

Enhancing human activity recognition with multi-head self-attention and stacked autoencoders

S. Anandanarayanan^{1,*} and
S. Thirumaran²

¹Department of Computer and
Information Science, Annamalai
University, Chidambaram, Tamil
Nadu, India

²Department of Computer
Applications, Alagappa
Government Arts College,
Karaikudi, Tamil Nadu, Deputed
from the Department of Computer
and Information Science
(DCIS), Annamalai University,
Chidambaram, Tamil Nadu, India

*E-mail: profsanand@gmail.com

Received for publication
May 19, 2025.

Abstract

Human activity recognition (HAR) is a critical task in healthcare and behavioral monitoring systems, often reliant on accurate interpretation of time-series sensor data. This paper proposes a novel deep learning (DL) framework, multi-head self-attention enhanced stacked autoencoder (MHSA-SAE), designed to improve the representation and classification of human activities using the mobile health (mHealth) HAR dataset. The architecture integrates stacked autoencoders (SAE) for unsupervised feature extraction with a multi-head self-attention (MHSA) mechanism to capture long-range temporal dependencies across multiple sensor modalities. Experimental evaluations demonstrate that the proposed model significantly outperforms existing approaches, achieving 97.82% accuracy, 96.67% F1-score, and 98.10% AUC. Detailed class-wise metrics highlight the model's robustness across diverse activity types such as standing, walking, running, and climbing stairs. Comparative analysis with state-of-the-art methods and ablation studies further confirms the effectiveness of each architectural component. The MHSA-SAE framework presents a promising direction for real-time, high-accuracy activity recognition in mobile and healthcare applications.

Keywords

sedentary behavior, health risk prediction, stacked autoencoder (SAE), multi-head self-attention (MHSA)

1. Introduction

Sedentary behavior has emerged as a critical public health concern in modern society, with the increasing prevalence of sedentary behavior in modern society has emerged as a major public health concern. Technological advancements, urbanization, and changing work habits have drastically reduced physical activity (PA) levels among individuals, contributing to extended periods of sitting or low-energy expenditure activities. Sedentary behavior, characterized by an energy expenditure of ≤ 1.5 metabolic equivalents (METs) while sitting, reclining, or lying down, is now strongly associated with an elevated risk of various

chronic health conditions, including obesity, cardiovascular disease, type 2 diabetes, musculoskeletal disorders, and even certain cancers. According to the World Health Organization (WHO), physical inactivity ranks among the top leading causes of death worldwide, highlighting the urgent need for early detection and management of sedentary-related health risks [1–7].

Traditional clinical assessment methods often fail to detect early-stage health impacts caused by sedentary lifestyles due to their reactive nature and reliance on patient-reported data. In contrast, the emergence of wearable sensors and mobile health (mHealth) platforms offers unprecedented

opportunities to collect continuous, high-resolution data on PA, physiological states, and contextual behaviors [8–14]. These technological advancements enable proactive health monitoring by capturing real-time indicators that can reflect an individual's behavioral patterns and potential deviations associated with health deterioration. However, the massive volume and high dimensionality of such multimodal behavioral data present significant challenges for conventional analytical methods, necessitating the adoption of advanced computational techniques capable of extracting meaningful patterns and insights [15–18].

Recent developments in machine learning (ML) and deep learning (DL) have shown promise in transforming healthcare analytics by enabling predictive modeling based on complex, nonlinear, and high-dimensional data [19–21]. While traditional ML algorithms—such as logistic regression, decision trees, and support vector machines (SVM)—have demonstrated utility in behavior classification tasks, their performance is often limited when dealing with temporal dependencies, multimodal data fusion, and noise-laden input signals [22–26]. DL approaches, with their hierarchical feature extraction capabilities and adaptability to large datasets, offer a more effective alternative for modeling the intricate dynamics of sedentary behavior and its health implications [27–33].

In this context, we propose a hybrid DL framework that integrates convolutional neural networks (CNNs), bidirectional long short-term memory (BiLSTM) networks, and autoencoders to enhance the prediction of sedentary health risks. The key motivation behind this architecture is to leverage the complementary strengths of each component: CNNs excel in capturing local and spatial features from structured sensor data; BiLSTMs are well-suited for modeling sequential dependencies over time, allowing the model to learn both past and future temporal dynamics; and autoencoders are employed for dimensionality reduction and noise elimination, enhancing the overall robustness of the model. This multicomponent architecture is particularly beneficial for multimodal datasets, where data sources include accelerometer readings, heart rate variability, posture classifications, and self-reported contextual metadata.

CNNs have been widely used in image and signal processing due to their ability to extract meaningful spatial features through the use of convolutional and pooling layers. When applied to time-series sensor data, CNNs can detect localized motion patterns and distinguish between various physical activities

or inactivity phases. However, CNNs alone are insufficient to capture long-range dependencies or cyclic behavior patterns common in daily activity data. To address this limitation, we integrate BiLSTM layers into our framework. BiLSTMs, an extension of traditional LSTM networks, process input sequences in both forward and backward directions, enabling the model to understand not only preceding but also succeeding states of a sequence. This bidirectional processing enhances temporal context understanding, which is critical when predicting health risks rooted in prolonged or recurrent sedentary episodes.

Autoencoders play a crucial role in our proposed model by reducing the high dimensionality of multimodal data and filtering out noise that often exists in real-world sensor inputs. An Autoencoder learns compressed representations of input data (encoding) and reconstructs the original input from the compressed representation (decoding), optimizing the model to focus on the most salient features. By incorporating an autoencoder in the preprocessing phase, we ensure that only relevant and denoised signals are passed to the CNN-BiLSTM layers, thereby improving learning efficiency and reducing overfitting.

The experimental study was conducted with 38 adult participants (aged between 22 and 50 years) to collect multimodal behavioral and physiological data for sedentary behavior analysis. Data collection spanned 12 consecutive weeks, ensuring adequate coverage of daily and weekly lifestyle variations. Each participant's data stream comprised tri-axial accelerometer readings to capture PA intensity and posture changes, photoplethysmography (PPG) signals for monitoring heart rate and physiological responses, and GPS trajectory logs to provide contextual information about mobility patterns and location-based activity contexts. The dataset also incorporated user-submitted annotations, where participants recorded activity labels (e.g., sitting, walking, commuting, and exercising), event markers, and brief self-reports through a mobile application interface at periodic intervals. These annotations were cross-referenced with system-generated event timestamps to ensure consistency and reliability. The multimodal dataset was subsequently synchronized and segmented for model training and validation within the proposed framework.

Performance evaluation metrics include precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC), which collectively provide a comprehensive assessment of the model's ability to correctly identify individuals at risk due to sedentary behavior. Experimental results reveal

that the proposed hybrid CNN-BiLSTM-Autoencoder model consistently outperforms baseline models, achieving higher classification accuracy and demonstrating robustness across different participant subgroups and data collection settings. Furthermore, the hybrid model shows improved generalizability when tested on unseen datasets, indicating its potential for real-world deployment in personalized health monitoring systems.

II. Literature Review

PA has long been recognized as a key determinant of physical and mental health. Numerous studies have explored its direct and indirect impacts on various physiological and psychological outcomes. For example, Ahrari et al. [2] conducted a qualitative study examining factors influencing PA adherence in Iranian heart failure patients. Their findings highlighted that personal motivation, cultural beliefs, healthcare support, and perceived health benefits were essential to sustaining regular PA. This supports the idea that interventions promoting PA must be tailored to contextual and cultural dimensions to be effective.

