Dual-Phase Neural Networks for Feature Extraction and Ensemble Learning for Recognizing Human Health Activities

Joy Dhar[a], Kapil Rana[a,g], Puneet Goyal[a,f], Azadeh Alavi[b], Rajib Rana[c], Bao Quoc Vo[d], Sudeepta Mishra[a], Sajib Mistry[e]

[a]*Department of Computer Science and Engineering, Indian Institute of Technology, Ropar, Punjab, 140001, India*
[b]*School of Computing Technologies, RMIT University, Melbourne, VIC, 3000, Australia*
[c]*School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield Central QLD, 4300, Australia*
[d] *School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Hawthorn, VIC, 3122, Australia*
[e]*School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA, 6102, Australia*
[f]*NIET, NIMS University, Jaipur, Rajasthan, 303121, India*
[g]*Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab, 147004, India*

Abstract

The integration of smart devices into healthcare has led to the creation of vast amounts of sensor data, which are crucial for advancing various healthcare applications such as elderly care, lifestyle enhancement, and health monitoring. Human Activity Recognition (HAR), which relies on these data, is essential for the success of these applications. While Deep Learning (DL) methods, particularly Convolutional Neural Networks (CNN) and Machine Learning (ML), have been somewhat successful in HAR, they often face performance limitations. These limitations arise from the challenges of extracting complex features from sensor-based HAR data and dealing with noise. Current methods often rely on a single-phase feature extraction process. In contrast, adopting a multi-phase feature extraction approach, which rigorously performs feature extraction across multiple distinct phases, could more effectively address these challenges. To overcome these challenges, we introduce a novel hybrid framework named Dual-Phase Fused Neural Networks with Ensemble Learning (DP-FusedNN-EL), designed to achieve robust fea-

ture extraction and enhanced human activity recognition tasks. This model operates in two main stages: dual-phase feature extraction and classification. Initially, it employs two neural networks for feature extraction: a novel Dual-Head Fused CNN for local features and a CNN combined with a Stacked Bidirectional Gated Recurrent Unit and Attention network for local-global features. Subsequently, it utilizes a Dual-Phase Ensemble Learning model for classification, aiming to reduce overfitting by leveraging the strengths of local-global features. We evaluated our DP-FusedNN-EL model on several HAR datasets, achieving remarkable performance with accuracies ranging from 87.47% to 99.66%. These results significantly outperform existing models, demonstrating the effectiveness of the DP-FusedNN-EL model in HAR tasks.

Keywords: Human Activity Recognition, Deep Learning, Ensemble Learning, Feature Extraction, Feature Fusion, Attention Mechanism

1. Introduction

Human Activity Recognition (HAR) has diverse applications, including research on human behavior, ubiquitous computing, and development of human-computer interfaces [1]. HAR has gained attention in smart home applications [2], healthcare [3], rehabilitation [4], surveillance [5], and gait analysis [1]. In healthcare, HAR plays a crucial role in classifying everyday human activities, ranging from simple tasks like walking to more complex ones like cycling, by analyzing various measurements [6]. The use of sensors in HAR facilitates early disease detection, improves fitness, enhances elderly care, and manages patient records, ultimately leading to improved patient health [6].

Sensory systems for data collection in HAR encompass diverse technologies like Wi-Fi, acceleration, audio, infrared, depth cameras, smartphones, and Bluetooth, each providing distinct advantages based on the application environment [1, 12]. Generally, activity recognition methods are categorized into visual-based, audio-based, and sensor-based approaches [1, 13]. Visual systems use cameras to record and detect human physical activities continuously, facing challenges related to confidentiality and cost [1]. Sensor-based systems directly interact with the body using devices like smartphones and Inertial Measurement Units (IMUs) to record activities [1]. Predicting human activities in HAR presents complex challenges but offers diverse bene-

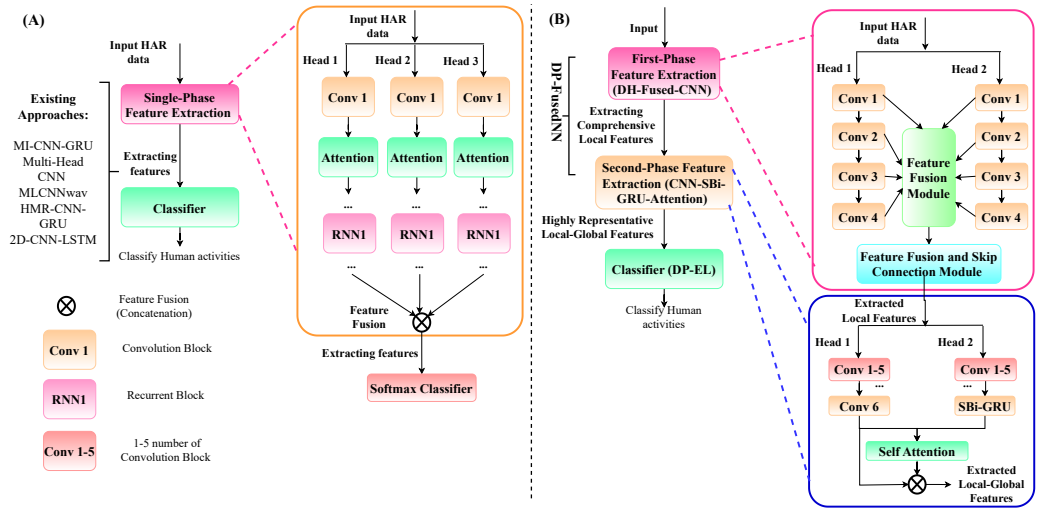

Figure 1: Comparison of feature learning in the HAR model: (A) Features extracted using existing single-phase methods, including MI-CNN-GRU [7], Multi-Head CNN [8], MLCNNwav [9], HMR-CNN-GRU [10], and 2D-CNN-LSTM [11]; (B) Our proposed dual-phase feature extraction approach (DP-FusedNN), which leverages DH-Fused-CNN (first phase) and CNN-SBi-GRU-Attention (second phase) to learn highly representative local and global features, enhancing human activity classification through the DP-EL method. Key components of the DH-Fused-CNN and CNN-SBi-GRU-Attention for improving performance in HAR tasks include: the convolution block, recurrent block (SBi-GRU), feature fusion module, skip connection with the feature fusion module, and attention mechanisms like self-attention. These elements are crucial for enhancing feature extraction and representation in the model.

fits across various applications [1]. Mobile sensors, including accelerometers and gyroscopes, are ideally suited for integration into building structures or portable devices [1]. Thus, these sensors convert motion into detectable signals crucial for HAR [1]. However, this approach has limitations in fully capturing all stance phases of the humanoid body, potentially affecting performance [1]. In industrial settings, employing multiple sensors enhances activity recognition accuracy and effectiveness [1, 14]. These sensors enable tracking vital metrics and mitigating health risks with cost-effective and precise recognition, driving interest in activity recognition [1, 6].

Various Deep Learning (DL) and Machine Learning (ML) algorithms are widely utilized across various industries, such as healthcare, excelling in various tasks including classification [15, 16] and object detection [17, 18]. These algorithms independently learn complex features from extensive large datasets without manual input. Smartphone sensors pose challenges for HAR tasks, where feature extraction plays a pivotal role and utilizes various strategies. Traditional ML models, such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT), Linear Discriminant Analysis (LDA), and Kernel Principal Component Analysis (KPCA), Group Feature Selection (GFS), Randomized SVM, Cooperative Genetic Algorithm (CGA), and Hybrid Tuple Selection approach (HTS) in [19, 20, 21, 22, 23, 24, 25, 26, 27] were explored and applied to the HAR problem. However, these approaches heavily used handcrafted features, which are time-consuming and generate lower performance in terms of accuracy [28]. For example, the methods in [25] and [27] obtained lower accuracies: 79.21% and 73.11% for the UCI HAR [29] and WISDM datasets [30], respectively.

Several DL models, including Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Bidirectional LSTM (Bi-LSTM), and Gated Recurrent Unit (GRU), have been utilized to address the lower performance issue in HAR. Despite promising results in activity recognition, these models still face ongoing difficulties, including complex feature extraction and dealing with noisy data [1]. For example, the methods described in [31] encountered challenges in classifying complex activities using shallow CNNs. It was primarily due to the extraction of limited handcrafted features, which relied on costly domain knowledge [32, 33].

To overcome these challenges, researchers have developed various hybrid methods [7, 11, 34, 10], CNN with feature fusion approaches [35, 36], and attention-based models [8, 37, 38, 39]. These methods aim to extract robust representative features and enhance model performance. However, existing

approaches still face challenges in capturing highly representative features due to their focus on single-phase feature extraction strategies. Instead, a multi-phase feature extraction strategy, which rigorously extracts features across multiple phases, can be more effectively capture highly representative features, as shown in Figure 1. As a result, these current methods have shown limited performance across diverse HAR datasets, such as UCI-HAR, WISDM, MHealth, and PAMAP2. In addition, among them some existing methods perform well on specific HAR datasets but still face challenges when generalizing across diverse HAR datasets due to difficulties in capturing highly desirable features. While some other works do not conduct extensive experiments on diverse popular HAR datasets, highlighting the need for more comprehensive experimentation.

The intuition highlights the need for a more comprehensive approach that rigorously extracts features multiple times to capture highly representative features, as shown in Figure 1. In addition, comprehensive experimentation also requires across various HAR datasets to this extensive approach performance. Given its significant impact on healthcare research and people's lives, achieving robust performance in HAR tasks is crucial. Motivated by these considerations, we propose a hybrid model known as the Dual-Phase Fused Neural Networks with Ensemble Learning (DP-FusedNN-EL) model. Specifically, the DP-FusedNN-EL model integrates DL and Ensemble Learning (EL) strategies to perform dual-phase feature extraction tasks, followed by classification tasks. Specifically, we develop Dual-Phase Fused Neural Networks (DP-FusedNN) to rigorously perform feature extraction in two distinct phases, thereby capturing highly representative local-global features, as shown in Figure 1.

Figure 1 highlights the differences between our proposed dual-phase feature extraction approach (DP-FusedNN) and existing single-phase feature extraction methods [8, 7, 10, 9, 11] in capturing highly representative features from HAR datasets to perform HAR tasks effectively. Our approach shows its effectiveness by rigorously extracting features across two phases, leading to human activity classification using the DP-EL model. It is achieved by designing two specific neural networks: the Dual Head Fused CNN (DH-Fused-CNN) network for learning comprehensive local dependencies and the CNN with Stacked Bi-GRU and Attention (CNN-SBi-GRU-Attention) network for extracting highly representative local-global features. In addition, we develop a Dual-Phase EL (DP-EL) model for HAR tasks. The proposed DP-FusedNN-EL approach undergoes extensive experimentation, aiming to

achieve high performance on the UCI-HAR [29], UCI-HAR-AAL [40], UCI-HAPT [41], MHealth [42], WISDM [30], and PAMAP2 [43] datasets.

The main contributions of this work can be summarized as follows:

1. We propose DP-FusedNN-EL, a novel framework that integrates a dual-phase feature extraction strategy with an ensemble learning model to achieve robust feature extraction and improved classification performance for human activity recognition tasks.
2. We propose the DH-Fused-CNN network to effectively learn comprehensive local features from sensor-based HAR datasets. To further refine these features and enhance their ability to capture both local and global dependencies, we design the CNN-SBi-GRU-Attention network.
3. We propose the DP-EL model, leveraging Stacking with Weighted Voting-based EL (SWV-EL) strategies for robust human activity classification tasks. It selects optimal base learners to create EL models using stacking with weighted voting strategies and combining these EL models to form the DP-EL model for activity classification.
4. While prior studies often use limited HAR dataset(s) for performance evaluation, we conduct extensive experiments on six well-known HAR datasets [29, 40, 41, 42, 30, 43] and compare the performance of our suggested approach with the state-of-the-art ML and DL approaches.

The rest of this study is organized as follows. Section 2 explores previous research works for HAR tasks. Section 3 introduces our proposed efficient method to perform HAR tasks. Section 4 presents experimental outcomes using well-known public HAR datasets. Section 5 concludes the study.

2. Related Works

Previous studies in HAR have extensively explored datasets such as UCI HAR, UCI HAR-AAL, UCI HAPT, WISDM, MHealth, and PAMAP2 [29, 30, 40, 41, 42, 43]. This section examines reported DL models, including CNNs with feature fusion, hybrid models, and attention mechanisms, applied to these HAR datasets.

2.1. CNN and Hybrid Models

In [44], researchers developed a HAR model using smartphone sensors, employing a CNN framework for activity classification. Another study [45]

proposed a DL framework combining CNN with statistical features for real-time activity classification, effectively retaining temporal information in time series data and enhancing HAR model performance. Inspired by this, recent research introduced lightweight CNN models, such as shallow CNN [31], layer-wise CNN [46], and Grouped Temporal Shift Networks (GTS-Net) [47], to perform HAR tasks. However, these methods achieved lower accuracies, ranging from 88.6% to 95.7% on UCI-HAR and WISDM datasets, as they prioritized computational cost over accuracy, limiting their feasibility in healthcare research where accurate classification is crucial with minimal computational costs [48, 49]. Because accuracy remains the primary focus for ML and DL researchers in healthcare applications, given its direct impact on patient well-being [48, 49]. Conversely, Recurrent Neural Networks (RNNs), integral to DL, excel in handling sequential data.

Traditional RNNs, however, faced challenges capturing extensive dependencies in HAR data [50]. LSTM and GRU, variants of RNNs, effectively address vanishing gradient issues, showcasing superior handling to capture long-term dependencies [1]. For instance, [51] stacked five LSTM cells to construct a robust classifier for human activities using smartphone sensor data. Based on this strategy, [52] proposed a bidirectional LSTM-based network for HAR, while [53] introduced a deep residual bidirectional LSTM framework, improving recognition rates in both temporal and spatial dimensions. Combining the strengths of CNN and LSTM, [54] developed a hybrid model to capture spatiotemporal patterns from raw sensor data, enhancing classification accuracy in HAR. Additionally, [55] presented a versatile hybrid CNN-LSTM architecture finely tuned for HAR with multimodal wearable sensors.

Inspired by influential research in [50, 51, 52, 53, 54, 55], several researchers have recently developed hybrid models to overcome lower performance issues by integrating CNN with RNN strategies. For example, Dua et al. [7] developed a Multi-Input CNN with a GRU network (MI-CNN-GRU) model, extracting local features through CNN and capturing long-term dependencies via GRU layers. This method achieved heightened accuracies of 96.2%, 97.21%, and 95.27% on UCI-HAR, WISDM, and PAMAP2 datasets, respectively. Nafea et al. [10] unveiled the Hierarchical Multi-Resolution CNN with a GRU technique (HMR-CNN-GRU) to perform HAR tasks. It extracts local and global features, achieving an accuracy of 94.5% and 99.38% on the UCI-HAR and MHealth datasets, respectively. Kosar and Barshan [11] proposed a unique 2D-CNN-LSTM hybrid model, differing from the stan-

dard 1D CNN-LSTM approach, by utilizing a fusion strategy and parallel branches. This approach achieved 95.66% accuracy on the UCI-HAR dataset. These prior methods [7, 10, 11] often struggled to attain high performance on HAR datasets. The main reason behind this is the challenges in capturing representative features from HAR datasets.

Tong et al. [34] developed a Bi-GRU with Inception model (Bi-GRU-I) to extract temporal and spatial features of human movements for HAR. Comprising a two-layer Bi-GRU and Inception module, this architecture optimized the deep neural network for HAR modeling. The Bi-GRU-I method achieved 95.42% and 98.25% accuracy on UCI-HAR and WISDM datasets, respectively. It performed well on specific HAR datasets but faced challenges when applied to a broader range of HAR datasets due to difficulties in extracting robust features. Helmi et al. [56] combined DL and swarm intelligence to create a robust HAR system. Their approach features a light extraction method using a residual convolutional network and a recurrent neural network (RCNN-BiGRU), alongside new feature selection techniques based on the marine predator algorithm (MPA) to optimize feature sets to perform human activity recognition tasks. Most recently, Al-Qaness et al. [57] developed the PCNN-Transformer, a parallel convolutional neural network and transformer architecture that utilizes parallel architecture and residual mapping to learn temporal features from sensor data for HAR tasks. Their PCNN-Transformer approach performed well on multiple HAR datasets. While Lalwani and Ramasamy [1] developed a multi-branched hybrid model, CNN-BiLSTM-BiGRU, to extract short-term patterns and long-term associations in sequential data by merging CNN, Bi-GRU, and Bi-LSTM components. The CNN-BiLSTM-BiGRU model [1] achieved classification accuracies of 99.32% and 96.10%, precisions of 92.82% and 79.65%, recalls of 93.1% and 84.57%, and F1-scores of 73.2% and 90.13% on WISDM and PAMAP2 datasets, respectively. This hybrid model has notable success in classification accuracy by extracting desirable short-term and long-range patterns from the input HAR datasets. However, this method exhibited shortcomings when assessing additional performance metrics like precision, recall, and F1 scores. These metrics play a critical role in dealing with the class-imbalanced HAR dataset [29, 30, 40, 41, 42, 43].

Drawing on EL, which has proven effective in ML and extends to DL, [58] introduced a novel training algorithm for LSTM models and an EL classifier combining multiple LSTM learners. This approach was further developed in [59], where a hybrid HAR model was created by parallelly integrating a

fully convolutional block and an LSTM block. This innovative fusion aimed to leverage the strengths of both components, enhancing the overall model's capabilities. Inspired by these breakthroughs in [58, 59], researchers have explored alternative strategies based on hybrid models merging DL and ML techniques. For example, the approach by Dahou et al. [60] integrated CNN for feature extraction, followed by a Binary Arithmetic Optimization Algorithm (BAOA) for feature selection tasks and SVM for classification tasks. This method achieved 95.23% and 99.5% accuracy on the UCI-HAR and WISDM datasets, respectively. However, it lacked comprehensive experiments and struggled to generalize its methods across diverse popular HAR datasets, highlighting the need for more extensive experimentation to address HAR tasks efficiently.

Some alternative studies to hybrid models utilize feature fusion techniques within CNN models to extract highly significant features from input HAR datasets to perform HAR tasks. Specifically, Wang et al. [35] developed the Adaptive Feature Fusion Network (AFFNet) based on a multi-scale fusion approach to combine temporal and distance features for human activity classification. It extracts multi-scale temporal features using a dynamic convolution network, while distance features are obtained through a prototype network. The AFFNet achieved 95.32% and 94.61% accuracy on UCI-HAR and WISDM datasets, respectively. This feature fusion approach still faced challenges in achieving high performance on HAR datasets due to difficulties in capturing essential features. Emphasizing the need for comprehensive experiments to generalize this strategy across diverse HAR datasets. Most recently, Liu et al. [36] introduced a UC Fusion method for feature extraction based on the strategy of features fusion technique. It merged the extracted unique features from each sensor with the common features shared across all sensors. This approach achieved high accuracies of 96.84% and 98.85%, precisions of 96.35% and 98.73%, recalls of 96.22% and 98.90%, and F1-scores of 96.27% and 98.83% on UCI-HAR and WISDM datasets, respectively, to perform HAR tasks.

