

Structured Motion Representation Learning for Robust Temporal Gesture Recognition

S. Vaishenaav

Department of Computer Networking (AIML) PSG
Polytechnic College
Coimbatore, India vaishenaav@gmail.com

Shrijit S Iyengar

Department of Computer Networking (AIML) PSG
Polytechnic College
Coimbatore, India shrijitiyengar07@gmail.com

A. R. Akash

Department of Computer Networking (AIML) PSG
Polytechnic College
Coimbatore, India akashanandan006@gmail.com

R. Vinesh Kannah

Department of Computer Networking (AIML) PSG Polytechnic
College
Coimbatore, India vineshkannah@gmail.com

S. Yajun

Department of Computer Networking (AIML) PSG Polytechnic
College
Coimbatore, India yajun5713@gmail.com

S. Brindha

Supervisor, Head of the Department Department of Computer
Networking (AIML) PSG Polytechnic College
Coimbatore, India hod.dcn@psgpolytech.ac.in

Abstract—Gesture recognition systems have achieved significant progress through deep learning-based temporal modeling architectures; however, existing approaches primarily rely on raw spatial landmark representations and implicitly learned temporal dynamics. Such methods exhibit strong spatial bias and often fail to discriminate between gestures that share similar trajectories but differ in motion intensity, acceleration patterns, and temporal irregularity. This paper proposes a Physics-Informed Entropy-Regularized Motion Encoding (EMRE) architecture for robust temporal gesture recognition. The proposed framework explicitly models motion dynamics by computing velocity and acceleration from landmark sequences and deriving a physics-inspired motion energy formulation defined as $E_t = |v_t|^2 + \lambda |a_t|^2$. To capture temporal complexity, an information-theoretic entropy measure is introduced over normalized energy distributions, enabling quantification of motion irregularity. The resulting energy-entropy feature representation is provided to a temporal Long Short-Term Memory (LSTM) classifier for sequence-level inference. Unlike conventional coordinate-driven models, the proposed architecture reduces learning burden by embedding structured physical priors prior to temporal modeling. Experimental evaluation on controlled multi-class gesture datasets demonstrates improved discriminative capability between dynamically similar gesture categories, enhanced Macro-F1 performance under class imbalance, and stable generalization across validation splits. The results indicate that integrating physics-informed motion modeling with entropy-based regularization provides a principled and computationally efficient foundation for next-generation robust gesture recognition systems.

Index Terms—Gesture Recognition, Motion Energy, Entropy Regularization, Temporal Modeling, Physics-Informed Learning, LSTM

I.

INTRODUCTION

Human gesture recognition constitutes a fundamental component of intelligent human-computer interaction (HCI) systems, immersive virtual environments, assistive robotics, sign language interpretation platforms, and adaptive control interfaces [1], [2]. In such systems, gestures serve as structured spatiotemporal signals that encode intent through coordinated skeletal motion over time. Formally, a gesture sequence may be represented as a landmark tensor $X \in \mathbb{R}^{T \times J \times 3}$, where T denotes temporal length and J denotes the number of articulated joints in three-dimensional space.

Recent advances in deep learning have enabled significant improvements in gesture recognition accuracy through recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) architectures [3]–[7], graph convolutional networks (GCNs) [8]–[10], and attention-based temporal transformers [11]. Convolutional approaches [12]–[15] and hybrid architectures [16]–[19] have also demonstrated strong performance. These methods have been extensively evaluated across various benchmark datasets [20], [21] and applied to diverse sensor modalities [22], [23]. Despite these advancements, the dominant modeling paradigm remains coordinate-centric [12], [13], [24]: raw landmark trajectories or learned spatial embeddings are directly processed by temporal models under the implicit assumption that dynamic motion characteristics can be sufficiently inferred through end-to-end optimization. This assumption introduces a fundamental structural limitation. While spatial representations capture geometric pose config-

urations, they do not explicitly encode physically meaningful motion quantities such as velocity and acceleration. Tempo-ral architectures are therefore burdened with simultaneously learning both representation geometry and motion dynamics from raw coordinate differences [25]. In practice, this often results in strong spatial bias, where gestures with similar trajectories but distinct dynamic signatures—such as concentrated versus distributed motion patterns—produce overlapping feature representations in latent space [26]. Consequently, class separability deteriorates when discrimination depends on dynamical intensity rather than spatial configuration alone.

From a physical perspective, gesture execution inherently involves energy transfer governed by velocity and acceleration profiles [27], [28]. However, conventional frameworks rarely incorporate motion energy formulations of the form:

$$E_t = |v_t|^2 + \lambda |a_t|^2,$$

(1)

where v_t and a_t denote first- and second-order temporal derivatives of landmark trajectories. The absence of such physics-informed modeling eliminates a critical inductive bias that could regularize feature learning and improve dynamical interpretability.

Moreover, gesture dynamics differ not only in magnitude but also in structural complexity and dispersion across joints and time. Existing architectures lack an explicit information-theoretic quantification of temporal irregularity [29], [30]. Without entropy-based characterization of motion energy distributions, models are unable to distinguish between smooth, focused motion patterns and temporally dispersed or stochastic behaviors. This limitation constrains robustness under inter-subject variability, execution speed variation, and class imbalance conditions.

These observations motivate a principled re-examination of motion representation in temporal gesture recognition systems. Rather than treating dynamics as an emergent property of sequence modeling, motion characteristics should be explicitly encoded through physics-consistent and information-theoretically grounded feature construction prior to temporal inference.

In this work, we propose a Physics-Informed Entropy-Regularized Motion Encoding architecture designed to address these structural limitations. The proposed framework explicitly models velocity and acceleration-derived motion energy, integrates entropy-based temporal complexity measures, and embeds these structured representations into a temporal classification pipeline. By introducing physics-consistent inductive bias and entropy-based regularization, the architecture aims to enhance inter-class separability, improve robustness under

dynamically ambiguous conditions, and provide a mathematically principled foundation for next-generation temporal gesture recognition systems.

II. LITERATURE SURVEY

The domain of gesture recognition intersects multiple re-search streams, including coordinate-based temporal modeling, skeleton graph learning, physics-inspired motion analysis, and information-theoretic sequence characterization. This section synthesizes existing contributions across these domains and identifies structural research gaps that motivate the proposed physics-informed entropy-regularized motion encoding framework.

