DOI: 10.61782/fa.2025.0157

FORUM ACUSTICUM EURONOISE 2025

PRELIMINARY RESULTS ON AUDIO EVENT CLASSIFICATION APPLIED TO A UNIVERSITY SQUARE IN MILAN (ITALY) BEFORE AN URBAN REGENERATION PROJECT

Andrea Potenza[1*] **Ester Vidaña-Vila**[2] **Andrea Afify**[3]
Roberto Benocci[1] **Rosa Ma. Alsina-Pagès**[2] **Giovanni Zambon**[1]
[1] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Italy
[2] Human Environment Research (HER), La Salle Campus Barcelona – Ramon Llull University, Spain
[3] Department of Physics, University of Milan, Italy

ABSTRACT

Biophonies, anthropophonies and geophonies characterize and shape an environment and contribute to the human appreciation of that place. Thus, sound event classification can be a useful tool to assess its quality and detect changes affecting it. In this study, different machine-learning models for multilabel sound classification are tested to monitor the ante-opera situation of the renewed square "Piazza della Scienza" of the University of Milano-Bicocca (Italy). The one-week monitoring was performed in May 2023 using 7 Song-Meter-Micros. The recordings were equalized to correct the devices' nonlinear frequency response. The paper is structured to: (a) test two sets of features in the Piazza's polluted soundscape by constant ventilation noise and other anthropogenic sources: YAMNet embeddings and classic audio features (such as MFCCs), (b) find the best algorithm between: decision tree, random forest, k-nearest neighbor and support vector classifier and (c) evaluate their performance when filtering the background ventilation noise to increase the datasets size. Preliminary results are presented with the final aim of optimizing the detection and applying it to describe the Piazza's soundscape, investigate differences in events spatial distribution, and evaluate the effects of the urban regeneration plan on the soundscape.

*__Corresponding author__: a.potenza@campus.unimib.it.

Keywords: *Sound event classification, YAMNet, MFCCs, soundscape, background noise.*

1. INTRODUCTION

Noise pollution is a major environmental health concern in Europe since noise made by traffic, trains and airplanes is the second most important cause of disease, behind air pollution caused by very fine particulate matter [1]. This issue is particularly impactful in urban areas where about 60% of the European population will live by 2050 [1, 2]. In urban centers there are multiple anthropogenic sound sources often overlapping with one another (e.g., engines idling, car horns, technical installations, sirens and alarms, nightlife) [3] and with biophonic sounds (e.g., birds vocalizations, dog barking, cicada sounds) and geophonic ones (e.g. wind, rain, thunder). The noise impact on humans does not depend solely on the overall sound pressure level but also on the typology of noise sources, the individual sensitivity and the social and cultural context [3–5]. This reasoning is also valid for faunal communities which are impacted by noise depending on its typology and on species' aural sensitivity [6, 7]. For these reasons, sound event classification is a topic that gained greater importance in these years and many studies have dealt with it. This technique applied to open environments is useful in many applications, from automatically discriminating against adverse meteorological conditions in large datasets [8,9], to detect specific noises [8] and to collect events for soundscape assessments [10] and bio/eco-acoustic analyses [8, 9, 11, 12]. Linked to urban settle-

ments, many works have been performed on event classification from real in-field recordings, focusing on single events or on overlapping multiple events [10–12]. In this work, sound event classification is applied to a university square in Milan (Italy) which is heavily impacted by constant technical installation noise. The study's final aim is to develop a classifier for different sound events (i.e., road traffic, tram passages, speech, sirens) which will be used to assess the changes in the square's sound environment after an architecture requalification [13] which will improve the public space usage and increase the surface occupied by trees, bushes and lawns instead of cement. Moreover, the classification will allow a better understanding of the ecoacoustic indices behavior in urban environments and of people's acoustic environment perception already carried out in the square. In this paper, the efficacy of two sets of features on the square recordings was tested: the YAMNet embeddings and more classical features, such as the Mel-frequency cepstral coefficients (MFCCs), using different machine learning classificators: Decision tree, Random forest, K-nearest neighbor e Support vector classifier.

