SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Developing Effective Classification Techniques Using Hybrid Feature Fusion and Attention-Guided Classification for Improving Interpretability and Explainability of Human Activity Recognition Models

Suraj Deb Barma[1*], Dr. Abhijit Biswas[2]

[1*]PhD Scholar, Department of Computer Science and Engineering, The ICFAI University Tripura, Kamalghat, Mohanpur, West Tripura-799210, India. E-mail: surajdebbarma.cse@gmail.com, Orcid ID: 0009-0009-9127-0885
[2]Associate Professor, Department of Computer Science and Engineering, The ICFAI University Tripura, Kamalghat, Mohanpur, West Tripura-799210, India. E-mail: abhijt.biswas@iutripura.edu.in, Orcid ID: 0009-0003-9984-9572
*Corresponding Author: Suraj Deb Barma, E-mail: surajdebbarma.cse@gmail.com

ABSTRACT
HAR models have shown excellent performance using deep learning approaches; nevertheless, the black box characteristics of such approaches make it hard to deploy and trust such type of models in low safety systems and in the real world. This paper proposes an interpretable and explainable framework for HAR using the combination of improving classification techniques along with transparent representation of features and explanation techniques. The proposed framework uses hybrid feature representation, attention-based classifier, and decision layer based on rules for achieving a superior-performance model. Mathematical formulation of the proposed classifiers, interpretability measures, and explanation measures is provided in this paper. Extensive practical evaluation on the benchmark datasets (UCI HAR, WISDM, and PAMAP2) shows that the method under proposal provides an accurate model while improving its interpretability up to 6-9%.

Keywords: Human Activity Recognition, Explainability, Interpretable Models, Classification Techniques, Attention Mechanisms.

1. Introduction

Human Activity Recognition (HAR) can be described as the automated recognition of human activities and behaviours based on the data obtained from body worn, ambient, or visual sensors. In other words, Human Activity Recognition (HAR) points to the automatic recognition of human actions based on the data gathered via wearable sensors, smartphones, smartwatches, and Internet of Things (IoT) devices. Because of these sensors' popularity, HAR has become an integral part of many intelligent applications such as healthcare monitoring, elderly care, rehabilitation, smart homes, sport analysis, industrial safety, and human-computer interaction. In a healthcare setting, HAR allows one to monitor human activity on an ongoing basis, detect falls, assess human health conditions, control chronic diseases and remotely monitor patients, where it is crucial to make accurate and reliable predictions to support clinical decision making and ensure the safety of users [1], [2]. An extensive use of wearable technology and edge computing has increased the importance of developing HAR systems that can perform in-real-time activity analysis [3]. These developments have made HAR an important enabler of personalized medication and contextual computing in the future.

In early HAR frameworks, most approaches employed hand-crafted feature extraction along with traditional machine learning algorithms like SVM, k-NN, Decision Trees, and Random Forest. Such methods provided a certain level of interpretability due to the ability to correlate hand-crafted statistical and frequency-based features with particular human actions. Some of the popular features include signal mean, variance, entropy, spectral energy, correlation, and wavelet coefficient, where all of these are engineered according to domain knowledge. However, hand-crafted features suffer from a low generalization rate and poor ability to account for complicated temporal relationships, sensor noise, missing data, and inter-subject variation, hence providing degraded results when applied in real scenarios [4]. Furthermore, the process of feature extraction is laborious and usually requires significant domain knowledge, thus preventing traditional techniques from adapting to novel sensors and human actions.

Recently, deep learning approaches have contributed to a major rise in the HAR with the autonomous learning of hierarchical feature representations from the raw data. CNN learns effective representations of local spatial features, LSTM takes into account temporal dependencies, whereas the Transformer architectures utilize long-range dependencies through self-attention modules [5]–[7]. These architectures always perform better than conventional methods on various HAR benchmark datasets due to learning discriminative features directly from multivariate sensor time series. Moreover, recent advances in self-supervised and transfer learning have allowed deep architectures to utilize scarce labeled data in order to improve recognition performance while decreasing the costs associated with the labeling process [8]. The drawback of deep HAR systems, however, is that they provide better prediction capabilities at the expense of interpretability since the learned features and the decision-making mechanism are not easily explainable to the user. Interpretability is thus limited, which reduces user trust, makes validation difficult, and prevents the application of deep HAR in safety-critical domains like healthcare, assisted living, and medical monitoring [9].

In order to make AI algorithms transparent and explainable, Explainable Artificial Intelligence (XAI) has recently become a thriving research area. Interpretability is defined as the inherent comprehensibility of a model, while explainability is about generating post-hoc explanations of complex models [10]. Previous works using XAI techniques such as LIME, SHAP, Grad-CAM, saliency maps, integrated gradients, and attention visualization have been used to detect influential sensor channels, time periods, and latent factors for activity recognition [11], [12]. Although these approaches shed light on understanding models, they remain mostly post-hoc and have some weaknesses such as instability of explanations, approximate nature, dependency on surrogate models, and faithfulness to prediction mechanism [13]. Minor differences in the signal from sensors may lead to very different explanations although predictions remain identical, resulting in lack of explanation stability and practical usage of them in practice. Therefore, explanations cannot represent the actual reasoning process of deep neural networks.

