FIA 2020/22
XII CONGRESSO/CONGRESO IBEROAMERICANO DE ACÚSTICA
XXIX ENCONTRO DA SOCIEDADE BRASILEIRA DE ACÚSTICA - SOBRAC
Florianópolis, SC, Brasil

Automatic Identification of city intrusive noises using Deep Learning

Borin, M.[1]; Holtz, M.[2]; Jacomussi, L.[3]; Monteiro, C.[4]; Weitbrecht, P.[5]; Jardim, C.[6]

[1–6] Harmonia, São Paulo, SP, Brazil, {marcel.borin, marcos.holtz, leonardo.jacomussi, carolina.monteiro, cecilia.jardim, paola}@harmonia.global

Abstract

Long-term noise monitoring is a powerful and useful approach for assessing the acoustic impact of a construction site on the neighbourhood. Its application may also help on evaluating the indoor noise of places such as schools, hospitals, and offices. In most cases, the site under inspection is located in cities, which is inherently prone to intrusive noises, especially those coming from traffic. Identifying intrusive noise for each situation is a challenging task for machine learning algorithms, and also a time saving procedure on long-term noise monitoring. This paper is about the creation of an audio dataset for some typical city intrusive noises, and its application on a low-cost monitoring system called OTOH. It will also be discussed some methods of automatically determining the acoustic labelling of each part during a long-term monitoring using Deep Learning algorithm.

Keywords: acoustics, machine learning, urban noise, deep learning.

PACS: 43.50.Rq.

1. INTRODUCTION

Over half of the world's population are currently living in urban areas [1] and until 2050, more than two thirds of the world will leave in urban agglomerations. Noise pollution has become the second-largest public health issue that has an impact on people [2], and cities have become the epicentre of this type of hazard. In Brazil, the city of São Paulo has an estimated population of over 12 million residents, and noise complaints has increased by almost 27% in the last couple of years [3]. One of its main sources, alongside traffic noise, are construction sites which are already an inherent part of this megacity.

Since 2016, with the approval of the new city's Master Plan [4], a transformation in São Paulo's skyline can be observed. As this five-year development plan encourages the construction of residential building close to the mains transport hubs, the verticalization process has been intensified. Only from March 2021 to February 2022, 721 new construction licenses have been approved [5] and 1363 demolitions permits have been issued [6]. As response to the noise annoyance caused by construction, a new regulation has been approved in an attempt to control the noise pollution. The Decree No. 60.581 [7] establishes sound pressure level limits for day and nighttime of noises produced by construction sites in the city.

Long-term noise monitoring is often a useful approach for investigating and mitigating the noise levels from a place under interest. However, urban soundscapes can be extremely complex due to the substantial number of sound sources that can exist in the same place. The correct detection of sound events is a helpful manner of understanding the soundscapes and possible solutions. Therefore, machine listening is a promising field to develop tools of automatic detection of noise sources for noise monitoring in complex environments.

Training a machine learning algorithm for sound classification demands a significant amount of carefully labelled data, once a good and representative dataset is the main factor for a high accuracy model performance [8]. However, current available audio datasets do not represent accurately the São Paulo's soundscape, since they lack in means of transportation and common animals such as birds and insects. SONYC [9] project provides a dataset with 23 categories with sound from New York city, and it is used as a reference for this work, by choosing more suitable classes for the Brazilian environment. The UrbanSound8K [10] is a benchmark in the field, on the other hand, it only presents 10 classes that are not enough for complex situations.

2. OBJECTIVE

This present work has the objective of developing an urban audio dataset with the main representative sound sources from the city of São Paulo, and the implementation of a neural network architecture able to identify their presence in real-life sound environments.

3. AUDIO DATASET

In this section, the dataset will be presented with the audio acquisition methods and tagging methodologies which is mandatory for supervised machine learning algorithms. In Figure 1, it is indicated the classes of sound sources that were gathered to compose the urban noise dataset. They represent the main sound sources found in the city of São Paulo, Brazil, and for this reason the data is almost fully acquired within the city area.