2. MATERIALS AND METHODS

2.1 Study Area

The area under study is the main square of the University of Milano-Bicocca called "Piazza della Scienza". It is set in the Bicocca neighborhood in the north outskirts of Milan, an area which went through a renovation process in the last fifty years, shifting from an industrial site to a university and third sector district. The Piazza covers a surface area of 8590 m^2 and is surrounded by 6-floor buildings and arranged on two levels: a fully cemented ground level and four lowered courtyards, three of them with a lawn cover. In 2024, the Piazza underwent a regeneration process thanks to the MUSA project (Multilayer Urban Sustainability Action) [13]. This project is set into the framework of the National Recovery and Resilience Plan (NRRP) and aims at proposing management strategies to face the environmental, social, and economic sustainability challenges of the metropolitan city of Milan. The Piazza renovation aims to increase its environmental sustainability in terms of reducing the heat island effect, increasing the water descent into the aquifer, reducing local air pollution, and improving the area's biodiversity. Moreover, it will allow a better use of the public space and increase students' quality of life in the Piazza. These results will be achieved through the depaving and greening of the ground level and the renewal of the lowered courts (Figure 1).

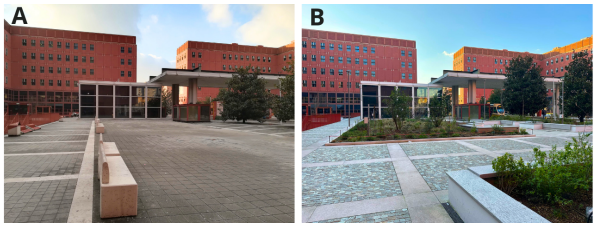

Figure 1. Piazza della Scienza before (A) and after (B) the regeneration process of MUSA.

2.2 Data collection and annotation

The monitoring campaign used to populate the database was performed in May 2023, from Monday 22th to Sunday 28th, in 7 sites of the Piazza (Figure 2). To capture the different sound sources present, 2 devices were placed at the ground level floor, 2 in the lowered courtyards and 3 at the first floor of the buildings.

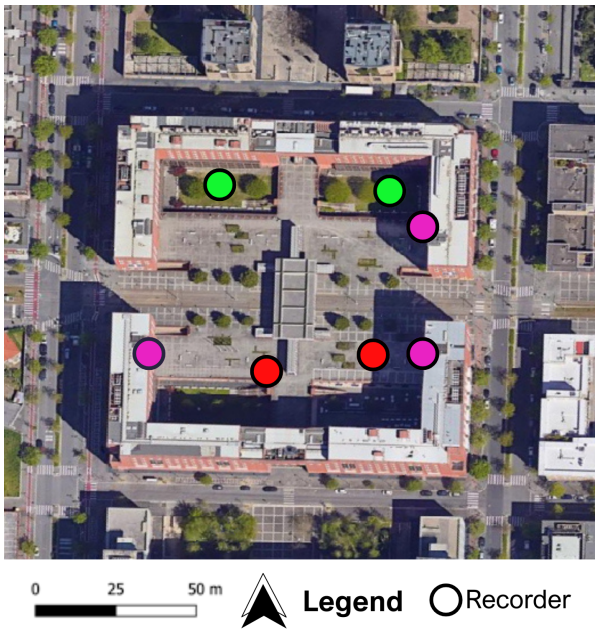

Figure 2. Piazza della Scienza (before the regeneration) and recording sites (pink: first floor, red: ground floor, green: lowered courtyards).

Table 1. Event labels present in the dataset, their description and total time duration.

Label	Description	Duration (min)
Road	Road traffic	17.55
Brake	Brake noise from vehicles and tram	5.65
Car horn	Car horn	0.24
Vpbor	Vehicle passing by on tram rails	0.89
Tram	Tram passing by, stopping and departing	12.41
Motorbike	Motorbike engine noise and passing by	3.79
Street cleaning	Street cleaning vehicles and operators	1.09
Siren	Ambulance and police sirens	6.10
SignalAlarm	Signals and Alarm noises	4.22
Construction	Noises from construction sites	3.29
TrolleyCart	Cart and trolley noises on pavimentation	5.39
Bird	Birds vocalizations	5.29
Dog	Dog barks	0.34
Insects	Insects sounds	4.33
Wind	Gusts of wind	1.66
Speech	Speech and human voices, exclamation, laughs	21.57
Ventilation	Technical installations noise	125.00
Complex	Unclassified sounds	3.23