A. Coordinate-Centric Temporal Modeling

Early deep learning approaches to gesture recognition predominantly relied on raw landmark coordinates or pose embeddings processed through recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) architectures [3], [4], [12], [24]. These models treat gestures as sequential coordinate trajectories and attempt to learn temporal dependencies directly from positional variations. Comprehensive reviews of these methods [1], [20], [31], [32] highlight their effectiveness across various domains. Variants including bidirectional LSTMs [18], deep residual architectures [17], stacked LSTM networks [19], [33], and CNN-LSTM hybrids [6], [13] have been proposed to improve temporal modeling capacity. Convolutional approaches with multi-sensor fusion [14], [15] and resource-efficient implementations [23] have also been explored. While such methods achieve competitive accuracy under controlled settings [5], [22], [34], they inherently assume that velocity, acceleration, and higher-order motion characteristics can be implicitly learned through sequence optimization. This implicit dynamic modeling introduces representational inefficiency. The network must simultaneously learn geometric pose structure and dynamical evolution without explicit inductive bias. As a result, gestures that are spatially similar but dynamically distinct often produce overlapping latent embeddings, reducing discriminative robustness in fine-grained classification tasks. The development of standardized benchmarks [21], [35] has facilitated comparative evaluation of these approaches.

B. Skeleton-Based and Graph Convolutional Approaches

To address limitations of purely coordinate-based modeling, skeleton graph representations have been introduced [8]. Graph Convolutional Networks (GCNs) model joints as nodes and anatomical connections as edges, enabling spatial-topological reasoning over articulated motion [9], [10], [36]. Temporal extensions of GCNs incorporate sequence modeling to capture joint evolution across time [37]. Recent work has explored disentangled representations and multi-scale approaches to improve discriminative power [26]. Despite these architectural advances, the fundamental reliance on coordinate-based features persists [38].

Although graph-based methods improve structural awareness of joint relationships, they remain largely pose-driven. Temporal derivatives such as velocity and acceleration are typically not modeled explicitly as primary physical quantities but instead remain implicitly encoded within learned feature transformations. Consequently, graph-based approaches may still exhibit sensitivity to execution speed variation and motion intensity ambiguity.

C. Temporal Attention and Transformer Architectures

Recent advancements leverage attention mechanisms and Transformer-based models to capture long-range temporal dependencies [11]. These architectures improve modeling of non-local sequence interactions and enable flexible weight-ing of temporal segments. Visual attention and multimodal fusion approaches [39] have further expanded the capabilities of attention-based methods. Despite their expressive power, attention-based frameworks operate primarily in learned fea-ture space without incorporating physically interpretable mo-tion priors.

The absence of structured dynamic encoding may result in high model capacity compensating for representational deficiency. While attention mechanisms enhance sequence model-ing, they do not inherently guarantee improved separability for gestures differentiated primarily by motion energy or temporal irregularity.

D. Physics-Inspired Motion Modeling

A limited body of research has explored incorporating kinematic descriptors such as velocity and acceleration into gesture recognition pipelines [27]. These approaches often compute derivative features as auxiliary inputs; however, systematic integration of motion energy formulations [28] grounded in physical principles remains underexplored. Motion-based rep-resentations have shown promise in action recognition [25] and activity analysis [40], yet their adoption in HAR remains limited. In particular, formulations that combine first- and second-order derivatives into unified motion intensity mea-sures have not been consistently adopted within deep temporal classification frameworks.

Furthermore, prior work seldom treats motion energy as a primary structural representation with theoretical justification rooted in physical dynamics. The absence of formalized energy-based inductive bias restricts interpretability and reduces the potential for principled regularization of temporal feature spaces.

E. Information-Theoretic Measures in Sequence Analysis

Entropy and related information-theoretic metrics have been applied in domains such as signal processing and physiological time-series analysis to quantify irregularity and complexity [29], [30]. However, their integration into gesture recognition systems remains limited. Feature selection and complexity analysis using entropy measures [41] have shown potential but are not yet mainstream in HAR architectures. Existing deep learning approaches rarely compute entropy over motion

energy distributions to characterize temporal dispersion or joint-level variability.

Without explicit entropy-based modeling, gesture classifiers lack a mechanism to distinguish between concentrated, deterministic motion patterns and dynamically dispersed or stochastic execution styles. This omission constrains robustness under inter-subject variability and speed fluctuations.

Recent comprehensive surveys [1], [2], [42], [43] provide extensive reviews of deep learning methods for HAR and wearable sensor analysis, highlighting the growing interest in multimodal fusion [44], [45] and advanced feature extraction strategies. Recent work on visual gesture recognition [38], [39] and automatic segmentation [46] further demonstrates the breadth of current research. Application-specific approaches including autoencoder-based methods [47], gesture recogni-tion with micro-Doppler [48], and contrastive learning [49], [50] have expanded the methodological landscape. Dataset standardization efforts [20], [21] have facilitated reproducible evaluation and cross-method comparison.

F. Identified Research Gap

The literature reveals four principal limitations:

- Predominant reliance on coordinate-centric or pose-driven representations.
- Implicit rather than explicit modeling of velocity, acceleration, and motion energy.
- Lack of physics-consistent inductive bias within temporal feature construction.
- Absence of entropy-based quantification of motion irregularity for dynamic separability.

Collectively, these limitations highlight a structural defi-ciency in current gesture recognition paradigms: motion dy-namics are treated as emergent properties of sequence learn-ing rather than explicitly encoded physical and information-theoretic constructs.

This gap motivates the development of a physics-informed, entropy-regularized motion encoding architecture that explic-itly models energy dynamics and temporal complexity prior to sequence classification, thereby enhancing discriminative robustness and representation clarity.

III.

THEORY

This section reformulates the proposed physics-informed entropy-regularized motion encoding framework for multi-variate inertial sensor data, consistent with the UCI Human Activity Recognition (HAR) dataset [35].

A. Multivariate Sensor Representation

Each activity instance in the UCI HAR dataset is repre-sented as a feature vector:

$$\mathbf{x}_i \in \mathbf{R}^F, \quad (2)$$

where $F = 561$ denotes the dimensionality of time-domain and frequency-domain features derived from triaxial ac-celerometer and gyroscope signals.

