陽明交通大學 百川學士學位學程

專題探索(一)專題期末書面報告

****
*　　*
*　　*
*　深度學習於人類活動行為辨識之研究　*
*　　*
*　　*
****

執行期間:112年9月11日 至 113年1月19日

學生姓名:許安

指導教授:曾新穆 講座教授

中 華 民 國　113　年　　1　　月　　19　　日

Abstract

This research project is dedicated to exploring techniques in Human Activity Recognition (HAR), specifically through sensor-based data obtained from wearable devices. The project will entail developing a neural network model to recognize and classify human activities efficiently.

We plan to deploy The HAR algorithms in smart healthcare, human-computer interfaces, and fitness with Artificial Intelligence of Things (AIoT) devices. This integration aims to enhance human well-being and introduce groundbreaking forms of entertainment.

As a measure of our model's success, this research has successfully achieved the introduction and implementation of cutting-edge neural network architectures combining Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and advanced attention mechanisms to enhance the recognition and categorization of human activities, with exceptional accuracy ranging from 93% to 95%. This research also conducts an extensive comparative analysis of the performance variations linked to our unique model design and the datasets' specificities.

Contents

1. Motivation and Background

Human Activity Recognition (HAR) is concerned with creating systems capable of identifying and categorizing human activities and behaviors through analyzing data procured from a diverse array of sensors. These include but are not limited to accelerometers, gyroscopes, and magnetometers, which are integrated into wearable devices, as well as environmental sensors that capture wifi, GPS, Bluetooth, and acoustic signals. By analyzing and classifying these time-series data, HAR systems aim to accurately discern and label a range of human behaviors, such as sitting, running, or ascending stairs, predicated on the sensor data acquired.

The applications of Human Activity Recognition are extensive and multifaceted, encompassing areas such as daily fitness monitoring, healthcare provision, and enhancing human-computer interfaces [9]. By capturing and evaluating the nuances of human kinetics, HAR permits the extraction of behavioral patterns that are pivotal in healthcare for both diagnostics and rehabilitative monitoring, aiding individuals in self-monitoring and lifestyle adjustments to preempt chronic conditions, including obesity, diabetes, and cardiovascular disorders [1].

Notably, HAR methodologies have been employed in creating frameworks for the surveillance and assessment of Parkinson's disease symptoms, exemplified by systems for cough detection [1]. Further, the integration of HAR into immersive technologies like augmented reality (AR) and virtual reality (VR) is redefining the bounds of human interaction with computational environments [1].

Human Activity Recognition (HAR) has burgeoned into extensive empirical research underpinned by many data sources that render the task both challenging and rich in potential applications. Among these sources are data obtained from video surveillance, environmental sensors, and object usage patterns, to name but a few [5]. However, a particularly compelling

subset within this domain is that derived from wearable and embedded sensors. The sensor-based approach to HAR boasts several advantages, primarily attributable to its hidden nature and the high granularity of data it provides.

Wearable sensors, such as accelerometers and gyroscopes, seamlessly integrate into the daily lives of individuals, thus facilitating continuous and real-time monitoring without intruding on the user's routine. The information captured by these sensors is inherently rich in context, offering intricate details of motion that are paramount for distinguishing between diverse human activities. Moreover, the sensor-based methodology inherently affords a level of privacy that visual methods may compromise, sidestepping concerns that often arise with video surveillance data.

2. Literature Review

In sensor-based Human Activity Recognition (HAR), deep learning approaches have gained significant attention due to their ability to extract intricate features from raw data, adapt to various contexts, and provide superior classification performance compared to traditional machine learning techniques. Key neural network architectures employed in this domain include:

- *Convolutional Neural Networks (CNNs)*

  CNNs are predominantly utilized due to their proficiency in processing time-series data, as they can automatically and adaptively learn spatial hierarchies of features from sensor data. Their layered structure allows them to detect local conjunctions of features and temporal dynamics, making them suitable for tasks where the spatial relationship between sensor readings is a discriminative feature [7].

  Advantages: CNNs can handle raw data input, reducing the need for manual feature

engineering. They are adept at capturing spatial and temporal features intrinsic to accelerometer, gyroscope, and magnetometer data streams.

