Recognizing Activities of Daily Living Using Audio and IMU Data from Commodity Smartwatches

Yvette Espinoza
Department of ECE
Purdue University
Los Angeles, CA
yespinoz@purdue.edu

Cameron Johnson
Department of ECE
Purdue University
Asheville, NC
john3096@purdue.edu

Jahangir Mollah
Department of ECE
Purdue University
Seattle, WA
jmollah@purdue.edu

***Abstract*—Human activity recognition is applicable to many fields, such as healthcare, but there are limitations on devices that can be used. The popularity of smart devices allow for comfort while also providing the data necessary to enable activity recognition. Smart watches prove especially practical for recording both acoustic and inertial data because the mic is at close proximity to the hand which is often close to the activity being performed. The IMU records useful motion on the wrist as opposed to a smartphone in the pocket or strapped to the chest. This project builds on previous activity recognition research and focuses on the effects of acoustic and motion noise on recognizing the activities of daily living.**

I. Introduction

Early research in activity detection aimed at finding differences in signal characteristics of the collected accelerometer data, and determining the ideal sensor placement. In [3], activities were grouped together into four categories by levels of physical activity, and sensors were placed throughout the body to determine the best location for each of the categories. Their results found each category performed best with a different sensor location. A sensor placed on the waist was best at detecting low level activities like eating, while a sensor on the chest or wrist was best at medium level activities like housework.

While different sensor placement based on an activity would be ideal for performance, it would not be practical for widespread use. An accelerometer by itself is not enough to determine an activity. More context is needed and different sensors on the body are an inconvenience for users. Human activity recognition is widely used in healthcare applications, with the elderly being a large part of the user demographics [6], to suit their needs the data collection would require an unobtrusive setup. To address the inconvenience of full body sensors and the need for context, [10] used a wrist-worn device equipped with an accelerometer and camera to recognize daily activities. The camera provided context for the activities, and the accelerometer provided characteristics of the body movement associated with different activities, both of which were used to train a model to predict the activity.

The advancement in smart devices, like smartwatches and conversational assistants, allow for more sensors in a user friendly device. Conversational assistants, like a Google Home, and smartwatches are used to train a model to recognize activities of daily living [1] [4].

A. Applications

Audio monitoring can also be used to detect hazardous situations such as loud noises, falls, or emergencies. By integrating audio analysis into a smartwatch, users can be alerted to potential dangers in real-time, providing an added layer of safety.

Human activity recognition has potential applications that can be leveraged in healthcare, dietary monitoring, and weight management. One application is to provide a user with a more holistic view of their lifestyle, providing details on how long they spend on activities such as cooking or browsing on a phone. This information can be used to reflect and better manage time. Another important application is for monitoring the progress of debilitating medical conditions, such as cognitive impairment.

We propose expanding on research recognizing activities of daily living using the IMU and audio data from a commodity smartwatch. We will be following existing research performed in [4], but adding extra noise while collecting data to better understand the effects of such noise in recognizing activities.

B. Motivation

The researchers of the original paper conducted two generally different experiments. In the first experiment, referred to as the semi-naturalistic experiment, an individual performs a series of activities for 30 seconds wearing a Fossil smartwatch which records inertial and acoustic data. There is a clear distinction between each activity and the watches are programmed to listen at frequent intervals. The authors implemented a variety of machine learning models to recognize the activity in this experiment with an accuracy of 94.3 %. The second experiment, referred to as the in-the-wild experiment, implements a less controlled, more realistic setting. One limitation of this paper was that their uncontrolled experiment had poor results due to background noise, mislabeling of activities due to the setup, and inconsistencies in how activities were performed. Since their in-the-wild experiment was not controlled, not all the activities they were interested in were

performed, in some cases activities overlapped, and combined with a limited sample size resulted in misclassifications [13].

