Activity Recognition in Nature using Sound and AI

Ruben Hessels
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
r.a.hessels@student.utwente.nl

ABSTRACT

The coronavirus pandemic has forced the Dutch government to put indoor activities to a hold. Therefore, many people seek alternatives outdoors. Especially, natural environments such as parks have become a popular place to visit. This left municipalities, park owners and nature conservers to wonder whether they could get concrete insight in the actual usage of the area. Based on this yet unknown data, parties responsible for the park's management, would be able to optimise how an area is being utilised. This paper evaluates a sound-based approach for human activity recognition in nature by using a convolutional neural network (CNN) with spectral image input. The current model is able to recognise four activities with 85% accuracy: *walking*, *running*, *cycling* and *null* (environmental noise only). Three phenomena that could affect the perceived acoustic data were investigated: sound attenuation, overlapping sounds and noise interference.

Keywords

Human Activity Recognition, Environmental Sound Classification, Convolutional Neural Network, Spectrogram

1. INTRODUCTION

Since the emergence of COVID-19, most people seek for alternative ways to spend time. Because of the government imposed regulations, these can only be performed individually or in a small group outdoors. Consequently, unorganised activities, which naturally enforce inter-person distance and maximum group size, have rapidly become more popular[19, 28, 11, 30]. Moreover, a recent survey[29], conducted by KRO-NCRV among 158 foresters across the Netherlands, suggests an increased amount of visitors in natural environments. Half of the surveyees even think that the areas reach their limits.

Although local organisations are aware of the change in popularity, there is no concrete data on the specific usage of outdoor environments. Municipalities, as well as park owners and nature conservers, are responsible for managing their natural environments and thus came to wonder whether their areas were optimally managed and utilised. By having an overview of the specific usage of parks, statistics derived from the measured data can be taken into account when decisions have to be made that affect the park. Statistics such as the most popular walking route or mountain bike trail, could be of great value to the responsible parties.

Various technologies including (infrared) camera systems[4] and wearables[33, 12] have already shown different approaches for recognising human activities with artificial intelligence (AI). Although many methods achieve high accuracies, most of them suffer from limitations concerning cost and privacy[20, 31]. To overcome these limitations, the acoustic method seems promising. It does not necessarily need to be positioned in line-of-sight, allowing it to be pervasive. Additionally, it is relatively low-cost[31] compared to other techniques. Its potential lies on aspects such as non-intrusiveness, unobtrusiveness and cost efficiency[31].

This research focuses on the performance of a monophonic (one activity recognised at a time) sound-based human activity classifier in nature with the use of deep-learning. Figure 1 shows a sketch of how such system would possibly be deployed in the future. The main research question is as follows:

How can a sound-based human activity classifier be used to monitor the usage of outdoor environments?

The challenge for the acoustic method will lie on aspects regarding the quality of the perceived data. The lower the quality of sound, the harder it will be to recognise the activity. Phenomena that possibly affect the quality of sound include sound attenuation, overlapping sounds and noise interference. In order to answer the main research question, the following three sub-questions can be formulated:

1. **Sound attenuation:** How do different distances between the system and the sound source affect the classification accuracy?
2. **Overlapping sounds:** How is the accuracy of a monophonic classifier affected by overlapping sounds from the same activity?
3. **Noise interference:** How does environmental noise affect the classification accuracy of the system?

First, this paper begins with the discussion of other literature related environmental sound classification (Section 2). Section 3 is about the methodology and starts off with the prototype setup. In contrast to the possible future deployment as illustrated in Figure 1, this prototype is solely used for data collection, because the evaluation of the AI's performance is not required to be done real-time on the prototype itself. The collected data is used to train

35th Twente Student Conference on IT July. 2nd, 2021, Enschede, The Netherlands.

a convolutional neural network (CNN) and to compose the test sets. The evaluation of the classifier is done with the prepared test sets from which the results should answer the three aforementioned research sub-questions. Details about the measurement environment are offered in Section 4. The results of the test sets are presented in Section 5. Section 6 provides possible explanations for the results. Finally, the findings of this research are summarised in Section 7.

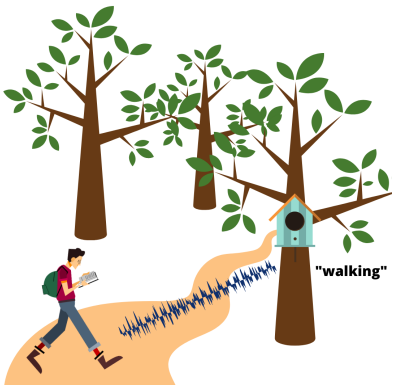

Figure 1. The system is embedded in a birdhouse unobtrusively observing sounds and recognising human activity

2. RELATED WORK

Wang et al.[31] performed a relevant comparative study surrounding environmental human activity recognition. Despite the fact that this study focuses on indoor environments rather than outdoor environments, there is still a lot of valuable information to acquire. As the study states, there are numerous technologies already existing that are able to automatically detect human presences and activities. For example, passive infrared sensors which are very cheap, but only detect the presence of stationary people. Also, the use of camera systems is mentioned. Despite these systems being very accurate, they make people feel spied upon[20]. This might even be more the case when it is used for actual surveillance in order to detect suspicious situations[6, 5].