The Song Meter Micros from Wildlife acoustic were used, set with a sampling rate of 48 kHz and an amplitude gain of +12 dB, recording for 1 minute and pausing for 1 minute. Furthermore, the recordings were equalized to correct the devices' nonlinear frequency response [14,15]. In the Supplementary Materials of [14] it is available the open source MATLAB script for equalizing the recordings and obtain linear-comparable audio files with other devices and between ecoacoustic studies. The recording campaign resulted in a total of 75 hours of acoustic data per sensor. The labelling process was performed on Audacity following previous works of coauthors [16]. Audacity is an audio recording and editing software which allows users to select part of an audio file and associate a text label with it. These labels can be exported in a text file (txt) as a list indicating the labels' name and their starting and ending time (time is expressed in seconds and calculated referencing to the audio beginning as the time zero). In this work, a single label was applied to each sound event independently of its time duration. The total labelled time is 2 hours and 5 minutes. The label taxonomy was adapted from [17] to fit the particular sound events present in the Piazza and discharge those not happening. In Table 1 are reported the 18 labels kept for this study area and their total time duration in the dataset. To be fed into the classifiers, the txt files from Audacity were converted into one-hot-encoded format.

2.3 Feature extraction

To detect the sound events in the Piazza two methods have been used to test their efficiency:

- YAMNET's embeddings.
- Features from the python package librosa.

2.3.1 YAMNet's embeddings

YAMNet is a pre-trained neural network model which has been used in many works in literature to perform sound event detection [9, 18] or to obtain features to be used as features for other algorithms [19]. The model relies on the MobileNet V1 architecture which consists of 27 convolutional layers, 1 global average pooling layer, and 1 fully connected layer; it employs ReLU activation functions, batch normalization and Softmax activation to obtain an

Table 2. Labels occurrences in each dataset.

Label	YAMNet OR	YAMNet filt	Librosa 1s OR	Librosa 1s filt	Librosa 20ms OR	Librosa 20ms filt
Road	2'211	6'633	1'115	3'345	5'435	16'305
Brake	738	2'214	397	1'191	1'866	5'598
Car horn	37	111	26	78	104	312
Vpbor	135	405	105	315	425	1'275
Tram	1'566	4'698	792	2'376	3'844	11'532
Motorbike	488	1'464	260	780	1'241	3'723
Street cleaning	137	411	67	201	334	1'002
Siren	761	2'283	373	1'119	1'842	5'526
SignalAlarm	536	1'608	285	855	1'384	4'044
Construction	426	1'278	235	705	1'095	3'285
TrolleyCart	692	2'076	368	1'104	1'760	5'280
Bird	755	2'265	467	1'401	2'074	6'222
Dog	51	153	37	111	158	474
Insect	556	1'668	279	837	1'370	4'110
Wind	240	720	143	429	640	1'920
Speech	2'750	4'864	1'457	2'733	6'931	12'649
Complex	487	1'461	347	1'041	1'430	4'290

output feature vector of size 1024 [20, 21]. The model predicts the 521 sound event classes of the AudioSet corpus [22]. Each audio file is first downsampled at 16 kHz and normalized to fit the AudioSet requirements before calculating the features (embeddings) on a time window of 0.96 seconds with a 50% overlap. The embeddings calculation is explained here [23] but, shortly, they are logarithmic Mels calculated in the range 0.125 - 7.5 kHz and fed to the MobileNet V1 model. As the final step, the model calculates the output scores for each of the 521 AudioSet classes. In our study, YAMNet was used as a feature extractor on each labelled audio file and each 0.96 seconds time frame is characterized by 1024 features.

2.3.2 *Features from the python package librosa*

To compare YAMNet's features performances, another set of features have been calculated. In particular, the python package librosa [24] was used. In this case, downsampling and normalization were not applied. The features were calculated on a 1-second time frame and a 20-millisecond time frame without overlapping; the double timeframes are kept to understand their influence on the classification. The following spectral features were calculated following previous works [25]:

- Mel-frequency cepstral coefficients (MFCCs), using a number of coefficients of 13.
- Spectral centroid (SC).
- Roll-off frequency, with a rollpercent of 10 and 90.
- Amplitude root-mean-square (RMS).
- Zero-crossing rate (ZCR).
- Static tempo.
- Flatness.
- Spectral bandwidth.

For each feature (except MFCCs and static tempo), 7 statistical descriptors were calculated: minimum, maximum, mean, median, standard deviation, skewness and kurtosis [25]. Thus, each 1-second time frame is characterized by 1272 values (of which 1222 are MFCCs), while each 20-millisecond time frame by 297 values (of which 247 are MFCCs).