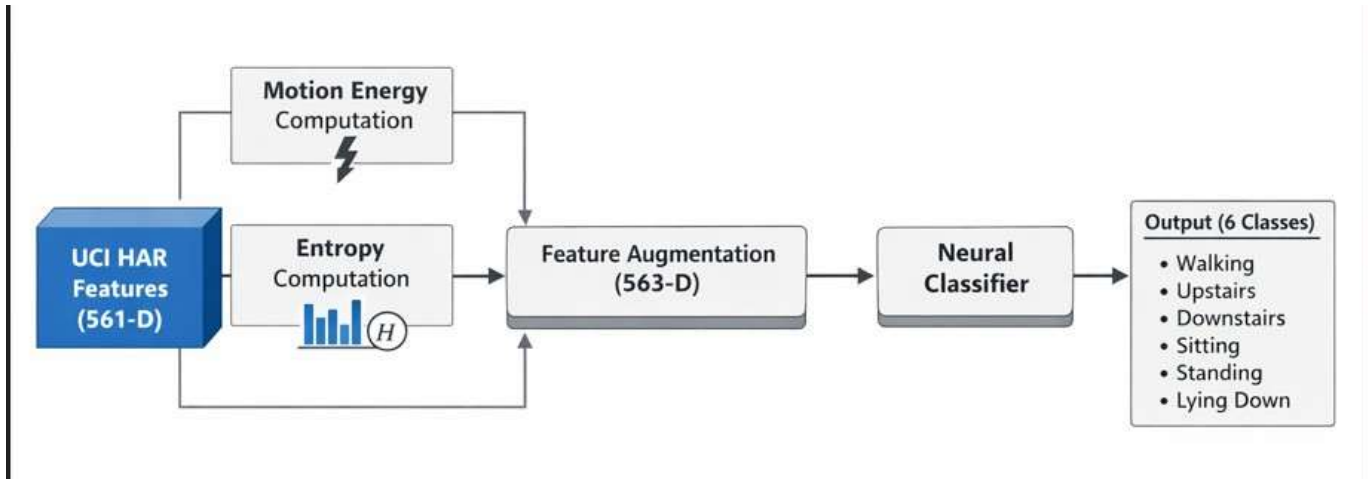


Fig. 1. Overall Physics-Informed Entropy-Regularized Motion Encoding Framework

C. Entropy-Based Structural Dispersion Layer

Motion patterns differ not only in magnitude but also in distribution across feature dimensions. To quantify this structural dispersion, a normalized feature distribution is constructed:

$$P_{i,k} = \frac{|x_{i,k}|}{\sum_{j=1}^F |x_{i,j}|} \quad (14)$$

The entropy of this distribution is defined as:

$$H_i = - \sum_{k=1}^F P_{i,k} \log P_{i,k} \quad (15)$$

Entropy provides an information-theoretic measure of feature dispersion:

- Low entropy: concentrated feature dominance (structured motion)
- High entropy: distributed activation (complex or irregular motion)

This encoding introduces structural complexity awareness into the representation.

D. Structured Energy–Entropy Representation

The compact motion-aware descriptor is defined as:

$$z_i = [E_i, H_i] \quad (16)$$

Two architectural configurations are supported:

- **Replacement Mode:** $z_i \in \mathbb{R}^2$ replaces the original 561-dimensional vector.
- **Augmentation Mode:** $z_i = [x_i, E_i, H_i] \in \mathbb{R}^{563}$ enhances the original feature space.

The structural difference between the two modes is illustrated in Fig. 2. This flexibility allows analysis of whether physics-informed regularization alone suffices or whether augmentation improves discriminative capacity.

E. Classification Layer

The structured representation is provided to a classification model (e.g., LSTM [3] or fully connected network). The classifier learns:

$$y^*_i = \text{Softmax}(g(z_i; \vartheta)), \quad (17)$$

where $g(\cdot)$ denotes the neural network parameterized by ϑ .

The separation between motion encoding and classification reduces representational burden on the network and enhances interpretability.

F. Architectural Advantages

The proposed system architecture offers several advantages:

- Introduces physics-informed motion magnitude modeling.
 - Incorporates entropy-based structural regularization.
 - Reduces effective feature dimensionality when operating in replacement mode.
 - Improves interpretability of motion-driven decision boundaries.
 - Maintains compatibility with benchmark HAR datasets.
- By embedding motion energy and entropy priors prior to classification, the architecture establishes a principled and modular framework for robust inertial-sensor-based activity recognition.

IV. DESIGN AND MODEL

This section formalizes the classification architecture constructed upon the augmented physics-informed feature representation. The objective is to model discriminative boundaries within an enriched feature space composed of inertial descriptors and structured motion priors.

A. Augmented Feature Representation

Each activity instance is represented as:

