Exploring Human Activity Recognition with Acoustic Data: A Comparative Study of CNN-LSTM, ViViT, and ResNet-Temporal Transformer Model

Alaa Humaidan*, Jeny Roy*, Sara Sharifzadeh*, Ruchita Mehta†, Andrea Tales‡, and Joe Macinnes*
*Department of Computer Science, Swansea University, Swansea, United Kingdom
2030685@swansea.ac.uk, jeny.roy@swansea.ac.uk, sara.sharifzadeh@swansea.ac.uk, william.macinnes@swansea.ac.uk
†Centre of Computational Science and Mathematical Modelling, Coventry University, Coventry, United Kingdom
mehtar8@uni.coventry.ac.uk
‡Centre for Innovative Aging, Swansea University, Swansea, United Kingdom a.tales@swansea.ac.uk

***Abstract*—This paper addresses the continuous Human Activity Recognition (HAR) problem using acoustic sensors, which finds application in aged population health and well-being monitoring. The challenging class imbalance problem has been studied using three main groups of time-series modelling strategies, including: 1) the local feature extraction based on Convolutional Neural Networks-Long Short-Term Memory Networks (CNN-LSTM), 2) feature learning based on global dependencies using Video Vision Transformer (ViViT), and 3) combination of the local and global features using ResNet-based frame-level feature extraction followed by Temporal Transformer (RNTT). The acoustic spectrograms have been augmented by adding noise, which has improved all models accuracy (83-86%). Research findings have also demonstrated the best level of resilience to noise condition using the proposed RNTT pipeline (93%), and then the ViViT model (86%).**

***Index Terms*—Keywords: HAR, CNN-LSTM, ViViT, ResNet, Temporal-Spatial Analysis, Acoustic**

I. Introduction

In recent years, the use of various digital sensor data for HAR has drawn considerable interest owing to its broad applications. In the context of elderly care, HAR is particularly valuable for promoting prolonged independent living by monitoring daily activities, detecting falls, and identifying accidents. Acoustic sensors that focus on identifying key actions, such as a call for ”help” or routine daily tasks, rather than private conversations, are well-suited for care home settings, where conventional wearable sensors might not be feasible.

Despite their potential, acoustic sensors have not been widely adopted for HAR for the elderly population. The data used for HAR is different in some aspects compared to other applications of acoustic modeling, such as speech recognition that usually uses data from a close microphone, or scene classification that mainly uses environmental data. For HAR application, the human-generated sounds are used not only based on their speech, but also as a result of their interaction with environment. Furthermore, the user is not necessarily close to the microphone.

Besides that, there are some methodological barriers that should be carefully addressed to design robust acoustic-based HAR models using Artificial Intelligence (AI) time-series analysis strategies. One of the important challenges is providing a large data set of continuous activity recordings, including scenarios similar to real life, with balanced class condition for Deep Neural Networks (DNN). The balance condition is necessary for fair AI modeling. Furthermore, given the diverse acoustic conditions in living environments, designing time-series HAR pipelines with strong features robust to noisy acoustic conditions is desired.

Traditional acoustic models utilized hand-crafted features such as PLP or MFCC [1] [2], which require more manual parameter tuning compared to DNNs. On the other hand, the efficiency of the more recent transformer-based pipelines compared to CNN-LSTM or cascading strategies to combine local CNN features with global dependencies based on transformers has not been well studied in the context of HAR. Considering the reports about the challenges of CNN-LSTMs in capturing long-range dependencies and generalization [3], it is reasonable to conduct comparative studies in HAR domain.

To address these challenges, this paper explores three different deep learning pipelines for HAR, using acoustic data: 1) a CNN-LSTM pipeline, 2) a ViViT pipeline and 3) a proposed cascaded pipeline designed based on ResNet for frame-level feature extraction followed by a Temporal Transformer (RNTT). We further address the class imbalance issue by augmenting the dataset with noise to improve fairness in model decisions across various classes of activities.

A newly collected dataset of acoustic signals, which includes recordings of specifically designed sequences of everyday activities, is utilized in this research. Then, based on an overlapped windowing strategy, the spectrograms have been computed at the window level. This dataset has been

This project is partly supported by Swansea University IAA funding scheme and Coventry University

augmented by applying various types of noises, and the robustness of the models under clean and noisy conditions has been explored. All models achieved good accuracy in clean condition with slightly better results of CNN-LSTM. In noisy test condition, the proposed cascaded RNTT and then, ViViT model demonstrated remarkable resilience to noise condition, offering better generalization compared to CNN-LSTM. To enhance reproducibility, the code used in this study is publicly available at: [4], we plan to release the data later, due to ongoing research involving the same data.

