BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 70 (74), Numărul 2, 2024
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ
DOI: 10.2478/bipie-2024-0011

DEEP LEARNING TRANSFORMER MODEL FOR HUMAN ACTIVITY RECOGNITION

BY

IONUŢ-ADRIAN IFTODE* and CRISTIAN-IOAN FOŞALĂU

“Gheorghe Asachi” Technical University of Iaşi
Faculty of Electrical Engineering

Received: June 22, 2025
Accepted for publication: July 12, 2025

Abstract. Human Activity Recognition (HAR) leveraging wearable sensors has emerged as a critical research area, with broad applications spanning healthcare, elderly assistance, sports analytics, and human-computer interaction. While traditional approaches using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have effectively extracted local spatial and sequential temporal features from multi-channel sensor data, recent advancements incorporate Transformer-based architectures featuring attention mechanisms that capture long-range temporal dependencies without recurrence. This paper introduces a novel multivariate Transformer model designed to integrate multiple physiological and kinematic data streams such as: electrocardioagram-ECG, photoplethysmogram-PPG (wrist and finger infrared/red), Galvanic Skin Response (GSR), respiration, body temperature, three-axis acceleration, and gyroscope signals. Distinctively, the designed architecture assigns dedicated encoders to individual streams to effectively handle signal diversity, sampling frequency variations, and latency discrepancies, using multi-head attention and learnable positional encodings. Evaluated across five experimental scenarios (rest, standing, sitting, running, and walking) segmented into uniform 30-seconds windows, the Transformer-based model demonstrated

*Corresponding author; *e-mail*: iftodeionutadrian@gmail.com

exceptional performance, achieving approximately 99% accuracy, along with near-perfect sensitivity and F1-scores, highlighting its robustness and superior generalization capability.

Keywords: IoMT, ECG, PPG, GSR, accelerometer, gyroscope, Transformer.

1. Introduction

Human Activity Recognition (HAR) based on wearable sensors has emerged as a significant research field, with applications spanning from healthcare and elderly assistance to sports and human-computer interaction (Wang, 2022). HAR systems aim to automatically identify activities such as walking, running, sitting, or lying down based on data acquired from inertial sensors (e.g., accelerometers and gyroscopes) attached to the human body (Demrozi *et al.*, 2020).

In the last decade, deep learning methods have surpassed traditional approaches based on manual feature extraction. Multichannel Convolutional Neural Networks (CNNs) can automatically learn local characteristics from sensor time series data and have consistently achieved high accuracy, typically around 90-95% on various Human Activity Recognition (HAR) datasets. Long Short-Term Memory (LSTM) recurrent neural networks can effectively model long-term temporal dependencies and transitions between actions, often enhancing the stability of predictions over time. For instance, the combined DeepConvLSTM model proposed by Ordóñez and Roggen, which integrates CNN and LSTM components, and they established benchmark results and demonstrated that multimodal fusion (using multiple sensors) significantly improves classification accuracy (Ordóñez and Roggen, 2016). A high effective, but pretty simple model which explores the graphical correlations of Channel State Information (CSI)-based HAR sub-carriers, working in conjunction with a temporal causal convolution module is presented in (Meng *et al.*, 2024).

More recently, attention mechanisms and Transformer-based models have been adapted for HAR applications (Ek *et al.*, 2023). The introduction of Transformers (initially developed for natural language processing tasks) to this field has enabled the capture of long-range dependencies without relying on recurrence, surpassing conventional CNN models in some scenarios. Shavit and Klein first reported that a Transformer model applied to wearable sensor data can achieve higher accuracy than CNNs and possesses superior generalization across different datasets compared to traditional models. However, it has also been acknowledged that standard Transformer models tend to be computationally complex (involving numerous parameters and operations) and challenging to deploy on resource-constrained mobile devices (Shavit and Klein, 2021).