Mushtaq et al.[15] and Piczak[17] used the CNN-spectrogram approach for environmental sound classification. Their methodology is very similar to the one in this paper. The main difference is that the datasets they used consist of various other types of environmental sounds relating to nature, animals and urban noises. It is not necessarily focused on identifying the activity a human is performing.

Another research presents an approach to accurately identify human activity based on the generated sounds of footsteps[32]. Even though this is very related, it only focuses on bipedal activities such as walking and running. This excludes the possibility of recognising cyclists.

Lastly, other studies show that human activity can be recognised using sensor-based wearables[12]. Especially the study making use of environmental background sounds for this purpose seems promising[33]. However, this research investigates the classification process while wearing the system. In contrast, the prototype will be installed at a fixed location. It also mainly focuses on power-consumption. This influenced the decision of the chosen classifier model.

3. METHODOLOGY

This section will elaborate upon the methods employed needed to conduct this research.

3.1 Setup

Ideally, recording sounds outdoors should be done in the same way as it would for future deployment. Therefore, a prototype was assembled mimicking the same conditions. Figure 2 shows the prototype inside a casing. The casing functions as a water-proof protective outer layer that prevents the system from being damaged by the environment. This will be especially useful when the system is deployed in nature for a longer period of time.

Figure 2. A Raspberry Pi 4 connected to a powerbank

An overview of the hardware used can be found in Table 1. The ReSpeaker 2-Mics Pi HAT is a dual-microphone expansion board for the Raspberry Pi. It is designed for AI and voice applications, which makes it very suitable for this research.

Table 1. Hardware overview

Hardware	Task
Raspberry Pi 4	Controlling microphones and storing recordings
ReSpeaker 2-Mics Pi HAT	Recording audio (2 channels)
Powerbank	Storing power
Micro USB cable	Transferring power

The microphones are located on top of the Pi HAT expansion board. In practice, the prototype is rotated in such a way that the microphones face downwards. Figure 3 reveals how the microphones are able to access the outside of the casing by two small holes. Two straws are inserted into these and perfectly fit around the small microphone units on top of the expansion board. In this way, the hardware is safe inside, while sounds can still be observed from the outside.

Figure 3. Straws sticking out of the bottom of the prototype

3.2 Data Collection

Datasets created with the purpose of environmental sound classification (ESC) exist and are used for experiments surrounding ESC. Examples of these include the UrbanSound8k[21] and the ESC-50[18] dataset. However, both do not fully cover the activity classes this research requires (see the *activity* column in Table 3). In contrast, the AudioSet[10] ontology contains 527 classes including the classes needed for this research. Although this seems perfect for this research, it should be noted that the AudioSet dataset is acquired from manually labeled YouTube video fragments. This means that the dataset has been generated with audio fragments from other environments, such as gameplay videos and indoor environments, as well. Therefore, it became evident that an own dataset had to be created, satisfying the conditions this research demands.

The data must be collected in the same way as the system will be eventually deployed. Meaning that the same microphones must be used and that the system must be positioned and oriented similarly while embedded in the casing. Using the same microphones as recorded with, is believed to improve accuracy as compared to using different microphones. The recording environment has been kept consistent during the data collection. Section 4 elaborates on this further. An overview of the audio output format can be seen in Table 2.

Table 2. Output format of the recordings

File type	WAVE
Format	16 bit little endian
Sampling rate	44100 Hz
Channels	2

Furthermore, all of the data has been collected in a clustered way. For example, the *walking* class has been recorded for a couple of minutes straight, recording multiple passages in one go. Figure 4 illustrates what this looked like in practice. The advantage of this is that all of the desired fragments in one recording are labelled naturally. The *original* column in Table 3 provides an overview of the amount of audio fragments that were extracted from the recordings (Section 3.3 will explain how).

Figure 4. Attached with an orange strip to the left tree, the prototype records the *walking* activity.

In order to capture ground truth, a camera was positioned in such a way that it could capture the entire path on which the data collection was conducted. The camera recorded audio as well, which opened the possibility of synchronising the video recording with the audio recordings made on the prototype. This was achieved by simply clapping at the start and end of an audio recording.

The *mountain biking* activity in Table 3 is displayed as a separate class, but is actually merged with the *cycling* class in order to increase the variation of data. As it is not the case that everyone owns the same bicycle, four different bikes used by four different people were involved to make the dataset more general. The same has been done with the other classes by involving four different people to perform all activities.