II. Related Work

In the field of acoustic signal classification, there are similar methods that can be employed for both HAR and environmental classification such as CNN-LSTM and unsupervised learning, where similar challenges of noise and variability arise. Although HAR deals specifically with human activities rather than environments, reviewing scene classification research, helps to inform the selection of models and data pre-processing strategies for HAR. Recent studies in acoustic scene classification have employed various machine learning models. Bae et al. Supervised strategies such as Hybrid CNN-LSTM models [5], leveraging both spectral and temporal features and unsupervised methods such as negative matrix factorization (NMF) [6] were employed for scene classification. Besides that Deep NN models were used [7] with emphasis on data augmentation. In the case of HAR, DNNs, CNN-GRU were used [8], [10]. Other studies [9] demonstrated the effectiveness of combining multiple sensors such as Radio-Frequency (RF) sensors and acoustic signals for detection of activities such as walking and falling. These studies collectively illustrate the importance of acoustic data in non-wearable HAR, with particular success achieved using CNN-based architectures and DNNs for sequence modeling.

On the other hand, transformers have recently shown promise in processing time-series data, particularly with their ability to capture global dependencies. The introduction of the Vision Transformer (ViT) [11] adapted the transformer architecture for image classification by treating image patches as sequential tokens. The idea was also extended for video data [12] by introducing ViViT model based on the spectro-temporal dimensions in parallel using tubelet embeddings. This concept is directly applicable to HAR using acoustic data, where spectrograms serve as sequences of frames similar to video [13].

III. Methodology

A. Data Acquisition and Pre-processing

This study utilized a dataset, collected at Coventry University using a 16-channel array-based UMA-16 acoustic sensor. A diverse set of young and elderly volunteers carried out the activities. Each activity series has been recorded 10 times by each of the 11 subjects. The raw audio data underwent several pre-processing steps as follows:

1) Manual Labelling, Windowing, and Spectrogram Computation: Each audio sequence was manually annotated and then segmented into windows of 400 ms with 50 percent overlap. This window size is chosen based on a grid search, matching the natural pace of human activities. A spectrogram has been computed for each 400 ms acoustic window. The spectrogram includes 35 bins (frames) of length 11.428 ms along time and 1025 bins along the frequency axis. A window contains frames that may represent the same or different activities. Due to this heterogeneous nature within some windows, a sequence-to-sequence prediction pipeline was developed. The activities classified in this study include ”silence”, ”coughing”, ”picking up food”, ”eating food”, saying ”hungry”, ”walking”, and ”dropping a spoon”.

2) Augmentation: Data augmentation has been applied to simulate diverse environmental conditions by adding white, pink, brown, and blue noise to the original audio files. The noise levels range from 0.007 to 0.013 with ($\mu = 0.01$, $\sigma = 0.002449$). This augmentation process was intended to expand the dataset size and increase model robustness against noisy environments.

3) Data Balancing: To mitigate potential class imbalance issues, data balancing based on over-sampling of minority classes and the under-sampling of majority classes has been performed. This has been conducted at the window level and since there are both homogeneous and heterogeneous windows, it wasn’t possible to generate the same number of frames for all the activities, though it has improved the balance condition of classes as a whole. Initially, a target of 600 samples (windows) per activity, comprising both clean and noisy signals, have been established. Then, the samples per class have been increased to 2500 to train models demanding more data.

B. Model Architectures and Training

1) 1st pipeline: CNN-LSTM:

- Spectrogram Pre-processing for CNN-LSTM Model

The input spectrograms were reshaped to (41, 25, 35) to match the input dimensions required by the CNN architecture. The process is shown in Figure 1.

This model architecture consists of three CNN layers and two LSTM networks to capture both frame-level spectro-temporal features and the dependencies across the frames. The model has not shown reasonable accuracy without augmented data. However, by leveraging both clean and noisy acoustics, the model demonstrated improved performance in recognizing activities under various environmental conditions.