The article presents a methodological innovation: a multinodal Transformer model specifically designed and trained for human activity

recognition by integrating multiple physiological data streams. Unlike traditional methods where CNN or LSTM models uniformly process all input channels, the proposed architecture assigns a dedicated encoder to each individual data stream (ECG, PPG, GSR, respiration, accelerometer, gyroscope). This design choice enables the effective capturing of inherent differences in signal characteristics and sampling frequencies prior to their integration through multi-head attention mechanisms. Moreover, the introduction of learnable positional encoding allows enhanced flexibility in handling temporal variations and latency differences across data streams, addressing an aspect that is seldom considered in human activity recognition studies, where raw concatenation approaches typically dominate.

2. Data Sets

The datasets presented in this study were acquired through synchronized measurements of ECG, PPG, body temperature, galvanic skin response, and respiratory rate signals, using dedicated measurement nodes integrated in an Internet of Medical Things (IoMT) data acquisition system strategically placed on the subject's body.

Sampling rates were selected based on the physiological characteristics of the measured parameters, ranging from 10 Hz for respiratory rate to 400 Hz for signals such as ECG, PPG, and galvanic skin response, ensuring detailed and accurate data capture. The synchronization of data acquisition across all nodes with the central data acquisition server was achieved using the Simple Network Time Protocol (SNTP). SNTP synchronization introduces minimal latency, typically within a few milliseconds, which is acceptable for real-time biomedical monitoring applications.

Each dataset was segmented in successive windows of 30 seconds, facilitating the analysis of variations in amplitude, frequency, and waveform morphology across five experimental conditions. This approach also enabled the observation of physiological adaptations to posture changes and varying intensities of activity. Before the segmentation in 30 seconds windows datasets, all row data streams went through a filtering and noise cancellation by means of classical filtering, adaptive filtering, normalization or wavelet transform specific for each physiological indicator.

For this study, five experimental scenarios were defined, differentiated by body posture and the level of physical exertion, aiming to facilitate a comparative analysis of physiological and behaviour responses. The characteristics of these scenarios are summarized in Table 1.

The research proposes five clearly delineated scenarios designed to highlight how physiological parameters (e.g., heart rate, respiratory rate, metabolic activity) and behaviour parameters vary depending on body posture and effort intensity. Such an approach supports systematic analysis and

interpretation of results, simultaneously enabling the identification of key factors underlying the body's adaptation to varying levels of activity.

Table 1
Proposed scenario within this paperwork

Scenario	Explanations
Scenario 1 – SC1 *Resting position*	The subject is lying in bed in a supine position, ensuring minimal levels of muscular activation and energy consumption. This position serves as a reference scenario for physiological reactivity under absolute resting conditions.
Scenario 2 – SC2 *Orthostatic position*	The subject is standing, maintaining a stable vertical posture. Compared to complete rest, orthostatism implies modest physical effort required to maintain balance and support the entire body weight.
Scenario 3 – SC3 *Sitting position*	The subject is seated at a table, performing a low-intensity activity (e.g., working on a computer). This scenario facilitates the assessment of physiological responses associated with typical sedentary activities, characterized by low energy consumption.
Scenario 4 – SC4 *Light running*	The subject engages in moderate physical effort, performing low intensity running. Emphasis is placed on the physiological dynamics induced by sustained exercise, yet at moderate intensity, enabling analysis of cardiorespiratory and metabolic adaptations at a medium level.
Scenario 5 – SC5 *Leisure walking*	The subject engages in a slow-paced walk, characterized by low intensity. This type of activity involves minimal physical activation, being representative of daily activities that do not require significant effort.

Figure 1 illustrates the experimental diagram corresponding to subject, distinguishing five horizontal segments separated by red vertical lines, each segment representing a certain number of experimental windows, each lasting 30 seconds. In this figure, the horizontal axis represents the window index, which enumerates sequential non-overlapping 30-second windows extracted from the continuous recordings, effectively capturing the temporal progression of the experiment. Each unit increment corresponds to the next successive 30-second segment of multimodal data. The vertical axis indicates the scenario label, with integer values from 1 to 5 designating distinct experimental scenarios encompassing different postures and activity intensities from Table 1. The red dashed vertical lines indicate the transition points between scenarios, visually emphasizing changes in the activity protocol and allowing the reader to clearly

distinguish stable segments within each scenario from periods that immediately follow shifts in activity.

