





Urbanphony-3-CNN: A Convolutional Neural Network for Identifying the Urban Soundscape Taxonomy in Spectrograms Generated from Audios of Historic Cities

Urbanofonía-3-RNC: una red neuronal convolucional para identificar la taxonomía del paisaje sonoro urbano en espectrogramas generados a partir de audios de ciudades históricas

Carlos A. Durán-Paredes ¹, Julián A. Grijalba-Obando ², Sebastián A. Cajas-Ordóñez ³, and Camilo Sánchez-Ferreira ⁴

ABSTRACT

Urban soundscapes are characterized by the overlapping of multiple sounds, posing challenges for automatic classification via deep learning. This study applies convolutional neural networks (CNNs) with transfer learning to identify diverse sounds in urban environments using visual representations called *Mel spectrograms* – time-frequency images of audio signals. We created the Urbanphony-3 (UP3) dataset from recordings in historic cities with significant sound overlaps and compared it against two established datasets with minimal overlap: UrbanSound8K (US8K) and Environmental Sound Classification 50 (ESC50). CNNs were trained with each dataset to develop the UP3-CNN, US8K-CNN, and ESC50-CNN models, enabling the automatic recognition of various urban sounds. Model performance was assessed through five-fold cross-validation, using accuracy and loss metrics, as well as confusion matrix analysis and ROC curves. The UP3-CNN model, which classified sounds from environments with frequent overlap, reached an accuracy of 75.2%. In contrast, ESC50-CNN and US8K-CNN, trained with less overlapped sounds, yielded better results (85.6 and 86.3%, respectively). These findings confirm that CNNs have great potential for classifying urban soundscapes, even under natural overlap. However, the performance gap between UP3-CNN and the other models indicates that CNNs are less effective when sounds overlap significantly. Thus, additional strategies are required to improve the results, including data augmentation, transformers, or optimization techniques. Future research should also extend automatic soundscape classification by considering other variables, such as emotional reactions, cultural preferences or contextual influence.

Keywords: soundscape studies, urban soundscape, convolutional neural networks, Mel spectrograms, supervised learning, cross-validation

RESUMEN

Los paisajes sonoros urbanos se caracterizan por la superposición de múltiples sonidos, lo que plantea desafíos para la clasificación automática mediante aprendizaje profundo. Este estudio aplica redes neuronales convolucionales (CNN) con aprendizaje por transferencia para identificar diversos sonidos en entornos urbanos por medio de representaciones visuales llamadas *espectrogramas de Mel* – imágenes tiempo-frecuencia de señales de audio. Se creó el conjunto de datos Urbanphony-3 (UP3) a partir de grabaciones en ciudades históricas con una notable superposición sonora y se comparó con dos conjuntos de datos consolidados con mínima superposición: UrbanSound8K (US8K) y Environmental Sound Classification 50 (ESC50). Las CNN se entrenaron con cada conjunto de datos para desarrollar los modelos UP3-CNN, US8K-CNN y ESC50-CNN, lo que permitió el reconocimiento automático de diversos sonidos urbanos. El desempeño de los modelos se evaluó mediante validación cruzada de cinco pliegues, usando métricas de exactitud y pérdida, así como análisis de matrices de confusión y curvas ROC. El modelo UP3-CNN, que clasificó sonidos de entornos con superposición frecuente, alcanzó una exactitud del 75.2 %. En contraste, ESC50-CNN y US8K-CNN, entrenados con sonidos menos superpuestos, obtuvieron mejores resultados (85.6 y 86.3 % respectivamente). Estos hallazgos confirman que las CNN tienen un gran potencial para la clasificación de paisajes sonoros urbanos, incluso bajo condiciones de superposición natural. Sin embargo, la brecha de desempeño entre UP3-CNN y los otros modelos indica que las CNN son menos efectivas cuando los sonidos se superponen de manera significativa. Por tanto, se requieren estrategias adicionales para mejorar los resultados, incluyendo aumento de datos, transformadores o técnicas de optimización. Asimismo, las investigaciones futuras deberían ampliar la clasificación automática de paisajes sonoros considerando otras variables, como las reacciones emocionales, las preferencias culturales o la influencia del contexto.