Table 3. Composition of the dataset in terms of the amount of samples

Activity	Original	Augmented	Total
Walking	47	53	100
Running	36	64	100
Mountain biking	28	29	57
Cycling	21	22	43
Null	72	28	100
Total	**205**	**195**	**400**

Additional recordings have been made that function as test sets. These test sets will be evaluated by the final deep-learning model in order to answer the research subquestions. Section 3.6 will elaborate on the evaluation of these sets.

3.3 Segmentation

Audio recordings of different activities have a different duration in which they are audible. For instance, the recordings of *walking* take longer than *cycling*, because walking the same distance takes longer than cycling it. Because of this difference, all audio fragments have been segmented into a fixed amount of seconds. A segmentation size of 4 seconds was chosen. It was thought that this length would capture the required information that was needed, because it proved to be sufficient for human ears to identify the activity during this short period. The number of recordings in Table 3 represent these 4 second fragments, not the entire event.

Audacity[23] was used to segment the raw recordings into 4 second fragments, while automatically labelling them. As the raw recordings could not always be evenly divided in fragments of 4 seconds, the ones shorter were omitted. It is important to maintain the same audio length for all fragments, because the deep-learning model (Section 3.5) expects a specific type of shape.

3.4 Data Augmentation

The *original* column in Table 3 shows that the initial dataset was quite small and imbalanced. Both of these factors have impact on the reliability of the results, because of a modeling error called 'overfitting'. The consequence of overfitting is that the model will be too biased for the given dataset by failing to predict in the same way for other datasets. Imbalanced datasets typically cause the model to have low error rates for the majority class (the class with the most samples) and a high error rates for the minority class (the class with the least samples)[3]. However, this is mainly true for severe imbalances (minority class $< 1\%$ of the total samples). Slight imbalances, like in the dataset used for this research, are not particularly problematic[9]. Although, this statement only holds

for big datasets. Therefore, 400 samples (Table 3) might not be enough to ignore this imbalance.

For this purpose, a data analysis technique called data augmentation was applied, which reduces[14, 7, 22] overfitting by increasing the size of the dataset, while adding more variation to it, and removes the imbalance by adding a set amount of augmented samples that even the total amount for all classes. Numerous methods can be used to achieve this. For this research, a time shift was applied to the original dataset. Multiple audio fragments were shifted with a random amount of time, while preserving their labels. Table 3 reveals the total effect of this on the dataset. The dataset was doubled (from 205 to 400 samples) and the classes were balanced to 100 samples per class.

3.5 Classification

A CNN is a deep-learning algorithm that is typically used for image classification[27]. Although not immediately obvious, the CNN can also be used for sound classification[13]. Other literature[26, 16, 2, 17, 15] demonstrate that CNNs are able to address audio-related tasks with high accuracies by converting audio clips to spectrogram images.

A spectrogram is an image in which the strength or intensity of different frequencies over time are represented. It can be acquired by applying the short-time Fourier Transform (STFT) on an audio clip. The spectrogram images of this research were fed to the CNN model displayed in Figure 5. The first two layers are preprocessing layers, which downsample the input to enable the model to train faster. After that follow two 2D convolutional layers. Vafeiadis et al.[25] showed that 2D CNNs outperform 1D CNNs with sound-based spectrogram input. All of the layers were provided by the Keras[8] library. Furthermore, a batch size of 32 was used.

```
Input shape: (1377, 129, 1)
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
resizing_4 (Resizing)        (None, 32, 32, 1)         0
_________________________________________________________________
normalization_4 (Normalizati (None, 32, 32, 1)         3
_________________________________________________________________
conv2d_8 (Conv2D)            (None, 30, 30, 32)        320
_________________________________________________________________
conv2d_9 (Conv2D)            (None, 28, 28, 64)        18496
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 14, 14, 64)        0
_________________________________________________________________
dropout_8 (Dropout)          (None, 14, 14, 64)        0
_________________________________________________________________
flatten_4 (Flatten)          (None, 12544)             0
_________________________________________________________________
dense_8 (Dense)              (None, 128)               1605760
_________________________________________________________________
dropout_9 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_9 (Dense)              (None, 4)                 516
=================================================================
Total params: 1,625,095
Trainable params: 1,625,092
Non-trainable params: 3
```

Figure 5. The CNN model

The model shown in Figure 5 originates from Tensorflow's[1] simple audio recognition implementation[24]. Their example focuses on detecting keywords from speech. As their CNN model was designed for audio recognition, it was found to be suitable for this research. Their code functioned as the basis for preprocessing and classifying the data. However, modifications had still to be made to make it work for the self-created dataset.

