adaptive weighting of temporally informative segments [34]. Qualitative inspection of attention distributions suggests increased emphasis on transition intervals and sensor regions associated with posture changes.

Rather than uniformly processing all time steps, attention mechanisms may enhance robustness by emphasizing discriminative temporal patterns while attenuating less informative fluctuations. However, these observations are based on qualitative analysis, and further quantitative investigation of attention behavior remains an important direction for future work.

5.1.4. Temporal Translation Invariance

Convolutional filters are applied consistently across time positions, introducing temporal translation invariance [30]. For posture recognition tasks, where characteristic patterns may occur at different temporal positions within a window, this property can improve robustness. In contrast, recurrent models may exhibit sensitivity to sequence ordering and temporal alignment, which can contribute to performance degradation under session-wise distribution shifts.

5.2. Modality Contribution Analysis

The ablation study indicates that temperature-only sensing (71.5% LOSO accuracy) outperforms pressure-only (62.3%) and vibration-only (58.7%) configurations. This observation differs from the predominantly pressure-centric focus in the posture recognition literature [3–6]. One possible explanation is that thermal contact patterns evolve gradually and reflect sustained body–surface interactions. Compared to pressure signals, which may fluctuate due to breathing or micro-adjustments, temperature measurements exhibit smoother temporal trajectories. This temporal stability may facilitate more robust feature extraction under cross-session conditions.

Additionally, relative thermal gradients across anatomical regions appear informative for distinguishing postural variations. For example, lateral leaning introduces asymmetric contact distributions, which may be reflected in temperature differences between symmetric sensor locations.

It is important to note that absolute temperature values are influenced by ambient conditions; however, relative thermal dynamics across regions appear more consistent across sessions, potentially contributing to improved generalization. While temperature-only sensing provides substantial discriminative capability, the full multimodal configuration (P + T + V) achieves the highest performance, improving LOSO accuracy by 13.7 percentage points compared to temperature-only configuration. This indicates that multimodal fusion provides complementary information that enhances robustness beyond any single sensing modality.

Vibration sensing contributes to the detection of dynamic behaviors such as periodic shifting and vibration-related activities. Although vibration classes are associated with vibration signals, classification requires spatiotemporal discrimination across multiple sensor locations rather than simple amplitude thresholding. In practice, vibration patterns interact with posture-dependent contact dynamics, meaning that the same vibration input may produce different pressure and temperature responses depending on body posture and contact distribution. This indicates that the model does not rely solely on vibration magnitude but instead learns multimodal spatiotemporal interaction patterns across pressure, temperature, and vibration channels.

Overall, these results suggest that posture recognition in seated environments is not determined by a single sensing modality, but rather by the interaction between spatial pressure distribution, thermal contact dynamics, and vibration-induced micro-movements.

5.3. Cross-Session Generalization Challenges

The LOSO results across the three sessions (87.2%, 84.5%, and 84.0%) indicate the presence of cross-session variability arising from multiple factors. First, ambient tempera-

ture variations across sessions (21–26 °C) introduce distribution shifts in thermal sensor readings. Although z-score normalization reduces baseline differences, residual variability in thermal dynamics may still influence classification consistency. Future work could explore domain adaptation strategies or incorporate ambient temperature as an auxiliary contextual feature to improve robustness.

Second, clothing differences and minor variations in seating position may introduce baseline shifts in both pressure and temperature measurements, affecting contact intensity and spatial distribution patterns. Data augmentation techniques, such as scaling and jittering, partially mitigate this variability. In addition, short calibration phases at the beginning of each session may further improve stability.

Third, long-term sensor drift—particularly in force-sensitive resistive pressure sensors—may affect measurement reliability over extended deployments [14]. Periodic recalibration or continual learning strategies may be required to maintain model performance over time [44].

The observed generalization gap of 4.66% (difference between holdout and LOSO accuracy) indicates a moderate sensitivity to session-level distribution shifts. While direct comparison across studies is not strictly equivalent due to differences in datasets, sensor configurations, and validation protocols, the relatively limited performance degradation observed in this study suggests that the combination of multimodal sensing and temporal convolutional modeling may contribute to improved robustness under session-wise variability.