2.4 Data filtering to increase the datasets

To increase the number of recordings populating the datasets, filtration of the ventilation noise (label "Ventilation") has been carried out. This operation have been achieved using the "Noise Reduction" function of Audacity. This function elaborates the noise profile of a recording and uses it to reduce that noise in other recordings.

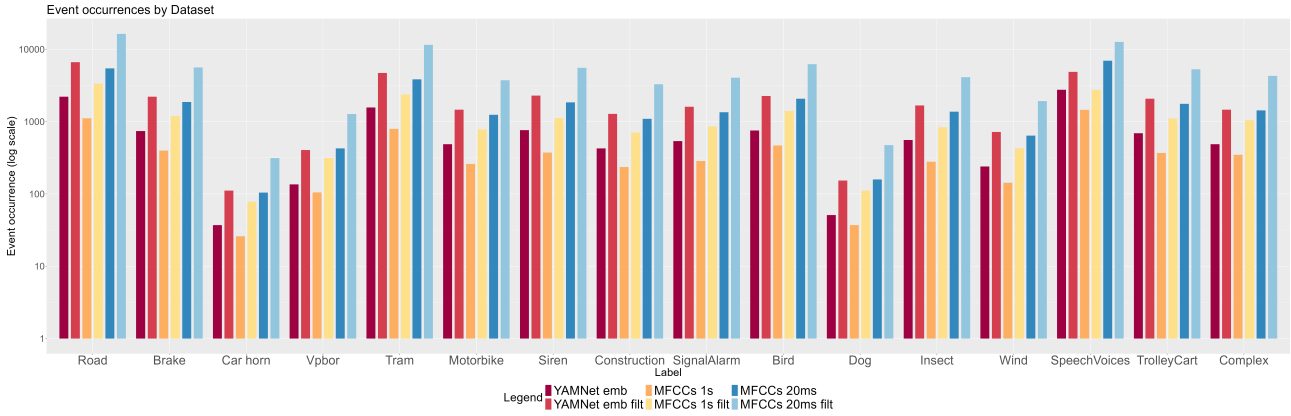

Figure 3. Labels occurrences in each dataset.

For each site, a midnight recording containing only ventilation noise was selected. Two settings were applied: T1 (Noise reduction: 6 dB, Sensitivity: 5, Frequency Smoothing = 4) and T2 (Noise reduction: 10 dB, Sensitivity: 10, Frequency Smoothing = 8). Furthermore, due to uneven occurrences of labels (see Table 2 and Figure 3), in the filtered recordings the label with the highest occurrence ("Speech") and the corresponding features were removed when the event was not overlapped with others. After these filtering and balancing processes, the datasets are shaped as reported in Table 2 and in Figure 3.

2.5 Detection

For the detection of sound events, different classification models have been implemented in the Python environment. A pipeline has been built to run the following classification models: decision tree (DT), random forest (RF), k-nearest neighbor (KNN), support vector classifier (SVC). These algorithms are executed to perform a multi-output classification to classify overlapping sound events. Different trials have been performed using three different scalers (standardscaler, minmaxscaler, robustscaler); standardscaler subtracts the mean and scales the data to unit variance, minmaxscaler uses the maximum and minimum values of the dataset to rescale the features in the range [0, 1], and robustscaler is based on the percentiles thus is less affected by outliers [26]. The datasets have been split into training (80%) and testing (20%) using the train_test_split command from sklearn. The best classifier will be chosen based on three metrics: Precision, Recall and F1-Score.

3. RESULTS AND DISCUSSION

In Table 3 and 4, the best values of Precision, Recall and F1-score are reported for each label and feature typology, indicating the relative dataset, classifier and scaler.

Looking at the "Dataset" columns of Table 3 and 4, we found that the classificators performed better with the augmented datasets (occurrences increased by filtering the Ventilation background noise). Regarding the librosa features timeframe, the 20-ms windowing allowed a better performance.

Focusing on the typology of Classifier and Scaler, the most frequent one are the Decision Tree (DT), followed by the Random Forest (RF), and the StandardScaler (Std).

In Table 3 (YAMNet embeddings), it is possible to notice that the majority of the F1-scores are above 60% with Street cleaning, Siren, Bird, Insect and Speech being over 70%. On the other hand, in Table 4 (Librosa features) events that have an F1-score higher than 70% are Street cleaning, Siren, Insect, Speech, Road, Tram, SignalAlarm, Construction and TrolleyCart. Dog barking obtained the lowest scores, probably due to its lower occurrence in the dataset (Table 2) and the masking caused by ventilation systems. Siren got the highest score, probably because of its marked distinguished spectrum.