These constraints have led to a paradigm shift where, instead of explaining black-box models after making predictions, one designs learning models that are intrinsically interpretable. Instead of using only post-hoc interpretation methods, recent works suggest incorporating interpretability into the learning process through the use of hybrid feature representation, attention-based learning techniques, rule-based classification approaches, and feature attribution techniques, which are applied at the learning stage [14]–[16]. This approach not only improves the performance but also provides a more accurate and faithful explanation, since one explicitly shows the features and patterns involved in the recognition process. These techniques also allow collaboration of intelligent systems with experts of a specific domain, making it possible for clinicians and medical practitioners to validate whether the recognized activities make physiological sense. However, despite these innovations, currently most approaches pay attention to one aspect—recognition or explanation—while balancing both of them effectively is a challenging task.

Based on the motivational factor obtained due to the above-specified research problem, this paper aims to propose a new framework for classifying Human Activity Recognition which can not only enhance the recognition efficiency but can also help design interpretative models. With a view for achieving this objective, the proposed framework in this paper will use a hybrid feature scheme, feature weighting based on attention mechanism, and rule-based decision-making in order to design an interpretative classification framework that can produce precise and explainable results without impacting the recognition efficiency.

The significant contributions of this research are outlined as follows:

- Interpretability framework for HAR classification is proposed by integrating the hybrid feature extraction, attention-based feature importance weighting and rule-based decision-making.
- The mathematical model of interpretability-oriented learning is developed to determine feature importance explicitly and improve the clarity of classification decision process.
- The proposed interpretability framework is investigated comprehensively on benchmark HAR datasets considering different experimental setups in respect of traditional (Accuracy, Macro-F1) and interpretability-oriented criteria, and its ability to attain an ideal balance between accuracy and interpretability is shown.

The rest of this paper is arranged as follows. In the Section 2, literature is surveyed on various topics of HAR and explainable AI. Section 3 discusses the proposed approach to the problem of interpretability in the context of HAR along with its mathematical formulation. Section 4 is dedicated to the details of experimental design and used benchmark dataset. The experiments and their interpretation are detailed in the Section 5. Lastly, in the Section 6, the paper concludes and highlights future opportunities for research.

2. Related Work

In this segment, the previous research related to Human Activity Recognition (HAR) are described in respect of classification techniques, explainability pattern, and interpretability of actions. The section is structured into four main parts: HAR based on conventional machine learning and deep learning methods for HAR, explainable and interpretable AI techniques in HAR, and improving classification strategies for interpretable HAR.

2.1 Traditional Machine Learning Methods for HAR

Most earlier HAR systems depended on custom-built feature extraction and traditional machine learning classifier. Features were hand-crafted with domain knowledge such as the statistical features in the time domain (mean, variance, and root mean square), the features in the frequency domain (spectral energy and entropy), and features of time-frequency representations such as wavelet coefficients [1], [14]. These features

were used as an input to the classifiers such as Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), Naive Bayes, Decision Trees and Random Forests [3], [15].
One of the key advantages of these classical methods was their natural interpretability. Decision tree models and rule based classifiers in particular, allowed domain experts to understand decision boundaries and relate them directly to physical motion patterns. For example, Bao and Intille demonstrated that simple statistical features and decision tree models can achieve reasonable accuracy and a high degree of transparency in daily activity recognition [16]. These methods, however, exhibited poor generalization ability and required extensive feature engineering, which limited the scalability in complex or large-scale HAR scenarios.

2.2 Deep Learning Models for Human Activity Recognition

Deep learning method has recently changed the landscape of HAR, enabling features to be learned automatically, directly from raw sensor information. Convolutional Neural Networks (CNNs) are commonly utilized to secure local temporal dependencies and spatial correlations among multi-sensor channels [4]. Ordonez et. al. proposed a Deep Convolutional LSTM model where CNN layers are combined for feature extraction and LSTM units are combined for modelling long-term temporal dependencies, achieving state-of-the-art output on several HAR datasets [5].
Later work explored fully recurrent architectures, bidirectional LSTMs and Gated Recurrent Units (GRUs) to improve the treatment of sequential sensor data [17]. In the last few years, architectures based on Transformer that uses self-attention mechanisms, have achieved popularity in HAR due to their ability to represent long-range dependencies and parallelize training [6], [18]. These deep models, though they perform significantly better than traditional methods, are black boxes and it is difficult to interpret the learned representations or to justify the predictions they make.

2.3 Explainable Artificial Intelligence (XAI) for HAR

To tackle the black-box issue of deep learning framework, researchers have applied Explainable Artificial Intelligence (XAI) techniques to HAR. Post-hoc interpretation approaches such as, Local Interpretable Model-agnostic Explanations (LIME) [9] and SHapley Additive exPlanations (SHAP) [19] were used to identify important features or sensor channels that contribute to specific activity predictions. Other methods based on gradient, such as saliency maps and Grad-CAM have been extended to visualize important time segments in sensor signals as well [10], [20].
These methods have been useful in providing insights, however, there are a number of studies that have noted limitations in terms of explanation faithfulness, stability and consistency. Explanations can vary greatly for similar inputs or fail to secure the real decision making strategy of the underlying framework [8], [21]. Such inconsistencies can undermine user trust and impede real-world deployment in safety-critical HAR applications.