2.2. Attention Models

Wang et al. [61] first introduced an attention method by combining traditional CNN procedures with attention sub-modules to assess the relationship between local and global features for HAR tasks on weekly labeled wearable sensor data. Similarly, a novel DL architecture proposed in [62] significantly altered the algorithm by replacing the baseline CNN's global loss with a

local loss, reducing memory requirements for sensor-based activity recognition. Based on the ideas in [61, 62], several researchers have introduced diverse attention strategies to combine them with variant CNNs or hybrid models. For example, Khan and Ahmed [8] introduced a Multi-Head CNN with attention mechanisms. This approach enhances CNN's representation ability by automating important feature extraction, which is crucial for performing HAR tasks. This attention model achieved accuracies of 95.38% and 98.18% on the UCI-HAR and WISDM datasets, respectively. In contrast, Yin et al. [37] introduced a combined strategy, a 1D Convolution-based Bi-LSTM Parallel Model with an Attention mechanism (ConvBLSTM-PMwA). This model aims to extract features, eliminate noisy data from the HAR dataset, and obtain good accuracies of 96.71% and 95.86% on UCI-HAR and WISDM datasets. Gao et al. [63] developed an advanced attention strategy: the Dual Attention Network-based HAR (DanHAR) method for multimodal HAR scenarios to enhance the CNN's representation power. This advanced strategy combines channel and temporal attention layers to extract channel-wise and temporal patterns, respectively. The DanHAR approach achieved outstanding accuracy rates of 98.85% and 93.16% in HAR tasks.

Tang et al. [38] also introduced a Triplet Cross-Dimension attention model for sensor-based activity recognition tasks. This model features three attention branches capturing cross-interaction features between sensor, temporal, and channel dimensions. Their approach achieved notable F1 scores of 93.2%, 96.77%, and 98.61% on PAMAP2, UCI-HAR, and WISDM datasets. Mim et al. [39] presented a hybrid approach incorporating an inception-attention-based method utilizing the GRU layer for effective temporal and spatial information extraction from sensor-based HAR data. Employing Inception with Convolutional Block Attention Module (CBAM) attention, their approach achieved commendable accuracies of 96.4%, 90.78%, and 99.13% on UCI-HAR, PAMAP2, and WISDM datasets [39].

Dahou et al. [9] introduced a wavelet transform strategy as an alternative to attention mechanisms for HAR tasks. [9] proposed a multilevel CNN, namely MLCNNwav, based on the Discrete Wavelet Transformation (DWT) strategy for global feature extraction without using attention mechanisms to perform HAR tasks. This approach obtained accuracies of 95.52% and 99.14% on UCI-HAR and WISDM datasets. Previous works [8, 37, 63, 38, 39, 9] lacked comprehensive experiments and faced challenges in generalizing their methods across diverse popular HAR datasets. It underscores the necessity for more extensive experimentation to address HAR tasks effi-

ciently. While Al-Qaness et al. [64] proposed Multi-ResAtt, a multilevel residual network that combines recurrent neural networks with attention mechanisms. This approach integrates initial blocks and residual modules in parallel, enabling effective extraction of time-series features and activity recognition across diverse HAR datasets. However, their approach achieved limited performance, with an accuracy of 87.82% on the PAMAP2 dataset. This limitation arises from an excessive focus on initial blocks and residual modules, rather than employing an extensive feature fusion process with a more integrated attention mechanism within these components.

Another study works as an alternative to CNN with an attention mechanism. For example, Essa and Abdelmaksoud [65] presented a transformer model, Convolution with a Self-Attention Network (CSNet) and Temporal-Channel CSNet (TCCSNet) for HAR tasks. These approaches capture local and global features and time- and channel-wise information from MHealth, PAMAP2, and WISDM datasets. However, this approach ultimately fell short despite attempting to extract robust features from various HAR datasets.

Knowledge distillation (KD) can enhance the performance of a compact student model by transferring knowledge from a more complex teacher model. However, many existing KD methods overlook the bias introduced by the teacher's logits during distillation, which can lead to sub-optimal results in student training. To address this problem, Xu et al. [66] designed a Contrastive Distillation framework with Regularized Knowledge (ConDRK), which enhances knowledge distillation by addressing biases from teacher logits. They leveraged a contrastive distillation approach using unbiased soft targets and contrastive learning to improve student model performance. This approach achieved a comparable performance of 96.57% on the UCI HAR dataset. However, it did not extensively explore diverse HAR datasets, limiting its effectiveness across broader applications in HAR tasks.

Summary of the Literature: Previous HAR studies mainly utilized CNN with feature fusion [35, 36], attention-based approaches [8, 37, 38, 39], and hybrid models [1, 7, 10, 11, 34] for robust feature extraction. However, many existing approaches, including hybrid models that combine CNN and RNN strategies with attention mechanisms, often face challenges to capture highly representative features. This limitation arises from their reliance on single-phase feature extraction processes, rather than adopting a multi-phase feature extraction strategy that rigorously extracts features across multiple phases from diverse HAR datasets to enhance performance. As a result, there is a need for a more comprehensive approach. In response, we pro-

pose DP-FusedNN-EL, a novel framework that integrates deep learning-based DP-FusedNN with ensemble learning (DP-EL) techniques for robust feature extraction and classification tasks.

One existing work [60] leveraged dual-phase feature extraction process, aimed to perform HAR tasks effectively. While it performed well on a specific HAR dataset [30], it still faced challenges when evaluated on other HAR datasets, such as [29]. This limitation arises from its reliance on CNNs for feature extraction and BAOA for feature selection from these learned features. It overlooks the potential of leveraging a hybrid model, such as our CNN-SBi-GRU-Attention network and misses the benefits of an extensive feature fusion strategy within a CNN like our DH-Fused-CNN network. Our approach addresses these gaps, resulting in the learning of more comprehensive features and improved performance across diverse HAR datasets. In addition, this existing work relied on traditional machine learning models, including multiclass SVM, while overlooking the potential of combining neural networks and ensemble learning model, such as ours DP-EL model.

3. DP-FusedNN-EL: Dual-Phase Fused Neural Networks with Ensemble Learning Model

In this section, we present the proposed DP-FusedNN-EL architecture, designed to enhance performance on HAR tasks, as shown in Figure 2. This architecture consists of two main components: (1) DP-FusedNN, which performs dual-phase feature extraction to capture highly representative local-global features from sensor-based HAR datasets, and (2) DP-EL, which leverages these features to effectively recognize human activities. Detailed descriptions of each component and their contributions to HAR performance are provided in the following sub-sections.

3.1. ***DP-FusedNN: Dual-Phase Fused Neural Networks***

Effective feature extraction is crucial for sensor-based HAR tasks to prevent overfitting and improve classifier performance [48]. Motivated by this concept, we design the DP-FusedNN method, which leverages a dual-phase feature extraction process: (1) The first phase employs the DH-Fused-CNN to extract comprehensive local features, and (2) the second phase uses the CNN-SBi-GRU-Attention network to capture highly representative local-global features, as shown in Figure 2.

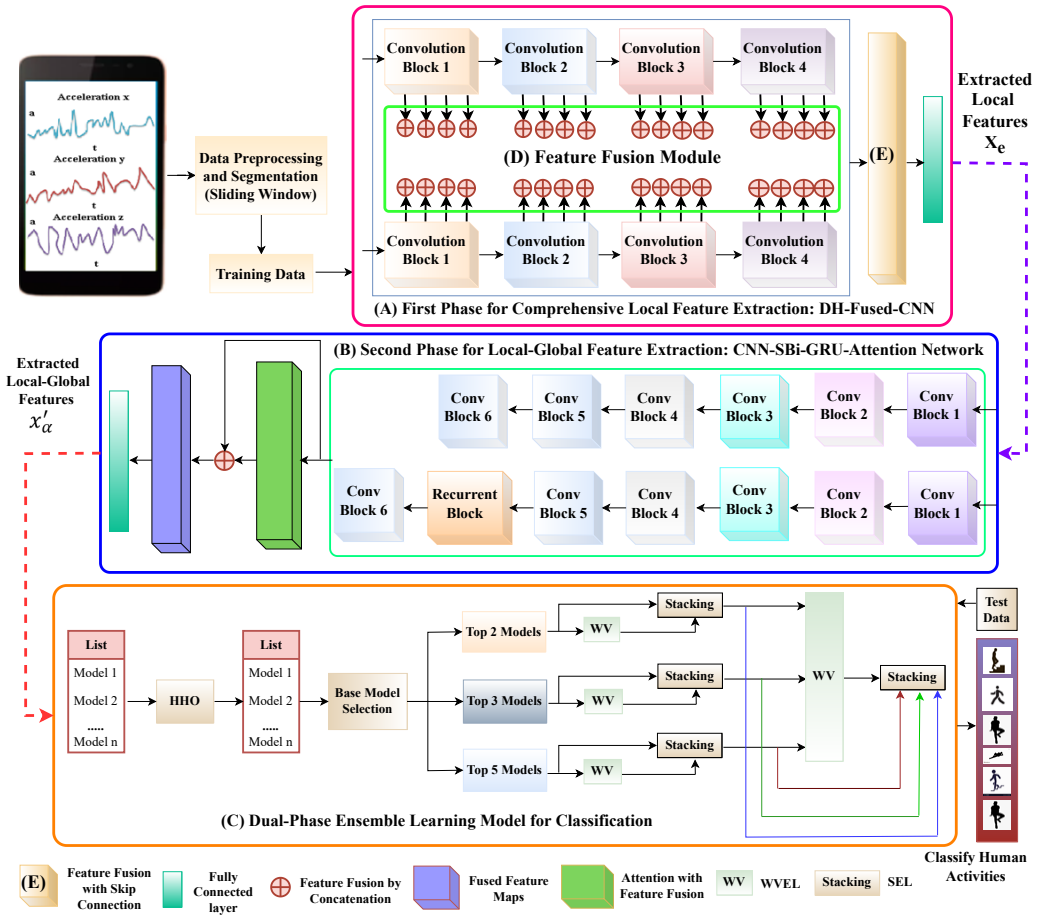

Figure 2: Architecture of the proposed Dual-Phase Fused Neural Network with Ensemble Learning (DP-FusedNN-EL) model for HAR tasks. The model comprises two components: DP-FusedNN (depicted in A and B) and DP-EL (shown in C). The DP-FusedNN employs a dual-phase feature extraction strategy: (1) DH-Fused-CNN extracts comprehensive local features X_e (A), leveraging a feature fusion module with intermediate concatenation layers to fuse feature maps from each block of one head with corresponding blocks from subsequent heads, capturing diverse local patterns (shown in D), and a feature fusion and skip connection module for comprehensive local feature extraction (exhibited in E); (2) CNN-SBi-GRU-Attention refines these comprehensive local features into highly representative local-global features X'_α using a convolution block, a recurrent block, and a self-attention-based feature fusion module (shown in B). (C) The key components of the DP-EL model include: Harris Hawk Optimization for hyperparameter tuning of machine learning models (e.g., Random Forest, SVM), model selection to choose the top 2, 3, or 5 models to form ensemble models using a stacking with Weighted Voting (WV) strategy (first phase), and combining these ensembles into a dual-phase ensemble model (second phase).

Existing methods [1, 7, 10, 11, 34, 60, 8] have employed hybrid architectures that combine CNNs and RNNs, often enhanced with attention mechanisms. These approaches are generally considered single-phase feature extraction networks. However, these approaches often struggle to capture comprehensive local-global features due to their inadequate focus on a rigorous feature extraction process. In contrast, DP-FusedNN performs feature extraction in two distinct phases, as shown in Figure 1, effectively capturing highly representative local-global features and thereby improving performance on HAR tasks. Detailed descriptions of the DH-Fused-CNN and CNN-SBi-GRU-Attention networks are provided in the following sub-sections.

3.1.1. First Phase Feature Extraction: DH-Fused-CNN

The Multi-Head CNN approach [8, 67, 68, 69] and feature fusion techniques [35, 36] effectively extract highly discriminative features for HAR tasks. These methods allow each network head to learn distinct patterns, which are then fused to capture diverse representations, thereby enhancing performance on HAR [35, 36, 8, 67, 68, 69].

Inspired by existing approaches [35, 36, 8, 67, 68, 69], the DH-Fused-CNN is designed to extract comprehensive local features from HAR data by leveraging CNNs' ability to learn filters for small sub-regions, thereby capturing local patterns and their variations [21, 45]. Unlike prior methods that rely on early or late fusion layers to capture diverse local information, DH-Fused-CNN employs multiple intermediate feature fusion layers within the feature fusion module, followed by a skip connection-based feature fusion module, as shown in Figure 2. This design enables the network to learn a more comprehensive local information. Specifically, it comprises a feature fusion module and skip connections to enhance local feature learning, thus improving the performance of the DP-EL model for HAR tasks.

The DH-Fused-CNN architecture comprises dual-head neural networks, where each head is composed of four convolutional blocks. Each block v contains two 1D point-wise convolutions p and one standard 1D convolution layer c. The c layer captures local features and generates feature maps, while the point-wise convolutions p serve dual purposes: the first p layer performs dimensionality reduction, and the second p layer generates output features for the feature fusion module. In this module, the intermediate feature fusion layers effectively fuse outputs from corresponding convolution blocks across the two heads to learn diverse local information, as depicted in Figures 3 and 4. To stabilize training, batch normalization η is applied after each c layer.

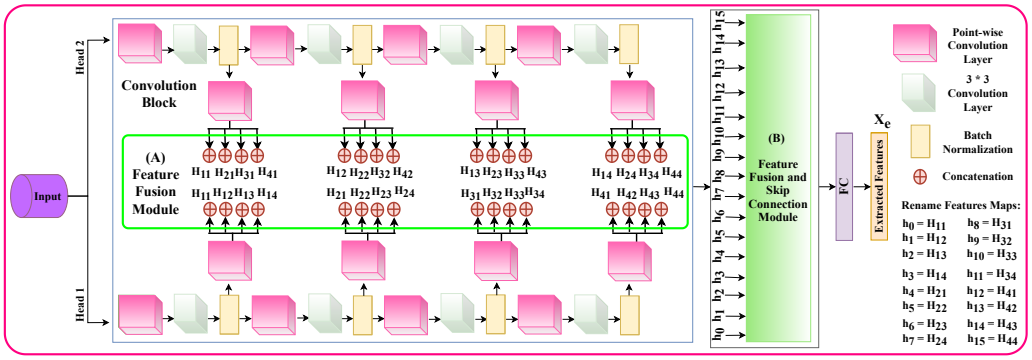

Figure 3: Architecture of DH-Fused-CNN approach as the first phase for feature extraction tasks. Key components include: (A) feature fusion module with intermediate feature fusion layers (concatenation) to fuse new feature maps from each block of one head with other feature maps from corresponding blocks of subsequent heads to learn diverse local information $h_{v\times v}$: $h_0 = H_{11}, h_1 = H_{12}, h_2 = H_{13}, \ldots, h_{13} = H_{42}, h_{14} = H_{43}, h_{15} = H_{44}$. Here v denotes convolution block. (B) feature fusion with skip connection module to extract comprehensive local features based on learning diverse local patterns.

A feature fusion with skip connections strategy (Figure 5) is then applied to extract comprehensive local features, which are further processed by the CNN-SBi-GRU-Attention network.

A. Convolution Block: In this study, we represent the input signal for the HAR dataset as α, where $\alpha \in \mathbb{R}^{T\times S}$, with T denoting the number of time steps and S represents the size of the feature set, serving as the input to the first convolution block v of both heads hd. The initial point-wise convolution layer p generates the j^{th} feature maps $x^{hd}_{v_{p_j}}$ after the first p layer in each v, where $1 \leq hd \leq q$ and the value of $q = 2$, as per the DH-Fused-CNN approach. A non-linear activation function ϕ_S is then applied to mitigate vanishing gradient or explosion issues, ensuring the generation of informative feature maps $x^{hd}_{v_{p_j}}$.

$$x^{hd}_{v_{p_j}} = \left[\left(\sum_{m=1}^{\text{Size}(\alpha)} \omega^{m}_{v_{p_j}} \alpha^{hd}_{(v_{p_j-1})m} \right) + bs^{hd}_{v_{p_j}} \right] \tag{1}$$

$$x^{hd}_{v_{p_j}} = \phi_S \left(x^{hd}_{v_{p_j}} \right) \tag{2}$$

where m denotes the feature map index at the $(p-1)^{th}$ layer, $\omega^{m}_{v_{p_j}}$ signifies the weight matrix for the first p of every block for both hd's.

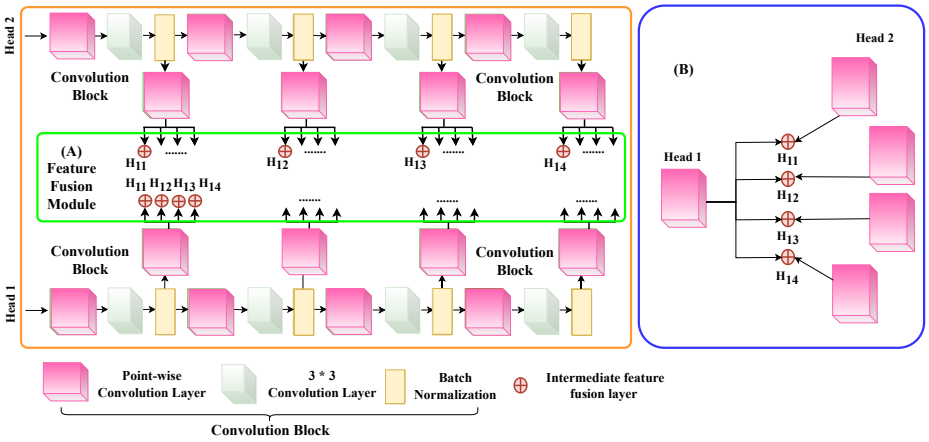

Figure 4: Detailed architecture to show how learned features are fused by the feature fusion module of the DH-Fused-CNN approach. Here (A) feature fusion module with intermediate feature fusion layers (concatenation) to fuse new feature maps from each convolution block of head 1 is fused with other feature maps learned from its corresponding blocks of next heads i.e., head 2, to learn diverse local information $h_{v\times v}$: $H_{11}, H_{12}, H_{13}, H_{14}$. Here v denotes convolution block. (B) Detailed analysis of how feature fusion module performs to learn diverse local information i.e., H_{11} to H_{14}.

Next, we perform a standard 1D convolution operation, as per equation (3), for every block of both hd's to facilitate the generation of informative feature maps $x^{hd}_{v_{c_j}}$. We then apply ϕ_S to it as per equation (2).

$$x^{hd}_{v_{c_j}} = \left[\left(\sum_{m=1}^{\text{size}(X)} \sum_{n=1}^{A} \omega^{m,n}_{v_{c_j}} x^{hd}_{v_{p_j} m} \right) + bs^{hd}_{v_{c_j}} \right] \tag{3}$$

where n denotes another feature map index at the $(c-1)^{th}$ layer for the kernel size A of the c^{th} layer at every block for both hd's.