Palabras clave: estudios en paisaje sonoro, paisaje sonoro urbano, redes neuronales convolucionales, espectrogramas Mel, aprendizaje supervisado, validación cruzada

Received: July 22th, 2024

Accepted: July 21th, 2025

¹ Eng., Universidad del Cauca, Colombia. Affiliation: Institución Universitaria Colegio Mayor del Cauca, Popayán, Colombia. Email: duranurbanphony@unimayor.edu.co

² Arch., Institución Universitaria Colegio Mayor del Cauca, Colombia. MSc, Universidade de Santiago de Compostela, España. Affiliation: Institución Universitaria Colegio Mayor del Cauca, Departamento de Arte y Diseño, Popayán, Colombia. Email: juliangrijalba@unimayor.edu.co

³ Eng., Universidad del Cauca, Colombia. MSc, Erasmus Joint Master Degree, Université de Bordeaux (Ubx), Pázmány Péter Catholic University, Autonomous

University of Madrid (UAM). Affiliation: Centre for AI (CeADAR), University College Dublin. Email: sebastian.cajasordonez@ucd.ie

⁴ Eng., Universidad del Cauca, Colombia. PhD, Universidade de Brasília, Brasil. Affiliation: Universidad del Cauca, Departamento de Física, Popayán, Colombia. Email: csanchez@unicuca.edu.co



Attribution 4.0 International (CC BY 4.0) Share - Adapt

Introduction

Noise, or unwanted sound, has a significant impact on the physical and mental health of urban residents worldwide, including the risk of hearing loss, an increased likelihood of suffering from cardiovascular diseases, sleep disturbances, stress, negative emotions, and social complaints, among others [1]–[5]. Consequently, the issue of urban sound, typically dominated by traffic and industries, is nowadays a critical topic in public and scientific debates in many cities suffering from acoustic pollution everywhere and at all times [6]. These discussions focus primarily on identifying effective strategies to mitigate the current and future production of disruptive sounds in urban environments, underscoring the urgent need for proactive actions by local governments and the civil society in general [7], [8].

Traditionally, acoustic environment evaluations in cities have relied on acoustic measurements that focus solely on the negative aspects of sound, particularly when sound levels exceed certain thresholds as measured in decibels (dB) [9]. However, this approach overlooks the fact that some sounds in noisy urban settings (e.g., water, birds, or cultural sounds) can have a positive impact on people [10], as they may trigger pleasant memories or help individuals to relax and recover from daily stress and cognitive fatigue [11]. Thus, a convenient pathway to overcome the classical notion of noise and its physical measurement is the emerging concept of *soundscape*, which emphasizes the importance of the subjective perception of sound. The International Organization for Standardization (ISO) defines *soundscape* as an “acoustic environment as perceived or experienced and/or understood by a person or people, in context” [56, p. 2]. This idea represents a paradigm shift in acoustic environment assessment, allowing for the recognition of both negative and positive sound experiences [12]. In urban contexts, adopting a comprehensive perspective of soundscape, encompassing both physical and perceptual factors, is essential for identifying accurate actions that contribute to building more livable and suitable cities. For instance, creating more green areas can offer desirable and relaxing natural sounds for urban dwellers [13].

Previous studies have focused on the assessment of urban soundscapes by identifying the different sounds that comprise them [14]–[18]. Some of the most recognized categories include: *anthropophony*, which encompasses sounds inherent in human actions, from everyday activities (walking, talking, or singing) to those assisted by technology (motorized vehicles or amplified music); *biophony*, consisting of sounds made by animals (bird songs or pets); and *geophony*, produced by natural elements constitutive of the geophysical environment, such as rain or wind [19]. This taxonomy is frequently employed for the manual classification of urban soundscapes, enabling the distinction of various sounds through audio recordings. This can reveal specific sounds and/or sound sets of high social, cultural, ecological, affective, and functional values amidst the generally prevailing and undesired urban bustle [20].

However, the manual classification of soundscape is highly time-intensive, limiting the task to relatively small datasets, and its performance is suboptimal for adequately describing the multiple sounds in urban environments [21], [22]. As an alternative to the limitations of manual classification, deep learning algorithms have gained traction for the automatic identification of city sounds, thereby reducing the time and effort involved [23].

It has generally been observed that the success rates of environmental sound classification (ESC) obtained with deep learning models outperform those of non-deep learning alternatives [24]. The primary reason for this is deep learning’s ability to utilize multiple hidden nonlinear transformations to automatically derive abstract representations of raw data (e.g., images or words) [25]. However, many parameters must be adjusted in designing deep learning models, making it almost impossible to design the best model for specific tasks, which includes classifying the various sounds originating in urban areas [26]. Therefore, it has been accepted that much remains to be done regarding the adaptation of deep learning models for classifying urban soundscapes [27]–[29].