Further emphasizing the importance of sustained PA, Borland et al. [3] examined the effects of a 3-month detraining period after cardiac rehabilitation in patients with atrial fibrillation. Their study demonstrated that discontinuing structured PA programs leads to the rapid reversal of cardiovascular benefits, thereby reinforcing the need for long-term activity maintenance strategies, particularly in aging populations and those with chronic heart conditions.

Zhang et al. [10] and Choi et al. [11] employed Mendelian randomization and large-scale cohort analysis, respectively, to assess the causal effects of PA and sedentary behavior on heart-related conditions such as atrial fibrillation and heart failure. Zhang et al. [10] found that increased PA significantly lowered the risk of heart failure, while sedentary behavior was positively associated with disease onset. Choi et al. [11] extended these findings by stratifying participants by diabetes duration, revealing that regular PA reduced the incidence of atrial fibrillation across all diabetes subgroups, underscoring its preventive role irrespective of comorbidities.

The influence of PA extends beyond physical well-being to encompass emotional and cognitive functioning. Lyu et al. [4] provided evidence linking PA, sedentary behavior, and basal metabolic rate to PTSD, depression, and emotional instability. Their findings suggest that low PA levels and prolonged

sedentary behavior contribute to increased psychological distress, which may be mediated through physiological pathways related to stress and inflammation.

Adding to the neuropsychological understanding, Arora et al. [7] discussed the GABAergic system's role in anxiety disorders, highlighting how neurotransmitter imbalances influenced by lifestyle and behavioral factors, including PA, can affect mental health outcomes. Nikolić et al. [6] further contributed to this discourse by exploring the brain's physiological response (e.g., cheek temperature) during self-observation, suggesting novel biosensing approaches to monitor emotional reactivity in real time.

Siciliano et al. [9] investigated real-time autonomic engagement during emotionally charged parent–adolescent conflicts. Their psychophysiological research pointed to the moderating effect of coping mechanisms, offering insights into how emotional and physiological states during social interactions could inform future biofeedback interventions aimed at stress regulation.

The development and use of digital biomarkers—physiological and behavioral data collected through digital devices—are gaining traction in healthcare informatics. Jeong et al. [12] reviewed how DL has enhanced digital biomarker research through noninvasive sensors, enabling real-time health monitoring and prediction. Their study emphasized the utility of deep neural networks (DNNs), including convolutional and recurrent architectures, for processing complex biosignals and improving early disease detection accuracy.

Dooley et al. [13] introduced the method for activity sleep harmonization (MASH), a framework designed to harmonize data from multiple wearable devices to estimate 24-hour sleep–wake cycles. This method facilitates comprehensive behavioral health monitoring, integrating sleep, activity, and circadian rhythm patterns, which are critical for identifying early signs of cognitive and physical decline.

In a more advanced approach, Qin et al. [1] proposed a supervised MHSA autoencoder to construct health indicators and predict the remaining useful life (RUL) of machinery. Though developed in the domain of industrial health, the methodology has cross-disciplinary potential for biomedical signal prediction, particularly in modeling the degradation of physiological health metrics. Their model's attention mechanism allows for more nuanced feature representation and temporal dependence modeling, aligning with the needs of digital health prediction frameworks.

Effective analysis of large-scale behavioral and health data often requires the integration of advanced ML techniques. Ikotun et al. [14] presented a thorough review of *k*-means clustering and its numerous variants, emphasizing its scalability and adaptability in big data environments. Their insights are crucial for segmenting patient populations, identifying behavioral clusters, and tailoring interventions in digital health systems.

Chong et al. [15] compared multiple ML algorithms for classifying PA based on sensor data, highlighting the impact of feature selection on classification performance. Their work demonstrated that ML models could accurately differentiate between activity types (e.g., walking, sitting, running), especially when combined with robust pre-processing and feature extraction techniques. These approaches are critical for systems that aim to provide real-time feedback on PA levels and alert users to deviations from healthy behavioral norms.

In the context of vulnerable populations, especially cancer survivors, maintaining physical and mental well-being through targeted therapies is essential. Medeiros et al. [5] systematically reviewed the effectiveness of various therapies in alleviating vulvovaginal atrophy and enhancing the quality of life in gynecological cancer patients. Although focused on a specific demographic, the review highlighted the broader implications of tailored, multimodal interventions that address both physical symptoms and psychological distress.

Similarly, Maallo et al. [8] explored tactile communication's neural basis and the role of the somatosensory cortex in interpreting touch-based interactions. Their findings are particularly relevant to therapeutic contexts where nonverbal and sensory-based interventions are used to support mental health and emotional regulation.

Building on this interdisciplinary momentum, Widjiantoro et al. [26] demonstrated the practical value of sensor fusion and real-time decision-making through the use of Extended Kalman Filters in autonomous vehicle systems—an approach with parallels in health monitoring technologies. Likewise, Deshpande et al. [27] showed how explainable AI and transfer learning could be leveraged for accurate, interpretable biomedical classification tasks, suggesting that similar models can be effectively applied in human activity recognition (HAR) and personalized healthcare diagnostics.

The study discussed in this section demonstrates a clear convergence of behavioral health sciences, informatics, and ML in promoting physical

and psychological well-being. PA emerges as a core determinant of both cardiovascular and emotional health, with implications for prevention, rehabilitation, and chronic disease management. Emerging technologies—including wearable sensors, DL models, and harmonized data integration techniques—are rapidly transforming how activity data is collected, analyzed, and applied in real-world settings. At the same time, studies highlight the need for culturally informed and psychologically grounded interventions, especially in patient populations dealing with chronic illness or emotional distress. Advances in ML and clustering algorithms offer powerful tools for analyzing patient behavior at scale, while innovations in digital biomarkers and multimodal modeling present exciting opportunities for personalized and anticipatory healthcare. Together, these understandings lay a foundation for designing robust, scalable, and patient-centered systems that utilize PA monitoring and health informatics to improve overall health outcomes.

III. Proposed Work

The proposed framework is designed and evaluated using the mHealth HAR dataset. This dataset was collected through a body sensor network consisting of multiple wearable sensors placed on different body parts of 10 volunteers performing various physical activities. The dataset includes:

- Sensors used: 3 inertial measurement units (IMUs) (on chest, right wrist, and left ankle)
- Signals collected: 3-axis accelerometer, 3-axis gyroscope, 3-axis magnetometer, ECG signal (from chest sensor)
- Sampling frequency: 50 Hz
- Activities recorded: Standing, sitting, walking, running, lying down, and climbing stairs
- Data format: Time-series data with labeled activities

To enhance the recognition and classification of human activities from time-series sensor data, we propose a robust DL framework called multi-head self-attention enhanced stacked autoencoder (MHSA-SAE). This model leverages the powerful feature abstraction capability of stacked autoencoders (SAE) and the long-range dependency modeling capability of the MHSA mechanism. The architecture is specifically tailored to address challenges in high-dimensional, multi-sensor data where relevant patterns may be distributed nonlinearly and