When comparing the F1-score between the datasets, the librosa features received higher scores for the majority of labels. This trend is also visible for Precision and Recall, even if not for the same sound events.

Table 3. Best results using the YAMNet embeddings

Label	Precision	Recall	F1-score	Dataset	Classifier and scaler
Road	0.64	0.67	0.65	Filter	DT - Std
Brake	0.59	0.62	0.61	Filter	DT - Std
Car horn	1.00	0.45	0.62	Filter	RF - Std
Vpbor	0.56	0.71	0.63	Filter	DT - Std
Tram	0.59	0.65	0.62	Filter	DT - Std
Motorbike	0.53	0.64	0.58	Filter	DT - Std
Street cleaning	0.71	0.77	**0.74**	Filter	DT - Std
Siren	0.83	0.83	**0.83**	Filter	DT - Std
SignalAlarm	0.55	0.61	0.58	Filter	DT - Std
Construction	0.64	0.63	0.63	Filter	DT - Std
TrolleyCart	0.66	0.67	0.67	Filter	DT - Std
Bird	0.96	0.61	**0.74**	Filter	RF - Std
Dog	0.45	0.45	0.45	Filter	DT - Std
Insect	0.73	0.72	**0.72**	Filter	DT - Std
Wind	0.48	0.56	0.52	Filter	DT - Std
Speech	0.91	0.68	**0.78**	Filter	RF - Std
Complex	0.56	0.56	0.56	Filter	DT - Std

Table 4. Best results using the Librosa features

Label	Precision	Recall	F1-score	Dataset	Classifier and scaler
Road	0.97	0.61	**0.75**	20ms Filter	RF - Std
Brake	0.58	0.59	0.58	20ms Filter	DT - Std
Car horn	0.40	0.57	0.47	20ms Filter	DT - Std
Vpbor	0.50	0.60	0.55	20ms Filter	DT - Std
Tram	0.98	0.59	**0.74**	20ms Filter	RF - Std
Motorbike	0.60	0.64	0.60	20ms Filter	DT - Std
Street cleaning	0.95	0.88	**0.91**	20ms Filter	SVC - Std
Siren	0.98	0.86	**0.92**	20ms Filter	RF - Std
SignalAlarm	0.99	0.56	**0.72**	20ms Filter	RF - Std
Construction	0.97	0.65	**0.78**	20ms Filter	SVC - Std
TrolleyCart	0.99	0.70	**0.82**	20ms Filter	RF - Std
Bird	0.57	0.64	0.60	20ms Filter	DT - Std
Dog	0.44	0.51	0.47	20ms Filter	DT - Std
Insect	0.93	0.74	**0.82**	20ms Filter	RF - Std
Wind	0.61	0.61	0.61	20ms Filter	DT - Std
Speech	0.94	0.55	**0.70**	20ms Filter	RF - Std
Complex	0.54	0.58	0.56	20ms Filter	DT - Std

4. CONCLUSIONS

Sound classification is a rich and growing research field in various acoustic branches nowadays and has been applied to numerous case studies around the world. This paper is focused on classifying the sounds occurring in the Piazza della Scienza in Milan, Italy.

These preliminary results provide important insights which show the better performance of the models using the librosa features instead of the YAMNet embeddings and the usage of a shorter time frame. Future analyses will be implemented on the librosa features dataset. They will consist of different steps like the implementation of confusion matrices and the application of a cross validation. Moreover, the GridSearch algorithm will be applied for tuning the models' hyperparameters. Furthermore, other algorithms will be tested (i.e., CNN-models) and the effects of the time window size on the performances will be studied. Finally, transfer learning will be taken into account to increase the dataset size, alongside the addition of audios recorded outside the Piazza to bypass the ventilation background noise.

5. ACKNOWLEDGMENTS

Investigation performed within the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D".

6. REFERENCES

[1] B. Lex, *Burden of disease from environmental noise. Quantification of healthy life years lost in Europe*. City: World Health Organization, 2011.

[2] U. N. D. of Economic and S. Affairs, *World Urbanization Prospects: The 2018 Revision*. City: United Nations, 2019.

[3] M. Raimbault and D. Dubois, "Urban soundscapes: Experiences and knowledge," *Cities*, vol. 22, no. 5, pp. 339–350, 2005.