2) 2nd pipeline: ViViT:

- Spectrogram Pre-processing for ViViT Model

Due to the requirement of square-shaped input patches by the ViViT model, the initial spectrograms of shape (1025, 35) were reshaped. [11] [13]. To achieve this, the last row of the spectrogram was removed using custom code, resulting in a new shape of (1024, 35). This reshaped data was further transformed into (32, 32, 35) to match the ViViT’s expected

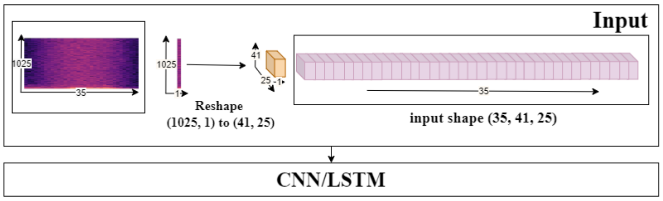

Fig. 1. Pre-processing Workflow for Patch Generation and Input Reshaping for CNN-LSTM Model

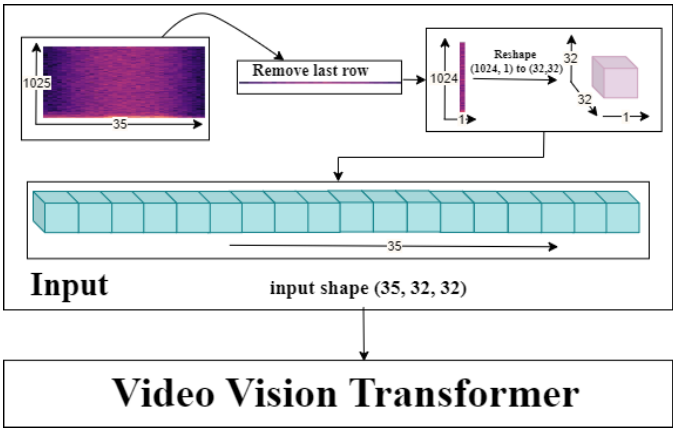

Fig. 2. ViViT architecture, input spectrograms are divided into 3D patches and processed by a Tubelet Embedding layer. Positional encoding keeps the track of sequential structure of the data, while the transformer layers capture both spectro-temporal dependencies

input format, the input shape processing is presented in Figure 2. The final shape for the reshaped window samples represents 35 frames of spectrogram data with a spatial resolution of 32×32 for each frame, which was fed into the ViViT model. The initial patch shape of (41,25) for CNN-LSTM did not yield results as effective as a (32,32) shaped ViViT. The model processed the reshaped spectrograms $35 \times 32 \times 32$ using the key components of ViViT [12]:

- Tubelet Embedding
- Positional Encoding
- Multi-Head Self-Attention

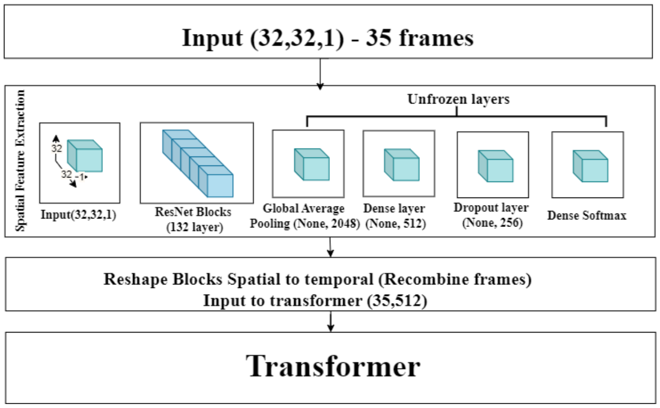

Fig. 3. RNTT architecture, cascading ResNet50 for frame-level spectro-temporal feature extraction with a Temporal Transformer for sequence modeling. The architecture concludes with a classification layer, making it highly effective for HAR tasks

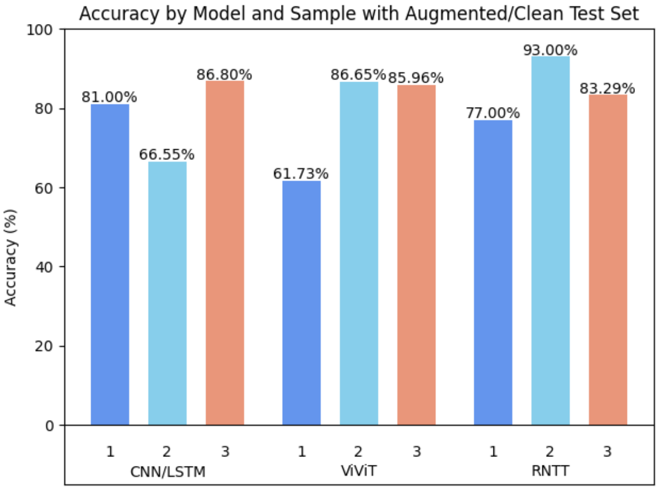

Fig. 4. In all scenarios, models were trained using both noisy and real data. In case 1, presented by the dark blue bars, M1 model of all pipelines trained on more than 3.5k training samples were used, and the test set contained a mix of augmented and real data. In case 2, presented by light blue, the M2 models of all pipelines trained using more than 15.5K samples were used and the test set also included mix data. However, in case 3, represented by orange bars, only noise-free clean test data was used. For CNN-LSTM pipeline, the M3 model was used that was trained using less augmented data (5k samples) due to its poor performance for highly augmented train set, while the same M2 models were employed for the other ViViT and RNTT pipelines.