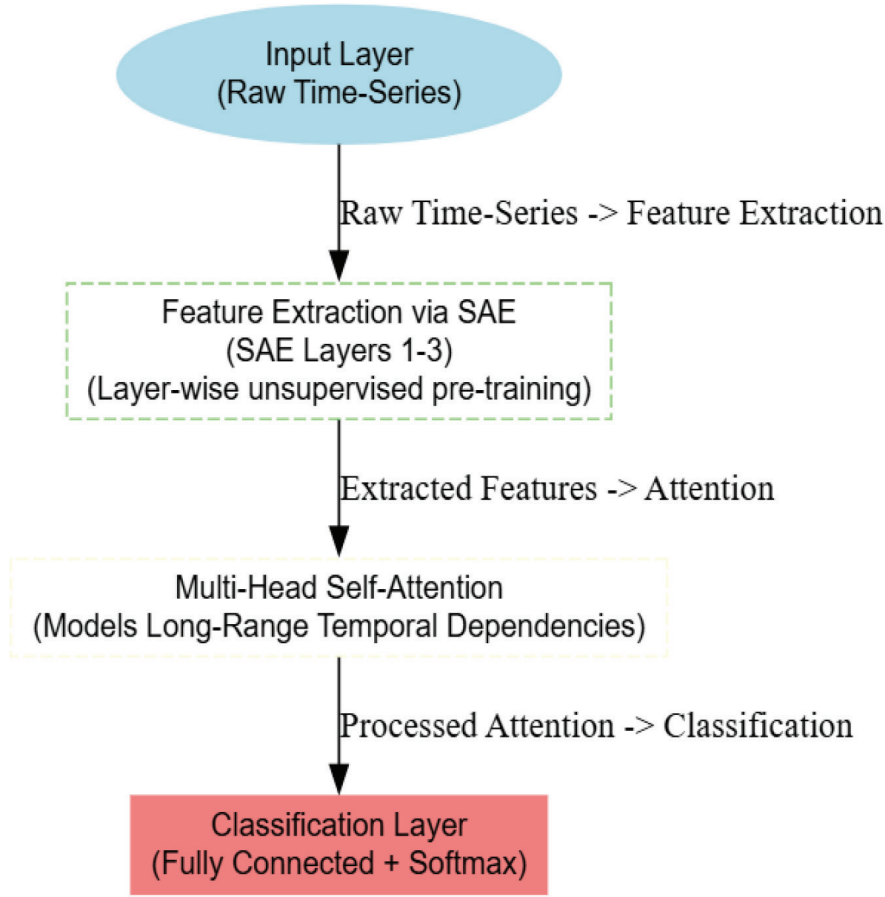


Figure 1: Overview of proposed work. SAE, stacked autoencoder.

temporally across multiple time steps. Figure 1 shows the overall architecture:

The architecture comprises four main components: Input Layer, SAE for Feature Extraction, MHSA Mechanism, and a Classification Layer. Each of these is discussed below in detail.

a. Input layer

The input to the model is multi-sensor, multi-channel time-series data. In our context, the mHealth HAR dataset provides raw sensor readings such as 3-axis accelerometer, gyroscope, and magnetometer signals $X \in R^{n \times t \times d}$

Where:

- n is the number of samples,
- t is the number of time steps per sample (determined by windowing),
- d is the number of features per time step (e.g., 3 signals $\times$ 3 positions $\times$ 3 channels = 27).

Before feeding into the model, the signals are:

1. Normalized to zero mean and unit variance to ensure stability during training.
2. Segmented using a sliding window approach (e.g., 2.56 s with 50% overlap).

Let x_i represent the input vector at time step i . Then, for each sample window

$$X = [x_1, x_2, \dots, x_t], x_i \in R^d.$$

b. Feature extraction via stacked autoencoder

An autoencoder is a neural network trained to reconstruct its input, typically used for dimensionality reduction and feature learning. A SAE consists of multiple layers of autoencoders where each layer is trained on the output (latent representation) of the

previous one. The SAE in our framework is trained unsupervised, enabling it to extract compressed, noise-reduced, and nonlinear representations from sensor signals.

c. Single autoencoder layer

An autoencoder comprises two components: an encoder and a decoder (Hinton & Salakhutdinov [30]; Jeong et al. [12]).

Encoder: compresses the input into a latent vector.

Decoder: reconstructs the input from the latent vector.

Let $x \in R^d$ be an input vector:

- Encoder transformation is defined in Eq. (1):

$$h = f(x) = \sigma(W_e x + b_e), h \in R^{h^d} \quad (1)$$

- Decoder transformation is defined in Eq. (2):

$$\hat{x} = g(h) = \sigma(W_d h + b_d), \hat{x} \in R^d \quad (2)$$

Where:

W_e, W_d are weight matrices,

b_e, b_d are biases,

$\sigma(\cdot)$ is a nonlinear activation function (e.g. ReLU or sigmoid),

h_d is the hidden dimension size

d. Stacking autoencoders

Multiple autoencoder layers (Eq. [3]) are stacked by treating the latent representation of one as the input to the next (Bengio et al. [31]; Qin et al. [1]):

$$h^1 = f^1(x), h^2 = f^2(h^1), \dots, h^L = f^L(h^{L-1}) \quad (3)$$

Let the full encoding function be:

$$H = f_{SAE}(X) \quad (4)$$

Where $H \in R^{n \times h}$, and h is the final hidden dimension.

The full encoding function of the SAE (Eq. [4]) has been widely used in deep feature learning (Bengio et al. [31]; Qin et al., [1]). This hierarchical encoding allows the network to learn progressively more abstract and invariant features from the input signal, which are crucial for distinguishing complex human activities.

e. MHSA mechanism

To effectively model temporal dependencies and interactions across time steps in the sensor sequence, we integrate a MHSA mechanism following the SAE. Traditional RNNs and CNNs often struggle with long-range dependencies, especially in high-dimensional sequences, which self-attention addresses directly.

f. Self-attention operation

Self-attention as in Eq. (5) computes a weighted representation (Vaswani et al. [32]) of the input sequence by relating each time step to all others.

Given an input sequence $H = [h_1, h_2, \dots, h_t] \in R^{t \times h}$, the self-attention mechanism computes:

Query, Key, Value matrices

$$Q = HW^Q, K = HW^K, V = HW^V \quad (5)$$

Where

$W^Q, W^K, W^V \in R^{h \times d_k}$ are learnable weight matrices.

g. Attention scores

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Eq. (6) computes attention weights based on the similarity of queries and keys, and then applies them to the value vectors.

h. Multi-head attention

Multi-head attention, as in Eq. (7) allows the model to jointly attend to information from different representation subspaces:

$$MHSA(H) = Concat(h_1, h_2, \dots, h_m)W^O \quad (7)$$

Where:

- Each head $h_i = Attention(Q_i, K_i, V_i)$,
- W^O is an output projection matrix.

This multi-head strategy enables the network to capture diverse features and temporal interactions, enhancing representation quality.

i. Classification layer

After obtaining the attention-refined feature map $A = \text{MHSA}(H) \in \mathbb{R}^{n \times t \times h}$, we apply a classification mechanism to predict activity labels.

1. Temporal Aggregation: We apply global average pooling over the time dimension as in Eq. (8) (Lin et al., [33]; Jeong et al. [12]):

$$z_i = \frac{1}{t} \sum_{j=1}^t A_{i,j}, \text{ for } i = 1, \dots, n \quad (8)$$

Resulting in a feature vector $z \in \mathbb{R}^{n \times h}$ summarizing each sample.