[4] S. Dreger, S. A. Schüle, L. K. Hilz, and G. Bolte, "Social inequalities in environmental noise exposure: a review of evidence in the who european region," *International journal of environmental research and public health*, vol. 16, no. 6, p. 1011, 2019.

[5] N. Singh and S. C. Davar, "Noise pollution-sources, effects and control," *Journal of Human ecology*, vol. 16, no. 3, pp. 181–187, 2004.

[6] R. J. Dooling and A. N. Popper, "The effects of highway noise on birds. report to the california. department of transportation, division of environmental analysis, sacramento, california," 2007.

[7] D. Waddington, M. Wood, B. Davies, and R. Young, "Habitats: Managing the ecological impacts of noise on wildlife habitats for sustainable development," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 265, pp. 5247–5251, Institute of Noise Control Engineering, 2023.

[8] A. Brown, S. Garg, and J. Montgomery, "Automatic rain and cicada chorus filtering of bird acoustic data," *Applied Soft Computing*, vol. 81, p. 105501, 2019.

[9] F. Terranova, L. Betti, V. Ferrario, O. Friard, K. Ludynia, G. S. Petersen, N. Mathevon, D. Reby, and L. Favaro, "Windy events detection in big bioacoustics datasets using a pre-trained convolutional neural network," *Science of the Total Environment*, vol. 949, p. 174868, 2024.

[10] D. Bonet-Solà, E. Vidaña-Vila, and R. M. Alsina-Pagès, "Prediction of the acoustic comfort of a dwelling based on automatic sound event detection," *Noise Mapping*, vol. 10, no. 1, p. 20220177, 2023.

[11] B. Swaminathan, M. Jagadeesh, and S. Vairavasundaram, "Multi-label classification for acoustic bird species detection using transfer learning approach," *Ecological Informatics*, vol. 80, p. 102471, 2024.

[12] Devos, Paul, "Birdsong of common birds in an urban soundscape as evaluated with recurrent neural networks," in *Forum Acusticum 2023 : 10th Convention of the European Acoustics Association, Proceedings*, p. 2, 2023.

[13] MUSA, "Multilayered urban sustainability action," 2023. Available at: `https://musascarl.it/`, Accessed on: 31-03-2025.

[14] A. Potenza, V. Zaffaroni-Caorsi, R. Benocci, G. Guagliumi, J. M. Fouani, A. Bisceglie, and G. Zambon, "Biases in ecoacoustics analysis: A protocol to equalize audio recorders," *Sensors (Basel, Switzerland)*, vol. 24, no. 14, p. 4642, 2024.

[15] T. Alvares-Sanches, P. E. Osborne, and P. R. White, "Impacts of the covid lockdown on the soundscape of an urban area: noise, psychoacoustic metrics and ecoacoustic indices," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 268, pp. 5700–5708, Institute of Noise Control Engineering, 2023.

[16] E. Vidaña-Vila, L. Duboc, R. M. Alsina-Pagès, F. Polls, and H. Vargas, "Bcndataset: Description and analysis of an annotated night urban leisure sound dataset," *Sustainability*, vol. 12, no. 19, p. 8140, 2020.

[17] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

[18] J. Mariscal-Harana, V. Alarcón, F. González, J. J. Calvente, F. J. Pérez-Grau, A. Viguria, and A. Ollero, "Audio-based aircraft detection system for safe rpas bvlos operations," *Electronics*, vol. 9, no. 12, p. 2076, 2020.

[19] N. H. Valliappan, S. D. Pande, and S. R. Vinta, "Enhancing gun detection with transfer learning and yamnet audio classification," *IEEE Access*, 2024.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[21] A. Avila and C. Pinzón, "Comparative analysis of vggish and yamnet models for welding defect detection," in *International Scientific-Technical Conference MANUFACTURING*, pp. 184–199, Springer, 2024.

[22] AudioSet, "Audioset, a large-scale dataset of manually annotated audio events." Available at: `https://research.google.com/audioset/index.html`, Accessed on: 31-03-2025.

[23] YAMNet, "Yamnet." Available at: `https://github.com/tensorflow/models/tree/master/research/audioset/yamnet`, Accessed on: 31-03-2025.

[24] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python.," *SciPy*, vol. 2015, pp. 18–24, 2015.

[25] R. Benocci, A. Afify, A. Potenza, H. E. Roman, and G. Zambon, "Toward the definition of a soundscape ranking index (sri) in an urban park using machine learning techniques," *Sensors*, vol. 23, no. 10, p. 4797, 2023.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.