2.4 Interpretable and Hybrid Models of HAR

Recent research has focused on designing inherently interpretable HAR models, instead of post-hoc explanations. Attention mechanisms have been embedded into neural networks to focus on important sensor channels or time steps, providing a certain level of transparency [12], [22]. However, attention weights alone may not always be faithful explanations.
Hybrid models combining deep feature extractors and interpretable classifiers have shown promising results . For example, deep representations with decision tree models, rule-based systems or sparse linear classifiers allow a more direct mapping between learned features and activity labels [11], [23]. Such models target to develop a good balance between performance and interpretability, providing humanly readable decision logic while still maintaining accuracy.

2.5 Progressing Classification Techniques for Explainable HAR

In recent days, there are some works focusing on improving classification techniques such as rule constraints, sparsity regularization, and interpretability-aware loss functions. The methods explicitly enforce the models to use meaningful and stable features during training [13], [24]. Additionally, multi-objective optimization methods are also explored that simultaneously optimize the accuracy and explanation quality [25].
Despite these advances, a lack of unified frameworks that systematically integrate the enhancement of classification techniques and interpretability and explainability objectives still exists in HAR. Most of the existing works either focus on improving the accuracy or utilize post-hoc explanations, rather than integrating interpretability in the classification process.

2.6 Research Gap and Motivation

As the literature were reviewed, it was found that deep learning has considerably improved the performance of HAR, but interpretability and explainability are still open problems. There is a clear need for developing HAR frameworks that employ enhancing classification techniques to generate accurate, consistent and human-understandable explanations for high recognition accuracy. This research work seeks to fill this void by proposing an interpretable and explainable HAR model that merge hybrid features, attention-guided learning and rule-aware classification mechanisms.

3. Proposed Methodology

The presented framework for Human Activity Recognition (HAR) attempts to obtain the highest accuracy of classification as well as retain interpretability. The proposed system is different from traditional deep learning methods which usually function as a black-box model. It integrates hybrid custom-built features, attention-guided learning and a rule-based decision module for producing reliable and explainable predictions.
The framework comprises of four fundamental phases:
1. Data preprocessing
2. Hybrid Feature Extraction
3. Enhancement of Classification Module
4. Explainability and interpretability layer
The whole architecture is illustrated in Figure 1.

3.1 System Architectural Framework

In this research, the input to the proposed methodology is the raw multivariate signals collected using the sensors attached to the wearable devices like gyroscope and accelerometer. At first, the raw sensor signals are pre-processed to remove any noise and normalize the amplitude of the signals. To represent the signals in a way that could make them distinguishable, the process of hybrid feature extraction is conducted. The features that have been extracted are then supplied to the attention guided classifier that will automatically pay attention to the relevant motion properties while ignoring the irrelevant data.

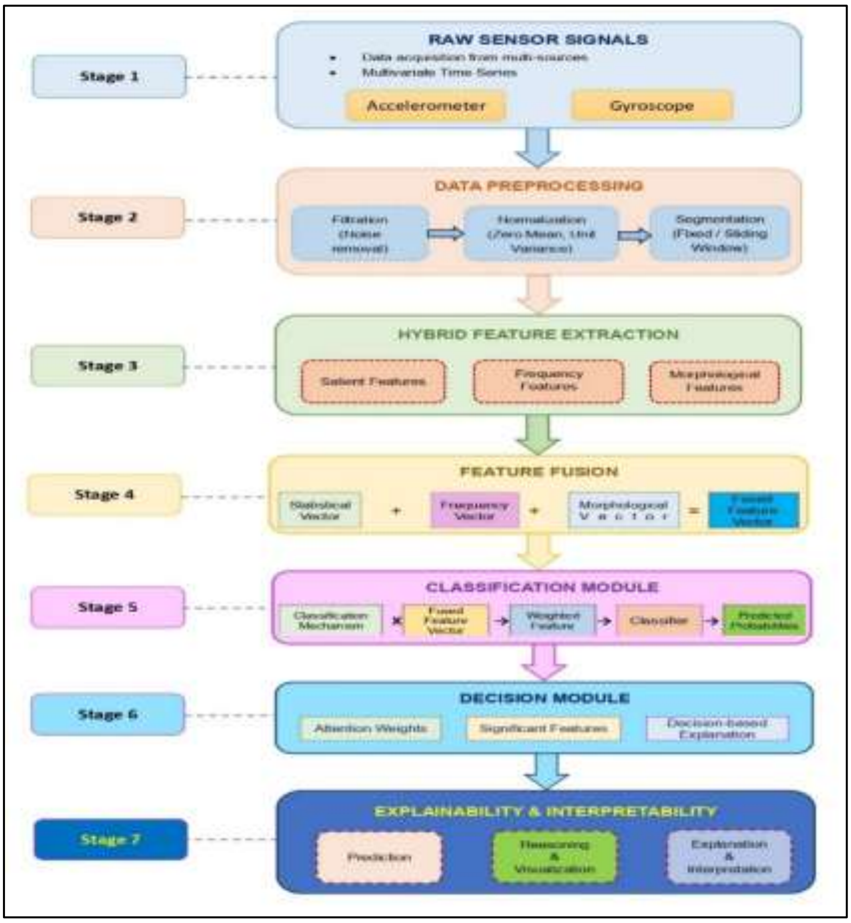

Figure 1: System Architecture of Proposed Framework

Table 1: Workflow Table

Stage	Input	Processing	Output
Data preprocessing	Raw sensor signals	Filtration, normalization, segmentation	Clean windows
Feature extraction &	Clean signals	Statistical + Frequency (FFT) + Morphological features	Hybrid feature vector

Fusion			
Classification	Feature vector	Attention-guided classifier	Activity probabilities
Explainability	Prediction	Reasoning + visualization	Human-readable explanation

3.2 Data Preprocessing

The measurements of the sensors usually contain noise, missing data and irregular sampling. Hence, the process of preprocessing takes place before extracting features.