5.4. Practical Deployment Implications

From a computational perspective, the TCN + Attn model achieves a mean inference latency of 14.2 ms and a throughput of approximately 70 samples per second, exceeding the 50 Hz sensor sampling rate. This indicates that real-time processing is feasible under the evaluated experimental conditions. With a memory footprint of approximately 180 MB, the model size is compatible with modern embedded computing platforms capable of supporting lightweight deep learning inference. However, actual deployment performance may vary depending on hardware-specific optimizations, runtime frameworks, and system integration constraints. Despite this, the reported computational characteristics suggest the potential suitability of the model for edge-based posture monitoring systems. Energy consumption was estimated under assumed computational conditions and may vary depending on hardware implementation and optimization.

From an ergonomic monitoring perspective, the achieved 85.23% LOSO accuracy indicates the potential for reliable detection of both static and dynamic sitting behaviors under cross-session conditions within the evaluated experimental setup. This level of performance enables practical applications such as posture awareness feedback, behavioral analytics, and detection of prolonged static sitting. Notably, dynamic posture classes—including periodic shifting and vibration-related behaviors—exhibit high precision, indicating reliable identification of movement-related adjustments. Such capability may be particularly useful for early detection of discomfort-related micro-movements, although further validation in real-world environments is required.

In addition, the proposed smart-skin sensing approach preserves user privacy by avoiding image or video capture. All sensing is performed through contact-based measurements, which may improve user acceptance compared to camera-based monitoring systems, particularly in workplace and personal environments.

5.5. Limitations

Despite the encouraging results, several limitations should be acknowledged. First, the dataset was collected from a single participant, which limits population-level generalization across diverse body morphologies, age groups, and body mass indices. Variations in height, weight distribution, and sitting habits may influence pressure and thermal patterns. Therefore, the results of this study should be interpreted as within-subject cross-session generalization rather than inter-subject generalization. Future studies should incorporate multi-subject datasets and adopt Leave-One-Subject-Out validation to evaluate population-level generalization.

Second, data acquisition was conducted in a controlled laboratory environment. Although cross-session validation was applied, real-world environments may introduce additional variability, including differences in chair types, clothing, ambient conditions, and environmental disturbances. Field validation in operational settings is therefore necessary to establish ecological validity.

Third, the posture taxonomy was limited to nine classes representing common sitting behaviors. While sufficient for initial ergonomic monitoring, this taxonomy does not cover more complex or transitional postures (e.g., crossed legs, perching, or asymmetric slouching). Expanding the classification framework would improve practical coverage but would require larger datasets and potentially more advanced modeling strategies.

Fourth, the data collection period spanned only seven days, limiting the ability to assess long-term effects such as sensor drift, behavioral adaptation, and seasonal environmental variations. Longitudinal studies are required to evaluate sensor stability and model reliability over extended periods.

Fifth, posture labels were assigned based on experimenter supervision and participant compliance without objective motion-capture validation. This may introduce label uncertainty, particularly during subtle posture transitions. Future work should incorporate motion-capture systems or synchronized video annotation to improve labeling reliability and assess inter-rater agreement.

Sixth, the holdout evaluation used overlapping sliding windows, which may introduce temporal correlation between training and testing samples and lead to optimistic performance estimates. Therefore, the LOSO cross-validation results are considered a more realistic indicator of model generalization across sessions.

Finally, computational efficiency metrics were evaluated on a high-performance GPU platform. Although inference latency suggests potential suitability for embedded deployment, hardware-specific benchmarking and optimization techniques (e.g., quantization or pruning) are required to validate real-time performance on resource-constrained devices.

Overall, these limitations highlight that, while the proposed framework demonstrates promising performance under controlled experimental conditions, the results should be interpreted as a proof-of-concept study, and further validation across subjects, environments, and hardware platforms is required to fully establish real-world applicability.

5.6. Positioning Within the State of the Art

Several prior studies report classification accuracies exceeding 90%, often obtained under within-session or random cross-validation protocols. While such evaluation strategies provide useful baseline comparisons, they may not fully reflect performance under temporal or contextual distribution shifts encountered in real-world scenarios. In contrast, session-wise validation offers a more conservative and practically relevant estimate of deployment performance.

In this context, the 85.23% LOSO accuracy achieved in this study reflects cross-session generalization under temporal variability within a controlled single-subject setting. Rather

than optimizing performance under random data splits, this work emphasizes robustness to session-level distribution shifts.