- *Recurrent Neural Networks (RNNs)*

  LSTMs are a special kind of RNN capable of learning long-term dependencies. They are specifically designed to avoid long-term dependency problems, making them effective for HAR tasks where understanding the temporal context is crucial [2].

  Advantages: LSTMs can capture temporal dependencies and sequences in time-series data, which is fundamental for activity recognition, where the sequence of sensor readings is important for identifying the activity.

- *Hybrid Models*

  These models combine CNNs and RNNs to harness the strengths of both architectures. For instance, a CNN can extract spatial features from raw sensor data, and an RNN can be employed to interpret temporal dynamics.

  Advantages: By leveraging the feature extraction capabilities of CNNs and the sequence modeling prowess of RNNs, hybrid models can offer robust performance in various sensor-based HAR tasks [1].

- *Attention Mechanisms*

  Recently, attention-based models, particularly those incorporating the Transformer architecture, have been explored for HAR tasks. They allow the model to focus on the most relevant parts of the sensor data sequence without the constraints of the sequential processing inherent to RNNs.

  Advantages: Attention mechanisms provide a way to handle longer sequences more effectively by focusing on the relevant parts, which can be particularly useful for complex activities that involve multiple stages or varying durations.

- *Graph Neural Networks (GNNs)*

  While less common, GNNs are being explored for HAR due to their ability to capture the relationships and interdependencies between different sensor positions or modalities.

  Advantages: GNNs can model the non-Euclidean structure of sensor networks, providing a way to incorporate the relational information between sensors, which can be essential for certain activities.

3. Materials and Methodology

3.1. Dataset Description

To assess the performance of the proposed human activity recognition technique, we utilized two widely recognized public HAR datasets. These include Opportunity [8] and WISDM [10], which offer a rich compilation of continuous sensor data across diverse sensors, encompassing a range of human activities conducted by various subjects.

- The WISDM dataset comprises a substantial collection of 1,098,207 samples from 29 participants. The data was acquired by having individuals carry Android-based smartphones in their pockets, recording tri-axial accelerometer data. This dataset records the activities performed by the participants, focusing on walking, jogging, climbing stairs, sitting, and standing. The data were collected using a dedicated smartphone application supervised by a dedicated researcher to ensure data quality. Following collection, the data was segmented into ten-second intervals, with features generated from 200 readings to provide valuable insights. The proportion of the overall samples linked to each activity was depicted with walking with 38.6%, jogging with 31.2%, upstairs with 11.2%, downstairs with 9.1%, sitting with 5.5%,

and standing with 4.4%. This dataset is valuable for research and analysis in various fields, particularly activity recognition and mobile sensing applications.

- The Opportunity dataset was collected with the participation of 4 subjects, with the use of body-worn sensors, object sensors, and ambient sensors. Data were collected while subjects performed daily morning routines in a simulated studio environment. Each subject undertook six different runs, five of which represented natural, daily activity scenarios, while the sixth run involved scripted sequences of activities. Annotated across various levels, including modes of locomotion, low-level actions related to objects, mid-level gesture classes, and high-level activity classes, this dataset serves as an ideal platform for evaluating classification, machine learning, automatic segmentation, sensor fusion, data imputation, sensor network research, transfer learning, feature extraction, classifier calibration, and adaptation.

3.2. Preprocessing Techniques

Several preprocessing techniques are commonly employed when working with the Opportunity and WISDM datasets.

1. Linear Interpolation [4,9]: As datasets are most often collected from the actions of participants, it's common to encounter missing data points. Linear interpolation methods are applied to address these gaps in the data. These methods involve using known data points to estimate and replace the missing values through calculation.
2. Scaling and Normalization [4,9]: Algorithms can process data most efficiently when the data are all in a similar format. However, not every sensor records data with the same sampling rates. To address this issue, scaling and normalization techniques are often used to standardize the scale of the samples within the dataset. These techniques

harmonize data from disparate sources and enhance the robustness and accuracy of data-driven analyses.

