

Lightweight Audio-Based Human Activity Classification Using Transfer Learning

Marco Nicolini^{*}, Federico Simonetta[†] ^a and Stavros Ntalampiras[†] ^b

LIM – Music Informatics Laboratory, Computer Science Department, University of Milan, Milan, Italy

Keywords: Audio Pattern Recognition, Machine Learning, Transfer Learning, Convolutional Neural Network, YAMNet, Human Activity Recognition.

Abstract: This paper employs the acoustic modality to address the human activity recognition (HAR) problem. The cornerstone of the proposed solution is the YAMNet deep neural network, the embeddings of which comprise the input to a fully-connected linear layer trained for HAR. Importantly, the dataset is publicly available and includes the following human activities: *preparing coffee, frying egg, no activity, showering, using microwave, washing dishes, washing hands, and washing teeth*. The specific set of activities is representative of a standard home environment facilitating a wide range of applications. The performance offered by the proposed transfer learning-based framework surpasses the state of the art, while being able to be executed on mobile devices, such as smartphones, tablets, etc. In fact, the obtained model has been exported and thoroughly tested for real-time HAR on a smartphone device with the input being the audio captured from its microphone.

1 INTRODUCTION

Human Activity Recognition (HAR) is the process of automatic detection and identification of physical human activities (Ramanujam et al., 2021). Its applications range from health care systems with real-time remote tracking of patients – e.g. medical diagnosis and tracking of elderly people –, to smart-home and safe-traveling systems, including the recognition of criminal human activity (Ntalampiras and Roveri, 2016) and activities in natural environments (Ntalampiras et al., 2012).


The main issue in HAR is to leverage motion signals to classify the type of action that is ongoing. The literature mainly focuses on motion and wearable sensors (Ramanujam et al., 2021) or vision sensors (Beddiar et al., 2020) and has recently embraced the deep-learning world (Chen et al., 2021). For instance, CNNs with inertial sensor data, such as accelerometers and gyroscopes, have been used to sample the acceleration and the angular velocity of a body (Bevilacqua et al., 2018). In such a context, real-time HAR has been achieved using deep learning models using information coming from sensors typically existing in smartphones (Ronao and Cho, 2016; Wan et al., 2020). However, real-time audio-based HAR is still


an open subject, and this work fills exactly that gap.

While multimodality is a key aspect in the HAR field (Chen et al., 2021) and every commercial smartphone device is equipped with one or more microphones, the specific modality has not been thoroughly explored in a stand-alone neither a multimodal setting, where information from multiple modalities is exploited. This work explores how existing audio-recognition models tailored to real-time classification can be applied to HAR.

A few previous works presented HAR models based on the corresponding acoustic emissions. One work focused on the feature extraction stage, consisting of a selection analysis via genetic search; the proposed method is tailored to low-power consumption devices, such as smartphones, and employs Random Forest (RF) and Neural Network (NN) models (Riboni et al., 2016). Another work focuses on transfer learning for data augmentation to reduce the imbalance bias and improve generalization (Ntalampiras and Potamitis, 2018). Finally, the audio channel has been used in multimodal online HAR systems (Chahuara et al., 2016).

Regarding approaches based on traditional machine learning, i.e. not based on deep neural networks, non-Markovian ensemble voting has been used for robotic applications (Stork et al., 2012). Social network analysis based on graph statistics has been applied by constructing networks between

^a  <https://orcid.org/0000-0002-5928-9836>

^b  <https://orcid.org/0000-0003-3482-9215>

windows of the audio fragments and comparing the graphs related to different activities (García-Hernández et al., 2017).

It should be mentioned that the spread of audio devices listening to users' speech, activities, etc. without transparently making it public has created serious privacy concerns (Lau et al., 2018). An existing work analyzes the impact of audio deterioration on speech intelligibility with the hope of finding privacy-friendly methods for HAR (Liang et al., 2020). To the best of our knowledge, this latter path is yet to be explored requiring more attention from the scientific community.

With the rise and ever-increasing adoption of Deep Neural Networks, the problem of data availability became of primary importance. While some datasets for audio-based HAR are available – see Sec. 2 –, various works have proven that using pre-trained models improves model performances and reduces training costs. As such, we adopted a Transfer Learning-based strategy, which is proven to be helpful for the generalization abilities of the resulting models (Pan and Yang, 2010; Zhuang et al., 2021).