3.1 Labelling from automatic collected audio

Low-cost acoustic monitoring systems, named OTOH, made of Raspberry Pi 4 B and MEMS microphones, were installed in some different spots throughout the city of São Paulo to monitor the noise from construction sites [11]. The monitoring time spanned from 24h to weeks of data collection. The sound pressure level was measured full time and a trigger was set for the audio recording. Usually, the sound level required for activating the audio recording was set to 85 dB(A) at day and 59 dB(A) at night. These levels are the noise limits established on São Paulo decree No. 60.581. The audios were split in 4 seconds length segments, recorded at a sampling frequency of 22050 Hz and converted into ".mp3" files for data saving. All audio data was

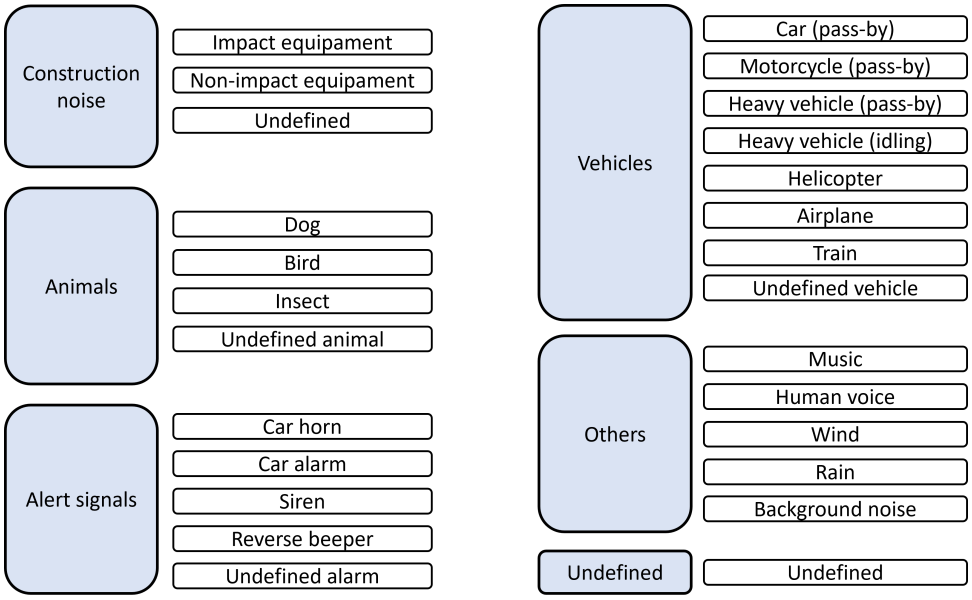

Figure 1: Scheme of sound sources classes used in the classification.

saved into an online platform where the team classification occurs.

An audio classification experienced team of 5 members performed the task of manually labelling the audio segments previously collected with the aid of a developed online interface as seen in Figure 2.

The upper side of the interface consists of the sound waveform in time domain, followed by its respective spectrogram. The visual clues given by the latter is a useful tool for aiding the audio classification [12], as well as the information of time and location of the sensor.

The classification requires a mandatory "Primary Class", which is the main acoustic source of the audio. An optional "Secondary Class" is also available in cases where there is a presence of more than one sound source in the audio segment. The classes follow the audio ontology presented in Figure 1. Both primary and secondary classes also receive a subjective "Presence" attribute, which can be "near", "far" or "undefined", to rate the distance between the sound source and the receiver.

Each member of the classification team is pretrained and familiarised with the audio set ontology and the platform, where they have the task to vote and rate each audio. The audios are individually presented and classified by each member. After three equal classification, the audio file class is considered a consensus and saved to the database with the corresponding metadata. The audios which do not receive consensus votes are considered undefined and are discarded.

Audio classification in complex environments such as cities can be a challenging task. The similarity of sound sources, or the lack of familiarity with them may cause even trained ears to misclassify its content, leading to an increase rate of undefined classifications, specially regarding sound sources coming from traffic and construction site. Furthermore, 4 seconds is sometimes insufficient for a listener to identify the correct sound source without context.