3. Segmentation [9]: This technique divides large data into smaller, more manageable units. Segmentation aims to partition the data into meaningful, homogeneous, or relevant parts, typically driven by specific attributes, patterns, or criteria. This division facilitates a more targeted and insightful analysis, processing, or interpretation of the data within each segment.

3.3 Neural Network Models

Building upon the distinct advantages of Convolutional Neural Networks (CNNs) in extracting spatial features and Recurrent Neural Networks (RNNs) in capturing temporal dependencies within time-series sensor data, our research endeavors to explore a hybrid neural network framework. This integrated model aims to leverage the convolutional layers of CNNs for their proficiency in detecting spatial patterns and the recurrent layers of RNNs, particularly Long Short-Term Memory (LSTM) units or Gated Recurrent Units (GRUs), to process the temporal sequences inherent in human activity data [7].

To further refine this model, we propose incorporating attention mechanisms, an innovation that has shown significant promise in sequence-to-sequence prediction tasks. The attention mechanisms are designed to selectively concentrate on critical portions of the sensor data sequence, thus enhancing the model's ability to discern and emphasize the most informative features pivotal for accurate activity classification.

This composite architecture, combining CNNs, RNNs, and attention modules, is engineered to extract a more holistic representation of the sensor signals, capturing both the fine-grained and coarse-grained patterns that define human activities. The fusion of these

neural network paradigms is expected to provide a granular understanding of the spatial-temporal data and improve the model's overall predictive performance.

The effectiveness of this hybrid, the attention-augmented model, will be rigorously assessed through a series of experiments and performance evaluations using standard metrics such as accuracy, precision, and recall. These experiments will be designed to validate the model's capability to handle the complexity of human movements and to ensure its applicability in real-world scenarios, to achieve computational efficiency suitable for deployment in resource-constrained environments.

3.4 Performance Metrics

1. Accuracy: Accuracy measures the proportion of correctly classified instances. This research aims to achieve the target of 90 percent accuracy.
2. Precision: Precision quantifies the ratio of true positive predictions to the total positive predictions. It is specifically useful when identifying false positives. This research aims to achieve the target of 80 percent precision.
3. Recall: Recall measures the ratio of true positive predictions to the total actual positive instances. It is important when true positives are costly. This research aims to achieve the target of 90 percent recall.
4. F1-Score: The F1-score is the harmonic mean of precision and recall. It balances the trade-off between precision and recall and is particularly useful in finding a balance between them. It is important when true positives are costly. This research aims to achieve the target of 80 percent F1-Score.
5. AUROC: Area Under Receiver Operating Characteristic Curve quantifies the ability of a model to distinguish between classes, regardless of the chosen classification threshold. This research aims to achieve the target of 85 percent of AUROC.

4. Results

4.1 Neural Network Model Overview

In Figure 1, our model seamlessly integrates convolutional layers designed to extract essential features. The architecture commences with a 1D convolution employing 64 filters, a kernel size of 3, a stride of 1, and a padding of 1. Subsequently, the output undergoes normalization, followed by 1D max pooling with a pool size of 2. To capture temporal dependencies, recurrent layers are introduced, utilizing Long Short-Term Memory (LSTM) with 64 hidden units and a specified number of layers. The output is then set to batch-first for further processing.

Attention layers are incorporated to augment the model's capability to capture pertinent patterns for classification. These layers implement a linear operation to compute attention weights, emphasizing specific temporal information within the input sequence. Following the attention layers, a fully connected layer handles the classification task. This layer utilizes a linear operation to map features extracted from the preceding layers to the output space, generating logits for classification. This step serves as the final stage in the model's architecture.

We implemented K5-fold cross-validation to systematically search for optimal parameters during training to enhance the model's performance. The strategic integration of convolutional, recurrent, attention, and fully connected layers ensures a profound understanding of the intricate patterns inherent in human activities.