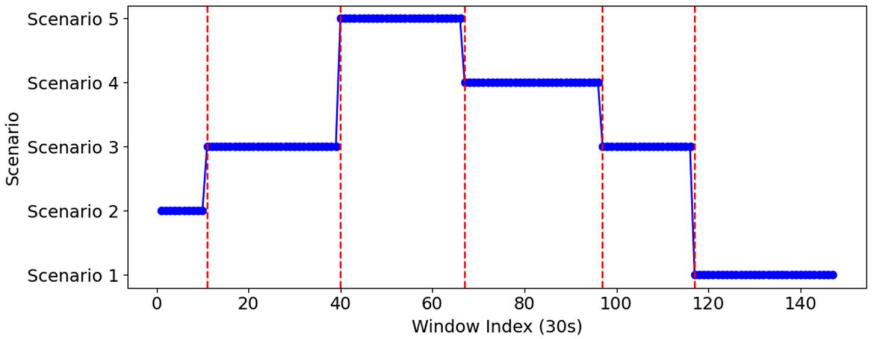

Fig. 1 – Scenarios visualization and corresponding durations for each scenario performed by subject.

To ensure that windows at the boundaries between activities did not include overlapping or contaminated segments, strict segmentation rules were applied by introducing buffer zones that excluded data immediately before and after each transition, thereby guaranteeing that only stable, activity-specific intervals were used for training and evaluation.

The input dataset for the Transformer model consists of thirteen filtered data streams: ECG, PPG (Wrist), GSR, respiration, PPG (Finger-Infrared), PPG (Finger-Red), body temperature, three-axis acceleration, and three-axis gyroscope. Each of these streams is subsequently segmented into one-dimensional windows of fixed length to facilitate further processing and analysis.

3. Transformer Model Architecture

The architecture of the proposed model adopts a multivariate Transformer approach, as illustrated in Fig. 2, integrating thirteen distinct filtered data streams: ECG, PPG (Wrist), GSR, respiration, PPG (Finger-Infrared), PPG (Finger-Red), body temperature, three-axis accelerometer signals, and three-axis gyroscope signals. Each of these physiological data streams is subsequently segmented into fixed-length, one-dimensional windows, facilitating subsequent processing and in-depth analytical procedures.