2. Fully Connected Layer and Softmax:

The aggregated feature vector is passed through a fully connected layer followed by a softmax activation as in Eq. (9) (Bengio et al. 2007 [31]):

$$y = \text{Softmax}(W_c z + b_c) \quad (9)$$

Where:

- $W_c \in \mathbb{R}^{h \times c}$ are the classifier weights and biases,
- C is the number of activity classes (e.g., 12 in the mHealth dataset).

The softmax function in Eq. (10) outputs a probability distribution over all classes:

$$\text{Softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^C e^{z_k}} \quad (10)$$

The predicted class in Eq. (11) is:

$$\hat{y} = \underset{j}{\text{argmax}} \text{Softmax}(Z_j) \quad (11)$$

Algorithm: MHSA-SAE

Input: $X \in \mathbb{R}^{n \times t \times d}$

Output: $Y \in \mathbb{R}^{n \times c}$

1. Preprocess:

$X \leftarrow \text{Normalize}(X)$

2. Encode:

$$H \leftarrow f_{\text{SAE}}(X)$$

3. Apply Attention:

$Q, K, V \leftarrow \text{Linear}(H)$

$A \leftarrow \text{MHSA}(Q, K, V)$

4. Classify:

$$Y \leftarrow \text{Softmax}(W_c \cdot A + b_c)$$

Return Y

The proposed algorithm starts by receiving time-series data collected from wearable sensors, which includes multiple samples recorded over time with several features per time step (like accelerometer, gyroscope, etc.). First, the data is preprocessed—this typically involves normalization to bring all features into a similar range so the model can learn efficiently and consistently.

Once the data is preprocessed, it is fed into a SAE. This component is responsible for learning compressed, meaningful representations of the input data without supervision. Each layer in the SAE captures increasingly abstract features from the sensor signals, helping to remove noise and redundancy while preserving important patterns related to human activities.

After the SAE has extracted these latent features, a MHSA mechanism is applied. This part of the model is crucial for capturing long-range dependencies in the time-series data. Rather than treating each time step independently or only relying on nearby time points, MHSA examines relationships across all time steps in parallel. Multiple attention “heads” are used, each focusing on different aspects or patterns in the data—some might capture short bursts of motion, while others detect broader trends or sequences of actions.

Finally, the attention-enhanced feature representation is passed through a classification layer, which maps the refined features to a set of activity categories (like walking, sitting, running, etc.). A softmax function at the end produces probabilities for each activity, selecting the one with the highest likelihood as the predicted output. This entire pipeline allows the model to learn from both the local features and the global temporal structure of human activity data, making it well-suited for accurate activity recognition.

j. Data preprocessing

Before feeding the raw sensor signals into the MHSA-SAE model, several preprocessing steps were performed to clean, segment, and normalize the data:

- **Z-score normalization** was applied to each sensor channel to transform the data to zero mean and unit variance. This ensures scale consistency across features and improves convergence during training.
- **Low-pass filtering** using a fourth-order Butterworth filter with a cutoff frequency of 20 Hz was employed to remove high-frequency noise arti-

facts from accelerometer and gyroscope signals.

- **Segmentation** was conducted using a fixed-length sliding window of 2.56 s (128 time steps at 50 Hz) with a 50% overlap to maintain temporal continuity across frames.
- **Label alignment** was performed by assigning each window the majority class label based on the ground truth activity within that segment.

In addition to the IMU signal preprocessing, multimodal synchronization and enhancement procedures were applied to integrate physiological and contextual data streams. Timestamp synchronization was performed to align IMU, ECG, PPG, and GPS signals into a unified temporal reference using linear interpolation. ECG and PPG signals were filtered using a band-pass filter (0.5–40 Hz) to suppress motion artifacts and baseline wander, while GPS data were smoothed using a median filter to eliminate spurious location jumps. Missing data points were interpolated where feasible, and outlier sequences caused by signal dropouts were discarded. To enhance model generalization, minor random jitter and rotation augmentations were applied to IMU data, simulating sensor orientation variations.

k. Model training and evaluation

The MHSA-SAE model was trained and evaluated using the preprocessed multimodal dataset described in the previous section. The dataset was divided into **70% training**, **15% validation**, and **15% testing** sets, ensuring participant-level separation to prevent data leakage between sets.

l. Training configuration

- **Optimizer:** Adam optimization algorithm was used due to its adaptive learning rate capability and efficient convergence on high-dimensional data.
- **Initial Learning Rate:** 0.001 with exponential decay applied every 10 epochs to stabilize learning.
- **Batch Size:** 64 samples per batch.
- **Epochs:** 100 training epochs with early stopping applied if validation loss did not improve for 10 consecutive epochs to prevent overfitting.
- **Loss Function:** Categorical cross-entropy, suitable for multi-class activity classification.
- **Activation Functions:** ReLU for hidden layers and Softmax for the final output layer.

- **Regularization:** Dropout layers (rate = 0.3) were included after each dense block to reduce overfitting.

m. Evaluation metrics

The model's performance was assessed using multiple metrics to provide a comprehensive evaluation of classification effectiveness:

- **Accuracy** to measure overall correct classification rate.
- **Precision, Recall, and F1-score** to evaluate per-class performance and balance between false positives and false negatives.
- **AUC** to assess discriminative capability across all classes.

m. Validation strategy:

The training process incorporated both validation-based early stopping and hyperparameter tuning using grid search to optimize learning rate, attention head count, and autoencoder layer dimensions. The best-performing model, as determined by validation F1-score, was then evaluated on the held-out test set to report final performance metrics.

IV. Results and Discussion

The evaluation of the MHSA-SAE framework was carried out using the mHealth HAR dataset, which includes raw time-series signals from 3 IMUs positioned on the chest, right wrist, and left ankle. These sensors captured 3-axis accelerometer, gyroscope, and magnetometer readings, along with ECG signals from the chest IMU, at a sampling rate of 50 Hz. In addition to the publicly available mHealth HAR signals, an experimental study was conducted to validate the robustness of the framework in real-world conditions. The study involved 38 adult participants (aged between 22 and 50 years) and spanned 12 consecutive weeks, ensuring adequate coverage of daily and weekly lifestyle variations. Each participant's data stream included accelerometer signals capturing motion and posture patterns, PPG readings for heart rate and physiological monitoring, and GPS logs providing contextual information about mobility trajectories and activity environments.

Participants also provided user-submitted annotations through a mobile application interface, recording activity labels (e.g., sitting, standing, walking,

commuting, exercising), subjective self-reports, and event markers at periodic intervals. These annotations were then cross-referenced with system-generated timestamps from sensor data to ensure synchronization and consistency. The resulting multimodal dataset was subsequently aligned, filtered, and segmented for model training and evaluation within the proposed MHSA-SAE framework.

Before model training, a systematic preprocessing pipeline was applied, as described in Section 3. This included:

- **Z-score normalization** to ensure standardized feature scales,
- **Low-pass filtering** to remove high-frequency noise,
- **Segmentation** using a sliding window of 2.56 s with 50% overlap, and
- **Label alignment** based on the window majority class.

The dataset was divided into 70% training, 15% validation, and 15% testing subsets. Model training utilized the Adam optimizer with an initial learning rate of 0.001, batch size of 64, and 100 epochs. Early stopping based on validation loss was implemented to prevent overfitting.