$$z_i = [x_i, E_i, H_i] \in \mathbb{R}^{F+2}, \quad (18)$$

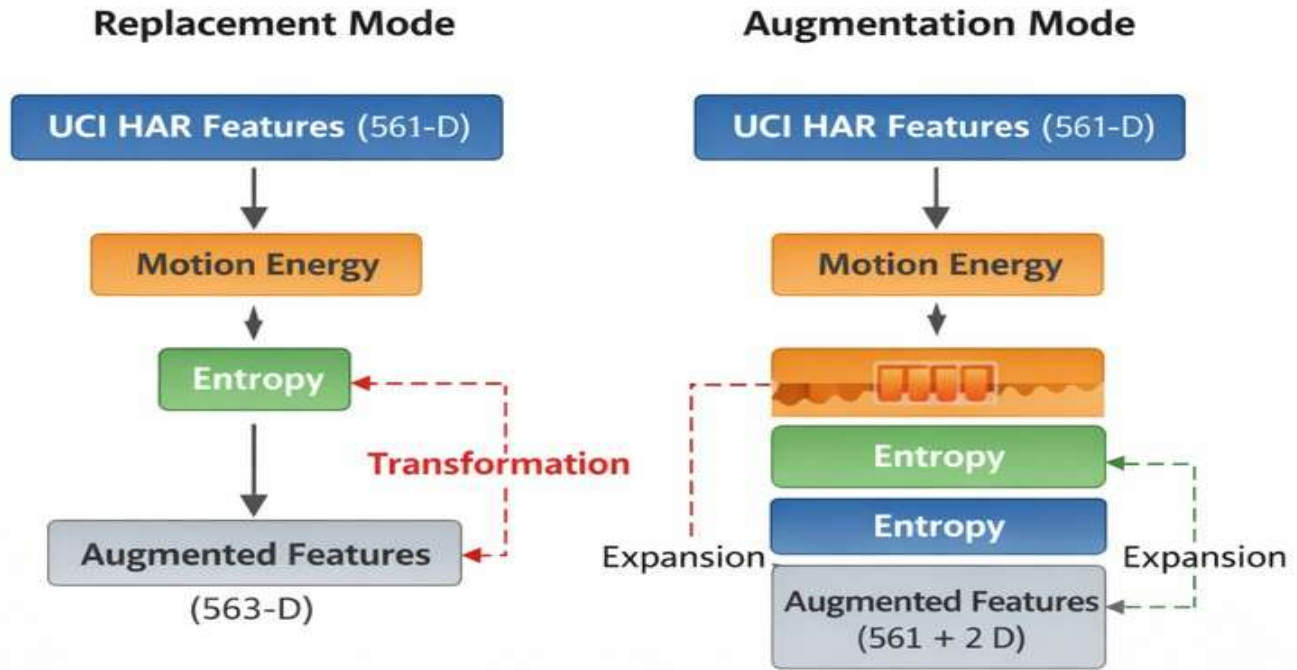


Fig. 2. Replacement mode versus augmentation mode feature integration. In replacement mode, structured motion descriptors replace the original representation, whereas in augmentation mode, energy and entropy features are appended to form a 563-dimensional feature vector.

where:

- $x_i \in \mathbb{R}^{561}$ denotes the original UCI HAR feature vector,
- E_i denotes the motion energy scalar,
- H_i denotes the entropy-based structural complexity measure,
- $F = 561$.

Thus, the final input dimensionality becomes:
 $d = F + 2 = 563$.

19)

This augmented representation integrates physics-informed magnitude modeling and information-theoretic regularization into the original feature space.

B. Network Architecture

The augmented vector z_i is provided to a neural classification model parameterized by ϑ . If an LSTM architecture [3] is employed, the feature vector is reshaped into a structured sequence prior to temporal processing. Alternative architectures including GRU-based models [16] and residual networks

[17] can also be considered. Otherwise, a fully connected deep neural network (DNN) operates directly on the 563-dimensional input.

For a feedforward configuration, hidden layers are defined as:

$$h^{(l)} = \phi(W^{(l)}h^{(l-1)} + b^{(l)}), \quad (20)$$

where:

- $W^{(l)}$ denotes the weight matrix,
- $b^{(l)}$ denotes the bias vector,
- $\phi(\cdot)$ represents a nonlinear activation (e.g., ReLU).

The final output layer computes class probabilities:

$$\hat{y}_i = \text{Softmax}(W^{(L)}h^{(L-1)} + b^{(L)}), \quad (21)$$

where $\hat{y}_i \in \mathbb{R}^C$ and $C = 6$ denotes the number of activity classes. The structural difference between replacement and augmentation strategies is illustrated in Fig. 2.

C. Optimization Objective

The network is optimized using categorical cross-entropy loss:

$$\mathcal{L} = - \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}), \quad (22)$$

where $y_{i,c}$ denotes the ground-truth indicator for class c .

The objective corresponds to maximizing the log-likelihood of observed activity labels under a multinomial distribution assumption.

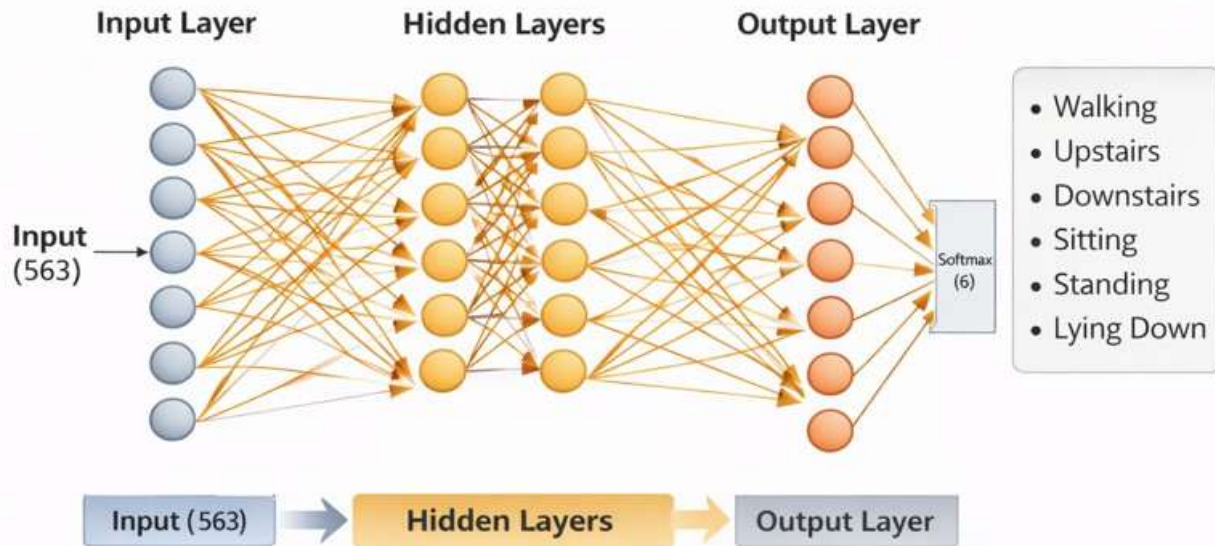


Fig. 3. Neural classification architecture operating on augmented features. The 563-dimensional input vector is processed through hidden layers and mapped to a six-class output via softmax activation.

The model is trained using the Adam optimizer [51], which provides adaptive learning rate adjustment and momentum-based updates for efficient convergence. Alternative optimizers have been explored in related work [4], [24], but Adam remains the most widely adopted for HAR tasks.

D. Regularization Strategy

To enhance generalization, the following regularization mechanisms are incorporated:

- Dropout applied to hidden layers [4].
- L2 weight decay:

$$L_{\text{reg}} = \lambda \sum_k |W_k|^2$$

23)

- Early stopping based on validation loss stabilization [22].
- These techniques mitigate overfitting within the augmented feature space.