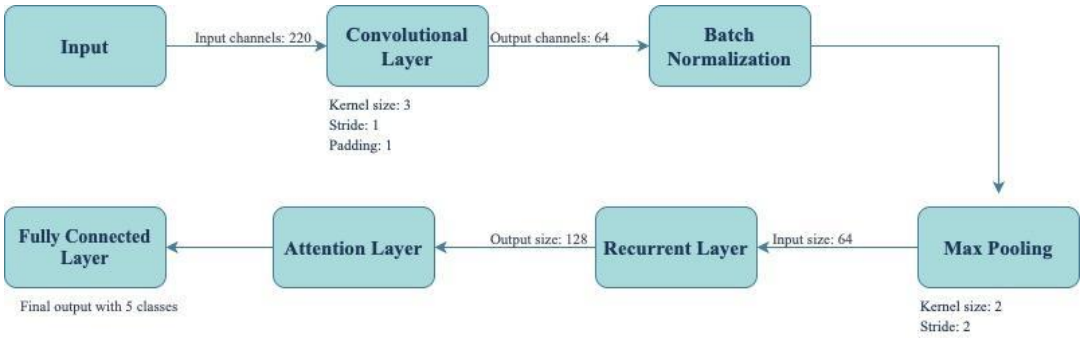

Fig. 1. Detailed Architecture of our Neural Network Model.

4.2 Neural Network Performance

The same hybrid neural network model was rigorously trained and tested on both the WISDM and OPPORTUNITY datasets in our experiments. As shown in Table 1, on the WISDM dataset, the model attained metrics including an accuracy of 93.64%, precision of 90.54%, recall of 90.95%, F1-score of 91.89%, and AUROC of 98.35%. In contrast, training with the OPPORTUNITY dataset, the model exhibited superior performance, reflected in the model of locomotion activities having an accuracy of 95.21%, precision of 95.25%, recall of 95.25%, F1-score of 95.19%, and AUROC of 99.45%, and the model of gesture having an accuracy of 85.42%, precision of 87.35%, recall of 85.42%, F1-score of 85.79%, and AUROC of 99.02%. The performance of the models over epochs can be visualized in Figures 2 to 4.

Table 1. Model Performance Comparison of OPPORTUNITY and WISDM dataset.

Dataset / Task	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUROC**
WISDM	93.64%	90.54%	90.95%	91.89%	98.35%
OPPORTUNITY (Gesture)	85.42%	87.35%	85.42%	85.79%	99.02%
OPPORTUNITY (Locomotion)	95.21%	95.25%	95.25%	95.19%	99.45%

As illustrated in Figure 5, the model performs better on the OPPORTUNITY dataset than the WISDM dataset. This enhanced performance can be attributed to the OPPORTUNITY dataset being collected using seven inertial measurement units and twelve 3D acceleration sensors, providing a more informative input space than the WISDM dataset, which relies solely on a 3-axis accelerometer. This richer sensory information in the OPPORTUNITY dataset likely contributes to its superior performance.

Despite being collected from a smaller number of participants (only four subjects), the OPPORTUNITY dataset exhibits greater diversity and comprehensiveness regarding the

activities and scenarios it covers. This diversity enhances the model's ability to generalize to a broader range of scenarios.

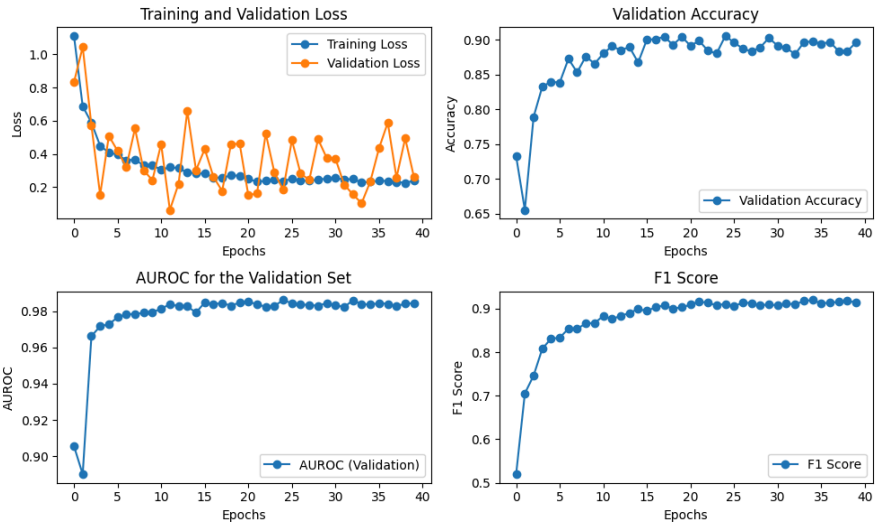

Figure. 2. Visualization of performance of the WISDM model.