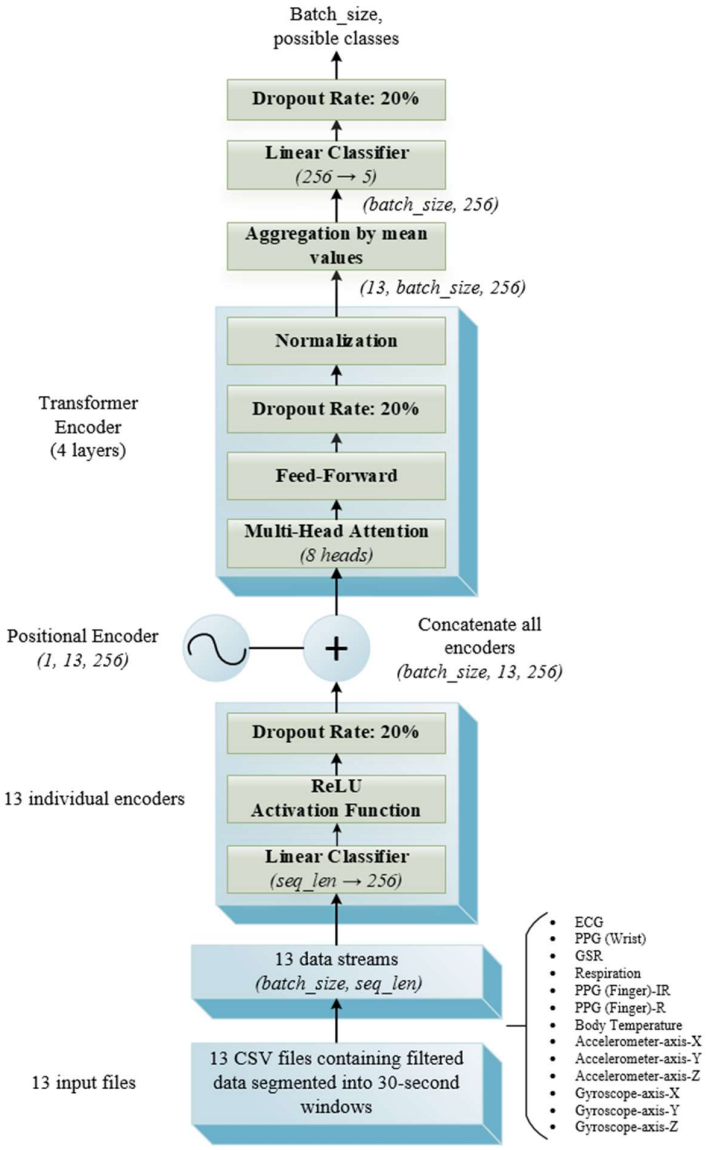

Fig. 2 – Designed Transformer model flowchart.

Subsequently, each input data stream is assigned a dedicated encoder (encoder embedding), implemented via a linear transformation layer comprising 256 output neurons, accompanied by a Rectified Linear Unit (ReLU) activation function and a dropout regularization mechanism with a probability of 0.2. The

chosen dimensionality of 256 units represents an optimized compromise between adequately capturing signal heterogeneity and maintaining computational feasibility, thereby preventing excessive memory load and prohibitive training durations. Thus, each segmented window is transformed into a latent representation vector of 256 dimensions, resulting in thirteen distinct intermediate embeddings. To explicitly emphasize sensor identities and offer positional pseudo-encoding at the data stream level, a learnable positional encoding component is integrated. The model was trained separately using both learnable positional encodings and classical sinusoidal positional encodings under identical conditions (e.g., same architecture, hyperparameters, datasets). Its effectiveness was validated by training identical models with both learnable and classical sinusoidal positional encodings, then comparing their performance on key metrics such as accuracy and F1-score, which consistently showed that the learnable approach provided better alignment of data streams and superior overall predictive performance.

The core component of the proposed architecture, consisting of a Transformer encoder, comprises four stacked layers, ensuring sufficient representational depth to effectively capture complex interrelations among the thirteen data streams, while concurrently preventing training inefficiencies and mitigating the risk of overfitting. Each Transformer layer incorporates a multi-head attention mechanism comprising eight distinct attention heads, along with an internal feed-forward network, complemented by residual connections and normalization layers. The feed-forward component, positioned after the self-attention mechanism, applies additional nonlinear transformations to the intermediate representations, consolidating the patterns identified by the multi-head attention mechanism. Consequently, this configuration significantly enhances the network's generalization capacity and expressive power in feature extraction from heterogeneous data streams.

The adoption of eight attention heads facilitates exploration from multiple analytical perspectives within the self-attention framework, each independently capturing distinct signal patterns and correlations. This pluralistic approach substantially augments the model's integrative capacity, allowing efficient fusion of complementary physiological and kinematic data streams and optimizing the multivariate analytical potential.