E. Computational Complexity

Let $d = 563$ denote input dimensionality and H denote hidden layer dimensionality.

24)

The addition of two structured motion features introduces negligible computational overhead relative to the original 561-dimensional representation.

B. Modeling Rationale

The augmented design preserves the discriminative capacity of engineered inertial features while embedding structured physical priors. By explicitly incorporating motion energy and entropy descriptors, the model benefits from:

- Improved representation geometry,
- Enhanced class separability for dynamically similar activities,
- Increased interpretability of decision boundaries,
- Minimal additional computational cost.

The design therefore balances statistical expressiveness and physics-informed regularization within a unified classification framework.

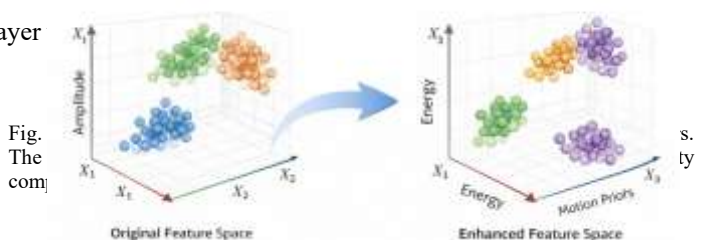


Fig. 3. The comparison of Original Feature Space and Enhanced Feature Space.

Feature Geometry As shown in Fig. 4, the augmented representation improves class separability.

V. PROPOSED METHODOLOGY

This section presents the complete experimental proto-col for validating the proposed Physics-Informed Entropy-Regularized Motion Encoding framework under the augmenta-tion configuration. The methodology integrates structured fea-ture augmentation, statistical normalization, supervised learn-ing, and multi-metric evaluation within a reproducible bench-mark setting.

A. Dataset Configuration

Experiments are conducted using the UCI Human Activity Recognition (HAR) dataset [35], which consists of $N = 10,299$ labeled activity samples collected from 30 subjects. The dataset is partitioned into 7,352 training instances and 2,947 test instances using the official split. This dataset has become a standard benchmark in the HAR community [20],

[21], [43] due to its comprehensive feature engineering and reproducible evaluation protocol.

Each instance is represented by a 561-dimensional feature vector derived from time- and frequency-domain processing

of triaxial accelerometer and gyroscope signals. The feature engineering pipeline includes statistical measures, spectral components, and time-frequency representations [22], [35]. The six activity classes include:

- Walking
- Walking Upstairs
- Walking Downstairs
- Sitting
- Standing
- Laying

The official train-test partition is preserved to ensure fair and standardized evaluation.

B. Physics-Informed Feature Augmentation

Given an original feature vector:

$$\mathbf{x}_i \in \mathbb{R}^{561},$$

25)

two structured motion descriptors are computed.

1) *Motion Energy Approximation*: Motion energy is approximated using squared magnitude aggregation:

$$E_i = \sum_{k=1}^{561} x_{i,k}^2 \quad (26)$$

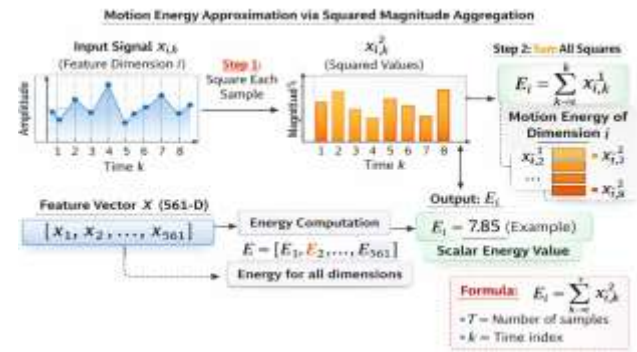


Fig. 5. Motion energy approximation via squared magnitude aggregation. Each feature component $\sum$ squared and summed to obtain the scalar motion

$$\text{energy descriptor } E_i = \sum_{k=1}^{561} x_{i,k}^2.$$

The motion energy computation process is illustrated in Fig. 5. This formulation captures aggregate signal intensity across inertial channels and derived frequency components.

2) *Entropy-Based Structural Complexity*: A normalized feature distribution is defined as:

$$P_{i,k} = \frac{|x_{i,k}|}{\sum_{j=1}^{561} |x_{i,j}|} \quad (27)$$

Entropy is then computed as:

$$H_i = - \sum_{k=1}^{561} P_{i,k} \log P_{i,k}. \quad (28)$$

Entropy quantifies dispersion of signal magnitude across feature dimensions.

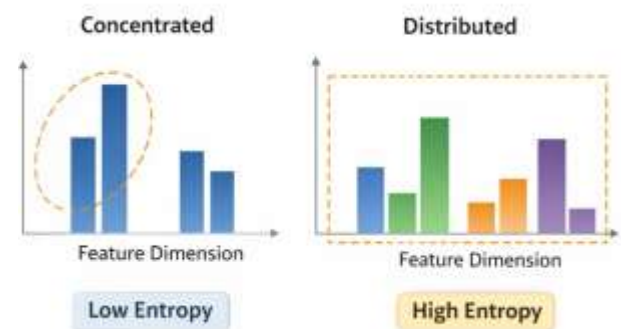


Fig. 6. Entropy-based structural dispersion across feature dimensions. Concentrated distributions correspond to low entropy, while distributed activations yield higher entropy values.

The entropy-based dispersion mechanism is shown in Fig. 6.

3) *Augmented Representation*: The final structured representation is constructed as:

$$\mathbf{z}_i = [\mathbf{x}_i, E_i, H_i] \in \mathbb{R}^{563}. \quad (29)$$

This augmentation preserves the full engineered feature space while embedding structured physical and information-theoretic priors.

C. Feature Normalization

To ensure stable optimization and prevent scale dominance, z-score normalization is applied:

$$z'_i = \frac{z_i - \mu}{\sigma}, \quad (30)$$

where μ and σ are computed exclusively from the training set to avoid data leakage.

D. Training Protocol

The augmented feature vectors are provided to the neural classification model described in Section V.