The contributions of this work are:

1. a proof-of-concept application for audio-based HAR operating in real-time on commercially available smartphones;
2. an exploration of the effectiveness of generic audio classification models for HAR-specific tasks;
3. a novel extension of an existing dataset for Audio-based HAR.

The rest of the paper is organized as follows: the following section describes the employed dataset and section 3 presents the proposed method. Subsequently, section 4 explains the experimental set-up and analyses the obtained results. Finally, section 5 demonstrates the developed prototype application and section 6 provides our conclusions and directions for future work.

2 DATASET

There are few audio datasets facilitating HAR based on audio data. In this work, we used an existing dataset (Riboni et al., 2016) composed of eight classes taken in various indoor environments, so that different background noises and acoustic conditions are well-represented. Specifically, the classes available are: *brewing coffee*, *cooking*, *using the microwave oven*, *taking a shower*, *dish washing*, *hand washing*, *teeth brushing* and *no activity*. Since this dataset suffers from imbalance issues, it was expanded to com-

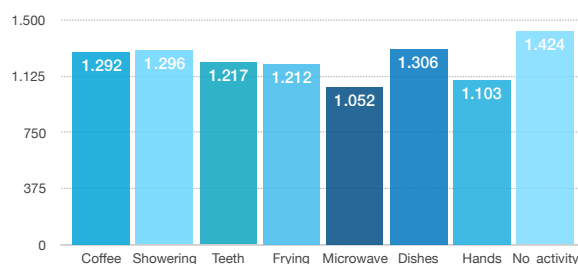


Figure 1: Total duration per each activity class in the used dataset, after expansion and balancing. Time is in seconds.

pensate the less represented classes. Namely, we manually selected smartphone and low quality microphones recordings from Freesound¹. The downloaded files were annotated using the existing directory structure of the dataset.

Since there were various file types with various sample rates (from 8 to 64 KHz) and channels (mono or stereo), we also converted all audio files to wave encoding format (.wav) with mono channel and 16 KHz sample rate. To improve the learning procedure, silence was removed from files using Reaper, a professional Digital Audio Workstation². Specifically, we used the dynamic split items tool with gate threshold set at -24dB, hence, the audio below that threshold the activity is considered silence – except for the “No_activity” class. Moreover, the standard deviation (21 m 25 s) and average duration (24 m 45 s) of the initial dataset clearly indicated a highly imbalanced situation.

Consequently, we reduced the cardinality of the classes having total duration above the average by iteratively removing one random audio file at a time from the dataset, until the total duration of the class was less than 24 minutes. In the resulting dataset the total average duration across classes is 20 m 44 s and the standard deviation between classes total duration is 1 m 57 s – see Figure 1.

3 THE PROPOSED METHOD

The proposed method is based on YAMNet³, a Neural Network for audio classification. YAMNet is fed with log-Mel spectrograms and outputs one tag among 521 classes from the AudioSet-Youtube⁴ corpus. YAMNet consists of 86 layers based on the MobileNet-v1 architecture (Howard et al., 2017), which is created

¹<https://freesound.org>

²<https://www.reaper.fm/>

³<https://github.com/tensorflow/models/tree/master/research/audioset/yamnet>

⁴<https://research.google.com/audioset/>

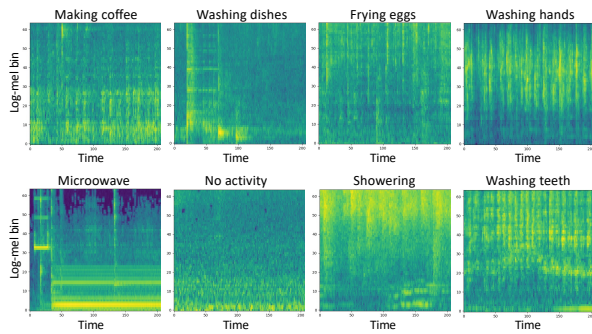


Figure 2: Mel-scaled spectrograms of representative segments of the considered human activities.

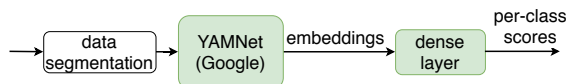


Figure 3: Block diagram illustrating the pipeline of the proposed method.

using depth-wise convolutions reducing the computational complexity of the network. Among those, only 28 layers have learnable weights, with 27 convolutional layers, and one fully connected layer.