In addition, the unbalanced sound source presence in the collected audios may resulted in the lack of samples of many classes. For example, considering only the data under consensus, 21.7% of the audio belonged to motorcycles, and 19.6% are from human voice. On the other hand, important classes such as cars and heavy vehicles represent less 1% and 2%, respectively.

Therefore, considering the challenging classifications and the unbalanced classes, another approach of data collection is considered in next section.

Figure 2: Interface of the classification platform.

3.2 Labelling from manual collected audio

As a solution for enhancing the data classification of unbalanced classes, multiple rounds of longer recording were carried out in specific locations, seeking the sound sources under study. Thus, a Tascam DR-05X mobile recorder was used for this methodology, configured in a sampling frequency of 44.1 kHz in mono mode and ".wav" file format. The duration of audio recordings were chosen to be around 5 minutes of various representative sites in the city of São Paulo, in different days. As a complementary material, a video was also recorded during the sound capture, to assist the classification of difficult data, such as vehicles noise. Figure 3 shows one location where the recording had been been carried out to collect data of traffic noise.

Figure 3: Setup for audio recording of traffic noise.

After the data recording, the audio files were examined and manually exported into ".wav" files format up to 4 seconds long to their corresponding path in the database by an expert of the team.

3.3 Audio dataset summary

The data manually collected corresponded to the major part of the dataset, and it successfully fulfilled most of the unbalanced classes. At the moment of this paper is being produced, some of the sound sources classes have already reached a minimum amount of files to be able to be used in the study. The number of audio files in for each class can be seen in Table 1. Only the classes with the number of files higher than 100 are shown.

4. CLASSIFICATION USING CNN

This session describes the neural network approach implemented for the audio classification task.

4.1 Pre-processing

The data chosen for this experiment consists of 8 classes: cars, motorcycles, heavy vehicles (pass-by), heavy vehicles (idling), background noise, human voice, birds and insects. All other classes were left behind, due to the lack of samples which can cause the model to fail to learn the underlying patterns to distinguish these classes [8].

Table 1: Number of audio files collected for each class

Class	# of files
Cars (pass-by)	347
Motorcycle (pass-by)	329
Heavy vehicles (pass-by)	372
Heavy vehicles (idling)	315
Background noise	240
Human voice	352
Birds	315
Insects	303
Rain	135
Helicopter	136

The CNN model requires fixed frames of audio to serve as input, therefore, all audio samples were chosen to be 4 seconds long at 22050 Hz sample rate. Audios smaller than 4 seconds were filled with zeros to complete the duration.

Then, each audio was converted into a time-domain representation. In this case, a logarithmic Mel Spectrogram with 64 bands and 1024 samples of hop size. So, each audio sample is converted to fixed frame of 87x64. Figure 4 shows the result of this operation for a sample file from the bird class.

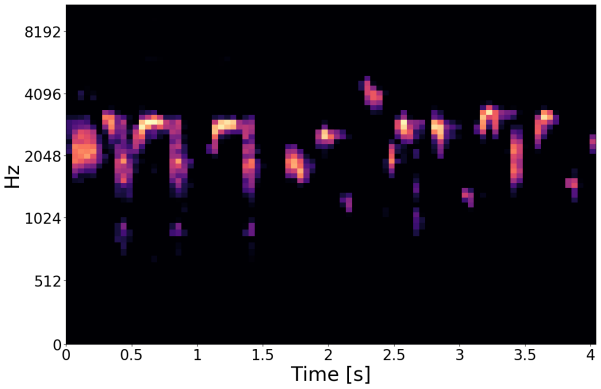

Figure 4: Mel Spectrogram of a bird class sample.

After the pre-processing, the dataset split was done as following: 60% for training, 25% for testing and 15% for validation.

4.2 Neural network architecture

The neural network architecture used in the project is schematically drawn in Figure 5 with the operations and its respective output shapes. A convolutional neural network (CNN) type was chosen due to its good performance on classification task for audio [13]. The model is composed by three sequential layers of convolutions and MaxPooling, before the fully connected dense layers. A batch normalization is done after every MaxPooling, and there is a Dropout before the final layer, with a L2 regularizer applied in the same layer as well. A "Relu" is used as activation function for both the convolutional and dense layers, except for the last one which a "Softmax" function is applied.