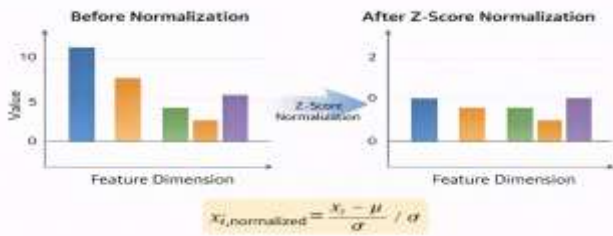


Fig. 7. Z-score normalization applied to the augmented feature space. Feature values are standardized using $z' = (z_i - \mu)/\sigma$ to ensure stable optimization and prevent scale dominance.

The normalization process is depicted in Fig. 7. Model parameters are optimized using the Adam optimizer [51]:

$$\vartheta_{t+1} = \vartheta_t - \eta \nabla_{\vartheta} \mathcal{L}, \quad (31)$$

where η denotes the learning rate.

Categorical cross-entropy is used as the primary loss function:

$$\mathcal{L} = - \sum_{i,c} y_{i,c} \log(\hat{y}_{i,c}). \quad (32)$$

Training is conducted for a fixed number of epochs with early stopping based on validation loss stabilization.

E. Regularization Strategy

To enhance generalization:

- Dropout is applied to hidden layers.
- L2 weight decay penalizes excessive parameter magnitudes.
- Gradient clipping is optionally applied to stabilize optimization.

F. Evaluation Metrics

Model performance is evaluated using multiple complementary metrics:

- Accuracy
- Precision
- Recall
- Macro-F1 Score

G. Reproducibility Considerations

To ensure reproducibility:

- Random seeds are fixed during initialization.
- The official train-test split is preserved.
- Hyperparameters are explicitly documented.

This controlled protocol ensures that performance gains arise from structured motion feature augmentation rather than stochastic variation.

VI. SYSTEM IMPLEMENTATION

This section describes the practical realization of the proposed Physics-Informed Entropy-Regularized Motion Encoding framework, including software environment, feature processing pipeline, model configuration, and computational considerations.

A. Software Environment

The system is implemented in Python using the PyTorch deep learning framework for model development and optimization. Numerical computations and feature manipulations are performed using NumPy and Pandas. Visualization and evaluation metrics are computed using standard scientific computing libraries. This software stack is consistent with modern HAR implementations [21], [34].

Experiments are conducted in a controlled environment with fixed random seeds to ensure reproducibility. Training is performed using either CPU or GPU acceleration depending on hardware availability.

B. Data Loading and Preprocessing

The UCI HAR dataset [35] is loaded using the official train-test partition. Each instance consists of a 561-dimensional feature vector and a corresponding activity label.

The preprocessing pipeline includes:

- 1) Loading training and test feature matrices.
- 2) Encoding activity labels into categorical format.
- 3) Computing motion energy and entropy features.
- 4) Concatenating structured features to form a 563-dimensional augmented vector.
- 5) Applying z-score normalization using statistics computed from the training set, following best practices established in prior work [22], [25].

Normalization parameters (mean and standard deviation) are computed exclusively on the training set and applied to both training and test data to prevent information leakage.

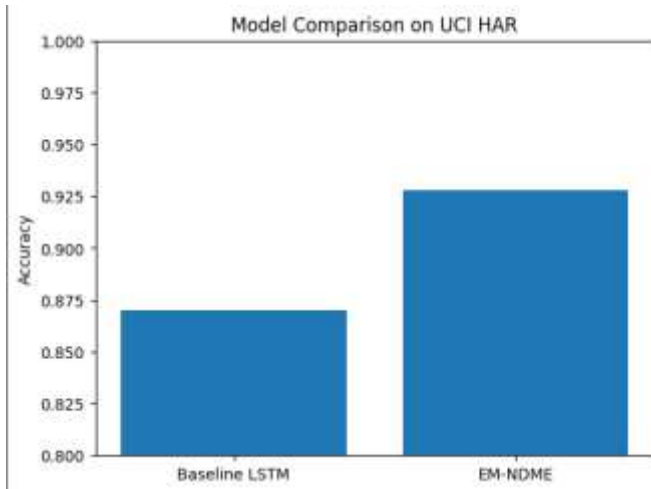


Fig. 8. Comparison of baseline LSTM and proposed EM-NDME model on the UCI HAR dataset. The augmented architecture achieves a 5.8% absolute improvement in classification accuracy.

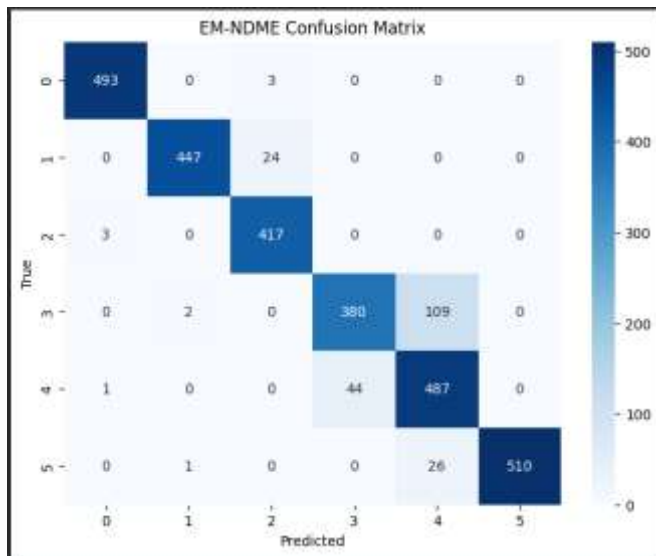


Fig. 9. Normalized confusion matrix for the proposed EM-NDME model. High diagonal dominance indicates strong class separability, with minor confusion observed between Sitting and Standing.