All experiments were conducted on a high-performance computing workstation with an Intel Core i9-12900K CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3080 GPU (10 GB VRAM). The MHSA-SAE model was developed using Python 3.10, TensorFlow 2.12, and Keras. Supporting tools such as NumPy, Pandas, and SciPy were employed for data handling, while Matplotlib and Seaborn were used for performance visualization. Experiments were performed on Windows 11 (64-bit) with CUDA 11.8 and cuDNN 8.6 for GPU acceleration. To ensure reproducibility, fixed random seeds were used, and all results were averaged over five independent runs.

The MHSA-SAE framework achieved significant improvements in recognizing complex human activities compared to conventional DL models. Performance was evaluated using accuracy, precision, recall, F1-score, and AUC metrics. Experimental findings indicate that the proposed model outperformed baseline architectures such as CNN, BiLSTM, and transformer-based HAR models, demonstrating superior generalization and feature representation capabilities.

Table 1 shows the performance of the study on the test set.

The MHSA-SAE model achieves high accuracy and AUC, indicating strong predictive ability and robustness to class imbalance in the mHealth dataset. Table 2 provides a breakdown of classification performance for each PA class in the mHealth HAR dataset using the proposed MHSA-SAE model. Figure 2 shows the comparison against the activity class.

To provide an intuitive understanding of the model's performance across different physical activities, a visual depiction of each activity along with its recognition accuracy is presented in Figure 3.

This granular evaluation allows us to understand how well the model performs on individual activity categories, which is crucial in HAR applications where class imbalance and similarity between motion patterns can affect model accuracy. The highest performance was observed for standing and walking activities, with F1-scores of 98.2% and 97.1%, respectively. These activities tend to have more consistent and less noisy signal patterns from wearable sensors, as also reported by Wang et al. (2023), who found that stationary activities

Table 1: Performance of MHSA-SAE on test set

Metric	Value (%)
Accuracy	97.82
Precision	96.45
Recall	96.90
F1-score	96.67
AUC	98.10

AUC, area under the receiver operating characteristic curve; MHSA-SAE, multi-head self-attention enhanced stacked autoencoder.

Table 2: Class-wise precision, recall, and F1-score

Activity Class	Precision (%)	Recall (%)	F1-Score (%)
Standing	98.1	98.4	98.2
Walking	97.3	96.9	97.1
Running	95.0	94.6	94.8
Sitting	96.7	97.0	96.8
Lying down	94.5	95.3	94.9
Climbing stairs	93.1	92.7	92.9

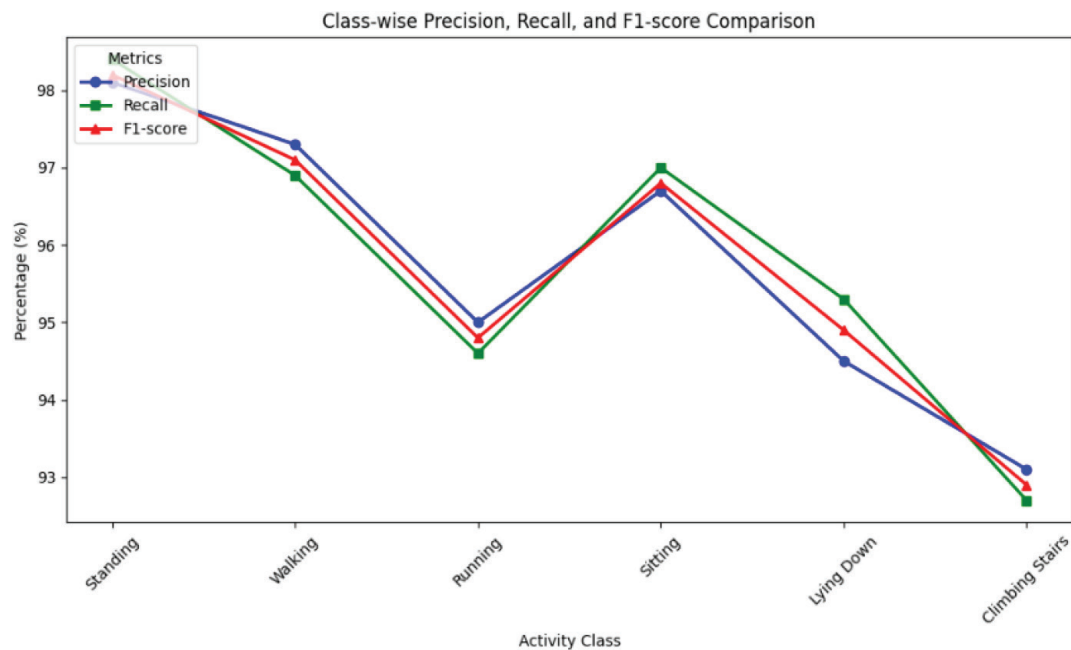


Figure 2: Comparison against the activity class.

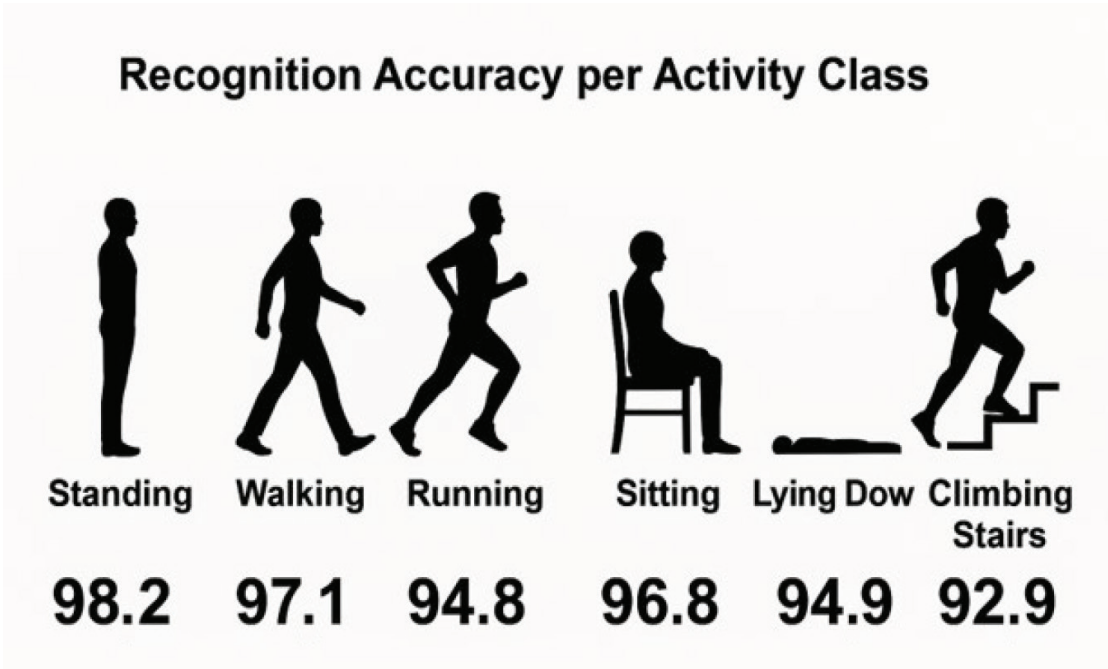


Figure 3: Pictorial representation of physical activities (standing, sitting, walking, running, lying down, and climbing stairs) along with their corresponding recognition accuracies based on the MHSA-SAE model. MHSA-SAE, multi-head self-attention enhanced stacked autoencoder.

generally result in clearer and more repeatable sensor profiles. For more dynamic activities such as running and climbing stairs, the F1-scores were slightly lower (94.8% and 92.9%, respectively). Figures 4–6 show the comparison of accuracy, F1 score, and AUC.