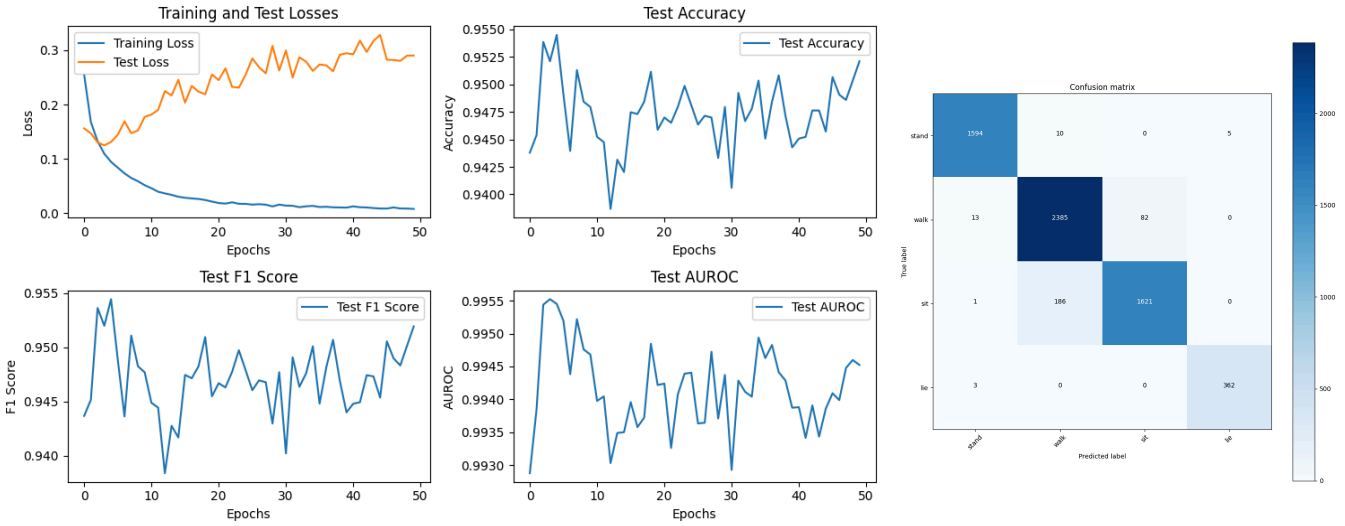

Figure. 3. Visualization of performance of the OPPORTUNITY Locomotion model.