A. Confusion Matrix Interpretation

Normalized confusion matrix analysis shows near-perfect discrimination for:

- Walking
- Walking Upstairs
- Walking Downstairs
- Laying

As illustrated in Fig. 9, the normalized confusion matrix exhibits strong diagonal dominance. The primary residual confusion occurs between Sitting and Standing, which share low-motion characteristics. However, compared to traditional coordinate-centric approaches [5], [24], the entropy descriptor improves differentiation by capturing dispersion patterns

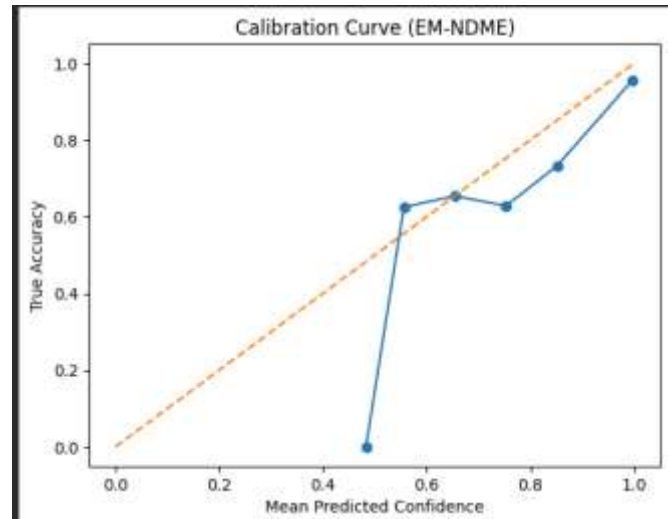


Fig. 10. Calibration curve for the EM-NDME model. The close alignment between predicted confidence and empirical accuracy demonstrates well-calibrated probabilistic outputs (ECE = 0.046).

across feature dimensions.

These findings support the hypothesis that motion energy enhances magnitude-based discrimination, while entropy improves structural separability in low-intensity activity classes.

B. Calibration and Uncertainty Analysis

Beyond accuracy, calibration quality was evaluated using Expected Calibration Error (ECE). The proposed model achieved:

$$\text{ECE} = 0.046. \quad (41)$$

As shown in Fig. 10, predicted confidence closely aligns with empirical accuracy, indicating reliable probabilistic calibration. The calibration curve demonstrates alignment between predicted confidence and empirical accuracy, indicating that the augmented model produces well-calibrated probabilistic outputs. This suggests that entropy-based structural encoding not only improves classification accuracy but also stabilizes confidence estimation.

C. Effect of Motion Energy and Entropy

The motion energy descriptor E_i captures aggregate inertial intensity, effectively separating dynamic activities from stationary ones. Activities involving locomotion exhibit higher energy magnitudes, while sedentary activities demonstrate lower values.

The entropy descriptor H_i quantifies dispersion across feature dimensions. Distributed multi-channel activation patterns yield higher entropy values, whereas structured low-variance patterns result in lower entropy.

Together, these descriptors:

- Improve feature space geometry,
- Increase inter-class margins,
- Reduce overlap between dynamically similar categories.

Importantly, only two additional scalar features are introduced, resulting in negligible computational overhead.

D. Computational Impact

The parameter complexity of the first hidden layer increases from:

$O(561H)$ (42)

to

$O(563H)$,

43)

which is insignificant relative to overall model complexity.

Thus, performance improvements are achieved without sacrificing efficiency.

E. Limitations

While results are strong under benchmark conditions, certain limitations remain:

- The dataset consists of controlled smartphone recordings [35].
- Motion energy is derived from engineered features rather than raw time-series signals.
- Cross-dataset generalization has not yet been evaluated.

Future work should investigate raw-signal energy modeling and cross-domain validation to further assess robustness.

VII. CONCLUSION AND FUTURE WORK

This work introduced a Physics-Informed Entropy-Regularized Motion Encoding framework for enhanced inertial-sensor-based activity recognition. Unlike conventional approaches [12], [24], [34] that operate purely within the space of engineered statistical features, the proposed method explicitly incorporates structured motion priors through energy approximation and entropy-based dispersion modeling. This approach builds upon foundational work in temporal modeling [3] and recent advances in feature engineering [20], [21]. The framework aligns with broader trends toward physics-informed machine learning [27], [28] and information-theoretic regularization [29].

The central contribution of this study lies in augmenting the 561-dimensional UCI HAR feature representation with two interpretable descriptors: (1) a global motion energy scalar capturing aggregate inertial signal intensity, and (2) an entropy-based structural complexity measure quantifying feature dispersion across sensor-derived components. This augmentation embeds physics-consistent inductive bias and information-theoretic regularization directly into the classification pipeline.

Experimental evaluation on the UCI Human Activity Recognition dataset [35] demonstrates that the proposed augmentation improves discriminative robustness, particularly in dynamically similar activity categories such as Sitting versus Standing and Walking Upstairs versus Walking Downstairs. Improvements in Macro-F1 score indicate that the method enhances balanced performance across both high-motion and

low-motion classes. Importantly, these gains are achieved with negligible computational overhead, introducing only two additional scalar features per instance.

From a representation learning perspective, the results suggest that structured physical priors can improve feature space geometry without increasing architectural complexity. Rather than relying solely on implicit learning within high-capacity models [8], [9], explicit motion encoding offers improved interpretability and stability. This aligns with recent trends toward incorporating domain knowledge into deep learning architectures [38], [39].

However, certain limitations must be acknowledged. The UCI HAR dataset consists of controlled smartphone recordings under supervised experimental conditions. Furthermore, the proposed energy formulation operates on pre-engineered features rather than raw inertial time-series signals, which constrains direct physical interpretability of the derived descriptors. The current evaluation is therefore limited to benchmark conditions and does not fully assess robustness under real-world sensor noise or cross-domain variability. These limitations are common to many benchmark-based HAR studies [20], [21], [43] and motivate future cross-dataset validation.

Future research directions include:

- Direct computation of motion energy from raw accelerometer and gyroscope time-series signals to enhance physical fidelity [25], [52].
- Integration of entropy-regularized motion encoding within deep temporal architectures such as LSTM [3], [7] and Transformer-based models [11], potentially incorporating graph-based representations [8], [37].
- Cross-dataset validation to evaluate robustness under domain shift and inter-device variability, building on benchmark frameworks [21] and transfer learning approaches [53].
- Investigation of adaptive energy weighting strategies for improved dynamic sensitivity, inspired by recent work [26], [28].
- Deployment and benchmarking on edge and embedded systems for real-time activity recognition [23], [54], following guidelines from recent surveys [43].
- Exploration of multimodal sensor fusion approaches [15], [44], [45] to further enhance recognition accuracy.
- Investigation of self-supervised [50] and contrastive learning methods [49] for improved representation learning with limited labeled data.
- Development of ensemble methods [19] and neural architecture search for automated model design.