This study contributes to the field in several aspects. First, it integrates pressure, temperature, and vibration sensing within a unified multimodal smart-skin framework, enabling the capture of complementary spatial, thermal, and dynamic interaction patterns. Second, it applies TCN-based temporal modeling with attention to posture recognition. While CNN − LSTM achieved slightly higher Macro-F1 under LOSO evaluation, TCN + Attn provided higher cross-session accuracy and significantly better computational efficiency, indicating a more favorable trade-off between generalization performance and computational cost. Third, it adopts a rigorous Leave-One-Session-Out validation protocol to provide realistic performance estimates under temporal variability. Fourth, it conducts a structured modality-level ablation study to quantify the relative contributions of each sensing modality and their combinations. Finally, it includes a comprehensive evaluation of computational efficiency, supporting the feasibility of real-time implementation.

Overall, this work advances multimodal posture recognition by emphasizing cross-session robustness, modality-aware analysis, and deployment-oriented evaluation, while acknowledging the need for further validation across subjects and real-world environments.

6. Conclusions

This study presented a multimodal smart-skin framework for sitting posture recognition, integrating pressure, temperature, and vibration sensing. Under Leave-One-Session-Out (LOSO) validation across 15,037 samples, the proposed system achieved 85.23% cross-session accuracy, providing a realistic estimate of temporal generalization performance under controlled conditions. These results should be interpreted as a controlled, single-subject proof-of-concept demonstrating the feasibility of multimodal smart-skin sensing for posture recognition.

Several key findings emerge. First, multimodal sensing improves robustness compared to single-modality configurations, yielding a 22.93% improvement over pressure-only sensing. Second, the TCN + Attn architecture demonstrates stronger cross-session generalization than CNN − LSTM (+2.56% LOSO accuracy) while reducing training time by 36.7% and inference latency by 33.9%, highlighting the efficiency of convolution-based temporal modeling. Although CNN − LSTM achieved a slightly higher Macro-F1 score under LOSO evaluation, TCN + Attn provided higher cross-session accuracy and significantly better computational efficiency, indicating a more favorable trade-off between generalization performance and computational cost. Third, ablation analysis identifies temperature sensing as the most discriminative single modality (71.5% LOSO accuracy), suggesting that thermal contact dynamics provide complementary information beyond pressure-based measurements. Finally, the observed inference latency (14.2 ms) and memory footprint (180 MB) indicate the potential feasibility of real-time processing in edge-based monitoring scenarios.

Beyond predictive performance, the proposed framework supports privacy-preserving ergonomic monitoring by avoiding camera-based sensing. The smart-skin approach enables continuous posture assessment using contact-based measurements, which may improve user acceptance in practical applications.

Despite these promising results, several limitations remain. The dataset was collected from a single participant, limiting anthropometric generalization. Validation in multi-subject and real-world environments is required to assess broader robustness. In addition, the posture taxonomy was limited to nine classes, and extended longitudinal studies are needed to evaluate sensor drift and long-term behavioral adaptation. Future work should also incorporate objective motion-based ground truth to improve labeling reliability.

Future research will focus on multi-subject validation, domain adaptation for improved environmental robustness, optimized sensor placement strategies, and scalable edge or federated learning approaches for privacy-preserving deployment. These steps are necessary to further validate the proposed framework under real-world deployment conditions.

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

Funding: This research was funded by BRIN RIIM, grant number 144/KS/IV/07/2025 and 236/SAM4/PPM/2025. The APC was funded by Directorate of Research and Community Service Telkom University.

Institutional Review Board Statement: The study protocol was approved by the Institutional Review Board of Telkom University. The approval code is 288/LIT07/PPM-LIT/2025, and the approval date is 1 August 2025.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to acknowledge the support provided by the Directorate of Research and Community Service (Direktorat PPM), Telkom University, and the Center of Excellence STAS-RG (Smart Technology and Applied Sciences—Rapid Research Generator), Telkom University. Their administrative, technical, and infrastructural support substantially facilitated the experimental setup, data acquisition, and overall research activities associated with this study.