The simulation is executed for 8 classes with more data in a total of 2560 audio samples. Furthermore, 50 epochs is used, with saving checkpoint each time the validation loss decreases, so, only the best model is stored considering this metric. Adam optimiser is used with a learning rate of 10^{-4}. Since the weight initialisation is random, the simulation is run 3 times, so the metrics can be averaged. The mean normalized confusion matrix is shown is Figure 6.

An average accuracy of 88.7% was found for the model. It is important to notice the difficulty of the model to discriminate among the means of transportation, once they present similar source of sound generation, which is mainly by combustion engine at low frequency. As expected, human voice, birds and insects present much more unique patterns in the spectrogram, reaching around 99% on the accuracy performance.

4.3 Application of data augmentation

Data augmentation is a common technique in machine learning to artificially generate more samples to small datasets, using modification on the already collected data. It is useful to increase the model's ability to generalize, decreasing overfitting and enhancing the accuracy performance prediction on unseen data [8].

In this study, a data augmentation is applied to the training data, and the chosen modification are as follows:

- **Gaussian noise addition** (amplitude range: 0.001 to 0.015) (p=0.5);
- **Time stretch** (rate range: 0.8 to 1.25)

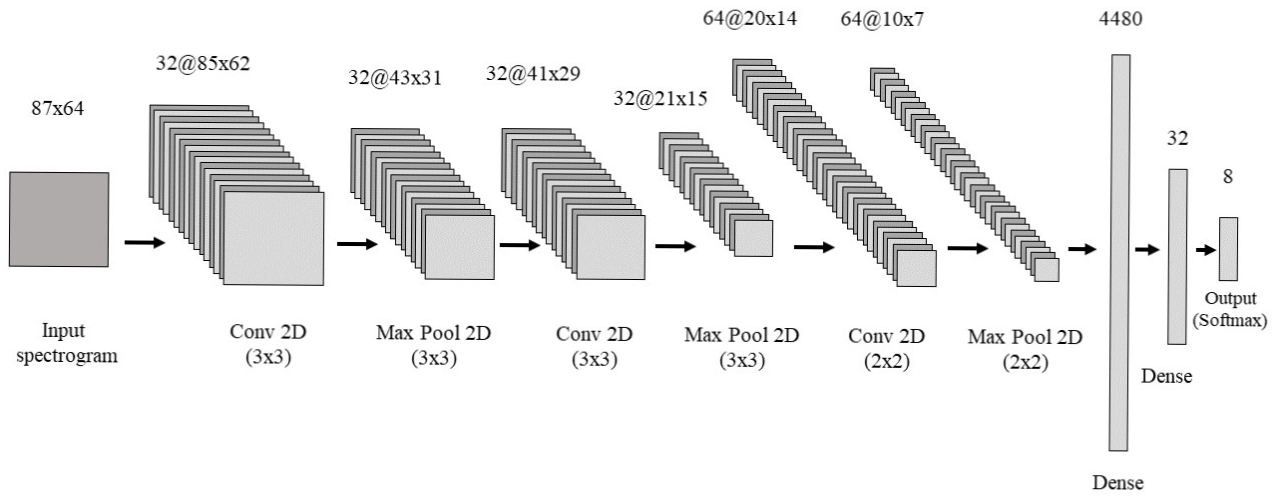

Figure 5: Neural network architecture implemented in the modelling.