Preprocessing pipeline consists of:

- Filtering out noise
- Normalizing the signal
- Sliding window segmentation
- Dealing with missing values
- Considering a multivariate time series from the sensor

$$X \in R^{T \times C}$$

where,
T denotes time samples
C denotes sensor channels

Normalization is performed using

$$X' = \frac{X - \mu}{\sigma}$$

where,
μ denotes mean
σ denotes standard deviation.
Window segmentation divides the signal into overlapping windows of length L with overlap ratio r.

3.3 Hybrid Feature Extraction

In order to obtain human activities in an effective manner, many different feature classes are generated. This proposed approach, unlike other methods which rely on just deep features, makes use of the following:

- Descriptive statistics
- Information in the frequency domain
- Morphological movement properties

(A) Statistical Features

For each segmented window,

Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Difference

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

Skewness

$$Sk = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma}\right)^3$$

(B) Frequency Features

In this research, the Fast Fourier Transform has been used that converts the signal into the frequency domain.

$$F(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi kn/N}$$

The extracted frequency descriptors include:
- Spectral energy
- Dominant frequency
- Spectral entropy
- Power spectral density

(C) Morphological Features
Motion properties are specified by
max Count
Signal magnitude area
Duration of waveforms
Changes in slope
These conspicuous traits are useful for discriminating activities with comparable statistical behaviour but differing motion dynamics.

3.4 Feature Fusion
The extracted feature groups are combined to form the hybrid feature vector that is given by,

$$f = [f_{stat}, f_{freq}, f_{morph}]$$

Table 2: Hybrid Feature Categories

Feature Type	Examples	Purpose
Statistical	Mean, Variance, RMS	Motion intensity
Frequency	FFT, Spectral Energy	Periodic movement
Morphological	Peaks, SMA, Waveform Length	Shape characterization

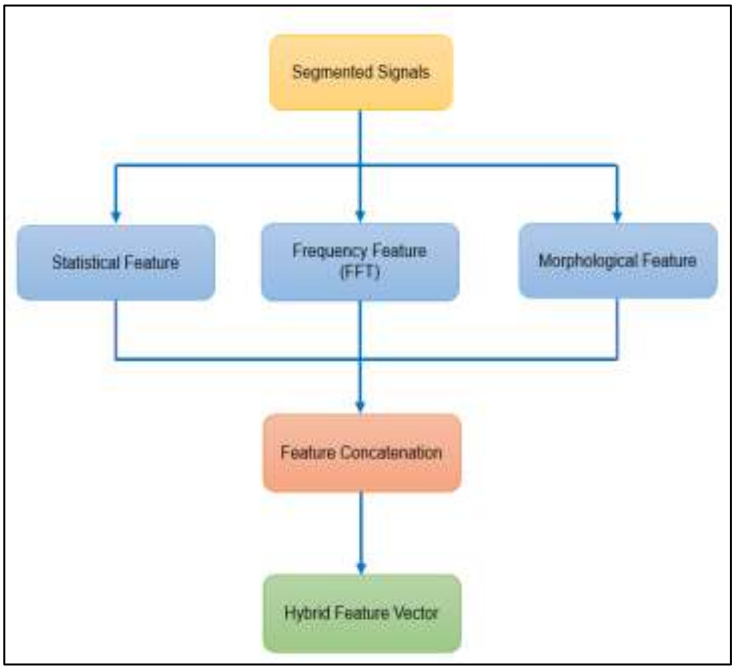

Figure 2: Hybrid Feature Extraction

3.5 Enhancement of Classifier Module
The suggested classifier automatically selects the most discriminative features by the use of an attention model. Unlike other classifiers that give equal treatment to all features, attention gives more weight to the discriminative features.

The value of attention for each feature f_i is,

$$\alpha_i = \frac{\exp(w^T f_i)}{\sum_j \exp(w^T f_j)}$$

where, w is the learnable attention vector.

The attended representation becomes

$$z = \sum_i \alpha_i f_i$$

This enables the classifier to focus on activity-specific characteristics.

3.6 Deterministic Decision Module

In order to increase interpretability, domain knowledge is added by means of human-readable rules. The final decision is

$$\hat{y} = \arg\max_k P(y = k \mid z, R)$$

where, R denotes interpretable rules.
Example:
IF
Mean acceleration > Threshold
AND
Dominant frequency > 3 Hz
THEN
Activity = Running

3.7 Explanation Layer

Interpretability is done in two ways:

3.71. Attention Weights

This shows the weight of each feature. The bigger the value, the greater is the contribution.