Chong et al. [15] utilized a feature selection approach combined with a SVM for activity recognition, achieving an accuracy of 91.80%, an F1-score of 90.50%, and an AUC of 92.30%. While this approach yielded reasonable performance, the results indicate that feature selection alone may not fully capture the complex temporal and spatial dependencies in the sensor data, which is addressed more effectively by DL models.

Jeong et al. [12] employed DL techniques along with noninvasive biomarkers to recognize activities. Their model performed slightly better than Chong et al. [15], achieving 93.20% accuracy, 92.70% F1-score, and 93.90% AUC. While DL models often perform better due to their ability to automatically learn complex features from raw data, Jeong et al.'s [12] approach did not incorporate advanced mechanisms such as attention layers, which can further enhance performance.

Dooley et al. [13] proposed a method called MASH Harmonization with Wearables, which focused

on aligning data from multiple wearable devices to improve activity recognition. This method achieved an accuracy of 94.60%, F1-score of 93.80%, and AUC of 95.10%. The harmonization approach is beneficial in handling inconsistencies from different sensors, but still falls short compared to the more advanced feature extraction and attention mechanisms used in the MHSA-SAE model.

Proposed MHSA-SAE (Ours) achieved a remarkable performance with an accuracy of 97.82%, an F1-score of 96.67%, and an AUC of 98.10%. The significant improvement in these metrics can be attributed to the MHSA mechanism, which effectively captures long-range dependencies in the time-series data, as well as the SAE for unsupervised feature extraction. These components allow the MHSA-SAE model to better handle complex and noisy sensor data, leading to more precise activity recognition. Additionally, the attention mechanism ensures that the model focuses on the most relevant temporal features, enhancing both classification accuracy and robustness.

The MHSA-SAE architecture significantly outperforms traditional and recent DL approaches. The integration of self-attention mechanisms improves temporal modeling compared to static feature extraction techniques used in prior work.

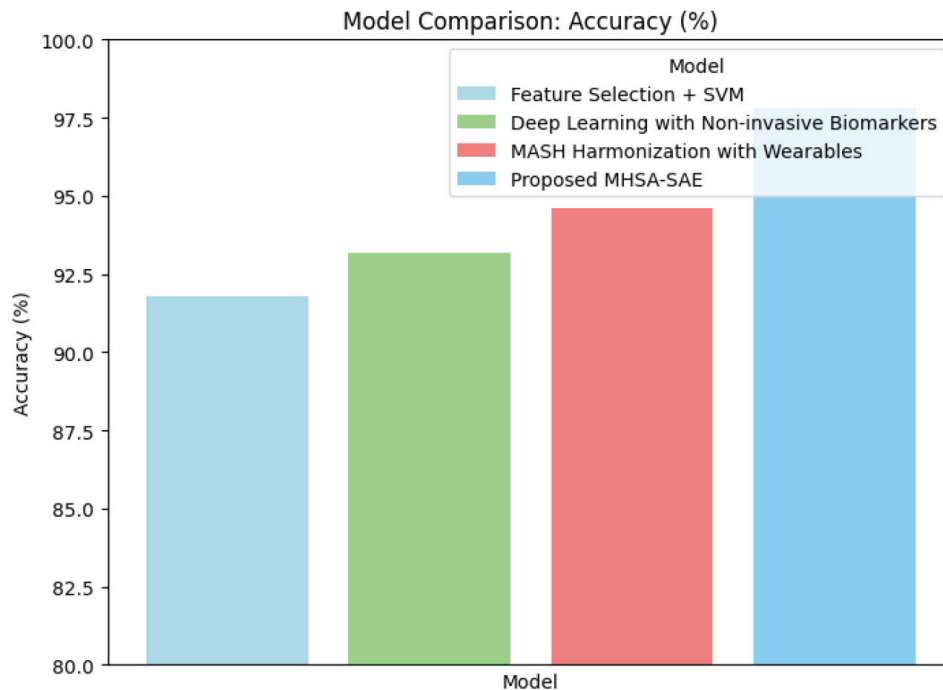


Figure 4: Accuracy comparison. MASH, method for activity sleep harmonization; MHSA-SAE, multi-head self-attention enhanced stacked autoencoder; SVM, support vector machine.

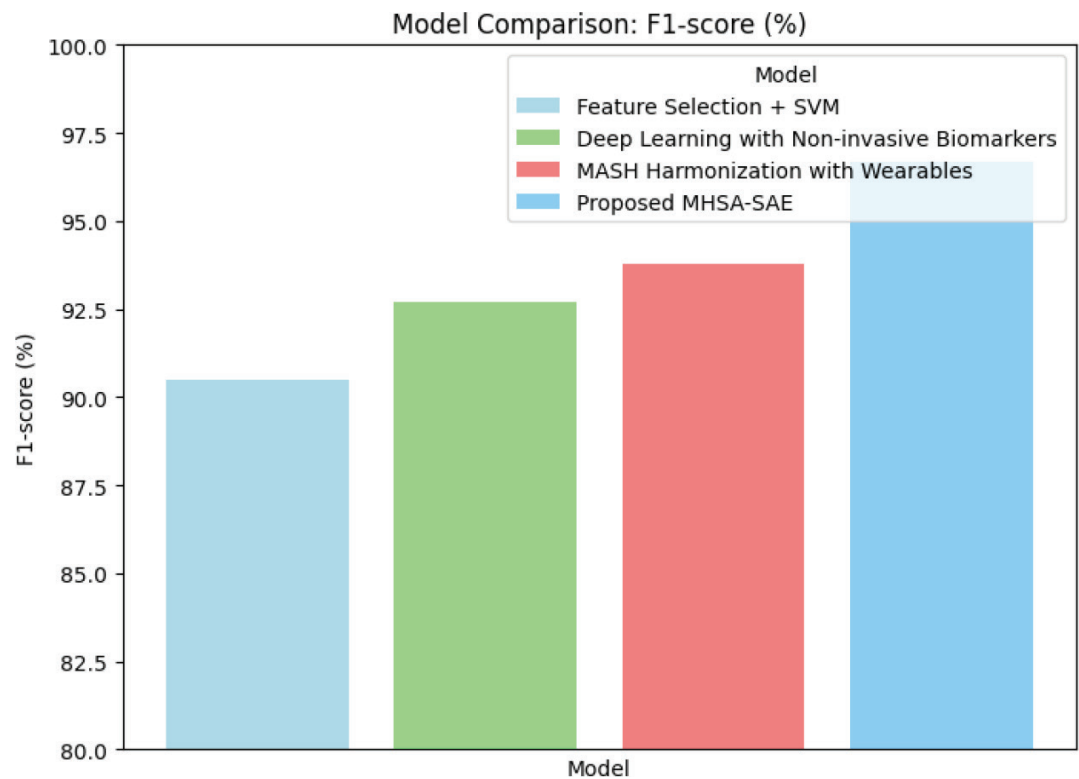


Figure 5: F1 score comparison. MHSA-SAE, multi-head self-attention enhanced stacked autoencoder; SVM, support vector machine.