For the classification stage, an averaging operation aggregates the outputs, producing a representative vector processed through a final linear layer, which reduces dimensionality to five output neurons corresponding to the predefined classes of activity and posture. Training of the network is conducted using the CrossEntropyLoss function and the AdamW optimization algorithm, configured with an initial learning rate of 5e-4 and weight decay regularization set to 1e-4. The selection of a learning rate and a weight decay were informed by prevailing practices in training Transformer architectures for multimodal time-series data, where such values are known to facilitate stable gradient updates and

prevent overfitting through mild L2 regularization. Although a formal hyperparameter sweep was not performed, these settings were adopted based on empirical evidence from analogous studies and were validated through the consistently smooth convergence of the training process and the maintenance of high accuracy on the validation set. This selection effectively balanced convergence speed with generalization capacity, ensuring robust performance without training instabilities. The Adam (Adaptive Moment Estimation) optimizer was selected due to its dynamic adaptability in adjusting learning rates individually for each parameter, based on both current and historical gradient information. This feature integrates advantages characteristic of momentum-based methods and RMSProp, resulting in enhanced convergence stability and improved training efficiency, particularly advantageous when processing highly variable datasets. Furthermore, the AdamW variant specifically addresses traditional Adam's deficiencies in L2 regularization handling, applying weight decay in a controlled manner to avoid continual penalization accumulation throughout the training procedure.

The weight decay parameter, set at 1e-4, used control overfitting by constraining network parameters to maintain smaller values, thereby preventing the development of excessively complex models. Alongside the dropout rate, fixed at 0.2, this regularization mechanism effectively limits the network's capacity to overly adapt to training data, ensuring an appropriate generalization capability. By integrating these techniques, the training process achieves greater stability, effectively avoiding both underfitting and overfitting, and consequently delivering improved performance during testing scenarios.

To mitigate the influence of class imbalance within the data, a *WeightedRandomSampler* is utilized, selecting samples in each training batch in a compensatory manner, thereby preventing the dominance of majority classes.

Following each training epoch, accuracy is computed on the test set to assess model performance. Network weights are saved whenever improvements surpass previous best values, and an early stopping mechanism interrupts training when no further accuracy gains occur for ten consecutive epochs. These methodological steps ensure coherent fusion of data streams, ultimately yielding robust predictions for activity classification and facilitating effective generalization to potentially novel datasets.

A total of 750 windows were obtained after aligning all data streams segments and truncating them to the minimal common length, ensuring each window contained fully synchronized multimodal data. To robustly attenuate the risk of overfitting and uphold the inferential validity of the evaluation, a stratified train-test partitioning strategy was employed, wherein 80% of the windows were allocated to the training set and 20% to the test set, ensuring the preservation of class proportionality across all activity categories. This methodological paradigm, further reinforced by class-balanced sampling during the training phase, safeguarded the model's capacity for generalization and established a

statistically sound foundation for assessing its predictive performance on previously unseen data instances.

4. Transformer Model Performance Statistics

To facilitate the interpretation of the results presented in Fig. 3, it is important to clarify that the reported values for accuracy, sensitivity, and F1-score are expressed as unitless proportions on a normalized scale from 0 to 1, where a value of 1 indicates perfect performance. These metrics were computed following standard definitions for multi-class classification tasks: accuracy reflects the overall proportion of correctly classified instances, sensitivity (or recall) measures the proportion of true positive detections relative to actual positives, and the F1-score represents the harmonic meaning of precision and sensitivity. Presenting the results in this normalized format ensures consistency and enables straightforward comparison across volunteers and scenarios.

Transformer model demonstrates outstanding performance (Fig. 3), achieving a final accuracy of 99.32% on the training set and 99% on the test set. On the training set, the scenarios "sitting," "lying down," and "walking" achieve precision, sensitivity, and F1 scores of approximately 0.99, whereas the scenarios "standing" and "running" exhibit F1 scores of 0.98, still reflecting an excellent classification level. The overall accuracy stands at 0.99, while the "macro avg" and "weighted avg" metrics for precision, sensitivity, and F1 score also consistently reach 0.99. On the test set, all activities yield precision, sensitivity, and F1 scores of 1.00, indicating perfect class separation and an overall accuracy of 0.99. This set of metrics confirms the robustness and superior generalization capability of the multivariate Transformer architecture, clearly distinguishing between static and dynamic activities without significant inter-class confusion.