The goal for this project is to expand on the previous work and better understand some of the limitations the researchers encountered. Since the main limitation was the performance degradation for their in-the-wild study, this project aims to have a semi-controlled environment, or semi-wild experiment, where activities are pre-defined but the noise from the in-the-wild study is added. The following are the areas of interest: measure the effect of wearing the watch on a passive instead of dominant hand, measure the effect of background noise while performing activities, and determine if more participants would lead to more accurate results.

II. Related Work

A. Sound Classification

A recent survey of hard of hearing people found that smartwatches, instead of smartphones, are the preferred device for sound awareness due to their social acceptability and support for both visual and haptic feedback [7]. Due to this demand, smartwatches are becoming more popular in sound classification research, but there are still limitations on performance due to acoustic variations, or background noise.

One study measured the effect of background noise on speech detection and found that a noisy environment caused performance to drop by about 20 % [9]. Their proposed solution to the performance degradation was to use new features, which did improve performance.

Other studies suggested having a user-centred data collection, where the user would record the same activity of interest in different scenarios, which would then be used to train a model to detect that activity [5] [7].

B. Motion Classification

There was also a study done where work proposes several techniques to improve the robustness of a Human Activity Recognition (HAR) system that uses accelerometer signals from different smartwatches and smartphones. When using smartwatches to recognize whole body activities, the arm movements introduce additional variability giving rise to a significant degradation in HAR[12]

Other research was done using Samsung Gear Smartwatch to collect data, then extract features, classify with H-SVM (Hierarchical Support Vector Machine) classifier and identify human activities classification. Experiment results show great effect at low sampling rate, such as 10 and 5 Hz, which will gives the energy saving. In most cases, the accuracy of activity recognition experiments was above 99 percent[14]

III. Machine Learning models

Classical machine learning models such as Random Forest, Naive Bayes, and AdaBoost have been used in previous studies and prove to be effective in human activity recognition, but they rely on heuristic handcrafted feature extraction. These models showed decent results, but accuracy was only in the range of 60 to 85 for classifying motion data and 40 to 70 when classifying audio data. In order to achieve more reliable classification, the authors of [4] experimented with two different deep neural network models and tried different fusion methods of the acoustic and motion data. These deep human activity recognition models allow input of raw sensory data captured by the smart watch instead of relying on expensive, handcrafted feature extraction and showed significant improvement to accuracy. The motion and acoustic data are trained separately using different models and then fused using a variety of techniques.

A. Inertial Model

Two different models were experimented with to classify the inertial data. One of the models used is a variation of a framework called DeepConvLSTM which is based on convolutional and LSTM recurrent units. The modification allows the model to extract accelerometer and gyroscope features separately and then concatenate each modality at the last layer before passing through the softmax classifier. This modification improved accuracy, but the other model which was used, called Attend and Discriminate, proves to be more effective. This framework incorporates cross-channel self-attention and temporal attention. Both gyroscope and accelerometer data are concatenated before processing through convolutional layers to learn the interaction between the two. The resulting feature maps pass through a recurrent neural network to model temporal dynamics.

B. Acoustic Model

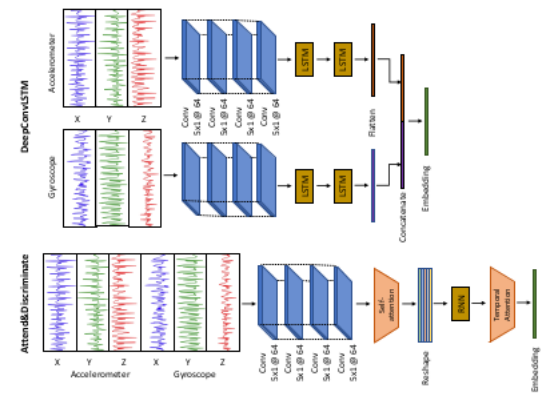
In order to classify the acoustic data, the audio clips required some pre-processing. Log-mel spectrograms are extracted from the clips by computing the short-time Fourier transform for each segment, using a Hanning window of 1024 samples and hop size of 320 samples, and then converted into a 64-bin log-scalled Mel spectrogram. Convolutional neural networks are applied to these log-mel spectrograms. The authors find a CNN architecture they refer to as CNN14 comprised of six convolutional blocks consisting of two convolutional layers and intermediary average pooling layers.