With the purpose of transferring knowledge learnt by YAMNet, we discard the last fully connected layer and substitute it with a new fully connected layer designed for our custom classification task. Overall, the portion of YAMNet we use has 3'195'456 trained parameters.

For creating the log-Mel spectrograms used by YAMNet, it is fundamental the understanding of frame and hop size. Indeed, long-time dependencies may be captured more easily with long frame sizes. At the same time, long frames may impact negatively the training of the network for short-time dependencies, because it could be saturated of information and could have issues in the identification of the most discriminatory features. Moreover, long frames can reduce the total number of frames. We empirically found that an optimal frame size was 15600×6 seconds with an overlap of the 50%. This value was chosen because 15600 (0.975 seconds) is the exact segment size required by YAMNet. Figure 2 illustrates log-mel spectrograms of representative segments of each class existing in the dataset.

After having segmented windows, in order to improve the generalization abilities of the model, we added Gaussian noise with mean $\mu = 0$ and standard deviation $\sigma = 0.2$. This technique was proven to make the model robust against realistic noise sources, e.g., previous work employ such a method in order to ignore the noise over time (Kahl et al., 2017).

The frames are then fed into the pre-trained YAMNet and the produced embeddings, after a ReLU ac-

tivation function, are passed to a fully-connected linear layer trained from scratch on the new dataset – see Figure 3. The output is then processed with SoftMax activation during training, while a simple *argmax* function can be used at inference time.

For training, we used “Adam” (Kingma and Ba, 2014) update algorithm and Categorical Cross-Entropy as multi-class loss function. Batch size was set to 10 and training was performed for 100 epochs. Both epochs and batch size values were chosen after preliminary exploration using cross-validation.

3.1 Comparative Analysis with k -nn

k -Nearest Neighbor Classifier (k -NN) despite its simplicity it is a suitable approach for multi-class problems (Hota and Pathak, 2018). The standard version of the k -NN classifier has been used with the Euclidean distance as similarity metric.

Feature Extraction. The short-term features feeding the k -NN model are the following: a) zero crossing rate, b) energy, c) energy’s entropy, d) spectral centroid and spread, e) spectral entropy, f) spectral flux, g) spectral rolloff, h) MFCCs, i) harmonic ratio, j) fundamental frequency, and k) chroma vectors. We opted for the mid-term feature extraction process meaning that mean and standard deviation statistics on these short term features are calculated over mid-term segments. More information on the adopted feature extraction method can be found in (Giannakopoulos and Piskakis, 2014).

Parameterization. Short- and mid-term window and hop sizes, have been discovered after a series of early experimentations on the various datasets. The configuration offering the highest recognition accuracy is the following: 0.05, 0.025 seconds for short-term window and hop size; and 1.0, 0.5 seconds for mid-term window and hop size respectively. Overall,

Table 1: Standard deviation of the obtained recognition accuracy per class and data division scheme.

Class	10-folds	3-folds
doing coffee	0.03188	0.01067
frying egg	0.03580	0.03124
no activity	0.01921	0.00790
showering	0.01381	0.00391
microwave	0.02805	0.00839
washing dishes	0.04032	0.01001
washing hands	0.01779	0.01384
washing teeth	0.02321	0.01221

Table 2: Average and Standard Deviation of Balanced Accuracy; * refers to the baseline model described in (Riboni et al., 2016); ** refers to the k -NN model

Folds	Avg
10	0.8617
10**	0.8091
3	0.8820
3*	0.8560
3**	0.8098

the both feature extraction levels include a 50% overlap between subsequent windows.

Moreover, parameter k has been chosen using test results based on the ten-fold cross validation scheme; depending on the considered data population, the obtained optimal values range in [3, 21]. The best k parameter obtained is $k=3$.

4 EXPERIMENTAL SET-UP AND RESULTS

For evaluating the proposed system, we designed two experiments, namely a 3 and a 10 fold cross-validation.