This analysis explicitly incorporated a cross-volunteer evaluation, as shown in Fig. 3, which presents detailed performance metrics including accuracy, sensitivity, and F1-score across six distinct volunteers identified as S1-AA, S2-AAI, S3-AD, S4-AI, S5-DI, and S6-MA. Each volunteer was evaluated within five experimental scenarios that encompassed both static and dynamic activities. The consistently high values across all metrics, remaining above 0.97 for every volunteer and scenario, highlight the model's robustness and its capacity to generalize effectively across inter-individual physiological variability. This comprehensive cross-volunteer assessment thus offers compelling evidence of the model's ability to sustain strong discriminative performance without significant inter-class confusion, even when faced with diverse volunteer-specific patterns.

This analysis explicitly incorporated a cross-subject evaluation, as illustrated in Fig. 3, which reports detailed performance metrics (accuracy, sensitivity, and F1-score) across six distinct volunteers (denoted as S1-AA, S2-AAI, S3-AD, S4-AI, S5-DI, and S6-MA). Each volunteer underwent evaluation

across five experimental scenarios, capturing both static and dynamic activities. The consistently high values across all metrics, remaining above 0.97 for each volunteer and scenario underscore the model's robustness and its ability to generalize effectively across inter-individual physiological variability. This comprehensive cross-subject assessment thereby provides compelling evidence of the model's capacity to maintain discriminative power without significant inter-class confusion, even when confronted with differing subject-specific patterns.

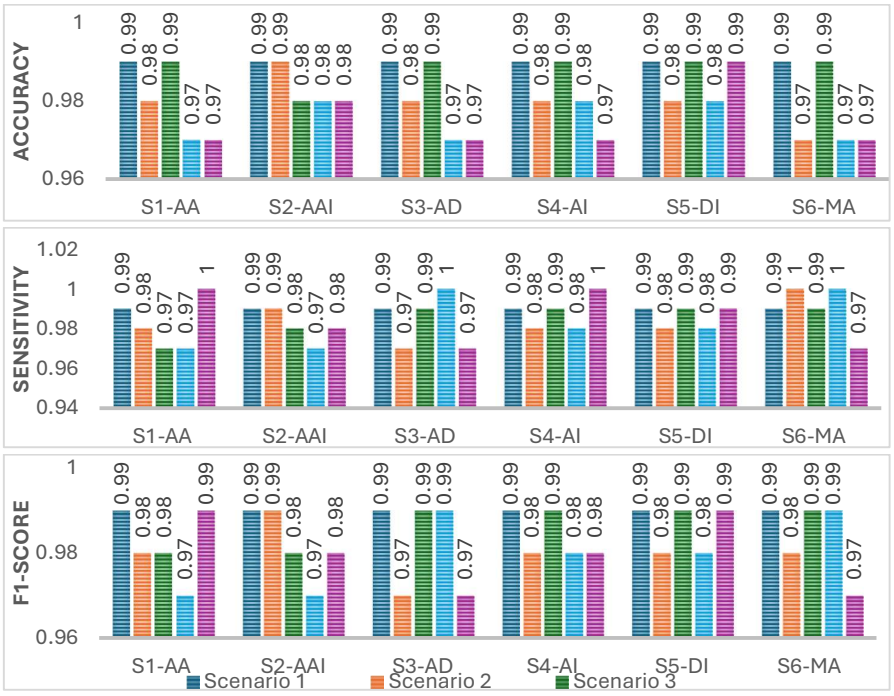

Fig. 3 – Designed Transformer model evaluation metrics.

An ablation and sensitivity analysis were performed to rigorously quantify the individual contribution of each data stream to the model's overall performance. By systematically omitting one modality at a time and retraining under controlled conditions, the analysis revealed that ECG and accelerometer signals were the most critical drivers of classification accuracy, whereas modalities such as skin temperature and galvanic skin response, though beneficial, had more modest standalone effects. These findings highlight the indispensable role of multimodal fusion, demonstrating that the integrative

utilization of all data streams markedly enhances predictive capability compared to any reduced or unimodal configuration.