C. Fusion Methods

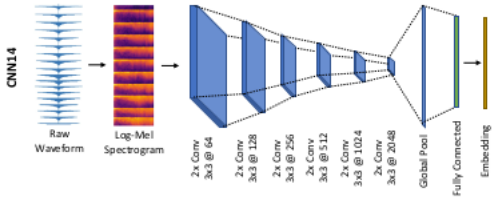
The inertial and acoustic data then need to be incorporated together. The model applies an aggregation method which concatenates the feature maps. There were other methods such as score-level fusion, but they did not perform as well. The concatenated feature maps are followed by a single classification head and joint training of both acoustic and inertial networks. The predicted class probabilities are averaged for a final probability value.

IV. Methodology

For the data collection, only a subset of the original activities was performed: writing, typing on a keyboard, brushing teeth, sweeping, and washing dishes. These activities were selected since they primarily rely on one measurement for detection, i.e. washing dishes is more easily identified by its



(a) Inertial Models.



(b) Acoustic Models.

Fig. 1. Architecture of Inertial and Acoustic Models, provided by [4]

acoustic data while writing is more identifiable by its motion data. A smartwatch was worn to collect the accelerometer and gyroscope data [2], while a phone was set nearby to collect the acoustic data. The motion and acoustic data were the inputs to a deep learning model which fuses the inertial data and acoustic data to determine the activity being performed, shown in Figure 1.

A. Hardware Setup

In the original paper the authors used the same watch, the Fossil Gen 4, for all of their experiments [4]. Their custom application, which collected motion and acoustic data, was developed on Android Wear OS 2.11, with inertial sampling at 50 Hz and acoustic sampling at 22.05 kHz.

For this project, two separate Samsung Galaxy watches were used with a custom app developed on Tizen OS 5.5. The inertial data was also sampled at 50 Hz, but since the smartphone recorded at 44.1 kHz, it was downsampled in post processing to match the rate of the original paper.

B. Watch App

The original intention was to collect both motion and acoustic data from the smartwatch, but while developing the app there were problems with writing to the watch's file system. The main problem seemed to be with the permissions configured on the native c app. Based on the available documentation, those problems could be addressed by switching to a web app instead, but due to lack of experience with web apps and time constraints, instead of storing the collected data to a file on the watch the inertial data was collected from the debug log and a smartphone was used to collect the acoustic data.

The final app only had one button to start and stop the motion data collection, shown in Figure 2.

Fig. 2. Watch app deployed on emulator

C. Schedule

The project was divided into four phases. More details on due dates and deliverables for each phase are provided in Table I.

V. Data Collection

The data collection was performed using a debug version of the watch app, which allowed us to access the debug log with the accelerometer and gyroscope data. Participants had two data collection sessions. For the first session the watch was worn on the dominant hand and there was no background noise, and for the second session either the watch was worn on the passive hand or background music was played while performing the activities. A phone was placed next to the participant during data collection to capture the activities acoustic data. Figure 3 shows the instructions for how to collect data. Figure 4 shows the Tizen Device Manager from which the debug log was exported.

A. Data Post-processing

To align the acoustic and inertial data, the first and last 5 seconds of the audio file were removed. The next step was to downsample the recordings, since the smartphones sampled at 44.1 kHz while the original paper sampled at 22.05 kHz. For the inertial data, the debug logs were parsed and only timestamps that collected both gyroscope and accelerometer data were used to generate the final csv.