As the recorded audio consists of two channels (stereo), two spectrograms were created per audio fragment. One for the left microphone and one for the right. Basically, this means that 800 spectrograms were created (2 times the total amount of 400 in Table 3) and used for the training, validation and test set. It was observed from the dataset that both microphones were still able to capture the activity separately, while differing slightly on the waveforms. In total, this could add more variation to the dataset, which potentially reduces overfitting.

Dataset has been split into an 80:10:10 ratio (train, test and validation respectively). Other ratios such as 70:15:15 and 60:20:20 have been experimented with as well. Yet for these, higher accuracies could not be attained. For reproducibility, different seeds for splitting the data were investigated. The results varied enormously depending on the seed. Eventually, the combination of reached accuracies by the test sets and the model's loss and validation loss during the training phase determined the seed continued with.

3.6 Evaluation

Approaching the research sub-questions will be done by evaluating additional test sets with the model. For each research sub-question, two test sets are compared. All test sets contain data that the CNN model is not familiar with, because it was not included in the training and validation set.

3.6.1 Sound Attenuation

The first research sub-question was concerned about the impact of different distances to the microphones on the classification accuracy. This can be examined by creating two additional test sets with different distances. The main dataset in Table 3 consists of audio recordings made from approximately the center of the path. While the additional recordings include distances from the closest and the farthest side of the path (see Figure 6 for the exact path dimensions). The difference in accuracies should illuminate the impact of sound attenuation.

3.6.2 Overlapping Sounds

The main dataset consists of monophonic activity data, which means that there are no overlapping sounds from the same activity. To study the effect, an additional test set that does contain overlapping sounds for all activities is composed. For this purpose, the activities *walking*, *running* and *cycling* were recorded for two people as well. There were two different pairs involved in this. The accuracy returned by the test set can be compared to the overall accuracy of the monophonic test set. In this way, the impact of overlapping sounds will be exposed.

3.6.3 Noise Interference

The impact of environmental noise on the classification accuracy can be studied by comparing two categories. One with severe environmental noise in the background (cars, trucks, wind) and one with almost nothing. The test sets were composed by listening to the recordings in the dataset and dividing them between the categories. The difference in accuracy should inform how well the classifier performs under the presence and absence of severe noise interference. Additionally, it could indicate to which category the model leans the most to.

4. MEASUREMENT ENVIRONMENT

The environment in which the data has been collected could have a great impact on the classification results. The type of surface or the weather conditions will have influence on how sounds will be generated and perceived.

All of the data has been recorded on one day under the same circumstances. It could therefore be very possible that the system's accuracy will decrease under different circumstances. For example, as all of the data has been recorded on a sunny day, a rainy day will yield different data. Consequently, making it harder for the CNN to make a correct prediction. An overview of the metadata stored can be seen in Table 4.

Table 4. Metadata

Date	1st of June 2021
Ground type	Dirt/Gravel
Orientation	Microphones facing downwards
Location	University of Twente campus
GPS coordinates	52.241634, 6.857067
Obtrusions	None
Weather	Sunny, no clouds
Wind	Little to nothing

The system was installed at a height of 3 metres. The reason for this is that during the eventual deployment, it prevents people from touching and possibly sabotaging the system. The tree on which the prototype had been attached is located 1.30 metre from the path. The width of the path is 2.50 metres. A sketch of the dimensions can be seen in Figure 6.

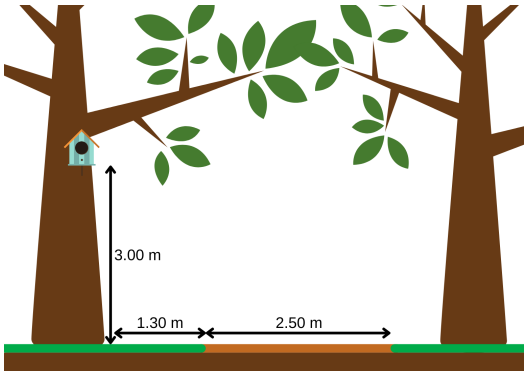

Figure 6. A side view displaying the dimensions of the measurement environment

The path on which the recordings have been made, locates at the University of Twente's campus (see Table 4 for GPS coordinates). Unfortunately, a lot of additional urban noise coming from cars and trucks has been observed in the dataset. This is because of a road block nearby, which caused all the traffic to use the road close to the recording spot. As for the type of surface the data was generated on, it was mostly dirt with a slight gravel texture. This is fortunate, because the additional gravel texture allows someone passing by to be standing out more clearly in the audio recording than for example a path made of concrete.

Very rarely, the recordings were disturbed by wind. The activity was inaudible during these events and was therefore omitted from the dataset. However, when the activity was actually audible during the severe interference of wind, it was included nevertheless. This would increase the variety of data. A continuous element of the background noise was the sound of singing birds. It would be very possible that the model behaves differently without the same continuous background noise. The recordings were taken at the 1st of June with many actively singing birds. This will probably not be the case in December.