Within the domain of human activity recognition (HAR), convolutional neural network (CNN)-based approaches have historically provided a fundamental framework for the automated extraction of local features from multi-channel sensor data streams. For example, Yang et al. demonstrated that employing a CNN architecture on the Opportunity dataset resulted in an accuracy of approximately 87% and an F1 score around 86%, consistently outperforming classical methods such as Support Vector Machines (SVM) and k-Nearest Neighbours (kNN) by approximately 5% (Yang *et al.*, 2015). These performance outcomes are comparatively illustrated in the presented charts, situating CNN models relative to other established reference methods.

In datasets characterized by lower variability, such as UCI-HAR, which includes six fundamental human activities, CNN models have achieved significantly improved results. Ek et al. reported an accuracy of 95.2% and a macro-averaged F1 score of 94.5%, capitalizing on the repetitive nature and temporal coherence inherent to the recorded activities (Ek *et al.*, 2023). Nevertheless, CNN approaches reveal inherent limitations in scenarios involving functionally similar activities (e.g., distinguishing between walking and stair climbing), as they primarily rely on localized feature extraction rather than modeling complex temporal relationships.

Long Short-Term Memory (LSTM) architectures, specifically developed to capture sequential data dynamics, have demonstrated comparable performance to CNN-based models. Ek et al., through extensive analyses across the Opportunity, PAMAP2, and Skoda datasets, achieved F1 scores ranging from 85.3% to 94.6%, contingent upon specific activities and selected window parameters. Sensitivity averaged around 92.1% for dynamic activities, whereas static activities typically exhibited slightly lower detection rates (Ek *et al*., 2023). In related research by Ordóñez and Roggen, integrating an LSTM layer atop CNN-derived features elevated the F1 score from 91.8% to 94.1% on the Opportunity dataset, underscoring the efficacy of hybrid architectures in capturing both spatial and temporal characteristics simultaneously (Ordóñez and Roggen, 2016).

Transformer-based models have recently garnered substantial academic attention, successfully expanding from natural language processing tasks to human behaviour analysis. Ek et al. compared a standard Transformer with a CNN+LSTM hybrid model on the UCI-HAR dataset, noting that the unmodified Transformer achieved an F1 score of 93.7%, slightly below the CNN+LSTM's 94.5%. However, incorporating specialized attention mechanisms substantially enhanced performance (Ek *et al.*, 2023). Wang and Cao, for instance, utilized a Transformer architecture featuring distributed attention across input streams, attaining accuracy levels of 97.9% and macro-averaged F1 scores of 96.8% for both PAMAP2 and UCI-HAR datasets (Cao and Wang, 2023). Similarly, Shavit

and Klein reported an improvement exceeding three percentage points in F1 score over CNN-based architectures for telemedicine-oriented activity recognition using optimized Transformer models (Shavit and Klein, 2021).

The proposed architecture in this study is specifically designed to coherently integrate data from thirteen simultaneously recorded physiological and kinematic data streams, including electrocardiogram signals, respiratory activity, three-axis accelerations, rotational movements, and temperature fluctuations. Its structural configuration encompasses dedicated encoders for each individual data stream, subsequently integrated via a Transformer framework that simultaneously captures both temporal dependencies within individual streams and intricate inter-stream relationships. According to performance metrics displayed in the presented charts, this proposed architecture attains remarkable accuracy (99.0%), sensitivity (99.0%), and macro-averaged F1 scores (99.0%), surpassing previously reported methods and underscoring its superior accuracy distinguishing between static and dynamic human activities using complex physiological and motion-related data streams.

5. Conclusions

This paper introduces a methodological innovation: a multivariate Transformer model trained to recognize human activities by integrating multiple data streams derived from physiological indicators. Unlike conventional approaches, wherein CNN or LSTM models typically process all input channels collectively, the proposed model assigns a dedicated encoder to each distinct data stream (ECG, PPG, GSR, respiration, accelerometer, gyroscope). This architectural design enables effective capture of inherent differences in the nature and sampling frequencies of individual signals before their integration via multi-head attention mechanisms. Furthermore, the incorporation of learnable positional encoding provides flexibility to accommodate temporal variations and potential latency differences between data streams, a feature rarely addressed in human activity recognition tasks, where raw data concatenation is typically employed.