We leverage η layer, which forms normalized feature maps $N^{hd}_{norm_v}$ for every block of both hd's. Based on these processes as specified in Equations 1 - 4, each convolution block is formed, and each output as generated feature maps of every v's is then forwarded to the following v's to form new feature maps $N^{hd}_{norm_v}$ until the following v's of both hd's are available.

$$N^{hd}_{norm_v} = \eta \left(x^{hd}_{v_{c_j}} \right) \tag{4}$$

B. Feature Fusion Module: The primary motivation for developing the feature fusion module is to effectively capture diverse patterns from different

branches of the network, enhancing the model's classification performance. Unlike existing approaches that use late or early fusion methods [8, 67, 68, 69, 35, 36], the DH-Fused-CNN architecture incorporates a feature fusion module with $v \times v$ intermediate feature fusion layers $\theta_{v\times v}(\cdot)$. This module fuses new feature maps $x^{hd}_{v_{p_j}}$ from each block of one head hd with feature maps $x^{hd+(q-1)}_{v_{p_j}}$ from corresponding blocks of subsequent heads to learn diverse local information $h_{v\times v}$, as shown in equation (5). An additional p layer is applied at the end of each convolution block to learn these new feature maps $x^{hd}_{v_{p_j}}$. This approach, exhibited in Figures 3 and 4 and detailed in Algorithm 1.

$$h_{v\times v} = \theta_{v\times v}\left(x^{hd}_{v_{p_j}}, x^{hd+(q-1)}_{v_{p_j}}\right) \tag{5}$$

Figures 3 and 4 visually represent about feature fusion module within the network's architecture. In each head, output feature maps $x^{hd}_{v_{p_j}}$ from each block are fused with feature maps $x^{hd+(q-1)}_{v_{p_j}}$ from subsequent heads $hd + (q-1)$ to learn diverse local patterns $h_{v\times v}$. Here, $v = 4$ and $h_{v\times v} = \{h_0, h_1, h_2, \ldots, h_{14}, h_{15}\}$, starting from the 0th position, resulting in $h_{4\times 4} = h_{16}$ diverse local information.

C. Feature Fusion and Skip Connection Module: The primary motivation for developing the feature fusion with skip connection module is to further fused features across different levels, specifically combining lower-level patterns with higher-level ones to capture comprehensive local features [70, 71]. This integration aims to enhance the classification model's performance. Inspired by this approach, the DH-Fused-CNN architecture employs the Feature Fusion and Skip Connection (FF-SC) module, as shown in Figure 5 and detailed in Algorithm 1. The FF-SC module is applied to each diverse local patterns h_o, where $o \in [0, v \times v - 1]$, aiming to learn comprehensive local information D^i_k at each iteration i. The indices i and k act as counters for the iterative fusion process, defined as $0 \leq i \leq 1$, $0 \leq k \leq \frac{(v\times v)-2}{2}$, and $i \leq o \leq (v \times v) - 1$, respectively. The FF-SC technique is designed to retain a comprehensive range of feature representations, from low-level to high-level, thus minimizing potential information loss during the fusion process.

The FF-SC technique is governed by the following conditions, as mathematically represented as Equation 6:

1. **Direct Skip Fusion:** When $o = i$, meaning the index o matches i, the diverse local information h^i_o is directly fused with another one h^i_{o+2} to learn comprehensive local information D^i_k. This approach allows the

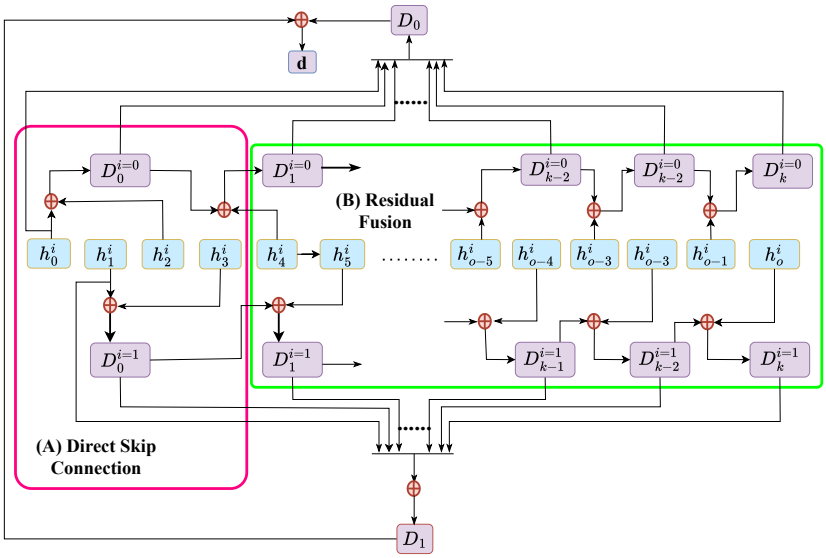

Figure 5: Detailed architecture of the feature fusion and skip connection module. This module processes each diverse local pattern $h_o^i : h_0^i, h_1^i, \ldots, h_{o-3}^i, h_{o-2}^i, h_{o-1}^i, h_o^i$, where $o \in [0, v \times v - 1]$, to learn comprehensive local information $D_k^i : D_0^{i=0}, D_1^{i=0}, D_0^{i=1}, D_1^{i=1}, \ldots, D_{k-2}^{i=0}, D_{k-2}^{i=1}, D_{k-1}^{i=0}, D_{k-1}^{i=1}, D_k^{i=0}, D_k^{i=1}$ at each iteration i. The key components include: **(A) Direct Skip Connection:** when $o = i$, the diverse local information h_o^i is directly fused with h_{o+2}^i to generate comprehensive local information D_k^i; **(B) Residual Fusion:** when $o \neq i$, the diverse local information h_{o+2}^i is fused with the previously learned comprehensive local information D_{k-1}^i to learn further comprehensive information D_k^i. Each learned local pattern D_k^i, $\forall\, i$ is subsequently fused with $h_{o \in i}$, where $o \in i = [0, 1]$. This process generates refined comprehensive local information, D_0 and D_1, which undergo another fusion step to form d, followed by a fully connected (FC) layer to extract comprehensive local features x_e. Indices i and k serve as counters in the iterative fusion process, defined as $0 \leq i \leq 1$, $0 \leq k \leq \frac{(v \times v) - 2}{2}$, and $i \leq o \leq (v \times v) - 1$.

network initially to skip certain intermediate learned local information to enrich the fusion process with diverse feature combinations.

2. **Residual Fusion:** Conversely, when $o \neq i$, the fusion operation involves combining the new diverse local information h^i_{o+2} with the previously learned comprehensive local information D^i_{k-1} to learn further comprehensive information D^i_k.

$$D^i_k = \begin{cases} \theta\left(D^i_{k-1}, h^i_{o+2}\right) & \text{if } (o \neq i) \\ \theta\left(h^i_o, h^i_{o+2}\right) & \text{if } (o == i) \end{cases} \tag{6}$$

By incrementally advancing o by 2 after each fusion operation, the FF-SC technique systematically fuses both newly learned information h^i_{o+2} with previously captured features D^i_{k-1}. This module ensures a comprehensive fusion of feature maps across different levels of abstraction, mitigating potential information loss and thereby improving the model's overall effectiveness in learning comprehensive local patterns.

Each learned local patterns D^i_k, $\forall\, i$ is further fused with $h_{o\in i}$ where $o \in i = [0, 1]$, as shown in Figure 5 and detailed in Equation 7. This process generates refined comprehensive local information D^i, which undergo another fusion step followed by a fully connected (FC) layer with δ to extract comprehensive local features x_e. These features x_e are subsequently used as an input to the CNN-SBi-GRU-Attention network to capture highly representative local-global features.

$$\left.\begin{array}{l} D_i = \theta\left(h_{o\in i}, D^i_k\right) \\ d = \theta\left(D_i\right) \\ x_e = \phi_S\left(\mathrm{FC}_\delta(d)\right) \end{array}\right\} \tag{7}$$

3.1.2. Second Phase: CNN-SBi-GRU-Attention Approach

Traditional CNNs excel at learning local features [45, 21], while self-attention mechanisms combined with RNNs (LSTM or GRU) effectively capture global patterns [72, 73, 74]. Hybrid approaches that integrate CNNs with RNNs and attention mechanisms [7, 10, 37] often achieve superior performance by addressing both local and global patterns. Some methods further enhance performance through multi-head neural network strategies [8]. Inspired by these approaches, we design the CNN-SBi-GRU-Attention method as the second phase in the feature extraction process of the DP-FusedNN framework, as shown in Figure 6 and detailed in Algorithm 2. This approach leverages self-attention with stacked Bi-GRU layer to capture

Algorithm 1 : DH-Fused-CNN Approach

1: **Input:** HAR dataset: training set = {x_train, y_train} and testing set = {x_test, y_test}.
2: **Output:** x_e;
3: **Procedure:**
4: **for** each hd **do:**
5: Perform all v according to equations (1 - 4);
6: **end for**
7: Perform feature fusion operations for all learnable feature maps for all heads to form fused feature maps according to equation (5).
8: Perform feature fusion with a skip connection strategy on all fused feature maps according to equations (6-7).
9: **return** x_e;

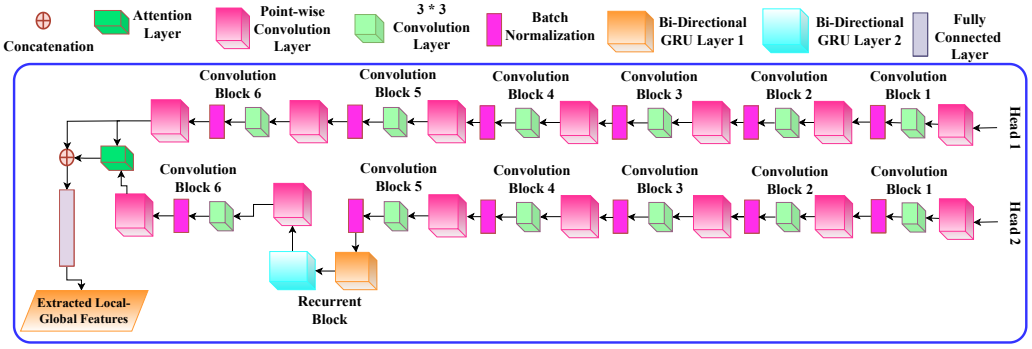

Figure 6: Architecture of the CNN-SBi-GRU-Attention approach as the second phase for feature extraction tasks to extract highly representative local-global features. Key components include: multiple convolution block, recurrent block (stacking-based Bi-GRU layer), and self-attention module with feature fusion strategy to refined comprehensive local features into highly representative local-global features.

global dependencies like [72, 73, 74], while CNNs focus on local patterns as [21, 45]. By combining these strategies, CNN-SBi-GRU-Attention effectively extracts highly representative local-global features from the comprehensive local features x_e provided by the DH-Fused-CNN. Unlike previous works [35, 7, 8, 38, 9, 34, 10, 39, 37, 11, 46, 36, 1] that utilize input HAR data, our CNN-SBi-GRU-Attention network refines x_e into highly representative local-global features, thereby further enhancing the performance of the DP-EL model for HAR tasks.

In this network, dual heads hd's are used, denoted as hd_1 and hd_2, where the value of hd lies between 1 and q as $1 \leq hd$ but $hd \leq q = 2$ as per this network. The first head, hd_1, consists of a CNN sub-network, while the second head, hd_2, comprises another sub-network: convolution with stacking based-Bi-GRU. These sub-networks consist of six convolution blocks v and one recurrent block, RB. The first five convolution blocks v's comprise point-wise and standard 1D convolution and batch normalization layers, accompanied by ϕ_S and generating normalized feature maps, $N^{hd}_{norm_V}$, as per equations 1 - 4. In blocks six and seven for each head, we implement the above-specified convolution layers with a batch normalization layer and add another point-wise convolution layer to form $x^{hd}_{v_{P_j}}$.

A. Recurrent Block Module: Incorporating the Recurrent Block (RB) module within the head $hd_{(q-1)}$, we apply a Stacking-based Bi-GRU (SBi-GRU) layer. This layer plays a crucial role in learning highly representative features. It uses forward and backward routes for each Bi-GRU layer g, leading to the generation of feature maps, as outlined in equation (8). These routes comprise two critical gates: the reset and update gates, denoted as R_t and U_t, where $U_t = \phi_U(\omega_U[hs_{t-1}; x_t] + bs_U)$, $R_t = \phi_R(\omega_R[hs_{t-1}; x_t] + bs_R)$ and $x_t \in N^{hd}_{norm_V}$ as the input tensor of the current unit for $g = 0$. While bs and $\phi(\cdot)$ represent the bias vector and element-wise logistic sigmoid function, ω_U, ω_R, and hs_{t-1} represent weight matrices, and h^s_{t-1} represents the hidden state of the current unit, respectively. The update gate controls the flow of information from both the forward and backward GRU cells to update the hidden state at a specific time step. This gate determines how much information from the past and future is combined to produce the current hidden state.

The reset gate controls the amount of data passed to the current candidate set $\overrightarrow{hs_t}$ within the previous stage, where $\overrightarrow{hs_t} = \tanh(\omega_{hs}.[(hs_{t-1} \otimes R_t), x_t] + bs_{hs})$ and $hs_t = (1 - U_t) \otimes hs_{t-1} + U_t \otimes \overleftarrow{hs_t}, t \in (1, m)$, which represents

the hidden state of the prior unit. While $\otimes$ and ω_h denote the element-wise multiplication and weight matrix, respectively. Consequently, these two routes yield distinct hidden layer states, enabling the extraction of diverse and informative information from the data.

$$X = \left[\theta\left(\overrightarrow{hs_t}, \overleftarrow{hs_t}\right)\right]^g \tag{8}$$

When generating an SBi-GRU layer, we follow the same process for the subsequent layer g, where we assign $x_t = X$. In this layer, we fuse each GRU layer's forward and backward routes' generated features for each state. This feature fusion process yields the output features X for all g, by synthesizing the forward and backward routes' of generated features. These features effectively enhance the inter-dimensional correlations in the sequential data. Subsequently, the resulting output X is directed to the seventh block to process further to generate feature maps $x_{v_{p_j}}^{hd(q-1)}$.

B. Attention Module with Feature Fusion Technique: We employ an attention technique $\psi(\cdot)$ to improve the quality of the learnable feature maps generated during our study. This attention mechanism focuses on salient features while effectively disregarding potentially confusing or unrelated information. Hence, it provides higher weights to related features and lowers the significance of unrelated ones. In our suggested approach, we apply $\psi(\cdot)$ to the output feature maps, $x_{v_{p_j}}^{hd}$ and $x_{v_{p_j}}^{hd(q-1)}$, obtained from the CNN and convolutional-stacking based-Bi-GRU sub-networks after taking them as inputs to it. These generated features are designated as queries, $hd == 1$ outputs and keys and values are as $hd + (q-1)$ outputs, while $\sqrt{y_z}$ denoting as the dimension of these inputs. Subsequently, estimate the alignment score $\vartheta(\cdot)$ using the dot product of inputs: $x_{v_{p_j}}^{hd=1}$ and $x_{v_{p_j}}^{hd=2}$ of $\psi(\cdot)$.

$$\vartheta\left(x_{v_{p_j}}^{hd=1}, \overline{x_{v_{p_j}}^{hd=2}}\right) = \vartheta\left(x^{0^T} \times \overline{x_{v_{p_j}}^{hd=2}}\right) \tag{9}$$

We then utilize the softmax activation ϕ_Π for normalizing these score values, thereby deriving the weight distribution for the associated values. Next, we use the feature fusion technique $\theta\cdot$ to fuse the first head's generated output $x_{v_{p_j}}^{hd=1}$ with the output of the $\psi\cdot$ to form a fused learnable feature map x_α. It is then forwarded to the FC layer to capture highly representative local-global features x'_α, as details in the following:

Algorithm 2 : CNN-SBi-GRU-Attention Approach

Input: x_e = TH-Fused-CNN
Output: x'_α;
Procedure:
for each hd **do:**
    Perform all v (up to 5) according to equations (1 - 4);
end for
if (v == 'six' and hd_2 == True) **then**
    Perform an SBi-GRU layer according to equation (8).
end if
if (b == 'six' and hd_1 == True or b == 'six' and hd_2 == True) **then**
    $N^{hd}_{norm_V}$ = Operate this block according to equations (1 - 4);
end if
Perform an attention mechanism on all heads' generated output feature maps according to equations (9-10).
Perform feature fusion operation and extract features according to the equations (11-12).
return x'_α;

$$\psi(Query, Key, Value) = \phi_{\Pi}\left(\frac{\vartheta\left(x^{0^T} \times \overline{x^{hd=2}_{v_{p_j}}}\right)}{\sqrt{y_z}}\right) \times x^{hd=2}_{v_{p_j}} \tag{10}$$

$$x_\alpha = \theta\left(x^{hd=1}_{v_{p_j}}, \psi(Query, Key, Value)\right) \tag{11}$$

$$x'_\alpha = \phi_S\left(FC_\delta\left(x_\alpha\right)\right) \tag{12}$$

Following the above-specified procedures, the DP-Fused-NN extracts crucial, highly representative local-global features x'_α from input local features x_e, forwarding them to the DP-EL model to recognize Human activities effectively.

3.2. DP-EL: Dual-Phase Ensemble Learning Model

Effective classifiers are crucial to enhance the performance of ML or DL models. Based on that fact, some researchers have developed EL methods that combine weaker and stronger models for improved outcomes [49]. Addressing this problem, we propose a DP-EL method, using the SWV-EL

technique for precise human activity classification tasks, as shown in Figure 2. The DP-EL model comprises two phases. Phase 1 considers the Nth classification models to form EL models, and Phase 2 combines these EL models. To develop EL models in Phase 1, it selects the bth number of best-performing models after estimating the threshold value based on their probability scores. It then forms the bth number of EL models based on two approaches: stacking EL and weighted voting EL techniques. Phase 2 combines these bth EL models through the SWV-EL strategy to form a robust classification model, as shown in Figure 2. Subsections provide more on DP-EL's specifics.

3.2.1. Phase 1: First-Phase Ensemble Learning Models

Phase 1 consists of three operations: firstly, generating base models and selecting the best performing models, and secondly, generating various EL models.

A. Generate Base Models and Select the Best Performing Models: This study uses N number of classifiers β indexed as $[1 : N]$ to actively contribute to form the DP-EL method, where these classifiers are initially arranged in a list L.

$$\beta = \{\beta_1, \beta_2, \beta_3, \ldots, \beta_{N-2}, \beta_{N-1}, \beta_N\} \tag{13}$$

There are some challenges in developing an EL model to achieve high performance. Because most classifiers struggle to generate effective performance to form an EL model. To address this challenge, we identify the best-performing classifiers $\text{Best}^b_{k_j}$ from the list, where k represents a list with j as the selected best classifiers and the value of b depends on the size of k. To achieve this, we initially estimate a probability score pr as an error score for misclassified examples for each β, as $pr(\beta_N) = [\text{Error}(\beta_N)]\,(\{\mathbf{x}'_{\text{train}\alpha}, \mathbf{y}_{\text{train}}\})$. In this case, $\{\mathbf{x}'_{\text{train}\alpha}, \mathbf{y}_{\text{train}}\}$ represents the extracted training features with the corresponding training labels. Subsequently, we compute a threshold value t that depends on the mean of all $pr_{\beta_N}, \forall\beta$ and compare it with each model's pr_{β_N}. This process helps us select the best-performing models from the available models.