5. RESULTS

The results that followed from the various test sets discussed in Section 3.6 will answer the formulated research sub-questions in Section 1. How do sound attenuation, overlapping sounds and environmental noise affect the classification accuracy?

Firstly, the overall test set returned an accuracy of 85%. Generally speaking for all test sets, *walking* and *running* scored the worst, while *cycling* and *null* scored the best. The corresponding confusion matrix of the *overall* test set is provided in Appendix A.1. Table 5 provides an overview of the prediction accuracy per activity for all sets.

Secondly, the *close* and *far* test sets concerned about the difference in distance returned a poor 17% and 47% accuracy respectively. The latter test set performed relatively better with a 30% difference, but the difference is still significant compared to *walking* of the *overall* test set (70%). The *close* test set predicted *running* instead of *walking* 50% of the times. For the *far* test set, the wrong predictions were equally spread over *running* (17%), *cycling* (17%) and *null* (20%). The confusion matrices are displayed in Appendix A.2.

Next, the test set containing overlapping sounds returned a 70% accuracy. Compared to the overall test set containing non-overlapping sounds, this is 15% less. Roughly speaking, the test set returned similar results as the *overall* test set. Except for the *walking* activity, which was misclassified for *running* 67% of the times. The confusion matrix is located in Appendix A.3.

Table 5. Summary of the resulting accuracies

Test set	Null	Walking	Running	Cycling	Average
Overall	100%	70%	80%	90%	**85%**
Close	-	17%	-	-	**17%**
Far	-	47%	-	-	**47%**
Overlapping	80%	27%	73%	100%	**70%**
Normal	87%	47%	53%	93%	**70%**
Noisy	80%	67%	67%	93%	**77%**

Lastly, the test sets with and without severe environmental noise returned a 77% and 70% accuracy respectively. The test set with environmental noise performed 7% better. Especially, the *walking* and *running* activities seemed mostly affected for both sets. For the *normal* test set, this was even worse. Their confusion matrices can be found in Appendix A.4.

The aforementioned accuracies are the result of combined predictions. One prediction comes from the left microphone channel, the other one from the right. The experiments showed that the combined predictions achieved higher accuracies than when judged separately (i.e. classification per audio channel). A difference that reached up to 9% in accuracy was observed.

Figure 7 illustrates the model's loss (blue line) and validation loss (orange line). A callback function was set to stop the training phase once no improvements were detected for 2 epochs (also called patience). As Figure 7

displays, the training stopped at 5 epochs, meaning the function was triggered at that time. Signs of overfitting were observed when the model was trained for more iterations (25 epochs) without the callback function. In this case, the validation loss continued to increase, while the model's loss did the opposite.

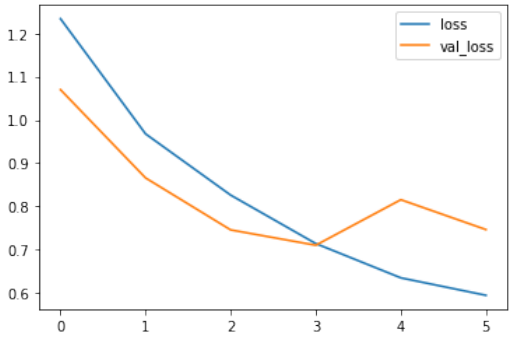

Figure 7. Epochs on the x-axis and loss values on the y-axis

6. DISCUSSION

The overall accuracy of 85% was higher than expected. In earlier stages of experimenting, it became clear that the dataset was too small due to low accuracies. After the data augmentation, these numbers increased substantially along with better (validation) loss values. This reduced the observed overfitting as well.

However, the model might still be slightly overfitted, because the dataset is still quite small, despite the data augmentation, and too specific. Figure 7 shows a gap between the two curves at the fifth epoch. This indicates that the current model has learned relationships that did not apply to the validation set. Consequently, the current model is probably more likely to perform worse on other test sets.

As is the case for all test sets, *walking* performs the worst (refer to Table 5). This could be explained by the fact that *walking* does not stand out in the dataset as much as the other activities. It is the slowest type of movement, meaning that less information of the movement itself can be captured in a 4 second audio fragment, and it generates the softest sounds, meaning that the environmental noise is relatively more dominant.

The bipedal gait cycle occurs for both *walking* and *running*. Both activities are a movements by foot. The only difference is the length of their cycles. Because these activities resemble each other the most, compared to other activities, this could explain why *walking* and *running* scored the worst. Furthermore, *running* performed better than *walking*, because it is less affected by the aforementioned factors. The movement is faster, so relatively more activity data is captured, and the generated sounds are louder, because of a greater impact with the surface.