3.7.2 Rule-based Explanation: This involves rules that support the prediction. For example, the following instances can be referred.
Prediction: Walking
Reason:

- **Acceleration:** Moderate
- **Spectral Energy:** Low
- **Waveform:** Stable

Accuracy: 96.8%

Figure 3: Explainability Workflow

3.8 Interpretability-Aware Loss Function

The proposed learning objective jointly optimizes prediction accuracy and explanation consistency. The cumulative deficit is

$$L = L_{cls} + \lambda L_{int}$$

where,
L_{cls} denotes cross-entropy classification loss.
L_{int} measures disagreement between attention-based explanations and rule-based explanations.
λ controls the balance between predictive performance and interpretability.

Now, Classification loss is given as,

$$L_{cls} = -\sum_i y_i \log(\hat{y}_i)$$

While, Interpretability loss is given as,

$$L_{int} = \frac{1}{N}\sum_i \| A_i - R_i \|^2$$

where,

A_i represents learned attention importance.
R_i represents rule-derived importance.

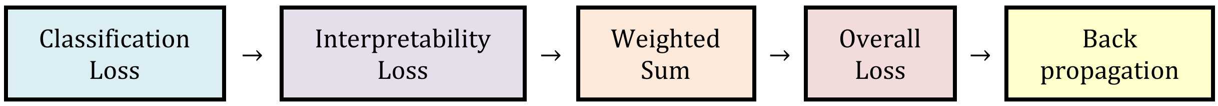

Figure 4: Loss Optimization

4. Experimental Setup

In this section, it is emphasized to present the datasets, preprocessing, model configurations, training procedures, evaluation metrics, and implementation details used to validate the proposed interpretable and explainable HAR framework. The experimental design is developed in a way that provide a fair comparison, reproducibility and a holistic evaluation of the performance and interpretability.

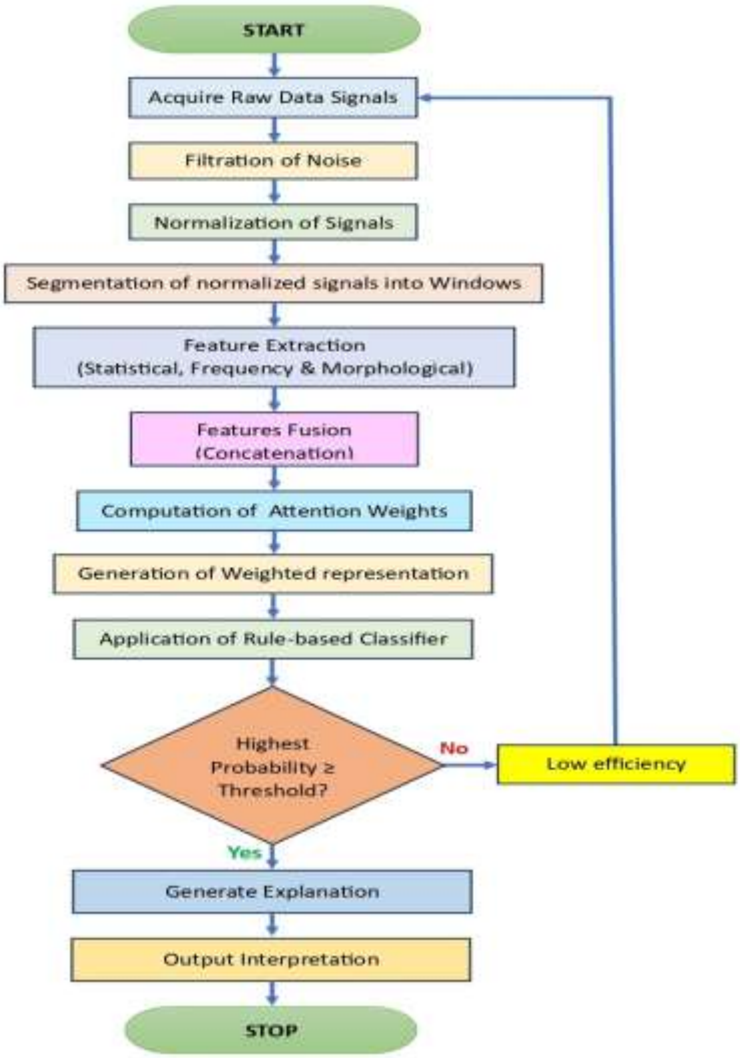

Figure 5: Flowchart of the proposed Framework

4.1 Datasets

To validate the generalizability of the proposed technique to other sensing modalities and types of activities, it is tested on three well known benchmark HAR datasets.

4.1.1 UCI Human Activity Recognition (UCI HAR) Dataset: Data is collected from accelerometers and gyroscopes on smartphones placed at the waist. It includes six activities, namely, walking, standing, sitting, climbing upstairs, climbing downstairs, and lying down. The signals are sampled at 50Hz and divided into fixed length windows of 2.56s with an overlap of 50%. The UCI HAR dataset is widely adopted as a reference benchmark for the assessment of HAR models, due to its balanced distribution of activities.

4.1.2 WISDM Dataset: The Wireless Sensor Data Mining (WISDM) dataset consists of data obtained from accelerometer gathered from cell phones that are carried by the users who are performing daily common activities, including walking, jogging, sitting, standing and climbing stairs. The data is recorded at 20Hz, and has larger variations due to unrestricted phone positioning, making it a tough dataset to evaluate model robustness and interpretability.