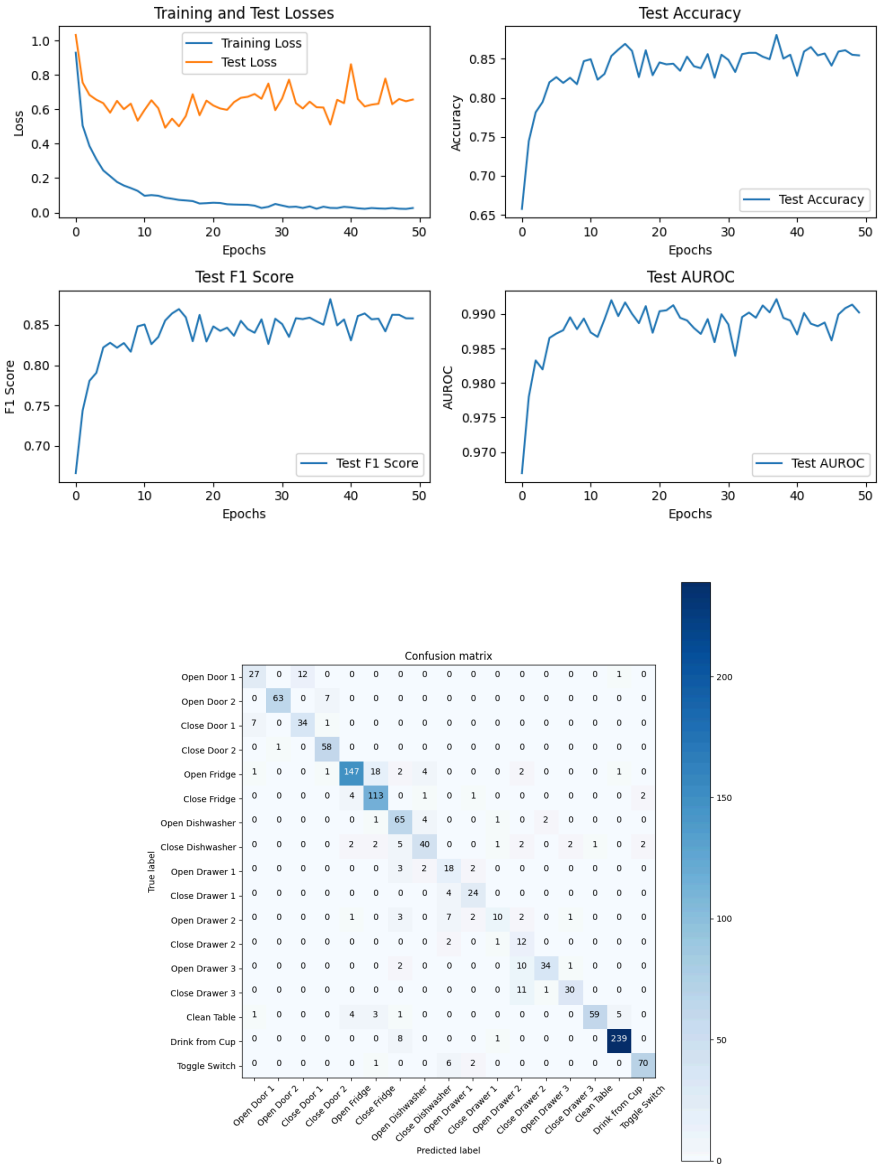

Figure. 4. Visualization of performance of the OPPORTUNITY Gesture model.

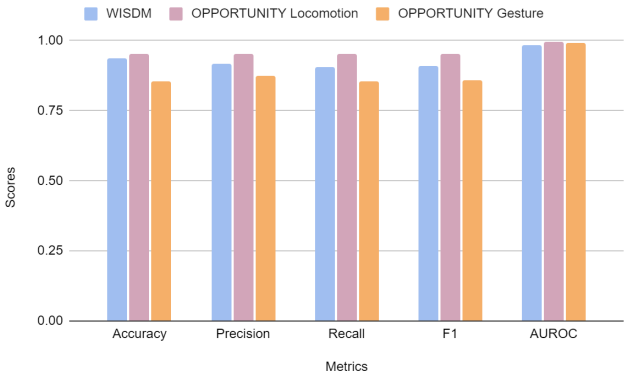

Figure 5. Comparative Analysis of Performance Metrics for the WISDM Dataset and Two Tasks within the OPPORTUNITY Dataset.

Furthermore, our results indicate that the model performed better on locomotion classification tasks within the OPPORTUNITY dataset. This enhanced performance can be ascribed to the dataset's inherent characteristics related to locomotion and gesture tasks. Specifically, the gesture classification tasks involve 17 distinct classes, whereas the locomotion classification tasks encompass only 4. However, there is a notable difference in the amount of data available for these tasks. The gesture classification tasks have training data comprising 14,258 instances and validation data of 1,221, whereas locomotion tasks have a more extensive dataset with training data amounting to 40,730 and validation data of 6,262. This divergence in both data quantity and class distribution may contribute to the model trained on locomotion tasks demonstrating greater robustness and confidence in predicting activities.