C. 3rd pipeline: RNTT

In the RNTT pipeline, ResNet50 extracts spectral features from each frame (size:32×32), treating temporal data as 35 separate frames. Optimal results were achieved with 5 epochs, 132 frozen layers, and a batch size of 32. Figure 3 illustrates the overall architecture.

Frame Sequence Embedding: The frame-level features from ResNet50 are reshaped into a time series sequence of shape (35, 512) for transformer input.

Temporal Transformer: The transformer captures temporal dependencies via multi-head self-attention, processing the sequence X_{seq} into X'_{seq}:

$$X'_{\text{seq}} = \text{Transformer}(X_{\text{seq}}) \tag{1}$$

Output Layer: The transformer's output is passed through a softmax classifier:

$$\hat{y} = \text{softmax}(W_o X'_{\text{seq}} + b_o) \tag{2}$$

where W_o and b_o are the weights and biases of the output layer.

IV. Results and Discussion

A. Results

The experiments were designed to evaluate the performance of the three different pipelines, in terms of two key factors: (1) the effect of low versus high levels of data augmentation

TABLE I
HYPERPARAMETERS FOR CNN-LSTM, VIVIT, AND RNTT MODELS

Model	Epochs	Batch Size	Learning Rate	Dropout	Activation	Optimizer
CNN-LSTM	300	20	1.00E-4	0.1	LeakyReLU, Softmax	Adam
ViViT	100	16	1.00E-5	0.3	ReLU	Adam
RNTT	50	32	1.00E-3	0.7	Mish	Adam

in training HAR model, and (2) HAR model accuracy in clean versus noisy environments. The results of these experiments are summarized in Table I and Figure 4.

The first models of all pipelines (M1) was trained using limited amount of augmented data as described in the methodology section, where 600 samples per activity have been initially used). These models were tested on a mix group of clean and noisy test samples. The second group of models (M2) for all three pipelines were trained using a higher level of augmented data, building upon the previously mentioned augmentation process by increasing the dataset to 2,500 samples per activity. These models have been applied on a test set including both clean and noisy samples. Since the performance for CNN-LSTM dropped with heavy augmentation, a third model M3 was developed for this pipeline using less augmented data. Then, in a third test scenario, including only clean acoustic test samples, the accuracy of M3 for CNN-LSTM and M2 models of ViViT and RNTT pipelines were evaluated.

Due to the significantly low accuracy observed in preliminary tests before augmentation, results for models trained only on clean data have been neglected from this section. The focus is instead on the impact of augmentation on performance and the robustness of each model in noisy and clean environments. The results of these experiments are presented in Figure 4 and detailed in Table I. Furthermore, in order to analyse the accuracy of these models for different groups of activities, the confusion matrices for all models are provided in Figure 5. While all models generally perform well for 'coughing', there is notable confusion between 'silence' and 'walking'. That can be explained based on the low levels of sound in walking activities in PC Lab surface.

B. Discussion

The experiments provide insights about the models performance, the effect of data augmentation, and resilience to diverse conditions. The CNN-LSTM model performed well with controlled augmentation in both clean and noisy condition, while showing sensitivity to over-augmentation. That can be due to its strong local frame level features compared to temporal features, because noise effect is dominant in spatial 2D frames. On the other hand, both the proposed RNTT and ViViT demonstrated good resilience to noise, maintaining high accuracy on both noisy and clean conditions. That is connected to the attention mechanism and tubelet embeddings which can capture strong global spectro-temporal features connected to the inherent acoustic features. That has made these models robust to noisy environments and also aligns with other research findings [14]. Since RNTT uses both frame level local features based on CNN, besides the global temporal features from transformer part, its accuracy has slightly dropped in clean condition (83.29%), when trained using augmented data. Because the CNN layers learn part of noise features, besides the inherent acoustic features.This proposed RNTT pipeline, due to its tolerance to noise, can be used in noise condition while also maintaining reasonable accuracy in clean condition.