TABLE I
SCHEDULE

Phase	Activities	Due Date
Planning	• Identify existing research data • Identify available hardware • Plan for how to capture and store data • Identify how to process data • Identify machine learning classification models	March 18
Data Collection	• Capture and process raw data • Capture audio data at all times • Review existing research data • Get volunteers for experiments	April 1
Analysis	• Fuse data collected from different sources • Label data collected • Run the classification models • Analyze results	April 15
Finalize Report	• Write up final report • Create presentation	May 6

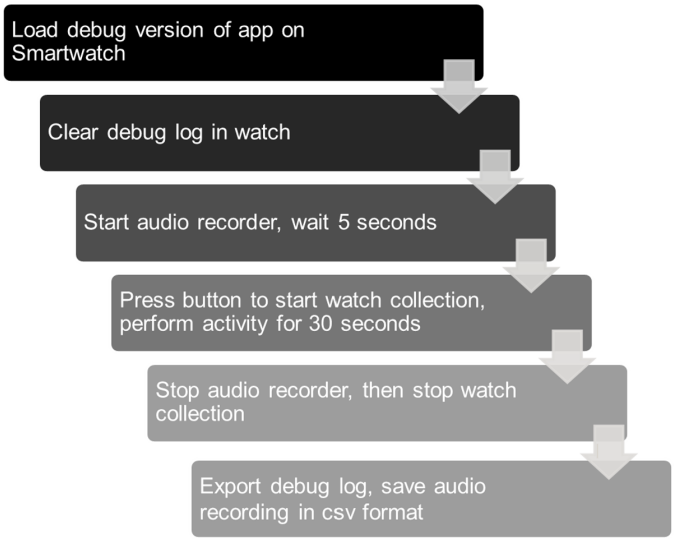


Fig. 3. Data collection steps provided to the participants

The final step was to rename the files and add them to the same file structure as used by the original researchers. Since only a subset of the original activities were performed, audio and motion data from other participants were used to create a full activity set that would be ready to test the models pre-trained on the available datasets.

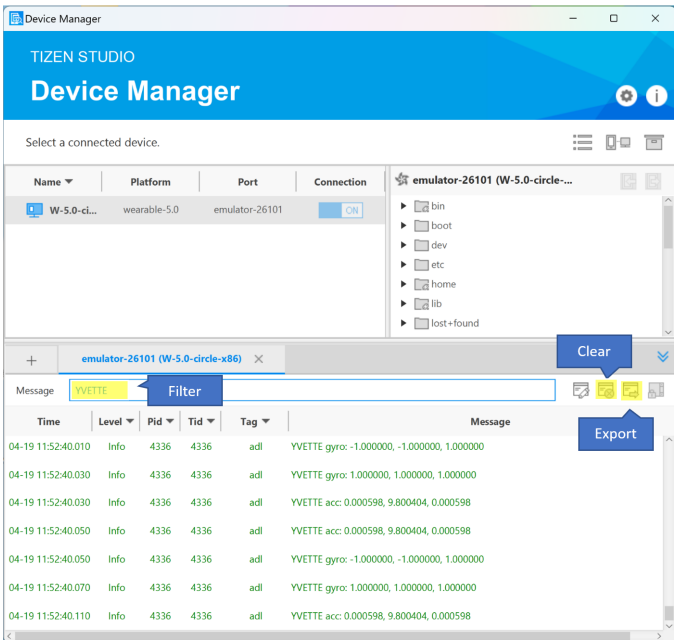


Fig. 4. Example of how to clear and export the debug log

VI. RESULTS

A. Running the original models

Although the machine learning models are publicly available, they required significant modification. The reason for these changes are mostly due to using different equipment. In order to increase our compute power, we tailored the model to use Google Colab which can run independently from our machines and employs more powerful hardware. In doing so, the pre-processing and training methods were still crashing due to insufficient memory. Specifically, the program stacked the audio and inertial data for all participants into arrays using a numpy method which required keeping the entire array in memory. We reconfigured the code to use numpy views to combine all of the data which proved much more efficient and ultimately cut the memory requirement in half. Other modifications to the model included changing plotting methods and various function parameters most likely due to the authors employing outdated libraries.