With exceptional performance across all metrics, our hybrid model is well-suited for diverse Human Activity Recognition (HAR) scenarios. The inherent adaptability of the hybrid model to varying sensor inputs showcases its versatility and positions it as a technologically resilient solution capable of addressing the dynamic demands of HAR scenarios in practical settings. The model's adaptability to varying sensor inputs makes it a robust choice for an extensive spectrum of real-world applications.

5. Conclusion

In this project, we delved into the realm of Human Activity Recognition (HAR), focusing on developing and comparing enhanced neural network models for activity classification using the WISDM and OPPORTUNITY datasets. Employing a hybrid model that integrates CNN, RNN attention mechanisms, and fully connected layers, the resulting

models demonstrated great results in each performance metric, with the model trained with the WISDM dataset having an accuracy of 93.64%, precision of 91.89%, recall of 90.54%, F1-score of 90.95%, and AUROC of 98.35%, and with the model trained with the OPPORTUNITY dataset having an accuracy of 95.21%, precision of 95.25%, recall of 95.25%, F1-score of 95.19%, and AUROC of 99.45%. However, a notable discrepancy in performance was observed, with the OPPORTUNITY dataset outperforming the WISDM dataset across all metrics.

This underscores the importance of considering dataset characteristics, such as sensory input richness, diversity, and annotation depth, in developing robust HAR models. The proposed neural network architecture, combining Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), attention mechanisms, and fully connected layers, proved effective in handling the complexities of activity recognition.

In future work, the exploration of larger and more diverse datasets, along with the incorporation of advanced techniques such as transfer learning and ensemble methods, could further enhance the generalization capabilities of HAR models. Additionally, real-world deployment scenarios and considerations for model interpretability and explainability should be explored to ensure the practical applicability of HAR systems in various domains, including healthcare, fitness, and immersive technologies.

6. References

[1] Zhang S, Li Y, Zhang S, Shahabi F, Xia S, Deng Y, Alshurafa N. Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. Sensors. 2022; 22(4):1476. https://doi.org/10.3390/s22041476

[2] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges, and Opportunities. ACM Comput. Surv. 54, 4, Article 77 (May 2022), 40 pages. https://doi.org/10.1145/3447744

[3] Jennifer R. Kwapisz, Gary M. Weiss and Samuel A. Moore (2010). Activity Recognition using Cell Phone Accelerometers, Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data (at KDD-10), Washington DC. doi: 10.1145/1964897.1964918

[4] K. Xia, J. Huang and H. Wang, "LSTM-CNN Architecture for Human Activity Recognition," in IEEE Access, vol. 8, pp. 56855-56866, 2020, doi: 10.1109/ACCESS.2020.2982225.

[5] M. A. A. Al-qaness, A. Dahou, M. A. Elaziz and A. M. Helmi, "Multi-ResAtt: Multilevel Residual Network With Attention for Human Activity Recognition Using Wearable Sensors," in IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 144-152, Jan. 2023, doi: 10.1109/TII.2022.3165875.

[6] Mekruksavanich S, Jitpattanakul A. LSTM Networks Using Smartphone Data for Sensor-Based Human Activity Recognition in Smart Homes. Sensors. 2021; 21(5):1636. https://doi.org/10.3390/s21051636

[7] Nafea O, Abdul W, Muhammad G, Alsulaiman M. Sensor-Based Human Activity Recognition with Spatio-Temporal Deep Learning. Sensors. 2021; 21(6):2141. https://doi.org/10.3390/s21062141

[8] R. Chavarriaga, H. Sagha, A. Calatroni, S.T. Digumarti, G. Tröster, J.d.R. Millán, D. Roggen, The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition, Pattern Recognit. Lett. 34 (15) (2013) 2033–2042

[9] Z. Zhou et al., "XHAR: Deep Domain Adaptation for Human Activity Recognition with Smart Devices," 2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Como, Italy, 2020, pp. 1-9, doi: 10.1109/SECON48991.2020.9158431.

[10] Gary M. Weiss, Kenichi Yoneda, Thaier Hayajneh, Smartphone and smartwatch-based biometrics using activities of daily living, IEEE Access 7 (2019) 133190–133202.