B. Generating Various EL Models: This study applies the SWV-EL strategy to form various EL models as the first-phase EL models. To achieve

this, we initially select the highest-performing classifiers $\text{Best}_{k_j}^b$ to serve as the base estimators μ_b^w for the weighted voting-enabled and stacking-enabled EL models, such that values of $w = [1, K_j]$ where $w = K_j$. We then train all these mu_μ^b to yield the respective predictive outcomes $out_{i,w}^b$ for these EL models. These predictive outcomes, in turn, contribute to creating a new feature matrix, often called the meta-features $(MF)^b$, for every b along with y_{train}. These are subsequently used to train individual meta-learners χ_b for all $\text{Best}_{k_j}^b$ to perform as every stacking-based EL model. We employ a weighted voting-enabled EL strategy to form these χ_b in this study. To achieve this, we determine the weight ω_w^b corresponding to every μ_w^b, where we subtract each probability score $pr(\mu_w^b)$ associated with every μ_w^b from the value of 1, as illustrated below.

$$\omega_w^b = \left[1 - pr\left(\mu_w^b\right)\right] \tag{14}$$

Next, allocate these weights ω_w^b to each μ_w^b, enabling the development of a weighted voting-enabled EL strategy. This process is pivotal for forming individual meta-learners χ_b, each tailored to be fitted on the new meta-features $(MF)^b$ corresponding to every b derived from $\text{Best}_{k_j}^b$ along with y_{train}. In this way, it generates the several EL models as the first-phase EL models $(\beta')^b$ for each selected highest-performing classifier.

$$\chi_b = EL_w^b\left(x'_{\alpha_i}\right) = \sum_w \mu_w^b \times \omega_w^b\left(x'_{\alpha_i}\right) \tag{15}$$

$$(MF)^b = (MF)^b \cup \left\{\left(\text{out}_{i,w}^b\right)\right\} \tag{16}$$

$$(\beta')^b = \left[\chi_b\left(\left\{(MF)^b, y_{\text{train}}\right\}\right)\right] \tag{17}$$

3.2.2. Phase 2: Second Phase Ensemble Learning Model

To form the second phase of EL model, β^F, we employ the $(\beta')^b$ models and following the same principles elucidated in equations (14-17). Each first-phase EL model serves as a set of base learners μ'_b , contributing their prediction results $out'_{i,b}$ to create new MF' . Then, form a new meta-learner χ' by utilizing the identical strategy demonstrated in equations (14-15). To achieve this, each base estimator, in conjunction with its associated weights $\omega'_b = \lfloor 1 - pr(\text{Best}_{K_j}^b)\rfloor$, contributes to constructing this new meta-learner χ'. It is then fitted to the MF' and y_{train} after employing the same strategy, as exhibited in equations (16-17). Through these well-defined procedures, we form a β^F, subsequently employed to evaluate its classification performance on the test data.

3.3. Network Settings

This part will explore our suggested DP-FusedNN architecture configuration outlined in Appendix 1. We employ multiple convolution blocks for all heads of dual phase feature extraction networks. Specifically, we use four v's for dual heads of the DH-Fused-CNN approach and six for dual heads of the CNN-SBi-GRU-Attention network. Additionally, the second head of the second phase of the CNN-SBi-GRU-Attention approach incorporates a recurrent block.

3.3.1. Network Settings for DH-Fused-CNN Network

In the DH-Fused-CNN network, v of every head comprises two p layers with a kernel size of 1 and one c layer with a kernel size of 3. The initial stride value for the first c of first v varies between 1 and 2 for both heads depending on the input HAR datasets used. Following this, we apply a η layer before applying another p layer for both heads in this network. Additionally, we perform 33 intermediate feature fusion processes, where 16 processes are employed to fuse features with other output feature maps for each head via feature fusion module, as explained in section 3.1.1(B). Furthermore, the rest of the operations involve feature fusion with a skip connection module, as demonstrated in section 3.1.1.(C). In this context, 14 operations perform the skip connections with feature fusion tasks. Rests combine the resulting fused features, as illustrated in Figure 4. Finally, we employ a FC layer to extract highly representative local features.

3.3.2. Network Settings for CNN-SBi-GRU-Attention Network

In this network, we employ six v's for both heads, with head 2 incorporating a recurrent block. Each of the first five blocks consists of a p layer, a c layer (kernel size: 3 for dual heads), and η layer, as depicted in Figure 5. For the sixth and seventh convolution blocks, respectively, in both heads, we introduce an additional p layer after these previously mentioned layers. In the case of head 2$'$s sixth block, we apply an SBi-GRU layer, as illustrated in Figure 5.

3.3.3. Network Settings for Attention Module with Feature Fusion Technique

Following the above-specified layers, the attention layer is applied to both heads' output of the CNN-SBi-GRU-Attention network. Subsequently, a feature fusion process is executed, and a single FC layer is employed for

Table 1: Description of HAR datasets, including the number of subjects, activities performed, window length, sampling rate, and the number of training and testing sets. The datasets covered are: UCI HAR [29], UCI HAR-AAL [40], UCI HAPT [41], MHealth [42], WISDM [30], and PAMAP2 [43].

Datasets	D1 [29]	D2 [40]	D3 [41]	D4 [42]	D5 [30]	D6 [43]
Subjects	30	30	30	10	29	9
Activities	6	6	12	12	6	12
Window Length	128	128	128	50	50	200
Sampling Rate (Hz)	50	50	50	50	50	50
Training Set	7352	4021	10692	4804	14723	7548
Testing Set	2947	1723	2673	2059	6310	2582

extracting the highly representative local and global features, which are then used for subsequent classification tasks.

4. Experiments and Results

This section has four sub-parts. The first part discusses this study's HAR datasets [29, 30, 40, 41, 42, 43] and outlines the data preprocessing strategies. The subsequent parts demonstrate how our suggested method performs on these datasets for human activity classification tasks.

4.1. Dataset Descriptions and Preprocessing Details

In this sub-section, we evaluate our proposed approach on six widely used HAR datasets: UCI HAR (D1) [29], UCI HAR-AAL (D2) [40], UCI HAPT (D3) [41], MHEALTH (D4) [42], WISDM (D5) [30], and PAMAP2 (D6) [43]. Table 1 provides a summary of these datasets, with implementation details outlined in the following sub-sections.

4.1.1. Dataset 1: UCI HAR dataset [29]

The UCI HAR dataset [29] was generated using a Samsung Galaxy S II smartphone worn by 30 volunteers aged 19 to 48. This smartphone, securely attached to the waist of each subject, was equipped with accelerometers and gyroscope sensors, allowing it to record axial linear accelerations and angular velocities at a sampling frequency of 50 Hz. Data preprocessing involved noise filtering and sampling using a fixed 128-width sliding window technique. Walking Upstairs, Downstairs, and many more activities were

captured through video and manual labeling. This dataset comprises 10,299 samples split randomly into 70% and 30% for training and testing. Appendix 4 presents a comprehensive tabular overview and graphical representation of the UCI-HAR dataset description.

4.1.2. Dataset 2: UCI HAR-AAL dataset[40]

The [29] dataset, aimed at enhancing the dataset's included in [40] performance in Ambient Assisted Living, gathered data from built-in accelerometers and gyroscopes in smartphones worn by 30 participants aged 22 to 79. The same activities as [29] were conducted for 60 seconds using identical sensors as specified in [29]. Noise reduction involved median and third-order low-pass Butterworth filters with 20 Hz cutoff frequency. Fast Fourier Transform (FFT) was used to estimate feature vector variables for each pattern. It resulted in 5744 samples, divided randomly into 70% and 30% for training and testing. Appendix 4 presents a comprehensive tabular overview of the UCI-HAR-AAL dataset description.

4.1.3. Dataset 3: UCI HAPT dataset [41]

The [41] dataset extends from [29] using smartphone data. Similar to [29], volunteers from the same age groups wore smartphones on their waists and recorded activities through video for manual labeling. Sensor signals in [41] were processed with a Butterworth low-pass filter to separate gravity and body motion. This dataset included six additional basic activities and postural transitions like stand-to-sit and sit-to-stand. In this dataset, we used 128 window lengths with a 50 Hz sampling rate, as shown in Table 1. We also used 50% overlapping for basic activities. We performed higher overlapping, 93.75%, to extract more data for postural transitions. This dataset was partitioned randomly: 10692 samples for training and 2673 samples for testing. Appendix 4 presents a comprehensive tabular overview of the UCI-HAPT dataset description.

4.1.4. Dataset 4: MHealth dataset

The MHealth dataset [42] comprises vital signs and body movement data collected from ten diverse volunteers performing 12 physical activities. These activities include standing still, sitting, lying down, walking, climbing stairs, waist bending forward, frontal elevation of arms, knee bending (crouching), cycling, jogging, running, and jumping front and back. Each activity is either performed for one minute or repeated 20 times. Data is collected

using three Inertial Measurement Units (IMUs) sensors placed on the left ankle, right wrist, and chest, capturing magnetic orientation, turn rate, and acceleration. This dataset generalizes well to common daily activities due to the diversity of body parts involved (e.g., frontal elevation of arms vs. knees bending), varying activity intensities (e.g., cycling vs. sitting and relaxing), and differences in execution speed or dynamicity (e.g., running vs. standing still).

In this study, we utilized the sliding window technique with the Time Series Feature Extraction Library (TSFEL) [75], applying a 50 Hz sampling rate and a window length of 50. This process generated 6,863 examples across 12 activities. The dataset was then randomly divided into training (70%) and testing (30%) sets, resulting in 4,804 training samples and 2,059 testing samples. Appendix 4 presents a comprehensive tabular overview of the MHealth dataset description.

4.1.5. Dataset 5: WISDM dataset

The Wireless Sensor Data Mining (WISDM) dataset [30] was collected by the WISDM lab using an Android mobile app equipped with a three-axial accelerometer in a controlled environment. Data were gathered from twenty-nine volunteers, each carrying a smartphone in their front leg pocket while performing six activities: walking, jogging, ascending stairs, descending stairs, sitting, and standing. The accelerometer continuously recorded data during these activities.

The dataset was augmented using the TSFEL strategy with a sliding window approach [75], applying a 50 Hz sampling rate and a window length of 50. This process generated 21,033 examples across 6 activities. The dataset was then randomly divided into training (70%) and testing (30%) sets, resulting in 14,723 training samples and 6,310 testing samples. Appendix 4 presents a comprehensive tabular overview and graphical representation of the WISDM dataset description.

4.1.6. Dataset 6: PAMAP2 dataset

In this study, the Physical Activity Monitoring Data Set (PAMAP2) [43] is utilized for its comprehensive collection of 27 signals captured under controlled laboratory conditions. PAMAP2 includes 52-dimensional data collected over 10 hours from nine subjects performing 18 daily activities. These activities encompass nine repetitive tasks (such as, walking, jogging, cycling), non-repetitive actions, and three postures (lying, sitting, standing). Most

activity instances lasted 4 min, with adjustments to fit building limits or prevent subject tiredness. The dataset features recordings from three IMUs and one heart rate monitor, positioned on each subject's dominant ankle, wrist, and chest. For data augmentation, a sliding window technique with a window length of 200 and a sampling rate of 50 Hz was utilized. Subjects 7 and 8 were designated for the testing set, while the remaining subjects were used for training. Appendix 4 presents a comprehensive tabular overview and graphical representation of the PAMAP2 dataset description.

4.1.7. Preprocessing Details

In this study, we employ several data pre-processing strategies: linear interpolation, scaling and normalization, and segmentation using the sliding window approach.

Linear Interpolation: The above-mentioned HAR datasets in subsection 4.1 are characterized by their real-world nature, involving wireless sensors worn by subjects [1]. As a result, data loss during collection is possible, often represented as NaN or zero values [1]. To address this, we use linear interpolation as noise reduction strategy [76] to fill in missing values and minimize their impact.

Scaling and Normalization: Normalising the input data within the range of 0 to 1 is imperative [1]. This approach mitigates the potential bias introduced during model training when employing substantial values directly from channels.

Segmentation: Smartphones and wearable sensors generate continuous signals, often in the form of time-series data, capturing activities as temporal sequences [1]. The initial step in HAR involves segmenting the sensor data using a sliding window approach, which divides the data into fixed-size windows [1]. For the UCI HAR, UCI HAR-AAL, UCI HAPT, and PAMAP2 datasets, we use the sliding window approach without applying the TSFEL strategy. In contrast, for the MHealth and WISDM datasets, we use the sliding window approach with the TSFEL strategy. Details of the segmentation process using the sliding window strategy, with or without TSFEL, for these diverse HAR datasets are provided in subsection 4.1.

4.2. Implementation Details and Performance Evaluation

We assessed the performance of our suggested DP-FusedNN-EL method for classifying human activities across six popular HAR datasets [29, 30, 40, 41, 42, 43], as shown in Figures 7−24. To enhance the generalization of

our DP-EL model training and address overfitting concerns, we employ a 5-fold cross-validation strategy. This strategy allows us to assess our proposed model performance by generating precision-recall curves based on the predictions made on the validation subsets within each fold of the cross-validation process. These precision-recall curves are presented in Appendix 2 for all utilized datasets. In addition, we use the Harris Hawk Optimization (HHO) technique [77] to optimize hyperparameters for each employed ML model, namely RF, DT, SVM, KNN, Logistic Regression (LR), NuSVM, XGBoost, Extra Trees (ET), and Light Gradient Boosting Machine (LGBM), for all utilized datasets, as exhibited in Appendix 3. We employed these ML models and various DL, including CNN, Multi-Layer Perceptron (MLP), CNN-Bi-GRU, CNN-Bi-LSTM, CNN-GRU, and CNN-LSTM, for performance comparison based on the performance assessment metrics, namely accuracy, precision, recall, and F1 score. The Python simulations were performed on a machine running Microsoft Windows 11 with an Intel Xeon 2.20 GHz CPU clocked at 2.20 GHz, 32 GB of RAM with 4 virtual CPU cores (Kaggle platform).

4.2.1. For UCI-HAR Dataset

The study demonstrates our proposed model's exceptional performance on the UCI-HAR dataset [29], achieving an overall accuracy of 96.97%, precision of 97.10%, recall of 96.76%, and F1 score of 96.85% across all human activities, as exhibited in Figure 7. Specifically, for each activity, accuracies range from 92.06% to 100%, with corresponding precisions, recalls, and F1-scores ranging from 93% to 100%, 92% to 100%, and 95% to 100%, respectively, as exhibited in Figure 7. The accompanying confusion matrix, as exhibited in Figure 8, highlights the significant results on the [29] dataset. Notably, walking and lying activities perform excellently, while walking upstairs and standing demonstrate strong results. However, sitting and walking downstairs activities require improvement.

Our proposed hybrid model consistently outperforms traditional ML and DL models. For ML models, accuracy improves by 1.14% to 10.92%, and F1 scores enhance by 1.02% to 11.09% for the dataset [29]. In the case of DL models, accuracy gains were 1.79% to 3.59%, and F1 score improvements were 1.63% to 3.37%. In contrast, DT, KNN, and CNN models exhibit comparatively lower performances, as exhibited in Figure 9.

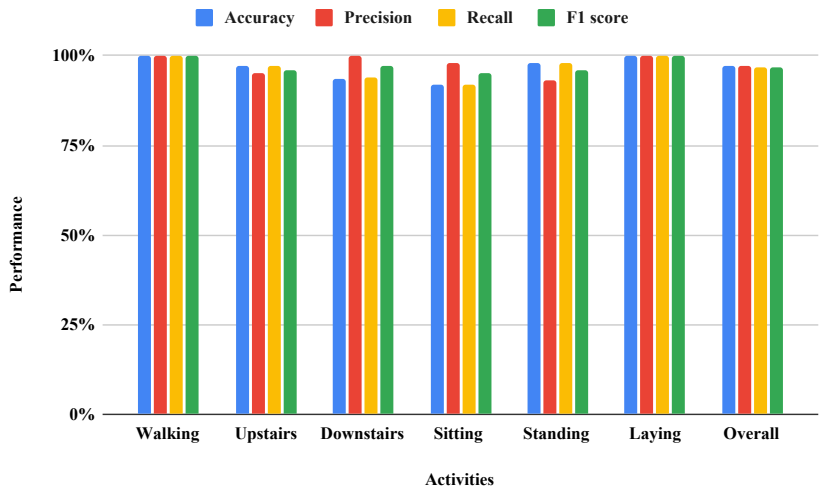

Figure 7: Performance of the proposed DP-FusedNN-EL model on the UCI-HAR dataset, evaluated across individual activities including walking, walking upstairs, walking downstairs, sitting, standing, laying, and overall performance.

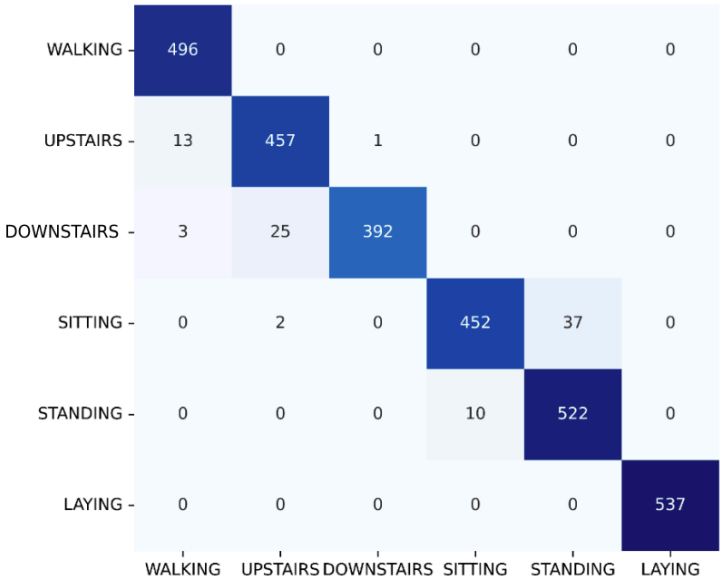

Figure 8: Confusion matrix of the proposed DP-FusedNN-EL model on the UCI-HAR dataset, evaluated across individual activities including walking, walking upstairs, walking downstairs, sitting, standing, and laying.

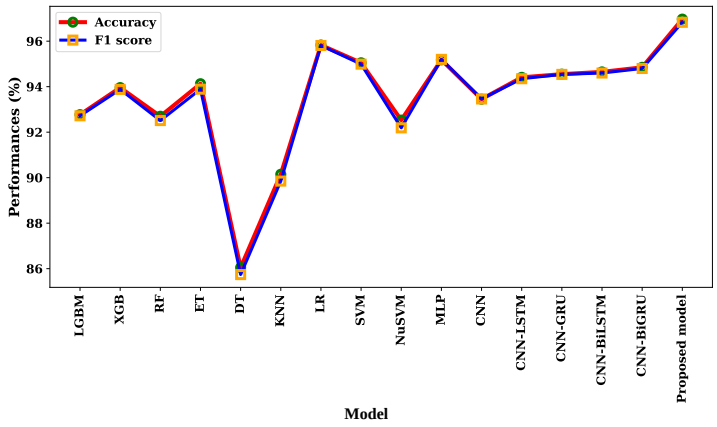

Figure 9: Performance comparison of the proposed DP-FusedNN-EL model on the UCI-HAR dataset (D1) [29] in terms of accuracy and F1 score, against several machine learning and deep learning models including LightGBM (LGBM), XGBoost (XGB), Random Forest (RF), Extra Trees (ET), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Nu-SVM, Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN), and hybrid models such as CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU.

4.2.2. For UCI-HAR-AAL Dataset

This study assesses the performance of our proposed model on the UCI-HAR-AAL dataset [40], achieving significant results. Specifically, we obtained an accuracy of 87.47%, precision of 87.54%, recall of 87.54%, and F1 score of 87.54% across all human activities, as shown in Figure 10. Remarkably, our proposed approach achieved highly desirable performances in terms of accuracies ranging from 79% to 93%, precisions ranging from 82% to 98%, recalls ranging from 82% to 98%, and F1-scores ranging from 82% to 98% for each activity on the UCI-HAR-AAL dataset. The accompanying confusion matrix, as exhibited in Figure 11, highlights that our proposed model improves the performance of the standing and laying activities. In addition, walking and walking downstairs activities still perform good results. However, the sitting activity still fails to achieve satisfactory performance.