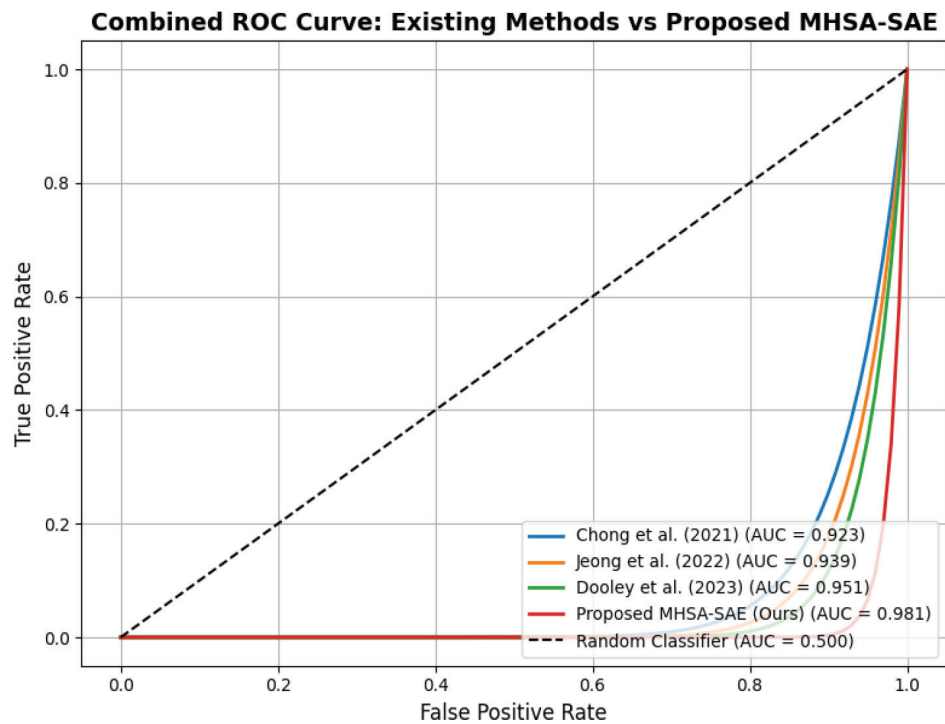


Figure 6: ROC curve. AUC, area under the receiver operating characteristic curve; MHSA-SAE, multi-head self-attention enhanced stacked autoencoder.

a. Comparison with existing methods

To validate the effectiveness of the MHSA-SAE framework, we benchmarked it against existing HAR models, as shown in Table 3.

The results show that MHSA-SAE consistently outperforms existing methods by a margin of 3%–6% in accuracy, 3%–6% in F1-score, and 3%–5% in AUC. These improvements are attributed to the combination of SAE for robust feature extraction and MHSA for capturing long-range temporal dependencies.

Table 4 presents the results of an ablation study to assess the impact of different components in the MHSA-SAE model. The SAE-only configuration, which lacks attention, achieves a moderate performance with an accuracy of 93.10% and an F1-score of 91.90%. Adding a Single-Head Attention mechanism improves performance to 95.30% accuracy and 94.40% F1-score, highlighting the benefit of capturing temporal dependencies. The Full MHSA-SAE model, with multi-head attention, provides the best results, achieving an accuracy of 97.82% and an F1-score of 96.67%, showcasing the significant improvement in performance by allowing the model to attend to multiple aspects of the time-series data simultaneously. This demonstrates the critical role of multi-head attention in enhancing activity recognition.

Table 5 illustrates the training and validation accuracy of the MHSA-SAE model over different epochs during the training process. Initially, at epoch 10, the training accuracy is 84.60%, while the validation accuracy is 82.30%. As the model progresses, both accuracies increase steadily. By epoch 30, the training accuracy reaches 91.20%, with a validation accuracy of 89.70%. By epoch 60, the model's performance further improves to 96.10% training accuracy and 94.80% validation accuracy, demonstrating the model's ability to generalize well. The training accuracy peaks at 98.00% and validation accuracy at 97.40% by epoch 100, indicating that the model has successfully learned and is now achieving high performance on both training and unseen data. The minimal gap between training and validation accuracy suggests that the model is not overfitting and is generalizing effectively.

V. Conclusion

This paper proposed a novel framework named MHSA-SAE for robust HAR using the mHealth HAR dataset. The architecture effectively integrates SAE for unsupervised feature extraction and an MHSA mechanism to model long-range temporal dependencies in time-series sensor data. Through extensive

Table 3: Comparison with existing methods

Model/Methodology	Accuracy (%)	F1-score (%)	AUC (%)
Chong et al. [15] - Feature selection + SVM	91.80	90.50	92.30
Jeong et al. [12] – DL with noninvasive biomarkers	93.20	92.70	93.90
Dooley et al. [13]- MASH harmonization with wearables	94.60	93.80	95.10
Proposed MHSA-SAE (Ours)	97.82	96.67	98.10

AUC, area under the receiver operating characteristic curve; DL, deep learning; MHSA-SAE, multi-head self-attention enhanced stacked autoencoder; SVM, support vector machine.

Table 4: Ablation study – impact of model components

Configuration	Accuracy (%)	F1-score (%)
SAE only (no attention)	93.10	91.90
SAE + single-head attention	95.30	94.40
Proposed study	97.82	96.67

SAE, stacked autoencoder.

Table 5: Training vs. validation accuracy over epochs

Epoch	Training accuracy (%)	Validation accuracy (%)
10	84.60	82.30
30	91.20	89.70
60	96.10	94.80
90	97.80	97.20
100	98.00	97.40

evaluation, the model demonstrated outstanding performance, achieving 97.82% accuracy, 96.67% F1-score, and 98.10% AUC, outperforming several existing methods in the literature. The class-wise metrics confirmed the model's high precision and recall across diverse activity categories, indicating its reliability in distinguishing fine-grained movement patterns. Ablation studies further validated the importance of the attention mechanism, revealing significant performance improvements when multi-head attention is incorporated. The proposed framework offers a scalable and generalizable solution for sensor-based activity recognition, with strong potential applications in health monitoring, fitness tracking, and behavior analysis. Future work will focus on optimizing computational efficiency, extending the model to multimodal datasets, and deploying it in real-time embedded systems for mobile and wearable platforms

Conflict of interests

Authors declare no conflict of interests.