The metrics used to evaluate the trained model are summarized in the following list:

- *Balanced accuracy per fold*: computes the mean of the true positive rate obtained on each class of a fold; it corresponds to the following formula:

$$\frac{\sum_{i=0}^{M-1} \frac{tp_i}{tp_i+fn_i}}{M}$$

where M is the number of classes, tp_i and fn_i are the number of true-positives and false-negatives of the i -th class;

- *Average balanced accuracy*: computes the mean of the balanced accuracy across folds;
- *Standard deviation of balanced accuracy*: computes the standard deviation of the average balanced accuracy;
- *Per-class standard deviation*: for each class, the true-positive-rate is averaged across the folds; then, the standard deviation is computed;
- *Normalized confusion matrix*: all the confusion matrices of all folds are summed; then, they normalized so that each row sums to one.

Comparing 10 and 3-fold cross-validation – see Table 2 – the model seems to suffer from the increase of data in the training set, leading to a decrease of performance in the 10-fold cross-validation. Table 1 shows the standard deviation across folds of

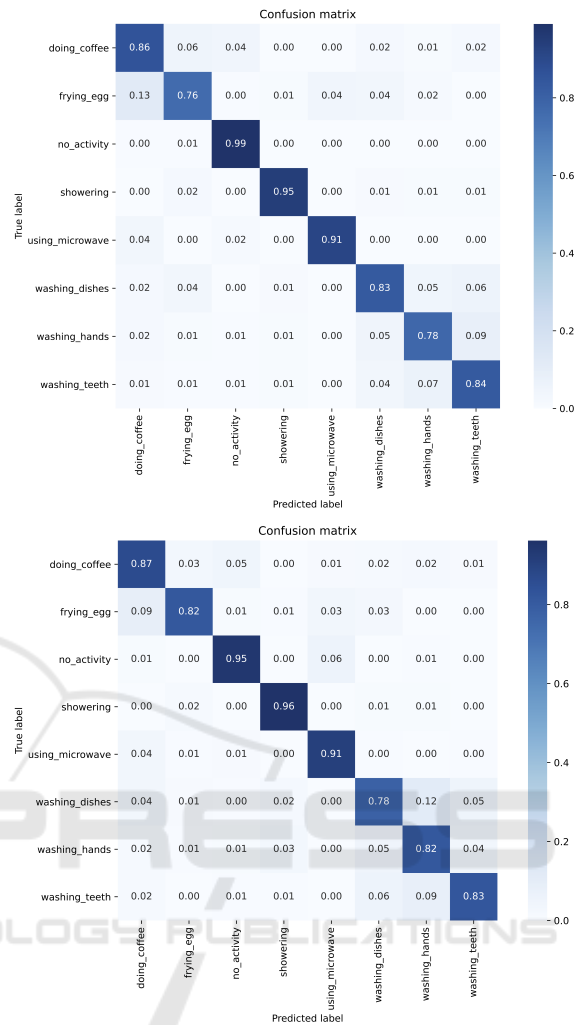


Figure 4: Confusion Matrix of 3-fold (top) and 10-fold (bottom) evaluation test. Matrices are created by summing the confusion matrix obtained in each fold and then normalizing so that each row sums to 1.

the true positive rates for each class. Since the effect is noticeable in all classes, we assume that it is not connected with specific samples. Therefore, the increased dataset size should be thoroughly assessed for possible negative effects, which may impact the learning ability of the model.

Overall, as shown in Figure 4, all classes are well-predicted. Miss-classifications are mainly concentrated in the differentiation among “washing dishes”, “washing hands”, and “washing teeth”, probably because of the water sound in the background. Other miss-classifications are between “frying egg” and “doing coffee”, probably because they include metal sounds (of pots and pans) and some parts with very low intensity sounds.

We compared the best-performing model to a

baseline work (Riboni et al., 2016) on the dataset without our expansion nor our balancing strategy. For the comparison, we used 3-fold cross-validation. The specific work focuses on the usage of low-power consumption devices. Even though the overall accuracy does not differ significantly from the baseline model, it is interesting to note that there are relevant differences in the way misclassifications are distributed across classes (see Figure 5):

- “doing coffee”: the proposed model produces less miss-classifications with “frying eggs” and “microwaves”;
- “no activity”: less miss-classifications with “washing teeth” using the proposed method;
- “using microwave”: the baseline model confused this class with “doing coffee”, while the proposed method does not;
- washing hands: similar, but our model performed 0.04 worse than the baseline in the 10-fold cross validation, while 0.08 worse in the 3-fold cross validation, miss-classifications in our classifier are in dishes and teeth (this could be because of the similar water environments);
- washing teeth: similar, but our model scored higher by about 0.03.