Our proposed hybrid model consistently outperforms traditional ML and DL models. Our approach improves accuracy by 0.34% to 11.62% compared to ML models and enhances F1 scores by 0.27% to 11.69%. Conversely, for DL models, accuracy improves within the 2% to 2.48% range, and F1 scores enhance between 1.98% and 2.54%. Notably, DT, KNN, and CNN show

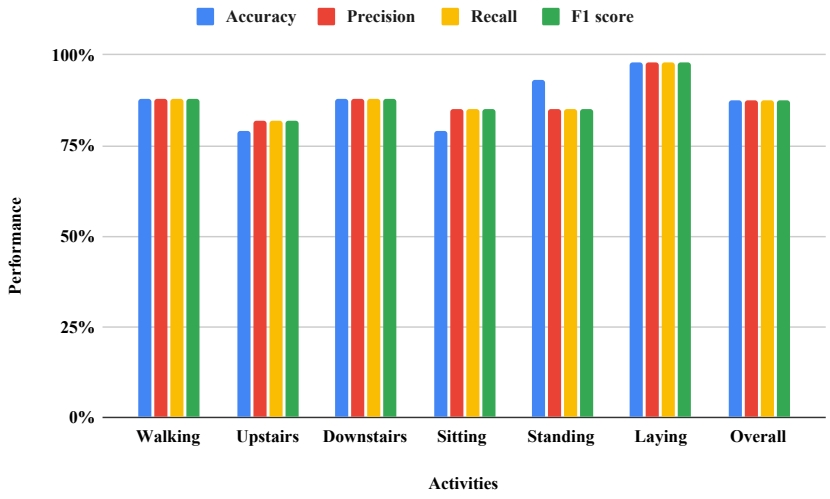

Figure 10: Performance of the proposed DP-FusedNN-EL model on the UCI-HAR-AAL dataset, evaluated across individual activities including walking, walking upstairs, walking downstairs, sitting, standing, laying, and overall performance.

relatively lower performances, as shown in Figure 12.

4.2.3. For UCI-HAPT Dataset

This study evaluates our proposed model's performance on the UCI-HAPT dataset [41], achieving notable results. Across all human activities, our model achieved an accuracy of 98.72%, precision of 98.99%, recall of 98.91%, and F1 score of 98.94%, as shown in Figure 13. Specifically, performances obtained in terms of accuracies ranged from 95.05% to 100%, precisions from 96% to 100%, recalls from 95% to 100%, and F1-scores from 96% to 100%, for each activity. The generated confusion matrix, shown in Figure 14, highlights outstanding performance in activities, such as walking upstairs, downstairs, lying, sit-to-stand, sit-to-lie, and stand-to-lie. While stand-to-sit, lie-to-stand, and standing activities seem to have good results. However, sitting and lie-to-sit activities need improvement to match the performance of other activities on this dataset.

In comparing the performance of our proposed hybrid model with traditional ML and DL models, our proposed approach consistently outperforms these models. Accuracy shows improvement from 4.73% to 17.7%, while F1 scores experience enhancements between 4.8% and 15.94% compared to ML models. In contrast, for DL models, accuracy shows improvement rang-

Figure 11: Confusion matrix of the proposed DP-FusedNN-EL model on the UCI-HAR-AAL dataset, evaluated across individual activities including walking, walking upstairs, walking downstairs, sitting, standing, and laying.

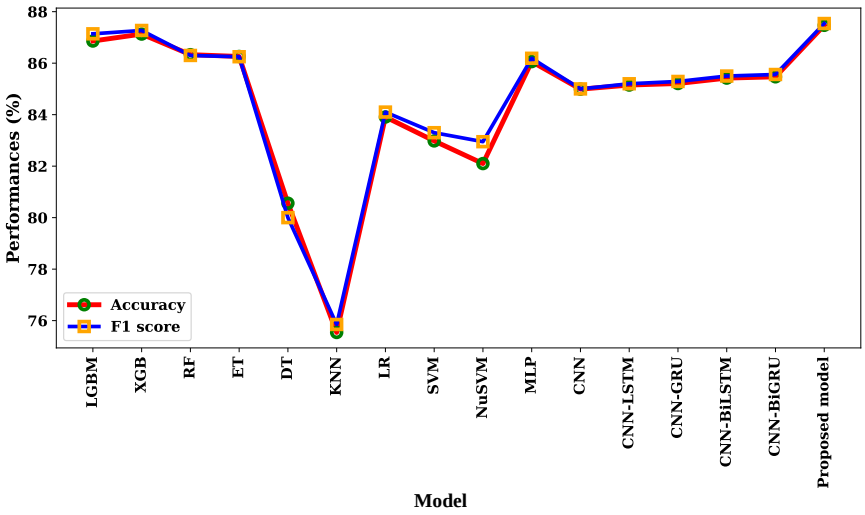

Figure 12: Performance comparison of the proposed DP-FusedNN-EL model on the UCI-HAR-AAL dataset (D2) in terms of accuracy and F1 score, against several machine learning and deep learning models including LGBM, XGB, RF, ET, DT, KNN, LR, SVM, Nu-SVM, MLP, CNN, and hybrid models such as CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU.

ing from 4.79% to 5.71%, while F1 scores experience enhancements between 4.14% and 4.8%. Notably, DT, KNN, and CNN exhibit comparatively lower performances, as illustrated in Figure 15.

4.2.4. For MHealth Dataset

Our proposed model exhibits exceptional performance on the MHealth dataset [42], achieving an accuracy of 99.66%, precision of 99.67%, recall of 99.67%, and F1 score of 99.67% across all activities, as shown in Figure 16. In particular, accuracies obtained range from 98% to 100%; precisions received range from 99% to 100%; recalls range from 99% to 100%; and F1 scores range from 99% to 100% for each activity. The generated confusion matrix, as exhibited in Figure 17, highlights the outstanding performance of all activities. Our proposed hybrid model consistently outperforms traditional ML and DL models, showing accuracy improvements ranging from 0.68% to 9.33% and F1 score enhancements between 0.7% and 9.85% compared to employed ML models. Conversely, for DL models, our proposed model exhibited accuracy improvements ranging from 1.67% to 5.71%, with F1 score enhancements between 1.33% and 1.77%. Notably, DT demonstrates the

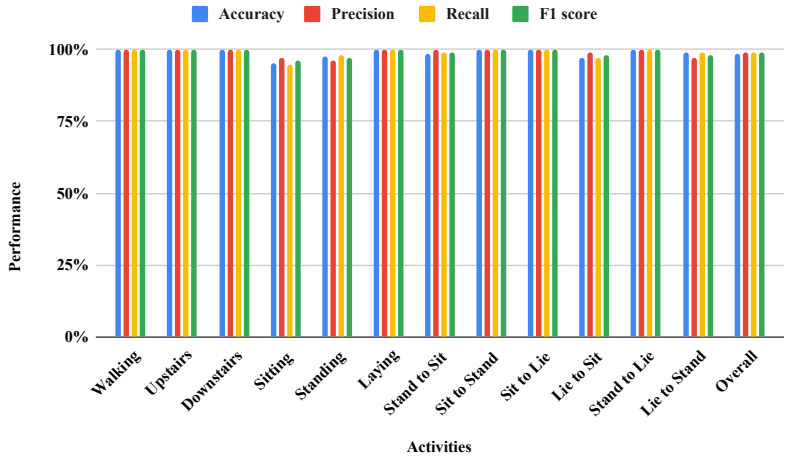

Figure 13: Performance of the proposed DP-FusedNN-EL model on the UCI-HAPT dataset, evaluated across individual activities including walking, walking upstairs, walking downstairs, sitting, standing, laying, stand to sit, sit to stand, sit to lie, lie to sit, stand to lie, lie to stand, and overall performance.

lowest performance, as shown in Figure 18.

4.2.5. For WISDM Dataset

The study evaluates our proposed model's performance on the WISDM dataset [30], achieving notable performances in terms of an accuracy of 97.78%, precision of 97.12%, recall of 96.67%, and F1 score of 96.89% across all human activities on this dataset, as exhibited in Figure 19. Particularly, our proposed approach obtained accuracies ranging from 92% to 99%, precisions ranging from 93% to 99%, recalls ranging from 93% to 99%, and F1 scores ranging from 93% to 99% for each activity, respectively. The generated confusion matrix, as exhibited in Figure 20, highlights the outstanding performance of walking, jogging, sitting, and standing activities. However, walking upstairs and downstairs activities need improvement in enhancing performance comparable to other activities on this dataset.

Our proposed hybrid model consistently outperforms traditional ML and DL models, with accuracy improvements ranging from 1.47% to 7.47% and F1 score enhancements between 1.8% and 9.77% for the dataset [30] compared to ML models. Conversely, DL models show accuracy improvements ranging from 1.74% to 2.61%, with F1 score enhancements between 2.11% and 3.51%.

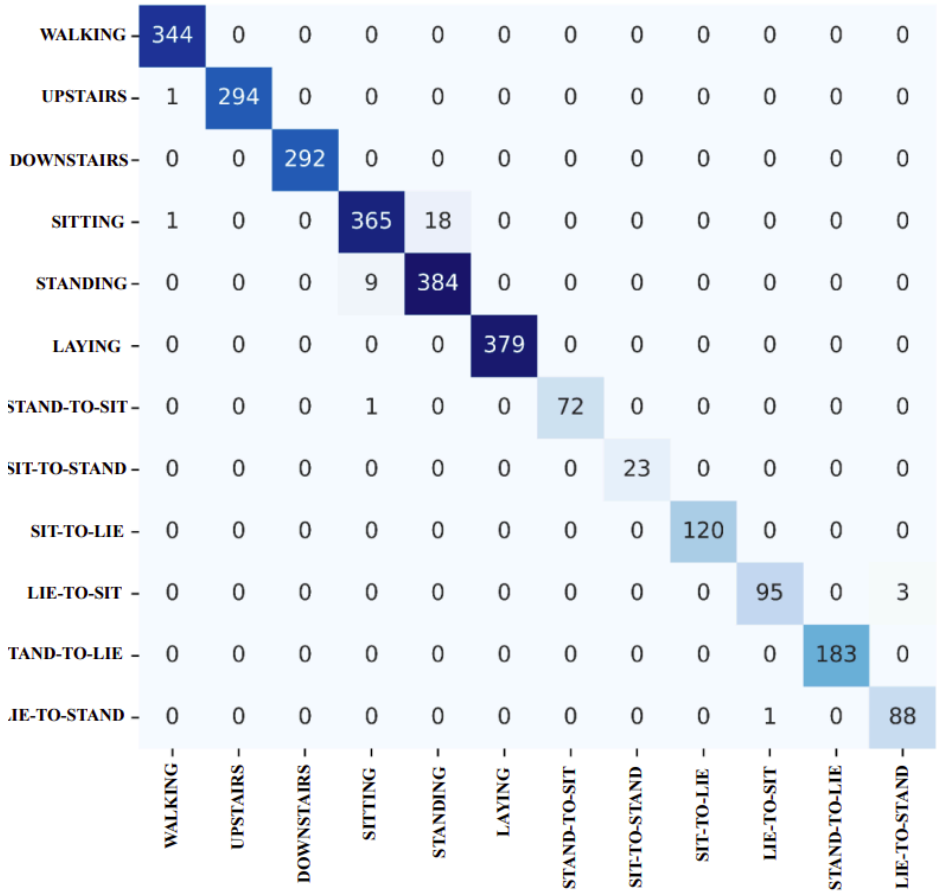

Figure 14: Confusion matrix of the proposed DP-FusedNN-EL model on the UCI-HAPT dataset, evaluated across individual activities including walking, walking upstairs, walking downstairs, sitting, standing, laying, stand to sit, sit to stand, sit to lie, lie to sit, stand to lie, and lie to stand.

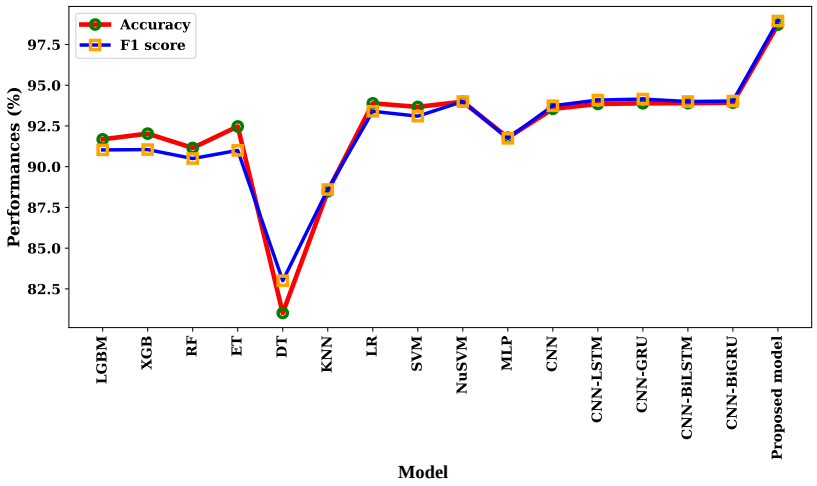

Figure 15: Performance comparison of the proposed DP-FusedNN-EL model on the UCI-HAPT dataset (D3) in terms of accuracy and F1 score, against several machine learning and deep learning models including LGBM, XGB, RF, ET, DT, KNN, LR, SVM, Nu-SVM, MLP, CNN, and hybrid models such as CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU.

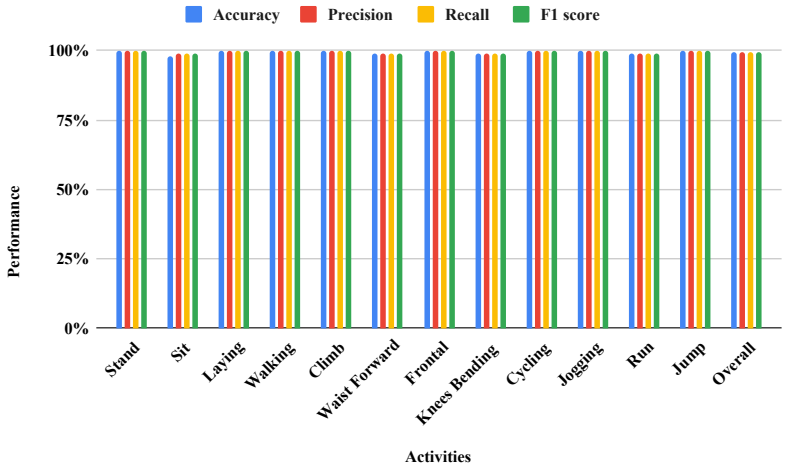

Figure 16: Performance of the proposed DP-FusedNN-EL model on the MHealth dataset, evaluated across individual activities including standing still (stand), sitting (sit), lying down (laying), walking, climbing stairs (climb), waist bending forward (waist forward), frontal elevation of arms (frontal), knee bending, cycling, jogging, running (run), and jumping front and back (jump), and overall performance.

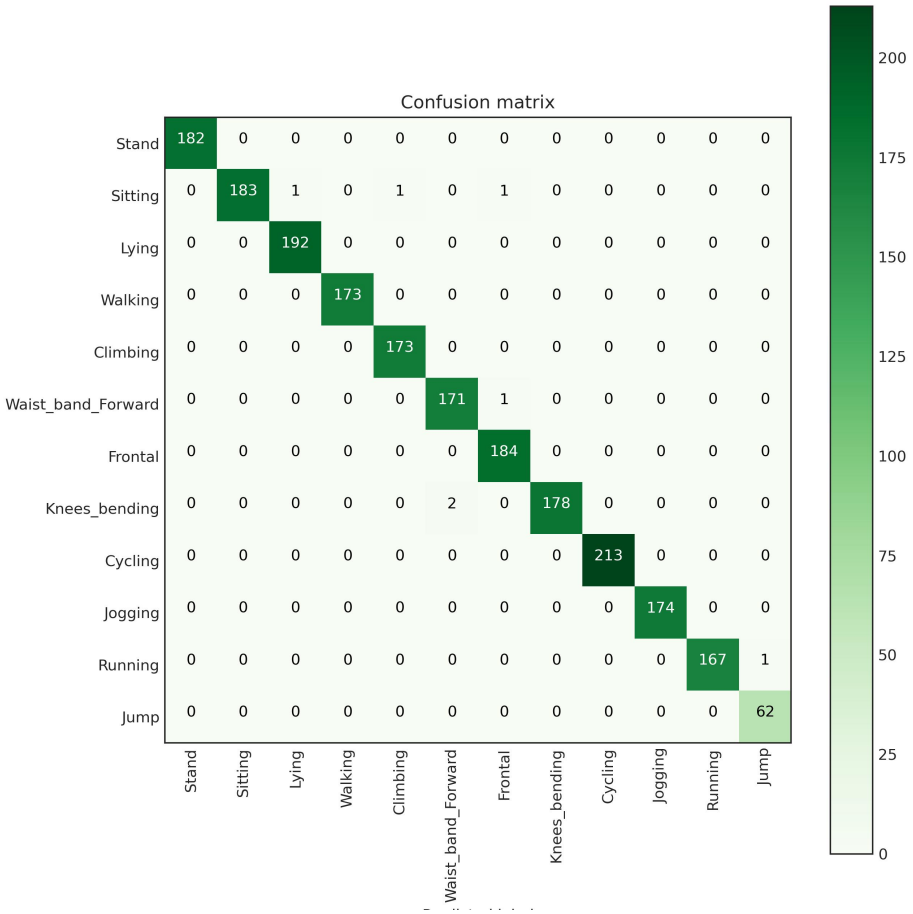

Figure 17: Confusion matrix of the proposed DP-FusedNN-EL model on the MHealth dataset, evaluated across individual activities including stand, sitting, lying, walking, climbing, waist bending forward, frontal elevation of arms (frontal), knee bending, cycling, jogging, running, and jump.

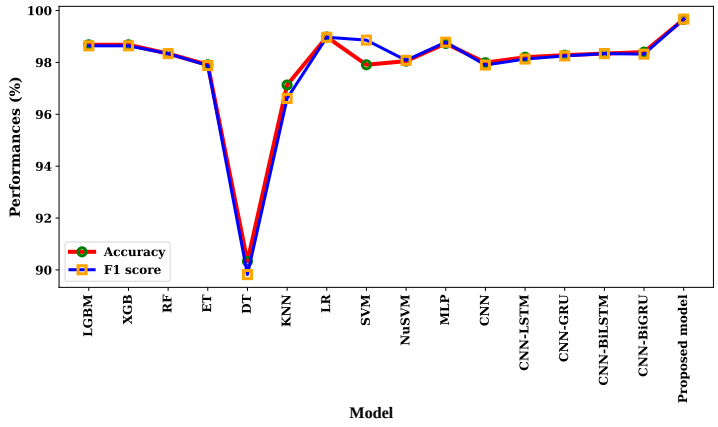

Figure 18: Performance comparison of the proposed DP-FusedNN-EL model on the MHealth dataset (D4) in terms of accuracy and F1 score, against several machine learning and deep learning models including LGBM, XGB, RF, ET, DT, KNN, LR, SVM, Nu-SVM, MLP, CNN, and hybrid models such as CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU.

Notably, DT exhibits the lowest performances among ML and DL models, as shown in Figure 21.