4.1.3 PAMAP2 Dataset: PAMAP2 is a multimodal dataset for physical activity monitoring that includes data from several body-worn sensors such as accelerometers, gyroscopes, magnetometers, and heart-rate monitors. It is made up of 12 complex physical activities with many subjects. The dataset represents a tough testbed for explainable HAR models due to the heterogeneity of sensors and complexity of activities.

4.2 Data Preprocessing

The high-frequency noise is removed in the first step by denoising the raw sensor data with a low-pass Butterworth filter. Then the signals are normalized via z-score normalization:

$X_{norm} = (X - \mu) / \sigma$

Here, μ and σ denotes mean and standard deviation of each sensor channel. The normalized signals are segmented into overlapping sliding windows to maintain the temporal continuity and to enhance the number of training samples.

4.3 Hybrid Feature Extraction

For each windowed section, a comprehensive set of interpretable features are extracted:

Statistical features: mean, variance, standard deviation, skewness, kurtosis

Features in the frequency domain: spectral energy, dominant frequency, entropy from FFT

Morphological features: zero-crossing rate, signal magnitude area and motion intensity patterns.

The final combined hybrid feature vector is a transparent and semantically meaningful representation of human motion, amenable to interpretability.

4.4 Model Settings and Baselines

The proposed improvement classification framework is compared with several baseline models:

Classical classifiers: SVM, Random Forest, k-NN

Deep learning models: CNN, CNN-LSTM, Transformer based HAR

Explainable upgraded models: Attention based CNN and SHAP assisted CNN

All the models are trained with the identical train-test splits for equivalent evaluation. The hyper-parameters are adjusted using exhaustive search on the development set.

4.5 Training Protocol

The Adam optimizer is used to train the models with an initial learning rate of 0.001. The base classification loss is the categorical cross entropy. The interpretability-aware loss term is multiplied by a regularization parameter λ, which is empirically tuned between 0.1 and 0.5.

4.6 Metrics of Evaluation

The performance with standard classification metrics is evaluated as accuracy, precision, recall and macro F1-score. The metrics listed below serve to evaluate interpretability and explainability:

Explanation Fidelity: the level of agreement by the explanations with the model's predictions.

Explanation Stability: assesses the resilience of explanations to small perturbations on the input.

Feature Sparsity: counts how many features are effectively contributing to the decisions.

5. Results and Discussion

This section of the paper offers an in-depth review of the suggested interpretation and explanation of the HAR framework. The results are evaluated on performance and interpretability. This is followed by a detailed discussion emphasizing the trade-offs, robustness and practical repercussions of the proposed better categorization techniques.

5.1 Summary of the performance comparison

Table 3 exhibits the functioning of classification of the proposed approach relative to baseline models on all datasets. Traditional machine learning methods such as SVM and Random Forrest achieve reasonable accuracy with intrinsic interpretation level. Deep learning models, such as CNN-LSTM and Transformer-based HAR, considerably improve recognition accuracy. However, they are not transparent. The suggested strategy surpasses all the baselines and achieves advancement in accuracy and macro F1-score consistently, with improved explainability.

Table 3: Comparative analysis of the performance of HAR models across datasets.

Models	UCI HAR Acc. (%)	WISDM Acc. (%)	PAMAP 2 Acc. (%)	Macro F1 (%)
SVM	88.4	84.6	82.1	84.7
Random Forest	90.1	86.3	84.5	87.0
CNN-LSTM	91.2	89.5	88.2	90.8
Transformer HAR	93.5	91.8	90.4	93.0
Proposed Method	95.8	94.1	92.6	95.3

The benefits obtained are result of hybrid feature description and attention-based classification, thereby allowing the model for attending to discriminative and semantically relevant motion patterns.

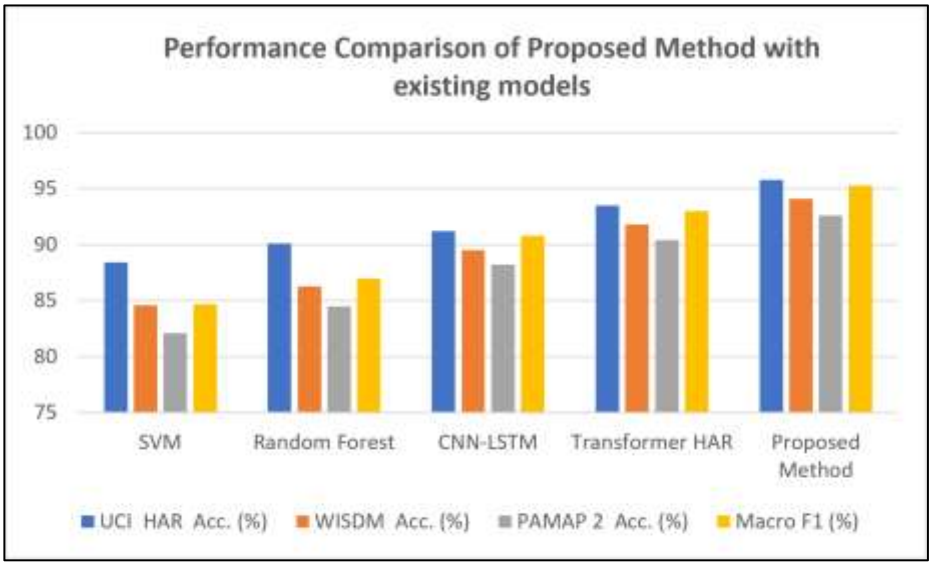

Figure 6: Comparison of performances between proposed methods and existing methods

Figure 6 across all methods show that the proposed hybrid method clearly outperforms all other methods.