In addition, we contrasted the best performing model with a k -NN classifier described in subsection 3.1 trained on the expanded dataset, by looking at table 2 k -NN is outperformed by both the baseline model of Riboni and by the classifier implemented with YAMNet. Figure 6 shows the confusion matrix of the 10 cross forld validation of the k -NN model.

The implementation of the proposed classifier along with the presented experiments is available at <https://github.com/LIMUNIMI/HAR-YAMNet> ensuring full reproducibility of the obtained results.

5 ANDROID APPLICATION

For experimental purposes, we have also built an Android application which uses the proposed model to classify real-life sounds. The application runs in real-time and performs a new inference every 500 milliseconds on the previous 500 ms of recorded audio, visualizing it on the screen. An example of the application running on Android mobile operating system is shown in Figure 7.

To improve the accuracy of the app, two filters were implemented. The first filter leverages the original YAMNet predictor – not the one we trained – to filter-out silence: if the original YAMNet predicts silence with a score >0.89 , no prediction is performed

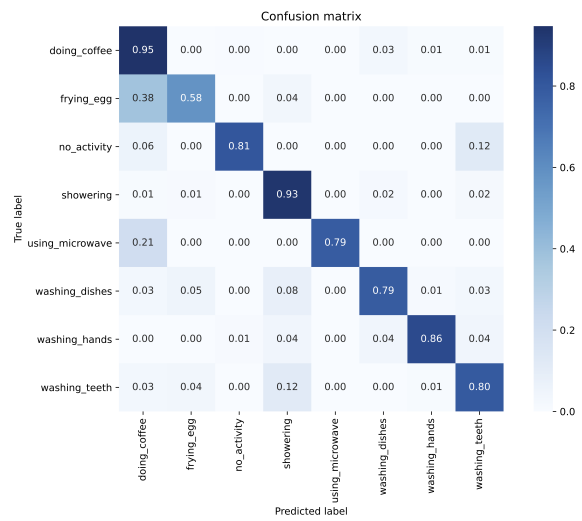


Figure 5: Confusion Matrix of best model obtained with the baseline model (Riboni et al., 2016).

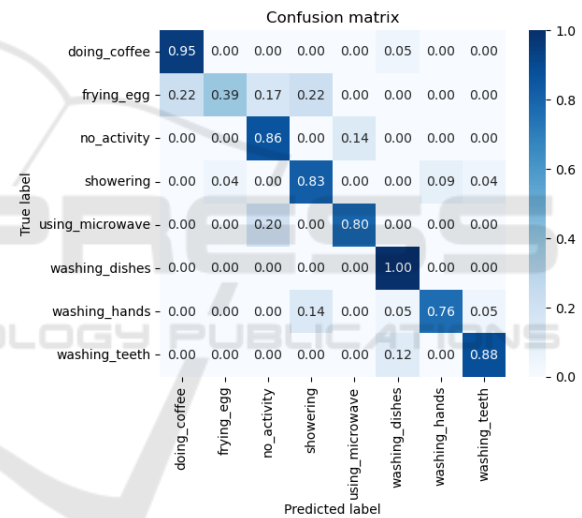


Figure 6: Confusion matrix of best k -NN configuration where $k=3$.

by our model and a corresponding message is visualized on the screen (“No activity from YAMNet”).

The second filter is a threshold of 0.3 for each class: it filters classification for low probability scores, so that if a classification is lower than the threshold its classification is disregarded and no tag is visualized on the screen.

6 CONCLUSION AND FUTURE DEVELOPMENTS

This article proposes a Deep Neural Network framework with transfer learning from a CNN (YAMNet)

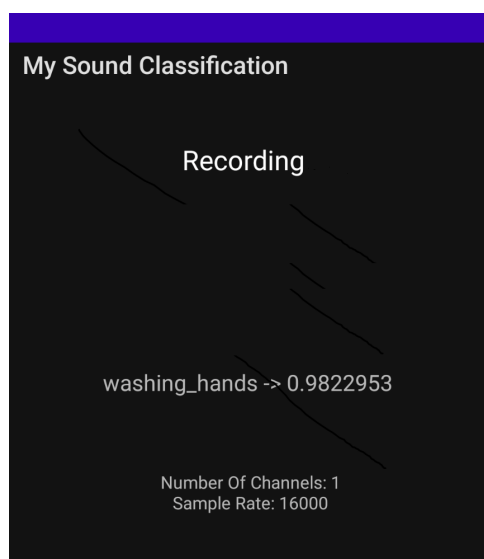


Figure 7: Screenshot of the developed prototype application running on Android mobile operating system. The winning class along and the associated probability is displayed to the user.