B. Number of participants

We wanted to verify whether adding more training data was likely to improve the accuracy of the model. To do this, we ran the model using different numbers of participants and recorded the final test accuracy to evaluate whether the improvement starts to decrease. Figure 5 shows the results of this experiment after training for 100 epochs. As you can see, training on just two participants performs poorly with an accuracy of around 40 percent. If you run the same experiment using more participant data, the accuracy improves significantly. Importantly, the upward trend shows no sign of leveling. The plot suggests that including more training data will further improve the model accuracy. Figure 6 shows the confusion

matrix for the original model trained on all the available semi-naturalistic data. The model performs well, but there is still room for improvement.

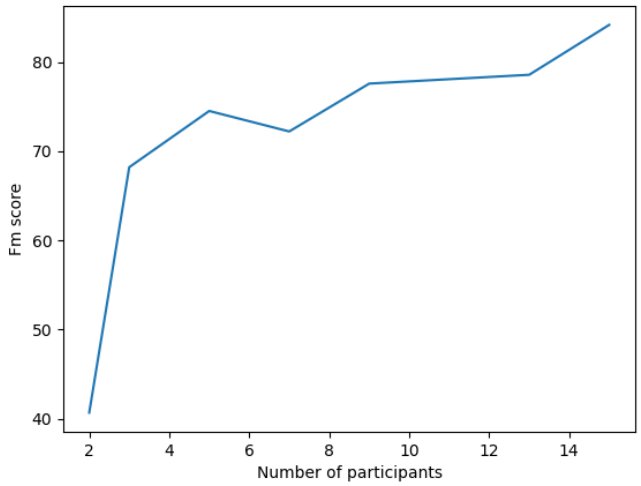


Fig. 5. F-score for training on increasing amount of participants

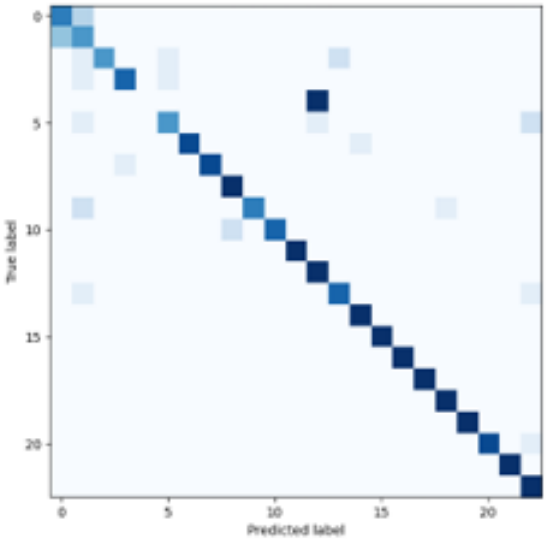


Fig. 6. Confusion matrix for model trained on Semi-Naturalistic dataset

C. Inference on trial data

To better define the data collection process a trial run was conducted where only three activities were performed: writing, typing and wiping the table. Each data collection lasted 15 seconds, and motion data was sampled at 10 Hz instead of 50 Hz as in the original paper. After post-procesing, the data was fed into a model that was pre-trained with the original dataset, the model output is shown in the Figure 7 confusion matrix.