The remaining two activities *null* and *cycling* performed best. These activities greatly differ from the others and each other. The *cycling* activity creates a very continuous grinding type of sound, because the wheels are in constant contact with the ground unlike *walking* and *running*. The *null* activity is actually present in all of the other activities, but still seemed easily distinguishable because it lacks any additional type of activity sound.

6.1 Sound Attenuation

The *close* and *far* test set yielded very poor accuracies. Especially the *close* test set, which unexpectedly returned a 17% accuracy and misclassified 50% of the files as *running*. Observations showed that the files in the *close* test set are generally louder than the ones in the *far* test set. Intuitively, one could think that the closer the recordings are taken, the clearer the recordings are and therefore the easier the classification is. However, this seems not to be the case.

A possible explanation why *walking* is confused so much with *running* can be reasoned with the following. As mentioned before, the difference between *walking* and *running*, besides a difference in movement speed, is that *running* generates louder sounds than *walking*. If one of these distinguishable aspects disappears (i.e. the difference in volume disappears), then it would be more difficult for the classifier to distinguish the two.

However, this theory only holds for the *close* test set. In a way, the *close* test set resembled the *running* activity the most. The same does not apply for the *far* test set, because the volume levels do not match the *running* activity like for the *close* test set. This possibly explains why the wrong predictions were equally spread over the incorrect classes. When the classifier incorrectly predicts, there is no class particularly close to it. In this case, the classifier seems not as convinced as for the *close* test set.

Lastly, both test sets only contained the *walking* activity. It can therefore not be concluded whether far recordings perform better than close recordings, because *running* and *cycling* might have returned totally different results. Though, what can be said is that both sets were severely affected by the change in distance for the *walking* activity. Especially if it is compared to the *walking* accuracy (70%) of the *overall* test set.

6.2 Overlapping Sounds

An average accuracy of 70% for the *overlapping* test set is quite adequate, considering that the model has not been trained with overlapping sounds. However, looking at the accuracies per activity, the *walking* activity stands out the most with a poor 27% accuracy. In 67% of the cases, *walking* was confused with *running*. A possible reason for this, could be that the sound of multiple people walking at the same time really resembles the *running* activity. The model perceives multiple footsteps in a short period of time, which from the model's point of view would indicate a faster bipedal movement like *running*.

Moreover, the other activities did not really seem affected by multiple people performing the same activity. They roughly returned the same accuracies as the *overall* test set. This can be explained by observing that the overlapping sound fragments of *running* and *cycling* did not really deviate from their single person activity recordings. Two people cycling still creates the same grinding type of sound as with one person cycling.

The *null* class was included in the test set to be able to compare it to the *overall* test set such that they have the same compositions. However, there is no such thing as overlapping environmental noise. So, assuming that the *null* class remained consistent for both the *overall* and *overlapping* test set, it could be argued that the accuracy is actually 75% instead of 70%. That is, when the accuracy for *null* is 100% instead of 80%, just like for the *overall* test set to which it is compared. Nevertheless, this does not change anything for the discussion held in this subsection.

6.3 Noise Interference

Unexpectedly, the noisy test set performed better (77%) than the one with barely any noise (70%). Intuitively, it could be reasoned that the important audio information would be masked by the noise and therefore reduce the classification accuracy. Apparently, the features extracted by the model are not heavily affected by this. Observations from various spectrograms revealed that most of the environmental noise lays in the lower frequency ranges (bottom of the spectrogram), while the actual activities are mostly represented in higher frequency ranges. This could suggest that the environmental noise does not interfere as much with the relevant range in the spectrogram as initially thought of.

As was the case with the previously discussed test sets, *null* and *cycling* do not seem affected. The presence and absence of severe noise interference is noticeable when looked at the *walking* and *running* activities. A theory could be that the training, validation and/or test set for the classifier might have been biased towards noisy data. After all, the data for these sets was divided without any regard for 'noisyness'. If more noisy data was present in the training set, then chances are higher that such imbalance has caused the model to perform better for these.

7. CONCLUSIONS

The prototype demonstrated that a sound-based human activity recognition approach for monitoring the usage of outdoor environments can be promising. It cannot be concluded whether this approach would work under different conditions than specified in this research. An overall prediction accuracy of 85% was achieved with the monophonic CNN-spectrogram model for four classes (*null*, *walking*, *running* and *cycling*). The activities *walking* and *running* were the most difficult to recognise for all test sets.

A difference in distance severely affected the classification accuracy for the *walking* activity. For the test set recorded more closely to the prototype, this was an unexpected 17%, and for the test set recorded farther away, this was a poor 47%. The closer test set misclassified 50% of the cases as *running*. Compared to the overall accuracy of *walking* (70%), these are massive drops.

Overlapping sounds were not found to be extremely problematic except for *walking*. Overall, the overlapping sounds test set scored a 70% in accuracy. Again, *walking* performed poorly with 27% and proved to be troublesome by being mislabeled as *running* 67% of the times. The rest of the activities scored similar to the *overall* test set.