to classify human activities using a reasonably-sized dataset. The obtained results demonstrate the superiority of the proposed system over the state-of-art based on supervised feature learning. Weaknesses of the model could emerge in case of scaling the number of classes with proportional number of instances; indeed, the model would need more data to learn a more complex problem for which the current neural architecture may not be enough accurate.

Future works include the use of artificial data augmentation to enlarge the dataset. Possibly YAMNet hyper-parameters could be fine-tuned if the dataset is sufficiently large. Moreover, the effectiveness of the smartphone application should be assessed thoroughly in terms of complexity along with the required resources. Finally, the developed application could be employed to enhance the capabilities of a wide range of systems including smart-home assistants, such as Amazon Alexa, Google Home, etc.

REFERENCES

- Beddiar, D. R., Nini, B., Sabokrou, M., and Hadid, A. (2020). Vision-based human activity recognition: A survey. *Multimedia Tools and Applications*, 79(41):30509–30555.
- Bevilacqua, A., MacDonald, K., Rangarej, A., Widjaya, V., Caulfield, B., and Kechadi, T. (2018). Human activity recognition with convolutional neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 541–552. Springer.
- Chahuara, P., Fleury, A., Portet, F., and Vacher, M. (2016). On-line human activity recognition from audio and home automation sensors: Comparison of sequential and non-sequential models in realistic smart homes 1. *Journal of Ambient Intelligence and Smart Environments*, 8(4):399–422.
- Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, 54(4):77:1–77:40.
- García-Hernández, A., Galván-Tejada, C., Galván-Tejada, J., Celaya-Padilla, J., Gamboa-Rosales, H., Velasco-Elizondo, P., and Cárdenas-Vargas, R. (2017). A similarity analysis of audio signal to develop a human activity recognition using similarity networks. *Sensors*, 17(11):2688.
- Giannakopoulos, T. and Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB Approach*. Academic Press, Inc., USA, 1st edition.
- Hota, S. and Pathak, S. (2018). KNN classifier based approach for multi-class sentiment analysis of twitter data. *International Journal of Engineering and Technology*, 7(3):1372.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.
- Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowanko, D., Ritter, M., and Eibl, M. (2017). Large-scale bird sound classification using convolutional neural networks. In *CLEF (working notes)*, volume 1866.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.
- Lau, J., Zimmerman, B., and Schaub, F. (2018). Alexa, are you listening? *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–31.
- Liang, D., Song, W., and Thomaz, E. (2020). Characterizing the effect of audio degradation on privacy perception and inference performance in audio-based human activity recognition. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '20*, pages 1–10, New York, NY, USA. Association for Computing Machinery.
- Ntalampiras, S. and Potamitis, I. (2018). Transfer learning for improved audio-based human activity recognition. *Biosensors*, 8(3):60.
- Ntalampiras, S., Potamitis, I., and Fakotakis, N. (2012). Acoustic detection of human activities in natural environments. *AES: Journal of the Audio Engineering Society*, 60(9):686–695.
- Ntalampiras, S. and Roveri, M. (2016). An incremental learning mechanism for human activity recognition. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

- Ramanujam, E., Perumal, T., and Padmavathi, S. (2021). Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*, 21(12):13029–13040.
- Riboni, D., Galván-Tejada, C. E., Galván-Tejada, J. I., Celaya-Padilla, J., Delgado-Contreras, J. R., Magallanes-Quintanar, R., Martínez-Fierro, M. L., Garza-Veloz, I., López-Hernández, Y., and Gamboa-Rosales, H. (2016). An analysis of audio features to develop a human activity recognition model using genetic algorithms, random forests, and neural networks. *Mobile Information Systems*, 2016:1784101.
- Ronao, C. A. and Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244.
- Stork, J. A., Spinello, L., Silva, J., and Arras, K. O. (2012). Audio-based human activity recognition using non-markovian ensemble voting. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 509–514. IEEE.
- Wan, S., Qi, L., Xu, X., Tong, C., and Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, 25(2):743–755.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