On the confusion matrix the left side indicates the input activity, and the bottom shows the predicted activity, with the

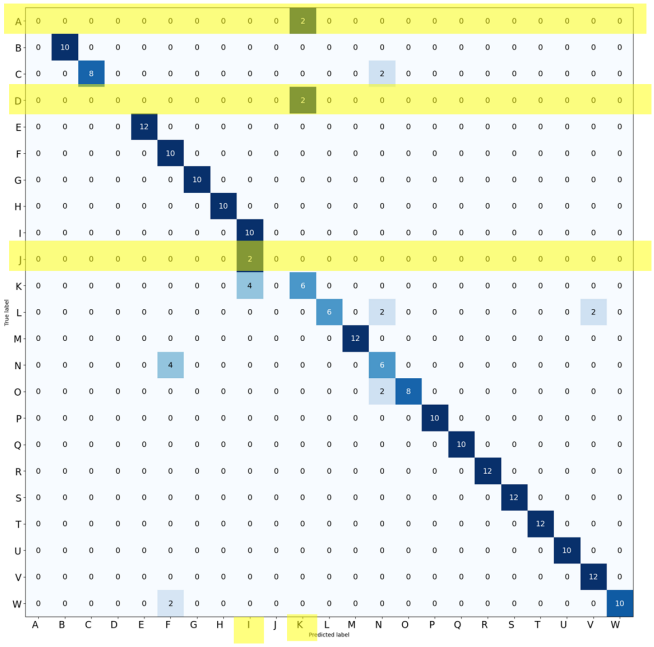


Fig. 7. Confusion matrix for trial data run from model trained with semi-naturalistic dataset

three activities performed highlighted: writing (A), typing (D), wiping table (J). The activities were all labeled incorrectly, writing and typing were mistaken for brushing hair (K) and wiping table was mistaken for scratching (I). After reviewing the results, the collection time was increased to 30 seconds, and the motion sampling rate was changed to 50 Hz. Due to time constraints the updated collection procedure was not able to be tested.

VII. Challenges

There were three main challenges encountered while working on this project.

A. Activity Selection

The first challenge was selecting the reduced activity list to analyze for the project. The original motivation behind the project was to apply the models and focus on detecting digital media usage, which is one of the applications for activity detection. Previous research showed the controlled environment results were accurate in detecting the different activities, but we found our semi-wild results were similar to the uncontrolled results, where differentiating between the different phone activities, typing on the phone and browsing on the phone, was difficult. The focus then became measuring the effect of noise that was causing the degraded performance between the controlled and uncontrolled experiments. Our motivation focused on testing with Motion noise and acoustic noise and evaluating the performance of the model.

B. Data Collection

The app development was a major challenge, mostly because none of the team members had any app development

experience. As previously mentioned, the original plan was for the watch to collect and locally store both acoustic and inertial data, and the files would later be moved from the watch to a computer. The older Samsung watches that were used had Tizen OS, and because it is not as popular as WearOS, there was a lack useful documentation while attempting to troubleshoot. Tizen OS has been replaced by WearOS in the newer Samsung watches, so future research shouldn't have these same problems.

C. Running the original models

The original models had to be modified before they could be used. Due to their complexity and memory requirements, only one member was able to get them to fully run. This was only after tailoring the model to run on Google Colab because none of us had the necessary hardware, specifically the memory and GPU processing. Even with upgrading to high-ram and GPU processing through Colab, the model required over five hours to train using 15 participants. Fortunately, we were able to save the weights after each training session which allowed the other members to continue work, but there was that added delay before more analysis could be conducted.

VIII. Conclusion

We were able to incorporate the data from the 15 participants whom the authors used to gather data performing various activities around the house. Using the Attend and Discriminate-CNN14 model and concatenating the acoustic and inertial data, we achieved fairly reliable results classifying human activities. Our preliminary results indicate that including more participant data will continue to improve accuracy. Given more time, we would have liked to experiment with more data and recording data in various noisy environments with factors such as dog barking or music in the background.

Acknowledgment

We re-implemented the following paper using their available models and datasets: [4].

References

[1] Rebecca Adaimi, Howard Yong, and Edison Thomaz. "Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5.1 (Mar. 2021). DOI: 10.1145/3448090. URL: https://doi.org/10.1145/3448090.