Environmental noise did not seem to interfere with the classification accuracy in a negative way. The experiments showed that the test set containing noisy data scored better (77%) than the test set without (70%). The difference between both sets was that *walking* and *running* dropped in accuracy for the test set without severe noise.

Lastly, further work is still required to increase the reliability of the results by expanding the dataset both with the amount and variation. Moreover, it should be verified whether different distances for other activities besides *walking* affect the classification accuracy as well. Additionally, it should be investigated whether the dataset was biased towards noisy data, which would validate the results for noise interference. Finally, the influence of different segmentation lengths could be explored by examining the differences in the results.

8. ACKNOWLEDGEMENTS

I want to thank Datacadabra for supplying me with the prototype (Section 3.1) and for helping me with the data collection (Section 3.2).

9. REFERENCES

[1] M. Abadi and A. Agarwal. Tensorflow: Large-scale machine learning on heterogeneous distributed systems, November 2015.

[2] O. Abdel-Hamid, l. Deng, and D. Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. 08 2013.

[3] R. Anand, K. Mehrotra, C. Mohan, and S. Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, pages 962–969, 1993.

[4] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger. Towards a sensor for detecting human presence and activity. pages 305–314, 02 2011.

[5] N. Bordoloi, A. Talukdar, and K. Sarma. Suspicious activity detection from videos using yolov3. 2020.

[6] D. Cavaliere and S. Senatore. Exploiting a multi-device knowledge meshing to agent-based activity tracking. pages 2576–2583, 2020.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets, 2014.

[8] F. Chollet. keras. https://github.com/fchollet/keras, 2015. Accessed: 02/06/2021.

[9] A. Fernández. *Learning from Imbalanced Data Sets*. 2018.

[10] J. F. Gemmeke, D. P., W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio, 2017.

[11] E. Heijnen and V. Dellas. Het sport- en beweeggedrag van young urban professionals tijdens de coronamaatregelen en hun verwachtingen voor de toekomst, 2020.

[12] A. Jalal, M. Quaid, S. Ud Din Tahir, and K. Kim. A study of accelerometer and gyroscope measurements in physical life-log activities detection systems. pages 1–23, 2020.

[13] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari. Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access*, pages 7717–7727, 2019.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, page 84–90, May 2017.

[15] Z. Mushtaq, S.-F. Su, and Q.-V. Tran. Spectral images based environmental sound classification using cnn with meaningful data augmentation. *Applied Acoustics*, 2021.

[16] H. Phan, L. Hertel, M. Maaß, and A. Mertins. Robust audio event recognition with 1-max pooling convolutional neural networks. 09 2016.

[17] K. J. Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2015.

[18] K. J. Piczak. Esc: Dataset for environmental sound classification. pages 1015–1018, 2015.

[19] H. v. d. Poel and I. Pulles. Monitor sport en corona, July 2020.

[20] F. Sadri. Ambient intelligence: A survey. 10 2011.

[21] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research, 2014.

[22] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. 2019.

[23] A. Team. Audacity. https://audacityteam.org/, 1999. Accessed: 16/06/2021.

[24] Tensorflow. Simple audio recognition: Recognizing keywords. https://www.tensorflow.org/tutorials/audio/simple_audio/, 2021. Accessed: 31/05/2021.

[25] A. Vafeiadis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui. Audio content analysis for unobtrusive event detection in smart homes. *Engineering Applications of Artificial Intelligence*, 2020.

[26] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen. A convolutional neural network approach for acoustic scene classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1547–1554, 2017.

[27] M. Valueva, N. Nagornov, P. Lyakhov, G. Valuev, and N. Chervyakov. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, pages 232–243, 2020.

[28] H. van der Poel, P. Nafzger, and P. van Eldert. Monitor sport en corona ii, November 2020.

[29] Y. Verkaik. Bijna driekwart boswachters en groene boa's ziet nadelig effect van toegenomen recreatiedrukte op de natuur, December 2020.

[30] Wandelnet. Wandelen in coronatijd, 2020.

[31] W. Wang, F. Seraj, N. Meratnia, and P. Havinga. Privacy-aware environmental sound classification for indoor human activity recognition. pages 36–44, June 2019.

[32] F. Xu and P. Li. Outdoor human footsteps event and environment joint recognition. pages 1211–1217, 2020.

[33] Z. Yi and K. Tadahiro. Wearable sensor-based human activity recognition from environmental background sounds. *Journal of Ambient Intelligence and Humanized Computing*, 02 2012.