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

References

1. Zemp, R.; Taylor, W.R.; Lorenzetti, S. Are pressure measurements effective in the assessment of office chair comfort/discomfort? A review. *Appl. Ergon.* **2015**, *48*, 273–282. [CrossRef] [PubMed]
2. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1297–1304. [CrossRef]
3. Ma, S.; Cho, W.-H.; Quan, C.-H.; Lee, S. A sitting posture recognition system based on 3 axis accelerometer. In Proceedings of the 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Chiang Mai, Thailand, 5–7 October 2016; pp. 1–3. [CrossRef]
4. Roh, J.; Park, H.; Lee, K.; Hyeong, J.; Kim, S.; Lee, B. Sitting Posture Monitoring System Based on a Low-Cost Load Cell Using Machine Learning. *Sensors* **2018**, *18*, 208. [CrossRef] [PubMed]
5. Zemp, R.; Rhiner, J.; Plüss, S.; Taylor, W.R.; Lorenzetti, S. Application of Machine Learning Approaches for Classifying Sitting Posture Based on Force and Acceleration Sensors. *Biomed Res. Int.* **2016**, *2016*, 5978489. [CrossRef]
6. Meyer, J.; Arnrich, B.; Schumm, J.; Troster, G. Design and Modeling of a Textile Pressure Sensor for Sitting Posture Classification. *IEEE Sens. J.* **2010**, *10*, 1391–1398. [CrossRef]
7. Mutlu, B.; Krause, A.; Forlizzi, J.; Guestrin, C.; Hodgins, J. Robust, low-cost, non-intrusive sensing and recognition of seated postures. In Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, Newport, RI, USA, 7–10 October 2007; pp. 149–158. [CrossRef]
8. Griffiths, E.; Saponas, T.S.; Brush, A.J.B. Health chair implicitly sensing heart and respiratory rate. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13–17 September 2014; pp. 661–671. [CrossRef]
9. Zhang, X.; Fan, J.; Peng, T.; Zheng, P.; Zhang, X.; Tang, R. Multimodal data-based deep learning model for sitting posture recognition toward office workers' health promotion. *Sens. Actuators A Phys.* **2023**, *350*, 114150. [CrossRef]

10. Liang, G.; Cao, J.; Liu, X. Smart cushion: A practical system for fine-grained sitting posture recognition. In Proceedings of the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kona, HI, USA, 13–17 March 2017; pp. 419–424. [CrossRef]
11. Sundaram, S.; Kellnhofer, P.; Li, Y.; Zhu, J.-Y.; Torralba, A.; Matusik, W. Learning the signatures of the human grasp using a scalable tactile glove. *Nature* **2019**, *569*, 698–702. [CrossRef] [PubMed]
12. Zhao, M.; Beurier, G.; Wang, H.; Wang, X. Driver posture monitoring in highly automated vehicles using pressure measurement. *Traffic Inj. Prev.* **2021**, *22*, 278–283. [CrossRef]
13. Zeng, W.; Shu, L.; Li, Q.; Chen, S.; Wang, F.; Tao, X. Fiber-Based Wearable Electronics: A Review of Materials, Fabrication, Devices, and Applications. *Adv. Mater.* **2014**, *26*, 5310–5336. [CrossRef]
14. Anwary, A.R.; Cetinkaya, D.; Vassallo, M.; Bouchachia, H. Smart-Cover: A real time sitting posture monitoring system. *Sens. Actuators A Phys.* **2021**, *317*, 112451. [CrossRef]
15. Wu, X.; Liang, Y.; Lu, L.; Yang, S.; Liang, Z.; Liu, F.; Lu, X.; Xiao, B.; Zhong, Y.; Xie, Y. Fabric-Based Flexible Pressure Sensor Arrays with Ultra-Wide Pressure Range for Lower Limb Motion Capture System. *Research* **2025**, *8*, 0835. [CrossRef]
16. Ren, Y.; Sun, X.; Liu, J. Advances in Liquid Metal-Enabled Flexible and Wearable Sensors. *Micromachines* **2020**, *11*, 200. [CrossRef] [PubMed]
17. Tsai, M.-C.; Chu, E.T.-H.; Lee, C.-R. An Automated Sitting Posture Recognition System Utilizing Pressure Sensors. *Sensors* **2023**, *23*, 5894. [CrossRef]
18. Xu, Y.; He, P.; Hu, H. Sitting posture detection system based on pressure sensors. In *Proceedings of the Seventh International Conference on Advanced Electronic Materials, Computers, and Software Engineering (AEMCSE 2024)*; SPIE: Bellingham, WA, USA, 2024; p. 16. [CrossRef]
19. Sri-ngernyuang, C. Evaluating Sitting Posture Using Pressure Sensors: A Lightweight Cnn Approach with Psychometric and Ergonomic Perspectives. *Test. Psychom. Methodol. Appl. Psychol.* **2025**, *32*, 421–433.
20. Chen, H.; Li, Y.; Fang, H.; Xin, W.; Lu, Z.; Miao, Q. Multi-Scale Attention 3D Convolutional Network for Multimodal Gesture Recognition. *Sensors* **2022**, *22*, 2405. [CrossRef] [PubMed]
21. Ordóñez, F.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* **2016**, *16*, 115. [CrossRef]
22. Ha, S.; Choi, S. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 381–388. [CrossRef]
23. Kang, Q.; Chen, E.J.; Li, Z.-C.; Luo, H.-B.; Liu, Y. Attention-based LSTM predictive model for the attitude and position of shield machine in tunneling. *Undergr. Sp.* **2023**, *13*, 335–350. [CrossRef]
24. Murad, A.; Pyun, J.-Y. Deep Recurrent Neural Networks for Human Activity Recognition. *Sensors* **2017**, *17*, 2556. [CrossRef]
25. Poulose, A.; Senouci, B.; Han, D.S. Performance Analysis of Sensor Fusion Techniques for Heading Estimation Using Smartphone Sensors. *IEEE Sens. J.* **2019**, *19*, 12369–12380. [CrossRef]
26. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [CrossRef]
27. Wang, B.; Tu, L.; Shi, Q.-Q.; Mo, S. Deep Learning-Based Sitting Posture Recognition from Pressure Distribution Across Hard and Soft Seat Environment. *Appl. Sci.* **2025**, *15*, 12744. [CrossRef]
28. Rashmi, M.; Guddeti, R.M.R. Multi-stream Multi-attention Deep Neural Network for Context-Aware Human Action Recognition. In Proceedings of the 2022 IEEE Region 10 Symposium (TENSYMP), Mumbai, India, 1–3 July 2022; pp. 1–6. [CrossRef]
29. Ding, Z.; Li, W.; Yang, J.; Ogunbona, P.; Qin, L. An attention-based CNN for automatic whole-body postural assessment. *Expert Syst. Appl.* **2024**, *238*, 122391. [CrossRef]
30. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]
31. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal Convolutional Networks for Action Segmentation and Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1003–1012. [CrossRef]
32. Ingolfsson, T.M.; Wang, X.; Hersche, M.; Burrello, A.; Cavigelli, L.; Benini, L. ECG-TCN: Wearable Cardiac Arrhythmia Detection with a Temporal Convolutional Network. In Proceedings of the 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), Washington, DC, USA, 6–9 June 2021; pp. 1–4. [CrossRef]
33. Guan, Y.; Plötz, T. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2017**, *1*, 1–28. [CrossRef]
34. Wei, X.; Wang, Z. TCN-attention-HAR: Human activity recognition based on attention mechanism time convolutional network. *Sci. Rep.* **2024**, *14*, 7414. [CrossRef] [PubMed]