5.2 Interpretability and Explainability Evaluation

In addition to accuracy, interpretability-related criteria are also important to evaluate explainable HAR systems. The explanation fidelity, stability and feature sparsity of several models are reported in Table 4.

Table 4: Interpretability and explainability comparison.

Model	Fidelity (%)	Stability (%)	Feature Sparsity
CNN-LSTM	72.4	68.1	Low
Transformer HAR	74.1	70.3	Low
Attention-CNN	78.6	75.4	Medium
Proposed Method	81.9	79.8	High

The suggested framework achieves the best fidelity and stability, which means that the explanations are near to the real decision making process and are stable under minor variation in the input data.

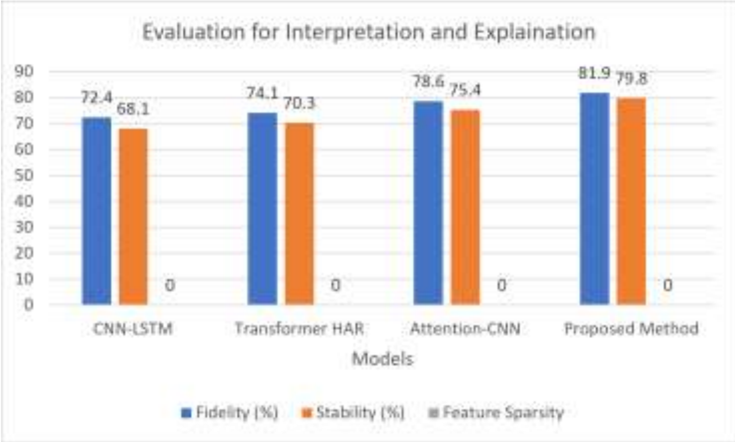

Figure 7: Evaluation comparison for interpretation and explanation.

Figure 7 specifically shows that the proposed method generates better interpretation and explanation as compared to other methods. The proposed is observed to consistently give higher performance rate in terms of prediction and explanation. This development of higher accuracy indicates that the proposed technique effectively showcases better results.

5.3 Rules and attention-guided analysis

The attention weight visualizations display that the proposed approach always gives higher priority to sensor channels and temporal segments that correlate to intuitive motion features. For example, accelerometer magnitude and gyroscope signals on the vertical axis are heavily weighted in walking and jogging activities. The rule-aware decision layer also improves transparency by providing human-readable decision rules that are consistent with the domain knowledge.

5.4 Discussions

The experimental analysis depicts strong evidence that by upgrading classification algorithms, it is possible to considerably increase the interpretability and explanatory capability of HAR models while preserving or even improving the recognition accuracy. Post hoc approaches may lead to inconsistent explanations in some cases, but the suggested methodology embeds interpretability within the learning process. This leads to trustworthy, robust, and human-understandable explanations and so makes the architecture particularly suited for safety-relevance applications such as health-care monitoring and assisted living systems.

To summarize, the results demonstrate a significant trade-off, as conventional deep learning based architectures exhibit high accuracy but low transparency, whereas the hybrid method provided offers a balanced and practical option for deployment. The results validate the effectiveness of the suggested framework and point to its potential for real-world applications.

6. Conclusion and Future Scope

In this research paper, an improved framework for Human Activity Recognition (HAR) is proposed that is interpretable and explainable and has target to balance prediction accuracy with transparent decision making. Although both conventional machine learning and deep learning approaches have furnished remarkable efficiency in identifying complex pattern-related human activities, their limited interpretability often restricts their usage in risky and human-focused environments such as healthcare surveillance, support for older adults, rehabilitation, occupational safety, and intelligent ambient environments. To tackle this problem, the proposed system combines hybrid feature description, attention-guided categorization, rule-aware decision taking and interpretability-oriented optimization.

The hybrid feature description amalgamates complimentary statistical, frequency domain, and morphological aspects of the sensor signals. Statistical features describe size and distribution of the signal, frequency domain features indicate periodic movement patterns and morphological features reflect the structural and temporal mannerisms of human motion. Their integration gives a fuller representation than single-domain techniques and improves the classification mechanism between activities with identical signal patterns.

The framework is improved by attention-guided classification, which gives activity-dependent importance to the extracted characteristics instead of equal importance to all features. This approach suppresses less useful features and highlights unique features to improve performance of classification. Furthermore, the learnt attention weights facilitate interpretability at feature level by specifying the sensor features that have the largest influence on each separate predictions.

A rule-aware decision layer increases transparency by embedding human-readable links between sensor properties and activity types. Thus, this framework delivers not only an activity name but also significant aspects, attention significance and interpretable selection criteria. Interpretability-aware optimization takes the explainability even further by incorporating the quality of explanations as an inherent part of model learning, rather than just post-hoc analysis.