APPENDIX

A. RESULTS

A.1 Overall

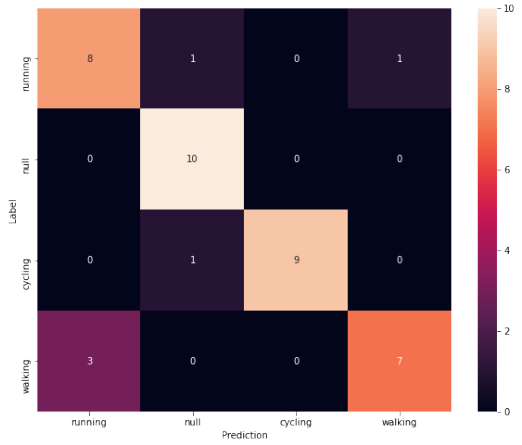

Figure 8. Confusion matrix of the overall test set

A.2 Sound Attenuation

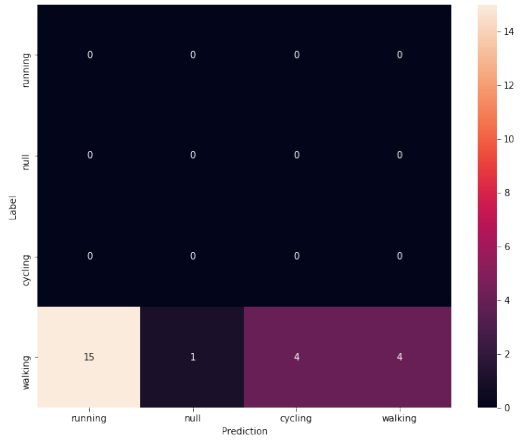

Figure 9. Confusion matrix of the close test set

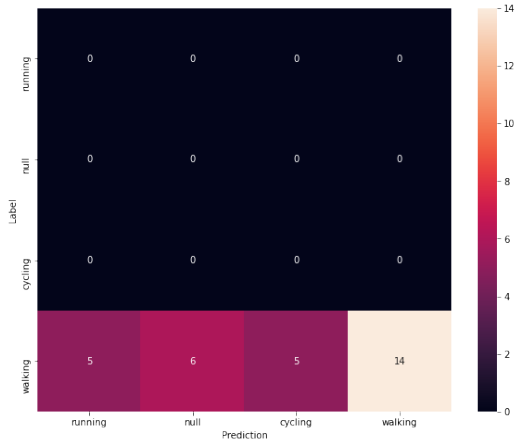

Figure 10. Confusion matrix of the far test set

A.3 Overlapping Sounds

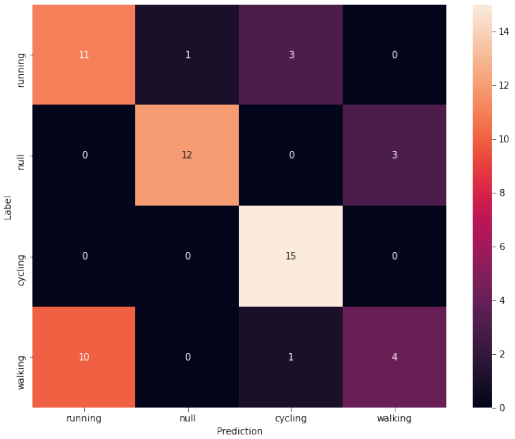

Figure 11. Confusion matrix of the overlapping sounds test set

A.4 Noise Interference

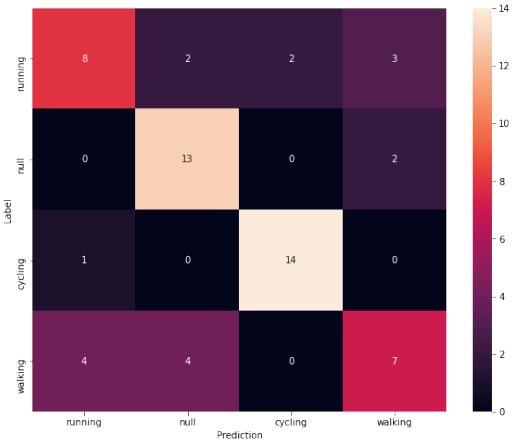

Figure 12. Confusion matrix of the normal test set

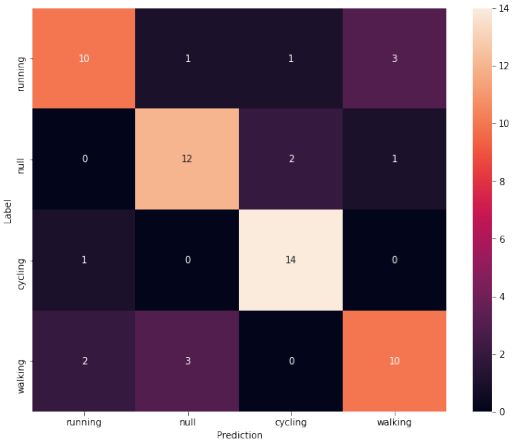

Figure 13. Confusion matrix of the noisy test set