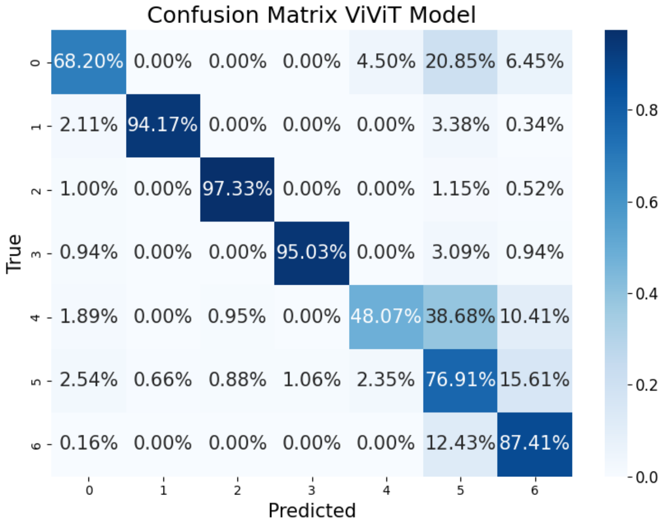

Fig. 5. Confusion matrix for the ViViT model when tested on clean samples (encoded labels. The encoded labels are as follows: 0 - coughing, 1 - drop spoon, 2 - feel bad, 3 - hungry, 4 - pickup/eat food, 5 - silence, 6 - walking)

These results emphasize the impact of augmentation levels on model performance, aligning with previous research on noise augmentation [2]. Besides that, Transformer-based models can outperform traditional CNN-LSTM models in noisy condition, but also demand more training data and computational resources, which is consistent with prior research [11], [15]. This makes them well-suited for real-world applications like elderly care monitoring in noisy environments.

CONCLUSION

This study evaluated three pipelines for acoustic HAR: CNN-LSTM, ViViT, and a proposed RNTT model. All models achieved comparable accuracy in clean condition, when trained using both clean and augmented data based on some types of noise. In the case of noisy condition, the proposed RNTT model achieved the highest accuracy, while ViViT and CNN-LSTM, were the second and third in that case. While these experiments highlight the positive effect of data augmentation on the accuracy of HAR models, they also demonstrate the challenges it poses for models generalization.

References

[1] Hermansky, H., ”Perceptual linear predictive (PLP) analysis of speech,” The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738-1752, 1990. DOI: 10.1121/1.399423

[2] W. Dai, ”Acoustic scene recognition with deep learning,” IEEE, 2016.

[3] E. M. Saoudi, J. Jaafari, and S. J. Andaloussi, “Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN,” *Scientific African*, vol. 21, p. e01796, 2023. doi: https://doi.org/10.1016/j.sciaf.2023.e01796

[4] A. Humaidan, ”HAR Acoustic Data: CNN-LSTM, ViViT, ResNet-Temporal Transformer Code,” GitHub Repository, 2024. [Online]. Available: , ” *GitHub Repository*,:https://github.com/Alaalhumaidan/HAR-AcousticData-CNN-LSTM-ViViT.git

[5] S. H. Bae, I. Choi, and N. S. Kim, ”Acoustic scene classification using LSTM & CNN,” IEEE, 2016.

[6] V. Bisot, R. Serizel, S. Essid, and G. Richard, ”Supervised non-negative matrix factorization for acoustic scene classification,” IEEE, 2016.

[7] D. Wei, J. Li, P. Pham, S. Das, and S. Qu, ”Acoustic scene recognition with DNN,” IEEE, 2016.

[8] S. Gupta, “Deep learning based human activity recognition (HAR) using wearable sensor data,” *International Journal of Information Management Data Insights*, vol. 1, no. 2, pp. 100046, 2021. doi: https://doi.org/10.1016/j.jjimei.2021.100046.

[9] M. Mohtadifar, M. Cheffena, and A. Pourafzal, “Acoustic- and Radio-Frequency-Based Human Activity Recognition,” in *Sensors*, vol. 22, no. 9, Article 3125, 2022. Available: https://www.mdpi.com/1424-8220/22/9/3125.

[10] M. Nicolini, F. Simonetta, and S. Ntalampiras, “Lightweight Audio-Based Human Activity Classification Using Transfer Learning,” in *Proc. of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, SciTePress, 2023. Available: https://www.scitepress.org/Papers/2023/116479/116479.pdf.

[11] A. Dosovitskiy et al., ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.

[12] A. Arnab et al., ”Video Vision Transformer,” ICCV, 2021.

[13] D. Rothman, Transformers for Natural Language Processing and Computer Vision: Explore Generative AI and Large Language Models with Hugging Face, Chat

[14] P. Kaur, Q. Wang, and W. Shi, ”Fall detection from audios with audio transformers,” arXiv, 2022.

[15] A. Vaswani et al., ”Attention is All You Need,” NeurIPS, 2017.