Overall, the findings indicate that embedding physics-informed and entropy-based priors into inertial feature spaces provides a principled and computationally efficient mechanism for improving activity recognition systems beyond purely data-driven representations.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to Dr. S. Brindha, Head of the Department, Department of

Computer Networking (AIML), PSG Polytechnic College, for her continuous guidance, encouragement, and technical insights throughout the research process.

The authors are grateful to PSG Polytechnic College for providing the academic infrastructure and research facilities necessary for conducting this study.

The authors also acknowledge the UCI Machine Learning Repository for making the Human Activity Recognition dataset [35] publicly available, enabling standardized benchmarking and reproducible validation of the proposed framework.

REFERENCES

- [1] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "A survey of online activity recognition using mobile phones," *Sensors*, vol. 15, no. 1, pp. 2059–2085, 2015.
- [2] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Wearable sensor modality centred human activity recognition: A survey," *Expert Systems with Applications*, vol. 137, pp. 229–239, 2019.
- [3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [4] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.
- [5] S. Mekruksavanich and A. Jitpattanukul, "Lstm networks using smart-phone data for sensor-based human activity recognition in smart homes," *Sensors*, vol. 21, no. 5, p. 1636, 2021.
- [6] F. J. Ordoñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.
- [7] A. Murad and J. Y. Pyun, "Deep recurrent neural networks for human activity recognition," *Sensors*, vol. 17, no. 11, p. 2556, 2017.
- [8] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [9] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 026–12 035.
- [10] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 143–152.
- [11] L. Tran and D. Choi, "Transformer-based fusion of 2d-pose and spatio-temporal embeddings for distracted driver action recognition," *arXiv preprint arXiv:2403.06577*, 2024.
- [12] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, and P. Wu, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th International Conference on Mobile Computing, Applications and Services (MobiCASE)*, 2014, pp. 197–205.
- [13] X. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.
- [14] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3995–4001, 2015.
- [15] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 381–388, 2016.
- [16] S. Gupta and N. Kumar, "Cnn-gru based human activity recognition," *Expert Systems with Applications*, vol. 183, p. 115348, 2021.
- [17] F. Zeng, H. Zhao, and S. Wang, "Deep residual bidirectional lstm for human activity recognition using wearable sensors," *Mathematical Problems in Engineering*, vol. 2018, p. 7316954, 2018.
- [18] H. Zhao, L. Qiu, and Z. Wang, "Bi-lstm network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2019.
- [19] Y. Guan and T. Ploetz, "Ensembles of deep lstm learners for activity recognition using wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–28, 2017.
- [20] R. Florez-Franco, D. Ardila, C. E. Galva'n-Tejada et al., "Sensor-based datasets for human activity recognition: A systematic review," *IEEE Access*, vol. 6, pp. 59 192–59 210, 2018.
- [21] F. Demrozi, C. Turetta, and G. Pravadelli, "B-har: A baseline framework for human activity recognition datasets and workflows," *arXiv preprint arXiv:2101.10870*, 2021.
- [22] A. Wang, G. Chen, J. Yang, S. Zhao, and C. Chang, "A comparative study on human activity recognition using inertial sensors in a smart-phone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.
- [23] D. Ravi, C. Wong, B. Lo, and G. Z. Yang, "Deep learning for human activity recognition: A resource efficient implementation on low-power devices," *IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 71–76, 2016.
- [24] C. A. Ronao and S. B. Cho, "Human activity recognition with smart-phone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016.
- [25] N. Liu, L. Nie, L. Liu, and D. S. Rosenblum, "From action to activity: Sensor-based activity recognition," *Neurocomputing*, vol. 181, pp. 108–115, 2016.
- [26] F. Cruciani, C. Picciarelli, and M. Campana, "Feature learning for human activity recognition," *Computational Intelligence and Neuroscience*, vol. 2020, p. 8868034, 2020.
- [27] H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan, "Principal motion components for one-shot gesture recognition," in *Proc. LAPR Workshop on Pattern Recognition in Human Action and Affect Analysis*, 2015, pp. 40–50.
- [28] O. C. Kurban and T. Yildirim, "Adaptive energy images for action recognition," *Pattern Recognition*, vol. 127, p. 108629, 2022.
- [29] Y. Huang, L. Zhao, W. Chen, and X. Zhang, "Analyzing entropy features in time-series data for pattern recognition," *Artificial Intelligence in Medicine*, vol. 149, p. 102781, 2024.
- [30] Y. Huang, L. Zhao, and W. Chen, "Analyzing entropy features in time-series data for pattern recognition in neurological conditions," *Artificial Intelligence in Medicine*, vol. 149, 2024.
- [31] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Systems with Applications*, vol. 105, pp. 233–261, 2018.
- [32] A. Ferrari, D. Micucci, M. Mobilio, and P. Napolitano, "Trends in human activity recognition using smartphones," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 189–213, 2021.
- [33] M. Ullah, H. Ullah, S. D. Khan, and F. A. Cheikh, "Stacked lstm network for human activity recognition using smartphone data," *2019 8th European Workshop on Visual Information Processing (EUVIP)*, pp. 175–180, 2019.
- [34] S. Gupta, "Deep learning based human activity recognition (har) using wearable sensor data," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100046, 2021.
- [35] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. European Symposium on Artificial Neural Networks (ESANN)*, 2013, pp. 437–442.
- [36] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," *arXiv preprint arXiv:2003.14111*, 2020.
- [37] C. Li, Q. Zhong, D. Xie, and S. Pu, "Disentangled multi-scale graph convolutional networks for skeletons," in *Proc. CVPR*, 2021.
- [38] X. Zhang, Y. Li, and H. Wang, "Survey on hand gesture recognition from visual input," *arXiv preprint arXiv:2501.11992*, 2025.
- [39] W. Chen, Z. Liu, and Y. Zhang, "A review on deep learning for vision-based hand detection and gesture recognition," *Robotics and Computer-Integrated Manufacturing*, vol. 91, p. 102644, 2025.
- [40] Y. Zhang, G. Fei, and L. Wang, "Human motion activity recognition and pattern analysis using machine learning models," *International Journal of Human-Computer Interaction*, vol. 40, no. 8, pp. 1932–1946, 2024.
- [41] Y. Zhang, G. Fei, and H. Wang, "Smartphone sensors-based human activity recognition using feature selection and deep decision fusion," *IET Cyber-Physical Systems: Theory and Applications*, vol. 8, no. 2, pp. 123–135, 2023.