[2] Rosa Andrie, Indrazno Siradjuddin, and Nofrian Hendrawan. "Improving Basketball Recognition Accuracy in Samsung Gear S3 Smartwatch using Three Combination Sensors". In: (Sept. 2020), pp. 386–390. DOI: 10.1109/ICOVET50258.2020.9230342.

[3] Louis Atallah et al. "Sensor Positioning for Activity Recognition Using Wearable Accelerometers". In: *IEEE Transactions on Biomedical Circuits and Systems* 5.4 (2011), pp. 320–329. DOI: 10.1109/TBCAS.2011.2160540.

[4] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. "Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches". In: 6.2 (July 2022). DOI: 10.1145/3534582. URL: https://doi.org/10.1145/3534582.

[5] Danielle Bragg, Nicholas Huynh, and Richard E. Ladner. "A Personalizable Mobile Sound Detector App Design for Deaf and Hard-of-Hearing Users". In: *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '16. Reno, Nevada, USA: Association for Computing Machinery, 2016, pp. 3–13. ISBN: 9781450341240. DOI: 10.1145/2982142.2982171. URL: https://doi.org/10.1145/2982142.2982171.

[6] Yi Chen et al. "Robust Activity Recognition for Aging Society". In: *IEEE Journal of Biomedical and Health Informatics* 22.6 (2018), pp. 1754–1764. DOI: 10.1109/JBHI.2018.2819182.

[7] Dhruv Jain et al. "SoundWatch: Exploring Smartwatch-Based Deep Learning Approaches to Support Sound Awareness for Deaf and Hard of Hearing Users". In: ASSETS '20. Virtual Event, Greece: Association for Computing Machinery, 2020. ISBN: 9781450371032. DOI: 10.1145/3373625.3416991. URL: https://doi.org/10.1145/3373625.3416991.

[8] Haik Kalantarian and Majid Sarrafzadeh. "Audio-based detection and evaluation of eating behavior using the smartwatch platform". In: *Computers in biology and medicine* 65 (Oct. 2015), pp. 1–9. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2015.07.013. URL: https://doi.org/10.1016/j.compbiomed.2015.07.013.

[9] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. "Content analysis for audio classification and segmentation". In: *IEEE Transactions on Speech and Audio Processing* 10.7 (2002), pp. 504–516. DOI: 10.1109/TSA.2002.804546.

[10] Takuya Maekawa et al. "WristSense: Wrist-worn sensor device with camera for daily activity recognition". In: *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. 2012, pp. 510–512. DOI: 10.1109/PerComW.2012.6197551.

[11] Adria Mallol-Ragolta et al. "harAGE: A Novel Multimodal Smartwatch-based Dataset for Human Activity Recognition". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. 2021, pp. 01–07. DOI: 10.1109/FG52635.2021.9666947.

[12] Rubén San-Segundo et al. "Robust Human Activity Recognition using smartwatches and smartphones". In: *Engineering Applications of Artificial Intelligence* 72 (2018), pp. 190–202. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2018.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0952197618300812.

[13] Sergey Smetanin and Mikhail Komarov. "Misclassification Bias in Computational Social Science: A Simula-

tion Approach for Assessing the Impact of Classification Errors on Social Indicators Research". In: *IEEE Access* 10 (2022), pp. 18886–18898. DOI: 10.1109/ACCESS.2022.3149897.

[14] Tao Tang et al. "Human Activity Recognition with Smart Watch Based on H-SVM". In: *Frontier Computing*. Ed. by Neil Y. Yen and Jason C Hung. Singapore: Springer Singapore, 2018, pp. 179–186. ISBN: 978-981-10-3187-8.

[15] Shibo Zhang et al. "Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances". In: *Sensors* 22.4 (2022). ISSN: 1424-8220. DOI: 10.3390/s22041476. URL: https://www.mdpi.com/1424-8220/22/4/1476.