35. Bulling, A.; Blanke, U.; Schiele, B. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* **2014**, *46*, 1–33. [CrossRef]
36. Aggarwal, J.K.; Ryoo, M.S. Human activity analysis. *ACM Comput. Surv.* **2011**, *43*, 1–43. [CrossRef]
37. Reiss, A.; Stricker, D. Introducing a New Benchmarked Dataset for Activity Monitoring. In Proceedings of the 2012 16th International Symposium on Wearable Computers, Newcastle, UK, 18–22 June 2012; pp. 108–109. [CrossRef]
38. Zhang, Q.; Jiang, F.; Wang, X.; Duan, J.; Wang, X.; Ma, N.; Zhang, Y. A Novel Method for Cross-Subject Human Activity Recognition with Wearable Sensors. *J. Sens. Technol.* **2024**, *14*, 17–34. [CrossRef]
39. Yang, J.; Liao, T.; Zhao, J.; Yan, Y.; Huang, Y.; Zhao, Z.; Xiong, J.; Liu, C. Domain Adaptation for Sensor-Based Human Activity Recognition with a Graph Convolutional Network. *Mathematics* **2024**, *12*, 556. [CrossRef]
40. Cook, D.J.; Crandall, A.S.; Thomas, B.L.; Krishnan, N.C. CASAS: A Smart Home in a Box. *Computer* **2013**, *46*, 62–69. [CrossRef]
41. Hammerla, N.Y.; Halloran, S.; Plötz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), New York, NY, USA, 9–15 July 2016; pp. 1533–1540.
42. Shen, J.; Chen, L.; He, X.; Zuo, C.; Li, X.; Dong, L. An Interactive Human-in-the-Loop Framework for Skeleton-Based Posture Recognition in Model Education. *Biomimetics* **2025**, *10*, 0431. [CrossRef]
43. Walid Talha, S.A.; Fleury, A.; Lecoeuche, S. Hierarchical classification scheme for real-time recognition of physical activities and postural transitions using smartphone inertial sensors. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 1243–1246. [CrossRef]
44. Li, Z.; Hoiem, D. Learning without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2935–2947. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.