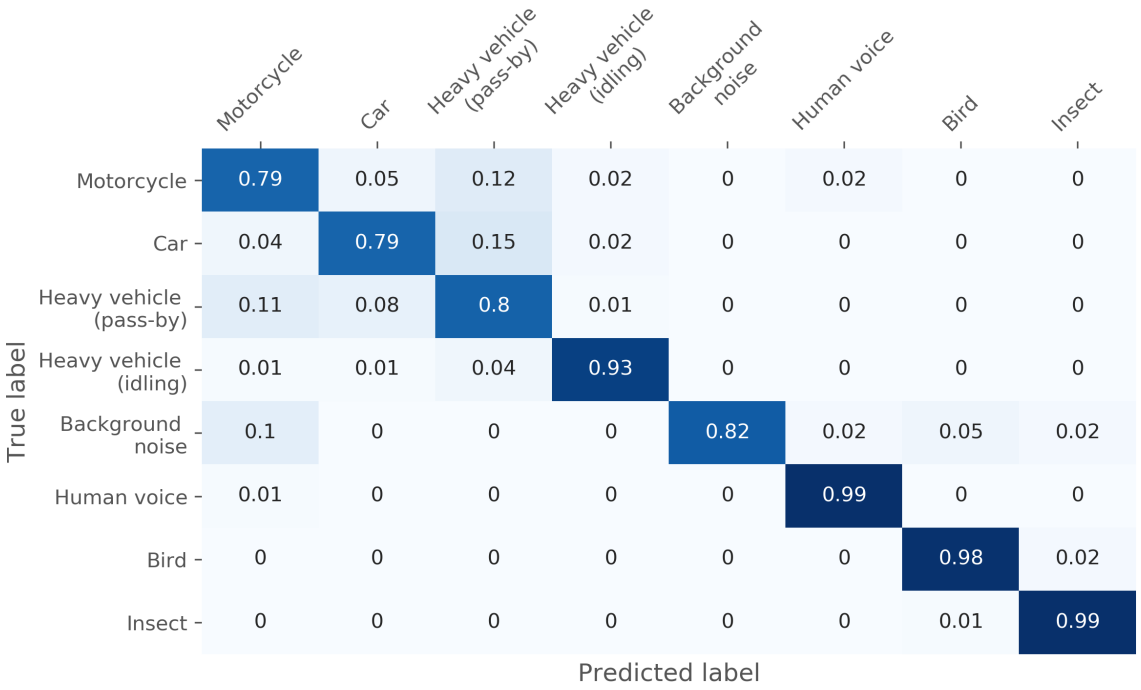

Figure 6: Normalized confusion matrix of the model

(p=0.5);

- **Pitch shift** (Semitone range: -2 to 2) (p=0.5);
- **Peak normalization** (p=1);
- **Volume gain** (range: -20 dB to 0 dB) (p=1).

The signal modification is applied in the presented order, and p is the probability of the respective augmentation to be executed. Peak normalization is always applied, so the gain volume can properly modify the data to a specific range of interest. These modification had been applied to all data classes, except for the background noise, where a range of lower gain was chosen for more realistic conditions. After the data augmentation, the dataset expanded to 4096 audio samples.

A simulation had also been run 3 times using the same settings as before, and the confusion matrix can be seen in Figure 7.