4.2.6. For PAMAP2 Dataset

The study evaluates our proposed model's performance on the PAMAP2 dataset [43], achieving notable performances in terms of accuracy of 96.04%, precision of 96.29%, recall of 94.93%, and F1 score of 95.53% across all human activities on this dataset, as exhibited in Figure 22. Notably, for each specific activity, our proposed approach obtained accuracies ranging from 84% to 100%, precisions ranging from 89% to 100%, recalls ranging from 89% to 100%, and F1 scores ranging from 89% to 100%. The confusion matrix, as exhibited in Figure 23, highlights the outstanding performance of walking, running, nordic walking, and laying activities. Sitting, standing, cycling, and ascending chair activities exhibit good results. However, rope jumping, descending chairs, vacuum cleaners, and ironing activities need to enhance performance comparable to other activities on this dataset.

In comparing the performance of our proposed hybrid model with traditional ML and DL models, our proposed approach consistently outperforms these models. Accuracy shows improvement ranging from 1.54% to 41.24%, while F1 scores experience enhancements between 0.94% and 42.34% for the

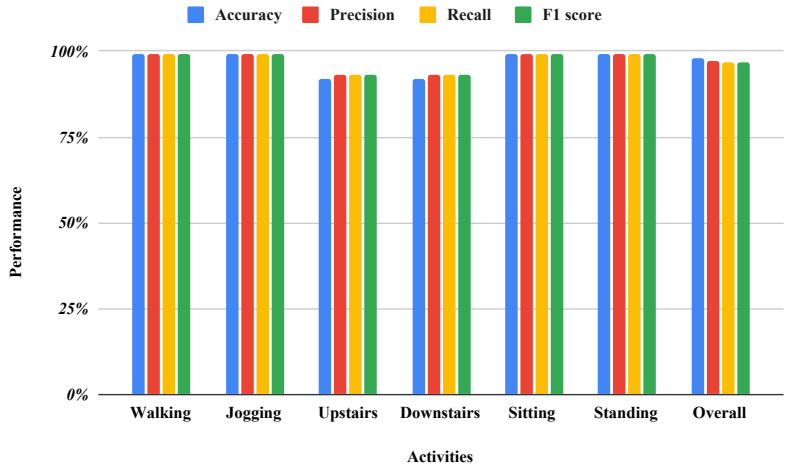

Figure 19: Performance of the proposed DP-FusedNN-EL model on the WISDM dataset, evaluated across individual activities including walking, jogging, upstairs, downstairs, sitting, standing, and overall performance.

dataset [43] compared to ML models. In contrast, for DL models, accuracy shows improvement ranging from 4.37% to 8.47%, while F1 scores experience enhancements between 3.99% and 8.63%. Notably, DT and KNN models exhibit lower performances among ML and DL models, as illustrated in Figure 24.

4.3. Performance Analysis on Different window Length

We extend our experimental analysis by evaluating the DP-FusedNN-EL model with different window lengths of 50, 100, 128, and 200. The experiments are conducted on the UCI HAR, UCI HAPT, and PAMAP2 datasets, as shown in Table 2. The results indicate that a window length of 128 provides the best performance for the UCI HAR and UCI HAPT datasets, with accuracy gains of 0.19% to 2.77% and F1 score improvements of 0.35% to 3.14%. For the PAMAP2 dataset, a window length of 200 yields the highest gains, achieving accuracy improvements of 2.61% to 7.46% and F1 score increases from 2.00% to 6.36%. These findings highlight the importance of selecting the optimal window length for each dataset to enhance performance.

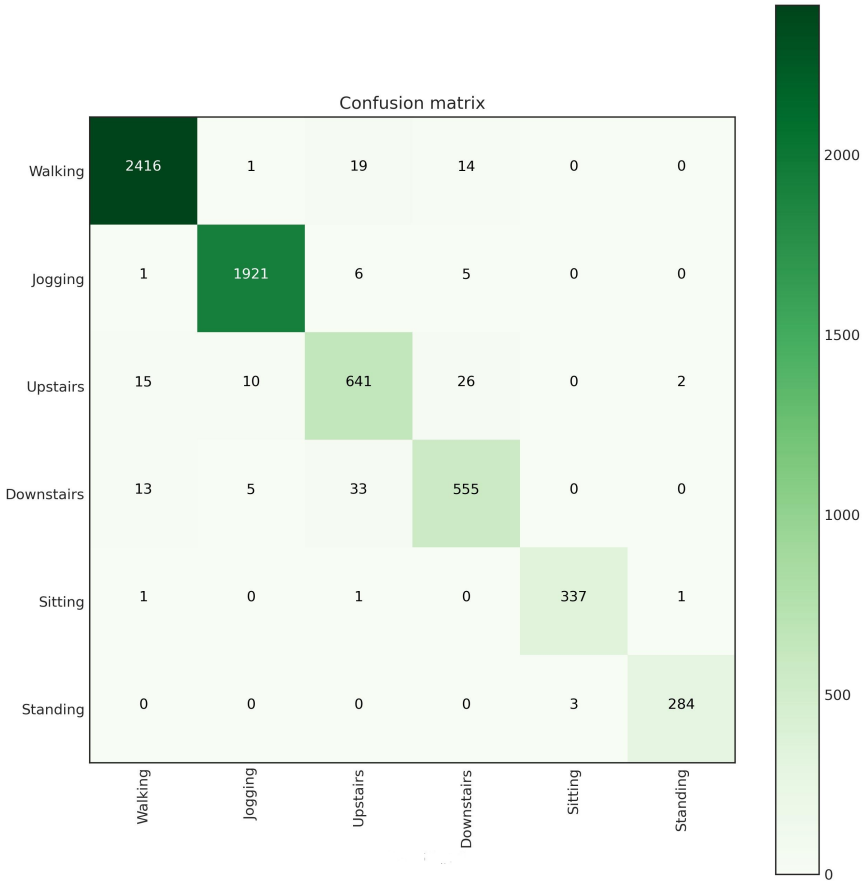

Figure 20: Confusion matrix of the proposed DP-FusedNN-EL model on the WISDM dataset, evaluated across individual activities including walking, jogging, upstairs, downstairs, sitting, and standing.

Table 2: Performance comparison of our proposed DP-FusedNN-EL model for different window length varying from 50, 100, 128, 200, conducted on the UCI HAR [29], UCI HAPT [41], and PAMAP2 [43] datasets.

	Window Length											
Dataset →	UCI HAR				UCI HAPT				PAMAP2			
Metrics ↓	50	100	128	200	50	100	128	200	50	100	128	200
Accuracy	96.08	96.30	**96.97**	96.78	95.95	97.28	**98.72**	98.44	88.58	92.67	93.43	**96.04**
F1 Score	96.10	96.26	**96.85**	96.50	95.80	97.40	**98.94**	98.37	89.16	92.84	93.52	**95.52**

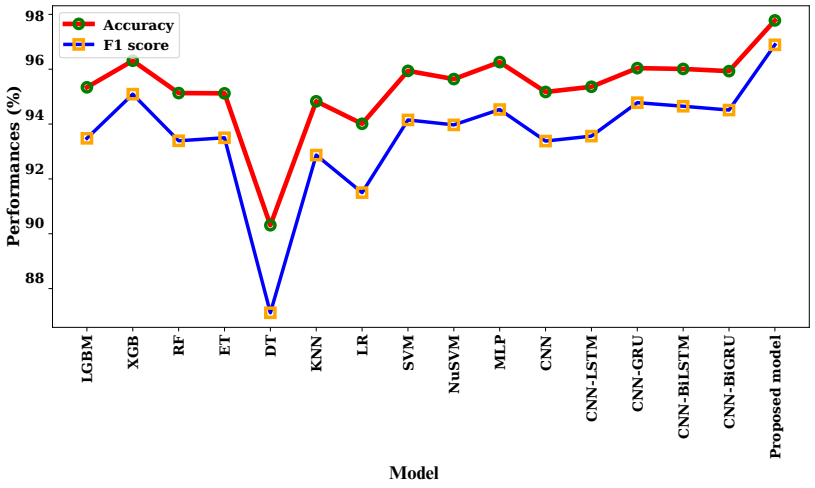

Figure 21: Performance comparison of the proposed DP-FusedNN-EL model on the WISDM dataset (D5) in terms of accuracy and F1 score, against several machine learning and deep learning models including LGBM, XGB, RF, ET, DT, KNN, LR, SVM, Nu-SVM, MLP, CNN, and hybrid models such as CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU.

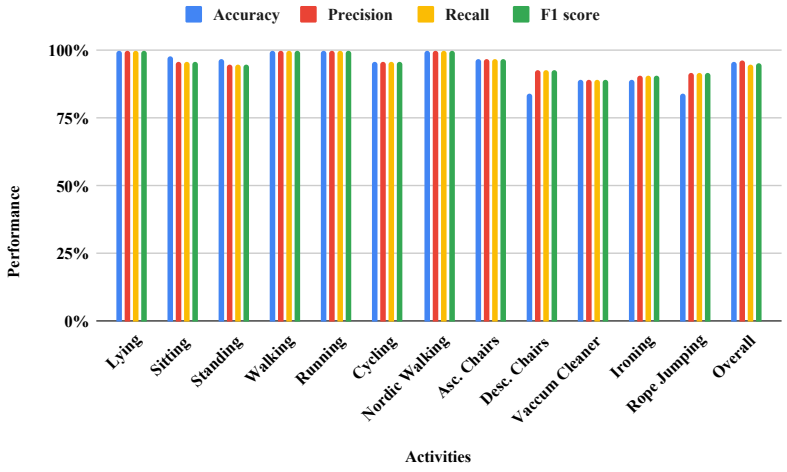

Figure 22: Performance of the proposed DP-FusedNN-EL model on the PAMAP2 dataset, evaluated across individual activities including laying, sitting, standing, walking, running, cycling, Nordic walking, ascending chairs, descending chairs, vacuum cleaner, ironing, rope jumping, and overall performance.

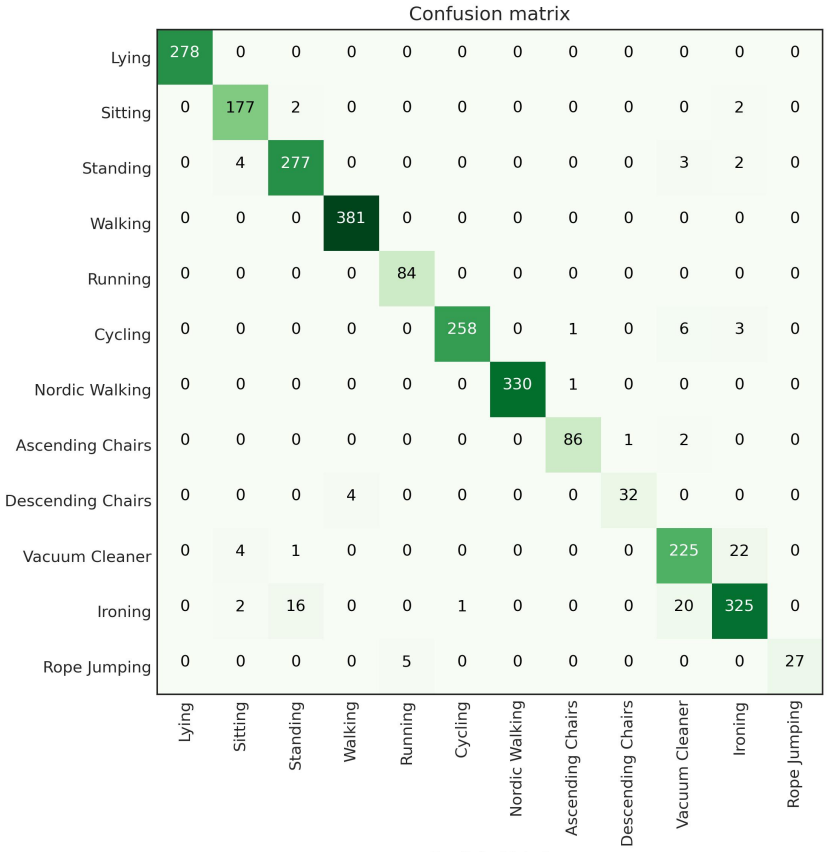

Figure 23: Confusion matrix of the proposed DP-FusedNN-EL model on the PAMAP2 dataset, evaluated across individual activities including laying, sitting, standing, walking, running, cycling, Nordic walking, ascending chairs, descending chairs, vacuum cleaner, ironing, and rope jumping.

Table 3: Performance comparison of our proposed DP-FusedNN-EL model with varying hyperparameter optimization techniques, such as Grid Search Optimization (GSO) [78], Random Search Optimization (RSO) [79], Bayesian Optimization (BO) [80], and Harris Hawk Optimization (HHO) [77] conducted on the UCI HAR [29], UCI HAPT [41], and PAMAP2 [43] datasets.

Dataset →	UCI HAR				WISDM				PAMAP2			
HPO →	GSO	RSO	BO	HHO	GSO	RSO	BO	HHO	GSO	RSO	BO	HHO
Accuracy	96.81	96.85	96.94	**96.97**	96.38	96.45	97.42	**97.78**	95.10	95.18	95.59	**96.04**
F1 Score	96.67	96.78	96.80	**96.85**	96.05	96.13	96.57	**96.89**	94.88	95.07	95.33	**95.52**

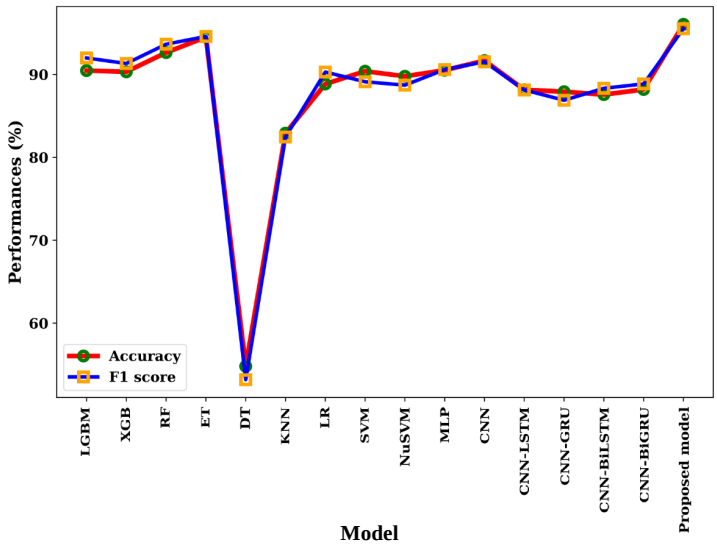

Figure 24: Performance comparison of the proposed DP-FusedNN-EL model on the PAMAP2 dataset (D6) in terms of accuracy and F1 score, against several machine learning and deep learning models including LGBM, XGB, RF, ET, DT, KNN, LR, SVM, Nu-SVM, MLP, CNN, and hybrid models such as CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU.

Table 4: Computational cost analysis of the DP-FusedNN-EL model, consisting of three approaches: DH-Fused-CNN, CNN-SBi-GRU-Attention, and DP-EL, evaluated on the UCI HAR [29], UCI HAR-AAL [40], UCI HAPT [41], MHEALTH [42], WISDM [30], and PAMAP2 [43] datasets. The table presents the number of parameters (in millions, M) and testing time per example (in milliseconds, ms) for each approach, along with the total parameters and testing time for the DP-FusedNN-EL model.

	Number of parameters required by each component (M)			Total parameters (in millions)	Test Time/Sample for Each Component (ms)			Total Test Time/sample (ms)
Dataset (↓)	DH-Fused-CNN	CNN-SBi-GRU-Attention	DP-EL	DP-FusedNN-EL	DH-Fused-CNN	CNN-SBi-GRU-Attention	DP-EL	DP-FusedNN-EL
UCI HAR	2.6	1.3	0.1	4.7	2.5	1.7	0.31	4.51
UCI HAR-AAL	2.6	1.3	0.12	4.72	2.3	1.45	0.3	4.05
UCI HAPT	1.3	2.6	0.08	3.98	4	10	0.6	14.6
MHEALTH	4.5	2	0.15	6.65	1.7	7	0.61	9.31
WISDM	3.7	2	0.2	5.9	2.2	5.4	0.38	7.98
PAMAP2	4.8	1.13	0.25	6.18	5	7	0.64	12.64

4.4. Performance Analysis with Different Hyperparameter Optimization

We extend the experimental analysis of the DP-EL approach by evaluating it with four hyperparameter optimization techniques: Grid Search Optimization (GSO) [78], Random Search Optimization (RSO)][79], Bayesian Optimization (BO) [80], and Harris Hawk Optimization (HHO) [77]. These experiments, detailed in Table 3, are conducted on the UCI HAR, UCI HAPT, and PAMAP2 datasets. The DP-FusedNN approach is used to extract highly representative local-global features for each variant of the DP-EL model with different optimization techniques. The results show that the DP-EL approach optimized with HHO consistently outperforms GSO, RSO, and BO. Specifically, HHO enhances accuracy by 0.18% to 1.4% and improves the F1 score by 0.18% to 0.76%.

The superior performance of HHO can be attributed to its efficient balance of exploration and exploitation, which is crucial for optimizing complex models in HAR tasks. Unlike GSO, which exhaustively searches all combinations, and RSO, which samples randomly without systematic exploration, HHO dynamically adjusts its search strategy based on previous results. This adaptability allows HHO to more effectively navigate high-dimensional hyperparameter spaces, leading to better overall performance.

4.5. Computational Cost Analysis

In this section, we briefly analyze the computational cost of the proposed DP-FusedNN-EL model across six HAR datasets: UCI HAR, UCI HAR-AAL, UCI HAPT, MHEALTH, WISDM, and PAMAP2, as summarized in

Table 4. The computational cost of the DP-FusedNN-EL model arises from three key approaches: DH-Fused-CNN, CNN-SBi-GRU-Attention, and DP-EL. DH-Fused-CNN is designed as the first-phase feature extractor, focusing on capturing highly discriminative local features from the input HAR data, as detailed in Section 3.1.1. CNN-SBi-GRU-Attention is employed as the second-phase feature extractor, capturing both highly representative local and global features from the initially extracted features, as described in Section 3.1.2. DP-EL is employed for the effective recognition of human activities, as outlined in Section 3.2.

The computational complexity analysis shows that the DH-Fused-CNN approach, comprising convolution block, feature fusion, and feature fusion with skip connections modules, often requires higher computational parameters across all datasets except UCI HAPT. This is due to the increased number of filters in each convolution layer and the additional burden of the concatenation layer used in both the feature fusion and feature fusion with skip connections modules. In contrast, the CNN-SBi-GRU-Attention approach, which extracts highly representative local-global features, demands fewer computational parameters. This is because it extracts these features already captured by the DH-Fused-CNN, thereby reducing the overall parameter count. The DP-EL approach, which employs multiple machine learning models such as support vector machines, logistic regression, and random forests, requires the fewest parameters among the three approaches for effective human activity recognition.

Overall, the DP-FusedNN-EL model's total parameter count ranges from 3.98 to 6.65 million. Despite this higher parameter count, the model's test time per sample remains competitive with state-of-the-art methods, as shown in Tables 12 - 15. Therefore, while the proposed model has a higher number of parameters, it achieves superior performance compared to existing models.

4.6. Ablation Study

This study extensively examined the impact of our DP-FusedNN-EL method's influence on enhancing the performance of classifiers. We initially investigated a DH-Fused-CNN approach to extract the highly discriminative local features from input HAR data fed into the DP-EL model to assess their effectiveness. Despite initial promise, its impact on generating performances was limited in capturing representative global features across specified HAR datasets in [29, 30, 40, 41, 42, 43], as shown in Tables 5-10. We then explored

Table 5: Experimental evaluation of the approaches utilized in the DP-FusedNN-EL model across the UCI HAR dataset [29]. This analysis examines the effect of various approaches combinations on model performance. Approaches include DH-Fused-CNN (first phase feature extraction), CNN-SBi-GRU-Attention (second phase feature extraction), DP-EL, Weighted Voting-based Ensemble Learning (WVEL), Stacking-based Ensemble Learning (SEL), and Harris Hawk Optimization (HHO).