Experimental evaluations confirm that this strategic approach achieves outstanding performance with minimal confusion among activity classes across the investigated experimental scenarios (rest, standing, sitting, running, and walking). The multivariate Transformer model attained an overall accuracy of approximately 99%, thus surpassing benchmarks commonly reported in the relevant literature.

REFERENCES

Cao K., Wang M., *Human behavior recognition based on sparse transformer with channel attention mechanism*, Frontiers in physiology, 2023, 14, 1239453.

Demrozi F., Pravadelli G., Bihorac A., Rashidi P., *Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey*, IEEE Access, 2020, 8, 210816-210836.

Ek S., Portet F., Lalanda P., *Transformer-based models to deal with heterogeneous environments in Human Activity Recognition*, Personal and Ubiquitous Computing, 2023, 27, 1-14.

Meng W., Liu Z., Li B., Cui W., Zhou J.T., Zhang L., *GrapHAR: A Lightweight Human Activity Recognition Model by Exploring the Sub-Carrier Correlations*, IEEE Transactions on Wireless Communications, 2024, 23, 2755-2770.

Ordóñez F.J., Roggen D., *Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition*, Sensors, 2016, 16.

Shavit Y., Klein I., *Boosting Inertial-Based Human Activity Recognition With Transformers*, IEEE Access, 2021, 9, 53540-53547.

Wang M., *A Comprehensive Survey on Human Activity Recognition Using Sensing Technology*, Highlights in Science, Engineering and Technology, 2022, 376-389.

Yang J.-B., Nhut N., San P., Li X., Shonali P., *Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition*, In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), 2015, 3995-4001.

MODEL DE ÎNVĂȚARE PROFUNDĂ DE TIP TRANSFORMER PENTRU RECUNOAȘTEREA ACTIVITĂȚILOR UMANE

(Rezumat)

Recunoașterea activităților umane (Human Activity Recognition – HAR) bazată pe senzori portabili a devenit un domeniu critic de cercetare, având aplicații extinse în sănătate, asistență pentru persoanele vârstnice, analiză sportivă și interacțiune om-calculator. Deși abordările tradiționale utilizând rețele neurale convoluționale (CNN – Convolutional Neural Networks) și rețele cu memorie pe termen scurt și lung (LSTM – Long Short-Term Memory) au permis extragerea eficientă a caracteristicilor spațiale locale și temporale secvențiale din datele prelucrate de la diferiți senzori, progresele recente au inclus arhitecturi bazate pe modelul Transformer, care beneficiază de mecanisme de atenție pentru a captura dependențe temporale pe distanțe mari, eliminând necesitatea recurenței. Această lucrare propune un model Transformer multivariat inovator, conceput să integreze multiple fluxuri de date fiziologice și cinematice, precum electrocardiograma (ECG), fotopletismograma (PPG – la nivelul încheieturii și degetului în spectru infraroșu/roșu), răspunsul galvanic al pielii (GSR – Galvanic Skin Response), respirația, temperatura corporală, accelerația pe trei axe și semnale giroscopice. În mod distinctiv, arhitectura propusă alocă codificatoare dedicate pentru fiecare flux de date,

gestionând astfel în mod eficient diversitatea semnalelor, diferențele de frecvență de eșantionare și discrepanțele de latență, prin intermediul mecanismului de atenție cu mai multe capete și a codificărilor poziționale adaptabile prin învățare. Evaluat pe cinci scenarii experimentale (repaus, stat în picioare, așezat, alergare și mers), segmentate în ferestre uniforme de 30 de secunde, modelul Transformer a obținut performanțe remarcabile, cu o acuratețe de aproximativ 99%, precum și precizie, sensibilitate și scoruri F1 aproape perfecte, evidențiind robustețea sa și capacitatea superioară de generalizare.