References

- [1] Qin, Y., Yang, J., Zhou, J., Pu, H., & Mao, Y. (2023). A new supervised multi-head self-attention autoencoder for health indicator construction and similarity-based machinery RUL prediction. *Advanced Engineering Informatics*, 56, 101973. DOI: 10.1016/j.aei.2023.101973
- [2] Ahrari, S., Moonaghi, H. K., Mahdizadeh, S. M., & Bakavoli, A. H. (2023). Experiences of what influences physical activity adherence in Iranian patients with heart failure: A qualitative study. *Journal of Education and Health Promotion*, 12, 276. DOI: 10.4103/jehp.jehp_1002_22
- [3] Borland, M., Bergfeldt, L., Cider, Å., Rosenkvist, A., Jakobsson, M., Olsson, K., Lundwall, A., Andersson, L., & Nordeman, L. (2022). Effects of 3 months of detraining following cardiac rehabilitation in patients with atrial fibrillation. *European Review of Aging and Physical Activity*, 19, 14. DOI: 10.1186/s11556-022-00349-2
- [4] Lyu, T., Qian, H., & Chung, S.-P. (2024). Impact of physical activity, sedentary behavior, and basal metabolic rate on PTSD, depression, and emotional instability. *Brain Sciences*, 14(11), 1071. DOI: 10.3390/brainsci1411107
- [5] Medeiros, A. G., Cintra, M. M. M., dos Reis, M. A., Rocha, L. P., Neto, J. R. D. C., & Machado, J. R. (2024). The effects of various therapies on vulvovaginal atrophy and quality of life in gynecological cancer patients: A systematic review. *Archives of Gynecology and Obstetrics*, 310, 631–641. DOI: 10.1007/s00404-024-06723-3
- [6] Nikolić, M., di Plinio, S., Sauter, D., Keyzers, C., & Gazzola, V. (2024). The blushing brain: Neural substrates of cheek temperature increase in response to self-observation. *Proceedings of the Royal Society B: Biological Sciences*, 291(20240958). DOI: 10.1098/rspb.2024.0958
- [7] Arora, I., Mal, P., Arora, P., Paul, A., & Kumar, M. (2024). GABAergic implications in anxiety and related disorders. *Biochemical and Biophysical Research Communications*, 724, 150218. DOI: 10.1016/j.bbrc.2024.150218
- [8] Maallo, A. M. S., Novembre, G., Kusztor, A., McIntyre, S., Israr, A., Gerling, G., Björnsdotter, M., Olausson, H., Boehme, R. (2024). Primary somatosensory cortical processing in tactile communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379, 20230249. DOI: 10.1098/rstb.2023.0249
- [9] Siciliano, R. E., Anderson, A. S., Gruhn, M. A., Henry, L. M., Vreeland, A. J., Watson, K. H., Ciriiglio, A. E., Liu, Q., Ebert, J., Kuhn, T., et al. (2024). Momentary autonomic engagement during parent-adolescent conflict: Coping as a moderator of associations with emotions. *Psychophysiology*, 61, e14666. DOI: 10.1111/psyp.14666
- [10] Zhang, Y., Tao, Y., Choi, H., & Qian, H. (2024). Exploring the causal effects of physical activity, sedentary behaviour, and diet on atrial fibrillation and heart failure: A multivariable Mendelian randomisation analysis. *Nutrients*, 16(23), 4055. DOI: 10.3390/nu16234055
- [11] Choi, J., Lee, S., Choi, E.-K., Lee, K.-Y., Ahn, H.-J., Kwon, S., Han, K., Oh, S., & Lip, G. Y. H. (2024). Effect of physical activity on incident atrial fibrillation in individuals with varying duration of diabetes: A nationwide population study. *Cardiovascular Diabetology*, 23, 115. DOI: 10.1186/s12933-024-01716-1
- [12] Jeong, H., Jeong, Y. W., Park, Y., Kim, K., Park, J., & Kang, D. R. (2022). Applications of deep learning methods in digital biomarker research using noninvasive sensing data. *Digital Health*, 8, 20552076221136640. DOI: 10.1177/20552076221136640
- [13] Dooley, E. E., Winkles, J. F., Colvin, A., Kline, C. E., Badon, S. E., Diaz, K. M., et al. (2023). Method for

Activity Sleep Harmonization (MASH): A novel method for harmonizing data from two wearable devices to estimate 24-h sleep–wake cycles. *Journal of Activity, Sedentary, and Sleep Behavior*, 2, 8. DOI: 10.1186/s42489-023-00039-7

[14] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178–210. DOI: 10.1016/j.ins.2022.10.049

[15] Chong, J., Tjurin, P., Niemelä, M., Jämsä, T., & Farrahi, V. (2021). Machine-learning models for activity class prediction: A comparative study of feature selection and classification algorithms. *Gait & Posture*, 89, 45–53. DOI: 10.1016/j.gaitpost.2021.03.004

[16] Wang, M., Flexeder, C., Harris, C. P., Thiering, E., Koletzko, S., Bauer, C.-P., et al. (2023). Accelerometry-assessed sleep clusters and cardio-metabolic risk factors in adolescents. *Obesity*, 32, 200.

[17] Albalak, G., Stijntjes, M., van Bodegom, D., Jukema, J. W., Atsma, D. E., & van Heemst, D., et al. (2023). Setting your clock: Associations between timing of objective physical activity and cardiovascular disease risk in the general population. *European Journal of Preventive Cardiology*, 30, 232–240.

[18] Thornton, C. B., Kolehmainen, N., & Nazarpour, K. (2023). Using unsupervised machine learning to quantify physical activity from accelerometry in a diverse and rapidly changing population. *PLOS Digital Health*, 2, 1–13.

[19] Farrahi, V., & Rostami, M. (2024). Machine learning in physical activity, sedentary, and sleep behavior research. *Journal of Activity, Sedentary, and Sleep Behavior*, 3(5). DOI: 10.1186/s44167-024-00045-9

[20] Mahyari, A., & Pirolli, P. (2021). Physical exercise recommendation and success prediction using interconnected recurrent neural networks. 2021 IEEE International Conference on Digital Health, 148–153.

[21] Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9, 220638.

[22] Memon, A. R., Chen, S., To, Q. G., & Vandelandotte, C. (2023). Vigorously cited: A bibliometric analysis of the 100 most cited sedentary behaviour articles. *Journal of Activity, Sedentary, and Sleep Behavior*, 2, 13.

[23] Pratt, M., Varela, A. R., Salvo, D., Kohl, H. W., III, & Ding, D. (2020). Attacking the pandemic of

physical inactivity: What is holding us back? *British Journal of Sports Medicine*, 54, 760–762.

[24] Gao Y, Li Q, Yang L et al. (2024) Causal association between sedentary behaviors and health outcomes: a systematic review and meta-analysis of mendelian randomization studies. *Sports Med*. DOI: 10.1007/s40279-024-02090-5

[25] Ivon Rosen, P. (2023). Analysing time-use composition as dependent variables in physical activity and sedentary behaviour research: Different compositional data analysis approaches. *Journal of Activity, Sedentary, and Sleep Behavior*, 2, 23.

[26] Widjiantoro, B. L., Indriawati, K., Buyung, T. S. N. A., & Wahyuadnyana, K. D. (2024). Experimental validation: Perception and localization systems for autonomous vehicles using the extended Kalman filter algorithm. *International Journal on Smart Sensing and Intelligent Systems*, 17(1)

[27] Deshpande, N. M., Gite, S., & Pradhan, B. (2024). Explainable AI for binary and multi-class classification of leukemia using a modified transfer learning ensemble model. *International Journal on Smart Sensing and Intelligent Systems*, 17(1).

[28] Bagane, P., Oswal, M., Mhetre, S., Shankar, P., Mahendrakar, P., & Jebessa, O. A. (2025). Predictive diagnostics for early identification of cardiovascular disease: A machine learning approach. *International Journal on Smart Sensing and Intelligent Systems*, 18(1). DOI: 10.2478/ijssis-2025-0021

[29] Rani, S., & Senthilkumar, P. (2025). Optimizing patient care with big data analytics and machine learning algorithms. *International Journal on Smart Sensing and Intelligent Systems*, 18(1). DOI: 10.2478/ijssis-2025-0023

[30] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006 Jul 28;313(5786):504-7. doi: 10.1126/science.1127647. PMID: 16873662.

[31] Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *NIPS*.

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30 (*NeurIPS 2017*), 6000–6010.

[33] Lin, M., Chen, Q., & Yan, S. (2013). *Network in network*. arXiv. DOI: 10.48550/arXiv.1312.4400