An average accuracy of 89.1% was found for the

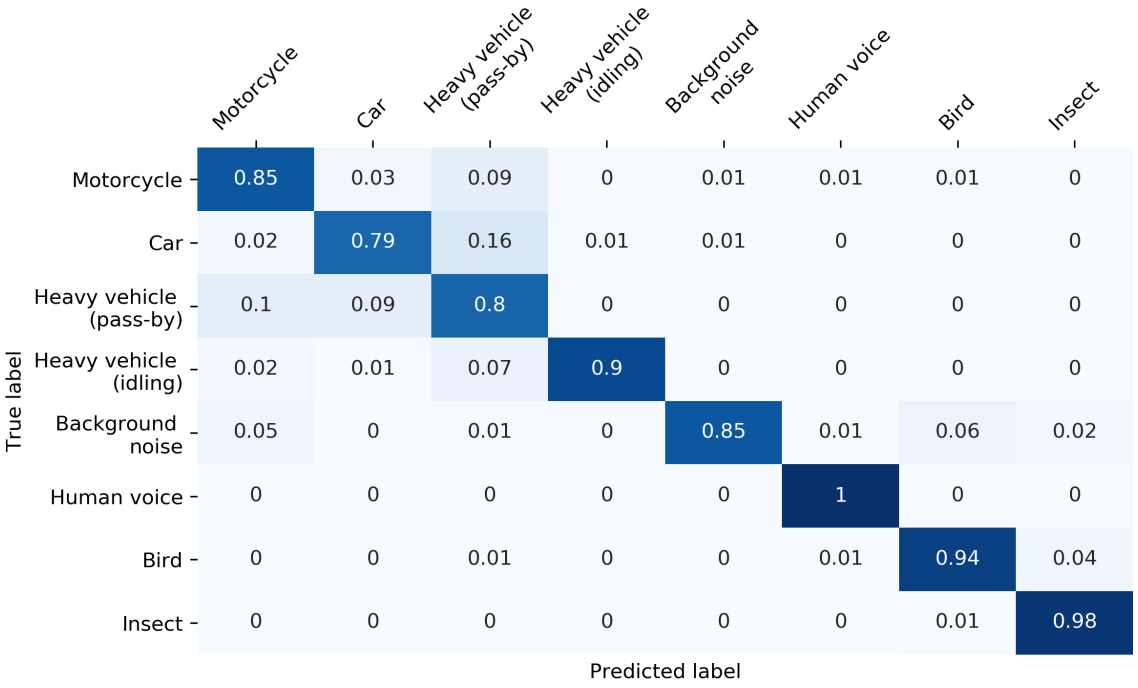

Figure 7: Normalized confusion matrix of the model with data augmentation

model, which is an insignificant increase compared to the model without data augmentation. The motorcycle accuracy performance raised 6%, however the other classes do not show any significant difference, and the variations could be explained by the random weight initialization. It is possible to conclude the data augmentation did not improve accuracy for this study. This technique should be optimised in further works, to achieve benefits for its application. One possible solution is to apply different processing for each class that can be more representative for their characteristics [14].

5. REAL LIFE APPLICATION

Model accuracy in a dataset controlled environment may be misleading when applied to real world, once the recorded audio generally presents classes that do not belong to the trained ones. Handling these data is a challenge, since they can be easily classified in an incorrect class. In this test, a 5 minutes continuous recording was carried out near to a street to capture vehicles noises, and its 4-seconds-long segments were labelled manually as a ground truth label. The trained model classified the continuous audio to acquire the predict label for each segment. For the classification, an onset trigger of 0.6 was used in order to handle classes that do not belong to the trained data. The samples that did not reach 0.6 of probability were classified as undefined, which was joined to the background noise class. The audio amplitude showed, during this study, to play a paramount role in the classification results, so a 10 dB gain was set to optimize them. Further investigation should be performed to investigate this matter.

A comparison between the predicted and the true labels classification can be seen in the timeline of Figure 8. Only 4 classes appeared in the test, so the rest were discarded from the legends. The model can predict well the vehicles pass-by when they are close to the microphone, however when they are approaching or moving away, there is a fade in/out in the sound level, which can either be classified as a vehicle or background noise. In an overview, the model can predict the main events occurrence, except for the motorcycle where only one event was correctly identified.

6. CONCLUSION

The construction of an audio data set is very time-demanding, and also requires caution and accuracy on the data labelling, once the data is