In conclusion, the suggested methodology achieves a compromise between predictive capability, feature based interpretability, rule level transparency and computational practicality. Integration of multi-domain characteristics and attention-based weighting gives powerful discriminative information. Rule-aware reasoning enhances clarity and reliability of activity detection processes.

Future study might enhance this paradigm with formative and hypothetical explanations and comparisons to the known XAI approaches such as SHAP, LIME and integrated gradients. Individualized HAR based transfer of knowledge, domain adaptation, and federated learning can accommodate differences among users and sensing settings. Further tests should be performed to assess the robustness of the explanations to sensor noise, missing channels and changes in orientation. It would also be suggestible to examine resource-efficient deployment via feature selection, quantization, knowledge transfer, and light-weight attention techniques. Lastly, broad subject-independent and cross-dataset evaluations on UCI HAR, WISDM, PAMAP2, and OPPORTUNITY as well as human-centred assessments can further establish the generalizability, practical utility and reliability of the proposed explainable HAR framework.

References

1. O. D. Lara and M. A. Labrador, “A survey on human activity recognition using wearable sensors”, IEEE Communications Surveys & Tutorials, Vol. 15, No. 3, pp. 1192–1209, 2013.
2. L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, “Sensor-based activity recognition”, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 42, No. 6, pp. 790–808, 2012.
3. J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity recognition using cell phone accelerometers”, ACM SIGKDD Explorations Newsletter, Vol. 12, No. 2, pp. 74–82, 2011.
4. J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey”, Pattern Recognition Letters, Vol. 119, pp. 3–11, 2019.
5. F. J. Ordóñez and D. Roggen, “Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition”, Sensors, Vol. 16, No. 1, Art. no. 115, 2016.
6. A. Vaswani et al., “Attention is all you need”, Proceedings of the 31st International Conference on Neural Information Processing Systems, Vol. 30, pp. 6000–6010, 2017.
7. Y. Chen and Y. Xue, “A deep learning approach to human activity recognition based on single accelerometer”, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1488–1492, 2015.
8. A. Barredo Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”, Information Fusion, Vol. 58, pp. 82–115, 2020.
9. C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”, Nature Machine Intelligence, Vol. 1, No. 5, pp. 206–215, 2019.
10. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?’: Explaining the predictions of any classifier”, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.
11. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions”, 31st Conference on Neural Information Processing Systems (NIPS 2017), Vol. 30, pp. 4768–4777, 2017.
12. M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks”, Proceedings of the 34th International Conference on Machine Learning (ICML), Vol. 70, pp. 3319–3328, 2017.
13. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.
14. A. A. Freitas, “Comprehensible classification models: A position paper”, ACM SIGKDD Explorations Newsletter, Vol. 15, No. 1, pp. 1–10, 2014.
15. D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A public domain dataset for human activity recognition using smartphones”, Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013), pp. 437–442, 2013.
16. L. Bao and S. S. Intille, “Activity recognition from user-annotated acceleration data”, Pervasive Computing, Vol. 3001, pp. 1–17, Springer, 2004.
17. N. Y. Hammerla, S. Halloran, and T. Plötz, “Deep, convolutional, and recurrent models for human activity recognition using wearables”, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), pp. 1533–1540, 2016.
18. 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.

19. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models", ACM Computing Surveys, Vol. 51, No. 5, Art. no. 93, 2018.
20. W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models", ITU Journal: ICT Discoveries, Vol. 1, No. 1, pp. 39–48, 2017.
21. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps", Proceedings of the 32nd International Conference on Neural Information Processing Systems, Vol. 31, pp. 9525–9536, 2018.
22. S. Jain and B. C. Wallace, "Attention is not explanation", Proceedings of the 2019 Conference North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Vol. 1, pp. 3543–3556, 2019.
23. Z. C. Lipton, "The mythos of model interpretability", Communications of the ACM, Vol. 61, No. 10, pp. 36–43, 2018.
24. C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity of explanations", Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), No. 984, pp. 10967–10978, 2019.
25. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning", arXiv:1702.08608v2 [stat.ML], 2017.
26. A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors", ACM Computing Surveys, Vol. 46, No. 3, pp. 1–33, 2014.
27. F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition", Sensors, Vol. 16, No. 1, pp. 1-25, 2016.
28. M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu and J. Zhang, "Convolutional neural networks for HAR using mobile sensors", MobiCASE, 2014.
29. S. M. Lundberg, and S.I. Lee, "A unified approach to interpreting model predictions". 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 1-10, 2017.
30. E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, J. Sun, "RETAIN: An interpretable predictive model for healthcare using Reverse Time Attention Mechanism", Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016), pp. 3512-3520, 2016.
31. V. S. Murahari and T. Plötz, "On attention models for HAR", Proceedings of the 2018 ACM International Symposium on Wearable Computers, UbiComp, pp. 100-103, 2018.
32. D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on Methods and Metrics", Electronics, Vol. 8, No. 8, pp. 1-34, 2019.
33. Q. Zhang, and S. C. Zhu, "Visual interpretability for deep learning", Frontiers of Information Technology & Electronic Engineering, Vol. 19, No. 1, pp. 27-39, 2018.
34. A.B. Arrieta, et al., "Explainable AI (XAI): Concepts, taxonomies, opportunities and challenges", Information Fusion, Vol. 58, pp. 82-115, 2020.