Model	Accuracy (%)	F1 (%)	Training Time (in sec)	Test Time (in ms)
DH-Fused-CNN+DP-EL	96.30	96.24	61.34	4.55
CNN-SBi-GRU-Attention+DP-EL	96.5	96.4	67.59	5.694
DP-FusedNN + WVEL	95.83	95.78	62.1	4.426
DP-FusedNN + SEL	95.18	95.14	60.11	**4.403**
DP-FusedNN-EL without HHO	96.78	96.64	**18.75**	4.494
DP-FusedNN-EL	**96.97**	**96.85**	67.34	4.51

Table 6: Experimental evaluation of the approaches utilized in the DP-FusedNN-EL model across the UCI HAR-AAL dataset [40]. This analysis examines the effect of various approaches combinations on model performance. Approaches include DH-Fused-CNN (first phase feature extraction), CNN-SBi-GRU-Attention (second phase feature extraction), DP-EL, Weighted Voting-based Ensemble Learning (WVEL), Stacking-based Ensemble Learning (SEL), and Harris Hawk Optimization (HHO).

Model	Accuracy (%)	F1 (%)	Training Time (sec)	Test Time (ms)
DH-Fused-CNN + DP-EL	87.06	87.01	41.5	4.052
CNN-SBi-GRU-Attention + DP-EL	85.85	86.00	57.59	4.85
DP-FusedNN + WVEL	85.93	87.00	45.3	3.957
DP-FusedNN + SEL	85.25	85.00	39.4	**3.935**
DP-FusedNN-EL without HHO	86.00	86.00	**13.75**	4.038
DP-FusedNN-EL	**87.47**	**87.54**	54.4	4.05

Table 7: Experimental evaluation of the approaches utilized in the DP-FusedNN-EL model across the UCI HAPT dataset [41]. This analysis examines the effect of various approaches combinations on model performance. Approaches include DH-Fused-CNN (first phase feature extraction), CNN-SBi-GRU-Attention (second phase feature extraction), DP-EL, Weighted Voting-based Ensemble Learning (WVEL), Stacking-based Ensemble Learning (SEL), and Harris Hawk Optimization (HHO).

Model	Accuracy (%)	F1 (%)	Training Time (sec)	Test Time (ms)
DH-Fused-CNN + DP-EL	98.30	98.40	102.36	14.36
CNN-SBi-GRU-Attention + DP-EL	98.10	98.10	147.99	2.3
DP-FusedNN + WVEL	96.80	96.80	111.88	14.3
DP-FusedNN + SEL	96.40	96.70	101.28	**14.2**
DP-FusedNN-EL without HHO	98.62	98.81	**18.75**	14.51
DP-FusedNN-EL	**98.72**	**98.94**	103.53	14.6

Table 8: Experimental evaluation of the approaches utilized in the DP-FusedNN-EL model across the MHealth dataset [42]. This analysis examines the effect of various approaches combinations on model performance. Approaches include DH-Fused-CNN (first phase feature extraction), CNN-SBi-GRU-Attention (second phase feature extraction), DP-EL, Weighted Voting-based Ensemble Learning (WVEL), Stacking-based Ensemble Learning (SEL), and Harris Hawk Optimization (HHO).

Model	Accuracy (%)	F1 (%)	Training Time (sec)	Test Time (ms)
DH-Fused-CNN + DP-EL	98.45	98.28	58.4	9.64
CNN-SBi-GRU-Attention + DP-EL	98.99	98.93	72.48	9.56
DP-FusedNN + WVEL	99.41	99.39	47	9.25
DP-FusedNN + SEL	99.41	99.39	49.8	**9.2**
DP-FusedNN-EL without HHO	99.41	99.41	**19**	9.25
DP-FusedNN-EL	**99.66**	**99.67**	60	**9.31**

Table 9: Experimental evaluation of the approaches utilized in the DP-FusedNN-EL model across the WISDM dataset [30]. This analysis examines the effect of various approaches combinations on model performance. Approaches include DH-Fused-CNN, CNN-SBi-GRU-Attention, DP-EL, WVEL, SEL, and Harris Hawk Optimization (HHO).

Model	Accuracy (%)	F1 (%)	Training Time (sec)	Test Time (ms)
DH-Fused-CNN + DP-EL	97.31	96.47	65.4	8.046
CNN-SBi-GRU-Attention + DP-EL	97.49	96.63	77.48	8.296
DP-FusedNN + WVEL	97.44	96.36	59.8	7.996
DP-FusedNN + SEL	96.78	95.55	57	**7.896**
DP-FusedNN-EL without HHO	95.52	95.1	**23**	7.94
DP-FusedNN-EL	**97.78**	**96.89**	68.4	7.98

the CNN-SBi-GRU-Attention strategy, occasionally outperforming the initial methods but not consistently outperforming prior works. We ultimately merged these methods into a Dual-Phase feature extraction approach, effectively capturing highly representative local and global features from the input HAR dataset. This integration led to notable performance improvements, as illustrated in Tables 5-10. Consistent enhancements in accuracy and F1-score, ranging from 0.26% to 3.3% and 0.41% to 4.93%, respectively, were observed across various HAR datasets [29, 30, 40, 41, 42, 43]. These findings underscore the effectiveness of our proposed approach compared to alternative methods.

Optimizing hyperparameters of ML models in our DP-EL model increased training time. In this study, the average training time is calculated based on the number of iterations (for optimizing hyperparameters) or epochs (to train dual-phase feature extraction to extract features) employed, and the results

Table 10: Experimental evaluation of the approaches utilized in the DP-FusedNN-EL model across the PAMAP2 dataset [43]. This analysis examines the effect of various approaches combinations on model performance. Approaches include DH-Fused-CNN, CNN-SBi-GRU-Attention, DP-EL, WVEL, SEL, and Harris Hawk Optimization (HHO).

Model	Accuracy (%)	F1 (%)	Training Time (sec)	Test Time (ms)
DH-Fused-CNN + DP-EL	95.05	94.51	77.3	12.72
CNN-SBi-GRU-Attention + DP-EL	95.59	95.00	86.79	0.733
DP-FusedNN + WVEL	94.88	95.15	66.68	12.56
DP-FusedNN + SEL	91.11	92.22	69.16	**12.42**
DP-FusedNN-EL without HHO	94.55	94.10	**14.16**	12.6
DP-FusedNN-EL	**96.04**	**95.52**	80.3	12.64

are presented in Tables 5-10. However, it yielded significant performance improvements. While neglecting optimization, it reduced training time but often led to inferior performance, as exhibited in Tables 5-10. Additionally, despite longer training times, the testing time per sample is often lower compared to state-of-the-art models, as shown in Tables 12 - 15. This highlights the effectiveness of our proposed model in performing HAR tasks efficiently.

Our ablation study also explores the DP-EL approach's impact on enhancing the performance of our proposed model. Two EL strategies, Stacking-based EL (SEL) and Weighted Voting-based EL (WVEL), were introduced but occasionally underperformed due to including lower-performing models, as exhibited in Tables 5-10. To address this, we introduced the DP-EL model, selecting top-performing models to develop several EL models based on the SWV strategy and combining them to form the DP-EL model. This approach demonstrated significant performance improvements in accuracy, increasing by 0.25% to 4.93% and, hence, surpassed the WVEL and SEL methods, as exhibited in Tables 5-10, asserting this proposed approach's superiority in developing our DP-FusedNN-EL model for performing human activity classification tasks.

Further experiments for ablation study: We extend the experimental analysis of the components employed in our proposed DP-FusedNN-EL model to evaluate its performance on the UCI HAR, UCI HAPT, and MHealth datasets, as summarized in Table 11. This table compares baseline approaches A1-A8, which explore different combinations of four key components: feature fusion (FF) and feature fusion with skip connections (FF-SC) from DH-Fused-CNN, and the recurrent block (RB) and attention module with feature fusion (AM-FF) from CNN-SBi-GRU-Attention. Specifically,

Table 11: Experimental evaluation of the components utilized in the DP-FusedNN-EL model across the UCI HAR, UCI HAPT, and MHealth datasets. This analysis examines the effect of various component combinations on model performance. Components include feature fusion (FF) and feature fusion with skip connections (FF-SC) from DH-Fused-CNN, and recurrent block (RB) and attention module with feature fusion (AM-FF) from CNN-SBi-GRU-Attention. The approaches are defined as follows: A1 employs a Dual-head CNN without any additional components; A2 and A3 incorporate either RB or AM-FF from CNN-SBi-GRU-Attention into the A1 approach; A4 represents the CNN-SBi-GRU-Attention approach; A5 signifies DH-Fused-CNN approach; A6 combines components from A2 and A5; A7 combines components from A3 and A5 similarly; A8 represents DP-FusedNN-EL model. All approaches use the convolution block module and DP-EL for consistency in evaluation. For performance comparison, the Dual-head CNN model is based on a multi-head CNN approach, utilizing multiple convolution blocks without attention or recurrent blocks.

	Components				UCI HAR		UCI HAPT		MHealth	
Approach	FF	FF-SC	RB	AM-FF	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
A1	✗	✗	✗	✗	93.80	93.84	95.80	95.75	96.57	96.50
A2	✗	✗	✗	✓	94.61	94.68	96.53	96.63	97.52	97.45
A3	✗	✗	✓	✗	95.32	95.35	96.97	96.90	98.05	98.10
A4	✗	✗	✓	✓	96.45	96.40	98.10	98.10	98.99	98.93
A5	✓	✓	✗	✗	96.30	96.24	98.30	98.40	98.45	98.28
A6	✓	✓	✗	✓	96.65	96.68	98.44	98.48	99.10	99.03
A7	✓	✓	✓	✗	96.61	96.59	98.50	98.53	99.05	99.05
A8	✓	✓	✓	✓	**96.97**	**96.85**	**98.72**	**98.94**	**99.66**	**99.67**

approach A1 uses a Dual-head CNN without additional components. Approaches A2 and A3 incorporate either RB or AM-FF from CNN-SBi-GRU-Attention into A1. Approach A4 represents the CNN-SBi-GRU-Attention model, while A5 signifies the DH-Fused-CNN model. Approaches A6 and A7 combine components from A2 and A5, and A3 and A5, respectively, to mimic the dual-phase feature extraction strategy of DP-FusedNN-EL; approach A8 integrates all components to form the DP-FusedNN-EL model. Consistency in evaluation is maintained across all approaches using the convolution block module and DP-EL. For comparison, the Dual-head CNN model serves as a baseline, leveraging only convolution blocks in each head of the network while excluding attention and recurrent components.

Results and Discussion: The experimental results in Table 11 demonstrate that baseline approaches A6 and A7, which incorporate three components: FF, FF-SC, and either RB or AM-FF, which outperform approaches A1-A5, which lack these combinations. The inclusion of these components in A6 and A7 facilitates the learning of highly representative local-global features, resulting in performance comparable to our proposed DP-FusedNN-

Table 12: Performance comparison of the proposed approach with prior state-of-the-art models, including AFFNet [35], MI-CNN-GRU [7], Multi-Head CNN [8], Triplet Attention [38], MLCNNwav [9], Bi-GRU-I [34], HMR-CNN-GRU [10], Inception+CBAM [39], ConvBLSTM-PmwA [37], 2D-CNN-LSTM [11], CNN-BAOA-SVM [60], GTS-Net [47], Layer-CNN [46], UC fusion [36] on the UCI-HAR dataset [29]

Authors	Methods	Acc (%)	Prec (%)	Rec (%)	F1 (%)	Training time (sec)	Test Time (ms)
Wang et al. [35]	AFFNet	95.32	-	-	-	-	-
Dua et al. [7]	MI-CNN-GRU	96.20	-	-	96.19	-	-
Khan and Ahmed [8]	Multi-Head CNN	95.38	95.48	95.42	95.37	-	25.62
Tang et al. [38]	Triplet Attention.	-	-	-	96.77	-	-
Dahou et al. [9]	MLCNNwav	95.52	96.2	96.13	96.11	-	-
Tong et al. [34]	Bi-GRU-I	95.42	95.47	95.56	95.45	-	0.7
Nafea et al. [10]	HMR-CNN-GRU	94.5	94.62	94.5	94.46	-	-
Mim et al. [39]	Inception+CBAM	-	96.4	96.27	96.27	-	4185
Yin et al. [37]	ConvBLSTM-PmwA	96.71	-	-	-	-	14.71
Kosar and Barshan [11]	2D-CNN-LSTM	95.66	95.65	95.67	95.62	-	1.60
Dahou et al. [60]	CNN-BAOA-SVM	95.23	95.33	95.33	95.33	-	-
Park et al. [47]	GTS-Net	-	-	-	95.7	-	3.84
Phukan et al. [46]	Layer-CNN	-	-	-	91.66	-	**0.541**
Liu et al. [36]	UC fusion	96.84	96.35	96.22	96.27	-	-
This study	**DH-FusedNN-EL**	**96.97**	**97.10**	**96.76**	**96.85**	**67.34**	4.51

Table 13: Performance comparison of the proposed approach with prior state-of-the-art models, including HMR-CNN-GRU [10], CSNet and TCCSNet [65] on the MHealth dataset [42]

Authors	Methods	Acc (%)	Prec (%)	Rec (%)	F1 (%)	Training Time (sec)	Test Time (ms)
Nafea et al. [10]	HMR-CNN-GRU	99.38	99.35	99.35	99.35	-	-
Essa and Abdelmaksoud [65]	CSNet	97.66	97.04	97.77	97.51	-	-
Essa and Abdelmaksoud [65]	TCCSNet	98.6	98.31	98.66	98.15	-	-
This study	**DH-FusedNN-EL**	**99.66**	**99.67**	**99.67**	**99.67**	**60**	**9.31**

EL model (A8), which integrates all four components. However, the DP-FusedNN-EL model (A8) still achieves superior performance over all baselines, including A6 and A7. These results highlight that incorporating all components in the DP-FusedNN-EL model leads to significant improvements over baseline models (A1 to A7), with performance gains ranging from 2.92% to 3.19% across all datasets.

4.7. State-of-the-Art Performance Comparison

In this section, we compare the performance of our proposed method against established state-of-the-art techniques in [1, 7, 10, 11, 34, 35, 36, 8, 37, 63, 38, 39, 9, 65, 60] on various HAR datasets [29, 30, 42, 43], as detailed in Tables 12-15. Focusing first on dataset [29], our proposed approach outperforms hybrid models in [1, 7, 10, 11, 34, 60], with performance improvements ranging from 0.66% to 2.47%, as shown in Table 12. Additionally, compared to attention-based CNN models [8, 37, 63, 38, 39], our suggested model

Table 14: Performance comparison of the proposed approach with prior state-of-the-art models, including AFFNet [35], MI-CNN-GRU [7], ConvBLSTM-PmwA [37], GTS-Net [47], CSNet and TCCSNet [65], and CNN-BiLSTM-BiGRU [1] on the WISDM dataset [30]

Authors	Methods	Acc (%)	Prec (%)	Rec (%)	F1 (%)	Training Time (sec)	Test Time (ms)
Wang et al. [35]	AFFNet	94.61	-	-	-	-	-
Dua et al. [7]	MI-CNN-GRU	97.21	-	-	97.22	-	-
Yin et al. [37]	ConvBLSTM-PmwA	95.86	-	-	-	-	12.11
Essa and Abdelmaksoud [65]	CSNet	91.21	82.56	82.37	84.14	-	-
Essa and Abdelmaksoud [65]	TCCSNet	92.51	85.18	85.37	86.2	-	-
Park et al. [47]	GTS-Net	-	-	-	88.6	-	**3.62**
Lalwani and Ramasamy [1]	CNN-BiLSTM-BiGRU	99.32	92.82	93.1	73.2	-	-
This study	**DH-FusedNN-EL**	**97.78**	**97.12**	**96.67**	**96.89**	**68.4**	7.98

Table 15: Performance comparison of the proposed approach with prior state-of-the-art models, including MI-CNN-GRU [7], Triplet Attention [38], Inception+CBAM [39], DanHAR [63], CSNet and TCCSNet [65], Shallow CNN [31], GTS-Net [47], and CNN-BiLSTM-BiGRU [1], on the PAMAP2 dataset [43]

Authors	Methods	Acc (%)	Prec (%)	Rec (%)	F1 (%)	Training Time (sec)	Test Time (ms)
Dua et al. [7]	MI-CNN-GRU	95.27	-	-	95.24	-	-
Tang et al. [38]	Triplet Attention	-	-	-	93.2	-	-
Mim et al. [39]	Inception+CBAM	-	90.78	90.3	90.3	-	-
Gao et al. [63]	DanHAR	93.16	-	-	-	-	14.71
Essa and Abdelmaksoud [65]	CSNet	88.43	84.09	85.34	83.83	-	-
Essa and Abdelmaksoud [65]	TCCSNet	89.1	86.42	87.95	87.82	-	-
Huang et al. [31]	Shallow CNN	91.93	-	-	-	-	-
Park et al. [47]	GTS-Net	-	-	-	76.2	-	**4.44**
Lalwani and Ramasamy [1]	CNN-BiLSTM-BiGRU	96.10	79.65	84.57	90.13	-	-
This study	**DH-FusedNN-EL**	**96.04**	**96.29**	**94.93**	**95.52**	**80.3**	12.64

achieves significant performance gains ranging from 0.08% to 1.74%. Furthermore, our proposed model demonstrates notable performance enhancements compared to feature fusion models [35, 36], with improvements ranging from 0.13% to 1.65%. Notably, our approach also exhibits faster testing times per sample, ranging from 0.225 milliseconds (ms) to 41.53 ms, outperforming previous works on the dataset [29], as exhibited in Table 12. In dataset [42], our approach surpasses the performance of hybrid and Transformer models in [10, 65], achieving performance improvements ranging from 0.28% to 1.52%, as exhibited in Table 13.

Moving to dataset [30], our approach outperforms hybrid, attention-based CNN, transformer, and feature fusion approaches in [1, 7, 35, 37, 65], with performance gains ranging from 0.57% to 14.56%, as presented in Table 14. Additionally, our model demonstrates faster testing times per sample, ranging from 3.236 ms to 250.5 ms, compared to previous research works on this dataset, as shown in Table 14. For dataset [43], our approach out-

performs hybrid models, attention approaches, and transformer models in [1, 7, 63, 38, 39, 65], achieving performance improvements ranging from 0.77% to 19.32%, as demonstrated in Table 15. Furthermore, our proposed model achieves faster testing times per sample, ranging from 3.801 ms to 94.851 ms, compared to prior research works on this dataset, as exhibited in Table 15. Therefore, our proposed approach has been thoroughly validated, demonstrating its clear superiority over previously mentioned methods in human activity classification tasks.

Our proposed method consistently outperforms other models, making it the preferred and most effective choice for classifying human health activities in healthcare applications. It underscores the significance of our research in advancing HAR tasks.