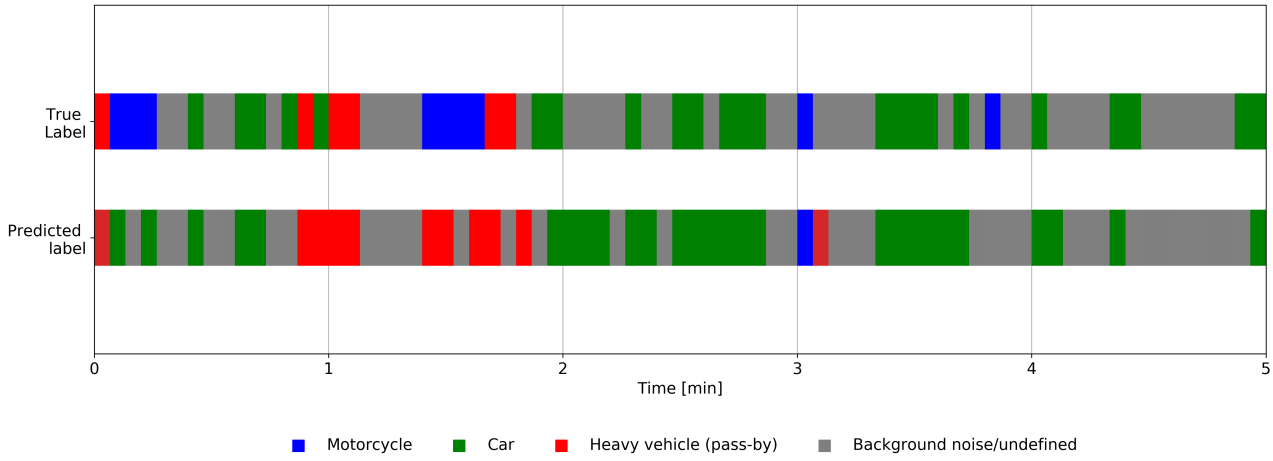

Figure 8: Predicted and True labels events timeline

the most important part for getting high accuracy models. The accuracy during the dataset modelling do not guarantee the same performance in real life applications due to complex environments. An onset trigger can be useful to minimize this situation, and other values should be studied. The dataset is still under development and further studies should be carried out when more data is accomplished. As a primary approach the model and dataset have responded satisfactorily well for traffic noise identification. Different data augmentation techniques should also be considered in order to raise the accuracy, and a special caution should be considered regarding the amplitude of the data, since the recording amplitude can leads to dramatic differences in the classification results, which is affected by the amplitude of the training samples. A careful data augmentation regarding the amplitudes can be useful on mitigating this situation.

REFERENCES

[1] Hannah Ritchie and Max Roser. Urbanization. *Our World in Data*, 2018. https://ourworldindata.org/urbanization.

[2] European Environment Agency. *Health risks caused by environmental noise in Europe*. Publications Office, 2020. doi:10.2800/524975.

[3] Prefeitura de São Paulo, 2021. URL http://dados.prefeitura.sp.gov.br/.

[4] Lei nº 16.050 - plano diretor estratégico do município de são paulo, 2014.

[5] Rafael Marko. Número de alvarás de empreendimentos concedidos em 2021 cai em São Paulo, 2022. URL https://sindusconsp.com.br/numero-de-alvaras-de-empreendimentos-concedidos-em-2021-cai-em-sao-paulo/.

[6] Camilla V. Mota. Demolições em alta apagam memória de bairros tradicionais de são paulo, 2021. URL https://www1.folha.uol.com.br/cotidiano/2021/10/demolicoes-em-alta-apagam-memoria-de-bairros-tradicionais-de-sao-paulo.shtml.

[7] DECRETO Nº 60.581, DE 27 DE SETEMBRO DE 2021. Regulamenta o controle de ruídos na execução das obras de construção civil no Município de São Paulo., 2021.

[8] Adrian Rosebrock. *Deep learning for computer vision with Python*. PYIMAGESEARCH, 1 edition, 2017.

[9] Mark Cartwright, Ana Elisa Mendez Mendez, Jason Cramer, Vincent Lostanlen, Graham Dove, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Bello. SONYC Urban Sound Tagging (SONYC-UST): A Multilabel Dataset from an Urban Acoustic Sensor Network. *Detection and Classification of Acoustic Scenes and Events 2019*, (October):35–39, 2019. doi:10.33682/j5zw-2t88.

[10] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A Dataset and Taxonomy for Urban Sound Research. doi:10.1145/2647868.2655045. URL http://dx.doi.org/10.1145/2647868.2655045.

[11] Paola Weitbrecht, Leonardo Jacomussi, Marcel Borin, Carolina Monteiro, and Cecilia Jardim. MEMS Based Low-Cost Urban Noise Monitoring: Tests and Case Study. *The 51st International Congress and Exposition on Noise Control Engineering - Internoise*, 2022.

[12] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer*

Interaction, 1(CSCW), 2017. ISSN 25730142. doi:10.1145/3134664.

[13] Marie A. Roch. How Machine Learning Contributes to Solve Acoustical Problems. *Acoustics Today*, 17(4):48, 2021. ISSN 15570215. doi:10.1121/at.2021.17.4.48.

[14] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 2016.