Discussion: Our method significantly outperforms existing approaches [35, 7, 8, 38, 9, 34, 10, 39, 37, 11, 60, 47, 46, 36] on the UCI HAR dataset, [65] on the MHealth dataset, [35, 7, 37, 65, 47, 1] on the WISDM dataset, and [7, 38, 31, 63, 39, 65, 1] on the PAMAP2 dataset. These methods often struggle to capture highly representative features due to their reliance on single-phase feature extraction. In contrast, our approach employs a dual-phase feature extraction policy coupled with a dual-phase ensemble learning model. This strategy rigorously extracts features across multiple phases from diverse HAR datasets, enhancing human activity recognition performance. Specifically, our method integrates an extensive feature fusion strategy within a CNN framework (DH-Fused-CNN) and a hybrid network approach (CNN-SBi-GRU-Attention network) to learn comprehensive local-global features. The DP-EL model then classifies human activities, leading to significant performance improvements across various HAR datasets.

4.8. Limitations, Challenges and Potential Solutions

The proposed DP-FusedNN-EL model requires more parameters than state-of-the-art approaches, including resource-constrained lightweight models. This can pose challenges for deployment on devices with limited resources, such as Internet of Things (IoT) devices. The DH-Fused-CNN phase, in particular, demands a higher parameter count for effective local feature extraction compared to existing models. The second phase also adds to this burden by requiring additional parameters to extract both local and global features. Consequently, this model may not be suitable for deployment on IoT devices.

To address these concerns, we need to redesign the DP-FusedNN-EL model to reduce its parameter count. This involves minimizing the number of filters in each convolution layer, replacing concatenation layers with addition layers in the FF and FF-SC modules, and reducing the number of neurons in the fully connected layers. These modifications aim to make the model more suitable for low-resource environments. However, they may lead to a performance drop of approximately 10% to 15% compared to the current configuration.

Alternatively, adopting a single-stage feature extraction strategy with a single-branch neural network, combined with a feature fusion approach and an optimal hierarchical attention mechanism, can address these issues. This strategy would enable the model to learn both refined local and global representations efficiently, potentially enhancing performance while remaining suitable for resource-constrained environments.

5. Conclusion and Future Tasks

We proposed an innovative hybrid approach based on DL and EL-based architecture, DP-FusedNN-EL, to automate HAR with applications in smart healthcare, such as early disease detection. Our approach features a dual feature extraction by dual neural networks operating in two phases to extract local and global features from diverse HAR datasets. Classification tasks are then performed using a DP-EL model. We conducted extensive experiments on multiple HAR datasets to validate and evaluate the effectiveness of our proposed model on diverse HAR tasks. In particular, we compared the performance of our proposed approach with many state-of-the-art ML and DL models. We demonstrated that the proposed approach can achieve significant improvements ranging from 0.08% to 19.32% over existing approaches across all employed datasets.

These performance improvements achieved by our model on various HAR tasks can contribute greatly to many important applications ranging from physical training to health management and early disease prevention. Given the significant performance improvements achieved by our model, we will investigate the applicability of this approach to other healthcare-related tasks, such as body language recognition for emotion and psychiatric symptom detection. Another important future work is to study the use of our model in online and mobile settings and address the associated challenges with these applications.

References

[1] P. Lalwani, G. Ramasamy, Human activity recognition using a multi-branched cnn-bilstm-bigru model, Applied Soft Computing 111344 (2024) 111344–111344. doi:10.1016/j.asoc.2024.111344.

[2] L. Yao, Q. Sheng, B. Benatallah, S. Dustdar, X. Wang, A. Shemshadi, S. Kanhere, Wits: an iot-endowed computational framework for activity recognition in personalized smart homes, Computing 100 (4) (2018) 369–385.

[3] M. Mousse, C. Motamed, E. Ezin, Percentage of human-occupied areas for fall detection from two views, Visual Computer 33 (12) (2017) 1529–1540.

[4] A. Mishra, S. Sharma, S. Kumar, P. Ranjan, A. Ujlayan, Effect of hand grip actions on object recognition process: a machine learning-based approach for improved motor rehabilitation, Neural Computing and Applications 33 (7) (2021) 2339–2350.

[5] D. Vishwakarma, C. Dhiman, A unified model for human activity recognition using spatial distribution of gradients and difference of gaussian kernel, Visual Computer 35 (11) (2019) 1595–1613.

[6] L. Cao, Y. Wang, B. Zhang, Q. Jin, A. Vasilakos, Gchar: An efficient group-based context—aware human activity recognition on smartphone, Journal of Parallel and Distributed Computing 118 (2018) 67–80.

[7] N. Dua, S. Singh, V. Semwal, Multi-input cnn-gru based human activity recognition using wearable sensors, Computing 103 (7) (2021) 1461–1478. doi:10.1007/s00607-021-00928-8.

[8] Z. Khan, J. Ahmad, Attention induced multi-head convolutional neural network for human activity recognition, Applied Soft Computing 110 (2021) 107671. doi:10.1016/j.asoc.2021.107671.

[9] A. Dahou, M. Al-qaness, M. Elaziz, A. Helmi, Mlcnnwav: Multilevel convolutional neural network with wavelet transformations for sensor-based human activity recognition, IEEE Internet of Things Journal (2023) 1–1doi:10.1109/JIOT.2023.3286378.

[10] O. Nafea, W. Abdul, G. Muhammad, Multi-sensor human activity recognition using cnn and gru, International Journal of Multimedia Information Retrieval 11 (2) (2022) 135–147. doi:10.1007/s13735-022-00234-9.

[11] E. Koşar, B. Barshan, A new cnn-lstm architecture for activity recognition employing wearable motion sensor data: Enabling diverse feature extraction, Engineering Applications of Artificial Intelligence 124 (2023) 106529. doi:10.1016/j.engappai.2023.106529.

[12] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[13] D. Madhuranga, R. Madushan, C. Siriwardane, K. Gunasekera, Real-time multimodal adl recognition using convolution neural networks, Visual Computer 37 (6) (2021) 1263–1276.

[14] R. Abdel-Salam, R. Mostafa, M. Hadhood, Human activity recognition using wearable sensors: review, challenges, evaluation benchmark, in: International Workshop on Deep Learning for Human Activity Recognition, Springer, 2021, p. 1–15.

[15] L. Chen, R. Wang, J. Yang, L. Xue, M. Hu, Multi-label image classification with recurrently learning semantic dependencies, Visual Computer 35 (2019) 1361–1371.

[16] D. Dewangan, S. Sahu, Rcnet: road classification convolutional neural networks for intelligent vehicle system, Intelligent Service Robotics 14 (2) (2021) 199–214.

[17] D. Dewangan, S. Sahu, Potnet: Pothole detection for autonomous vehicle system using convolutional neural network, Electronics Letters 57 (2) (2021) 53–56.

[18] D. Dewangan, S. Sahu, Deep learning-based speed bump detection model for intelligent vehicle system using raspberry pi, IEEE Sensors Journal 21 (3) (2020) 3570–3578.

[19] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. Reyes-Ortiz, Human activity recognition on smartphones using a multiclass hardware-friendly support

vector machine, in: International Workshop on Ambient Assisted Living, Springer, 2012, p. 216–223.

[20] M. Uddin, M. Billah, M. Hossain, Random forests based recognition of human activities and postural transitions on smartphone, in: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), IEEE, 2016, p. 250–255.

[21] A. Ignatov, V. Strijov, Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer, Multimedia Tools and Applications 75 (12) (2016) 7257–7270.

[22] G. De Leonardis, S. Rosati, G. Balestra, V. Agostini, E. Panero, L. Gastaldi, M. Knaflitz, Human activity recognition by wearable sensors: Comparison of different classifiers for real-time applications, in: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), IEEE, 2018, p. 1–6.

[23] M. Hassan, M. Uddin, A. Mohamed, A. Almogren, A robust human activity recognition system using smartphone sensors and deep learning, Future Generation Computer Systems 81 (2018) 307–313. doi:10.1016/j.future.2017.11.029.

[24] F. Tang, L. Adam, B. Si, Group feature selection with multiclass support vector machine, Neurocomputing 317 (2018) 42–49. doi:10.1016/j.neucom.2018.07.012.

[25] I. Razzak, K. Zafar, M. Imran, G. Xu, Randomized non-linear one-class support vector machines with bounded loss function to detect of outliers for large scale iot data, Future Generation Computer Systems 112 (2020) 715–723. doi:10.1016/j.future.2020.05.045.

[26] R. Guha, A. Khan, P. Singh, Cga: a new feature selection model for visual human action recognition, Neural Computing and Applications 33 (2021) 5267–5286. doi:10.1007/s00521-020-05297-5.

[27] A. Panja, A. Rayala, A. Agarwala, S. Neogy, C. Chowdhury, A hybrid tuple selection pipeline for smartphone based human activity recognition, Expert Systems with Applications 217 (2023) 119536. doi:10.1016/j.eswa.2023.119536.

[28] D. Thakur, S. Biswas, An integration of feature extraction and guided regularized random forest feature selection for smartphone based human activity recognition, Journal of Network and Computer Applications 204 (2022) 103417. doi:10.1016/j.jnca.2022.103417.

[29] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. Reyes-Ortiz, A public domain dataset for human activity recognition using smartphones, in: 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2013.

[30] J. Kwapisz, G. Weiss, S. Moore, Activity recognition using cell phone accelerometers, ACM SIGKDD Explorations Newsletter 12 (2) (2011) 74–82. doi:10.1145/1964897.1964918.

[31] W. Huang, L. Zhang, W. Gao, F. Min, J. He, Shallow convolutional neural networks for human activity recognition using wearable sensors, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–11. doi:10.1109/TIM.2021.3091990.

[32] Y. Wang, H. Xu, L. Zheng, G. Zhao, Z. Liu, S. Zhou, M. Wang, J. Xu, A multidimensional parallel convolutional connected network based on multisource and multimodal sensor data for human activity recognition, IEEE Internet of Things Journal 10 (16) (2023) 14873–14885. doi:10.1109/JIOT.2023.3265937.

[33] D. Jha, Z. Chen, S. Liu, M. Wu, J. Zhang, G. Morgan, R. Ranjan, X. Li, A hybrid accuracy- and energy-aware human activity recognition model in iot environment, IEEE Transactions on Sustainable Computing 8 (1) (2023) 1–14. doi:10.1109/TSUSC.2022.3209086.

[34] L. Tong, H. Ma, Q. Lin, J. He, L. Peng, A novel deep learning bi-gru-i model for real-time human activity recognition using inertial sensors, IEEE Sensors Journal 22 (6) (2022) 6164–6174. doi:10.1109/JSEN.2022.3148431.

[35] T. Wang, Z. Liu, T. Zhang, S. Hussain, M. Waqas, Y. Li, Adaptive feature fusion for time series classification, Knowledge-Based Systems 243 (2022) 108459. doi:10.1016/j.knosys.2022.108459.

[36] K. Liu, C. Gao, B. Li, W. Liu, Human activity recognition through deep learning: Leveraging unique and common feature fusion in

wearable multi-sensor systems, Applied Soft Computing 151 (2024) 111146–111146. doi:10.1016/j.asoc.2023.111146.

[37] X. Yin, Z. Liu, D. Liu, X. Ren, A novel cnn-based bi-lstm parallel model with attention mechanism for human activity recognition with noisy data, Scientific Reports 12 (1) (2022). doi:10.1038/s41598-022-11880-8.

[38] Y. Tang, L. Zhang, Q. Teng, F. Min, A. Song, Triple cross-domain attention on human activity recognition using wearable sensors, IEEE Transactions on Emerging Topics in Computational Intelligence 6 (5) (2022) 1167–1176. doi:10.1109/TETCI.2021.3136642.

[39] T. Mim, M. Amatullah, S. Afreen, M. Yousuf, S. Uddin, S. Alyami, K. Hasan, M. Moni, Gru-inc: An inception-attention based approach using gru for human activity recognition, Expert Systems with Applications 216 (2023) 119419. doi:10.1016/j.eswa.2022.119419.

[40] K. Davis, E. Owusu, Smartphone dataset for human activity recognition (har) in ambient assisted living (aal), UCI Machine Learning Repository (2016).

[41] J. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, D. Anguita, Transition-aware human activity recognition using smartphones, Neurocomputing 171 (2016) 754–767. doi:10.1016/j.neucom.2015.07.085.

[42] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, C. Villalonga, mhealthdroid: a novel framework for agile development of mobile health applications, in: Ambient Assisted Living and Daily Activities: 6th International Work-Conference, IWAAL 2014, Belfast, UK, December 2-5, 2014. Proceedings 6, Springer, 2014, pp. 91–98.

[43] A. Reiss, D. Stricker, Introducing a new benchmarked dataset for activity monitoring, in: 2012 16th International Symposium on Wearable Computers, 2012.

[44] S. Wan, L. Qi, X. Xu, C. Tong, Z. Gu, Deep learning models for real-time human activity recognition with smartphones, Mobile Networks and Applications 25 (2020) 743–755.

[45] A. Ignatov, Real-time human activity recognition from accelerometer data using convolutional neural networks, Applied Soft Computing 62 (2018) 915–922.

[46] N. Phukan, S. Mohine, A. Mondal, M. Manikandan, R. Pachori, Convolutional neural network-based human activity recognition for edge fitness and context-aware health monitoring devices, IEEE Sensors Journal (2022). doi:10.1109/JSEN.2022.3206916.

[47] J. Park, W. Lim, D. Kim, J. Lee, Gtsnet: Flexible architecture under budget constraint for real-time human activity recognition from wearable sensor, Engineering Applications of Artificial Intelligence 124 (2023) 106543. doi:10.1016/j.engappai.2023.106543.

[48] J. Dhar, An adaptive intelligent diagnostic system to predict early stage of parkinson's disease using two-stage dimension reduction with genetically optimized lightgbm algorithm, Neural Computing and Applications 34 (6) (2021) 4567–4593. doi:10.1007/s00521-021-06612-4.

[49] J. Dhar, N. Ayele, Multi-tier ensemble learning model with neighborhood component analysis to predict health diseases, IEEE Access 9 (2021) 138677–138715. doi:10.1109/ACCESS.2021.3117963.

[50] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Transactions on Neural Networks 5 (2) (1994) 157–166.

[51] M. Ullah, H. Ullah, S. Khan, F. Cheikh, Stacked lstm network for human activity recognition using smartphone data, in: 2019 8th European Workshop on Visual Information Processing (EUVIP), 2019, p. 175–180.

[52] F. Hernández, L. Suárez, J. Villamizar, M. Altuve, Human activity recognition on smartphones using a bidirectional lstm network, in: 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), 2019, p. 1–5.

[53] Y. Zhao, R. Yang, G. Chevalier, X. Xu, Z. Zhang, Deep residual bidir-lstm for human activity recognition using wearable sensors, Mathematical Problems in Engineering (2018) 1–13.

[54] R. Mutegeki, D. Han, A cnn-lstm approach to human activity recognition, in: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, p. 362–366.

[55] F. Ordóñez, D. Roggen, Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.

[56] A. M. Helmi, M. A. Al-qaness, A. Dahou, M. Abd Elaziz, Human activity recognition using marine predators algorithm with deep learning, Future Generation Computer Systems 142 (2023) 340–350.

[57] M. A. Al-qaness, A. Dahou, M. Abd Elaziz, A. M. Helmi, Human activity recognition and fall detection using convolutional neural network and transformer-based architecture, Biomedical Signal Processing and Control 95 (2024) 106412.

[58] Y. Guan, T. Plötz, Ensembles of deep lstm learners for activity recognition using wearables, Proc. ACM Interact., Mob., Wearable Ubiquitous Technol. 1 (2) (2017) 1–28.

[59] F. Karim, S. Majumdar, H. Darabi, S. Chen, Lstm fully convolutional networks for time series classification, IEEE Access 6 (2017) 1662–1669.

[60] A. Dahou, M. Al-qaness, M. Abd Elaziz, A. Helmi, Human activity recognition in ioht applications using arithmetic optimization algorithm and deep learning, Measurement 199 (2022) 111445. doi:10.1016/j.measurement.2022.111445.

[61] K. Wang, J. He, L. Zhang, Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors, IEEE Sensors Journal 19 (17) (2019) 7598–7604.

[62] Q. Teng, K. Wang, L. Zhang, J. He, The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition, IEEE Sensors Journal 20 (13) (2020) 7265–7274.

[63] W. Gao, L. Zhang, Q. Teng, J. He, H. Wu, Danhar: Dual attention network for multimodal human activity recognition using wearable sensors, Applied Soft Computing 111 (2021) 107728. doi:10.1016/j.asoc.2021.107728.

[64] M. A. Al-Qaness, A. Dahou, M. Abd Elaziz, A. Helmi, Multi-resatt: Multilevel residual network with attention for human activity recognition using wearable sensors, IEEE Transactions on Industrial Informatics 19 (1) (2022) 144–152.

[65] E. Essa, I. Abdelmaksoud, Temporal-channel convolution with self-attention network for human activity recognition using wearable sensors, Knowledge-Based Systems 278 (2023) 110867. doi:10.1016/j.knosys.2023.110867.

[66] Q. Xu, M. Wu, X. Li, K. Mao, Z. Chen, Contrastive distillation with regularized knowledge for deep model compression on sensor-based human activity recognition, IEEE Transactions on Industrial Cyber-Physical Systems (2023).

[67] H. Zhang, Z. Xiao, J. Wang, F. Li, E. Szczerbicki, A novel iot-perceptive human activity recognition (har) approach using multihead convolutional attention, IEEE Internet of Things Journal 7 (2) (2020) 1072–1080. doi:10.1109/JIOT.2019.2949715.

[68] D. Thakur, A. Guzzo, G. Fortino, Attention-based multihead deep learning framework for online activity monitoring with smartwatch sensors, IEEE Internet of Things Journal (2023) 1–1doi:10.1109/JIOT.2023.3277592.

[69] M. Islam, S. Nooruddin, F. Karray, G. Muhammad, Multilevel feature fusion for multimodal human activity recognition in internet of healthcare things, Information Fusion 94 (2023) 17–31. doi:10.1016/j.inffus.2023.01.015.

[70] G. Huang, Z. Liu, L. Van Der Maaten, K. Weinberger, Densely connected convolutional networks (2017). doi:10.1109/CVPR.2017.243.

[71] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[72] Z. Lin, M. Li, Z. Zheng, Y. Cheng, C. Yuan, Self-attention convlstm for spatiotemporal prediction, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 11531–11538.

[73] Y. Zhou, H. Zhao, Y. Huang, T. Riedel, M. Hefenbrock, M. Beigl, Tinyhar: A lightweight deep learning model designed for human activity recognition, in: Proceedings of the 2022 ACM International Symposium on Wearable Computers, 2022, pp. 89–93.

[74] N. T. H. Thu, D. S. Han, Hihar: A hierarchical hybrid deep learning architecture for wearable sensor-based human activity recognition, IEEE Access 9 (2021) 145271–145281.

[75] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, H. Gamboa, Tsfel: Time series feature extraction library, SoftwareX 11 (2020) 100456. doi:10.1016/j.softx.2020.100456.

[76] S. Ramani, P. Thévenaz, M. Unser, Regularized interpolation for noisy images, IEEE Transactions on Medical imaging 29 (2) (2010) 543–558.

[77] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, H. Chen, Harris hawks optimization: Algorithm and applications, Future Generation Computer Systems 97 (2019) 849–872.

[78] P. Liashchynskyi, P. Liashchynskyi, Grid search, random search, genetic algorithm: A big comparison for nas. arxiv 2019, arXiv preprint arXiv:1912.06059 17 (2023).

[79] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization., Journal of machine learning research 13 (2) (2012).

[80] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. De Freitas, Taking the human out of the loop: A review of bayesian optimization, Proceedings of the IEEE 104 (1) (2015) 148–175.