To address this gap, initial attempts in ESC research have shown that convolutional neural networks (CNNs), a type of deep learning architecture, are capable of classifying urban sounds based on spectrogram images generated from audio recordings in cities [30], [31]. For example, [32] showcased the capability of a hybrid CNN-LSTM model to classify some urban sounds according to predefined and generic sound classes in two benchmark datasets: UrbanSound8k (US8K) and Environmental Sound Classification 50 (ESC50). Another notable contribution is that of [30], who verified the capability of a CNN to classify anthropophonies, biophonies, and geophonies based on Mel spectrograms – a time–frequency visual representation of an audio signal using the Mel scale – from 5396 audio recordings collected along an urban-rural gradient in Sonoma County (USA) and auxiliary data from the Freesound dataset. Finally, it is worth mentioning the study by [33], who developed a deep neural network system to classify underwater soundscapes in an urban shore, mainly identifying biophonies, such as cetacean, fish, and marine invertebrates. However, these previous efforts do not have recourse to deep learning models for establishing the soundscape taxonomy of urban areas, and they tend to rely on public datasets (e.g., US8K and ESC50) with sounds recorded in a cleaner manner that does not register the common dynamics of overlapping sounds in these contexts.

This study investigates the application of a deep learning framework to recognize a wide range of sounds produced in real urban contexts, highlighting its value for soundscape assessment, planning, and design [34], [35]. We hypothesize that a deep learning solution is well-suited for the automatic classification of urban soundscapes, as it can recognize the visual patterns of sounds generated by real acoustic environments in spectrograms, enabling the identification

of each of those urban sounds [36], [37]. Thus, the aim is to evaluate the performance of CNNs in classifying the urban soundscape of datasets with overlapping and non-overlapping sounds.

This paper is organized as follows. First, we present three different datasets, including audio data collected by us in three different cities (overlapping scenario) and two well-known public-access audio datasets (non-overlapping scenarios), which were used to fine-tune CNNs for urban soundscape classification through a five-fold cross-validation approach. Next, we analyze the models' performance based on their classification accuracy. From this perspective, we discuss the potential of artificial intelligence (AI) models for classifying urban soundscapes. Finally, we outline our conclusions regarding automatic urban soundscape classification, as well as potential lines of research in the field.

Materials and methods

Datasets

This study utilized a self-collected dataset, referred to as *Urbanphony-3* (UP3). Additionally, the ESC50 and US8K datasets were employed, as both are extensively used in literature. To obtain UP3, soundscape data were collected in the historic areas of three cities, with the following sampling point grids: in Popayán, a 100 x 100 m grid with 50 points in five periods throughout the day; in Santiago de Compostela, a 100 x 100 m grid with 62 points in three periods throughout the day; and, in Venice, a 200 x 200 m grid with 102 points in two periods throughout the day. These sampling points were confined to historical city boundaries. In Popayán and Santiago de Compostela, the historic areas constitute only a small part of the city, whereas, in the case of Venice, the historic area coincides with the entire urban area (Fig. 1). The sampling point distributions focused on historic areas, as these locations are characterized by a variety of commercial,

residential, cultural, and recreational activities, which ensured the collection of highly varied acoustic information. At each sampling point, stereo audio recordings were taken using outdoor microphones, namely 250 recordings lasting 10 min in Popayán (Sennheiser MKH 416), 186 recordings lasting 5 min in Santiago de Compostela (Tascam DR-05X), and 204 recordings lasting 3 min in Venice (Sony ECM-XM1). In total, 640 recordings, amounting to 4054 min (67.5 h) of audio, were collected and saved in a waveform audio file format (.wav) with a sampling frequency of 48 kHz.

Regarding the other datasets, the ESC50 dataset consists of 2000 audio recordings (lasting 5 s) that are equally distributed across 50 predefined classes (40 audio samples for each class), sampled at 16 and 44.1 kHz. On the other hand, the US8K dataset is one of the most extensive urban sound samples available, containing 8732 audio clips (lasting 4 s) across ten classes, recorded at 22.05 kHz. Both datasets were recorded under controlled acoustic conditions, representing distinct sounds with minimal overlap. We utilized these two classical datasets as reference points of non-real acoustic environments in urban areas, while our main UP3 dataset featured the typical sound overlap of a real urban context.

Deep learning classification of urban soundscape

A wide range of recorded city sounds were identified and represented through spectrogram images. The distinction of each sound was made according to the urban soundscape taxonomy proposed in a previous work [38], referred to as *urbanphony*. This categorization was defined through several sound classes, grouped under the main categories of anthropophony (human sounds) and ecotophony (natural sounds) (Fig. 2). The geophony class, a type of ecotophony, was excluded due to the small number of records in the datasets. Following the adjusted urbanphony taxonomy, CNNs were trained with datasets composed of spectrograms. These spectrograms were labeled to indicate the presence or absence of specific sound subclasses, e.g.,

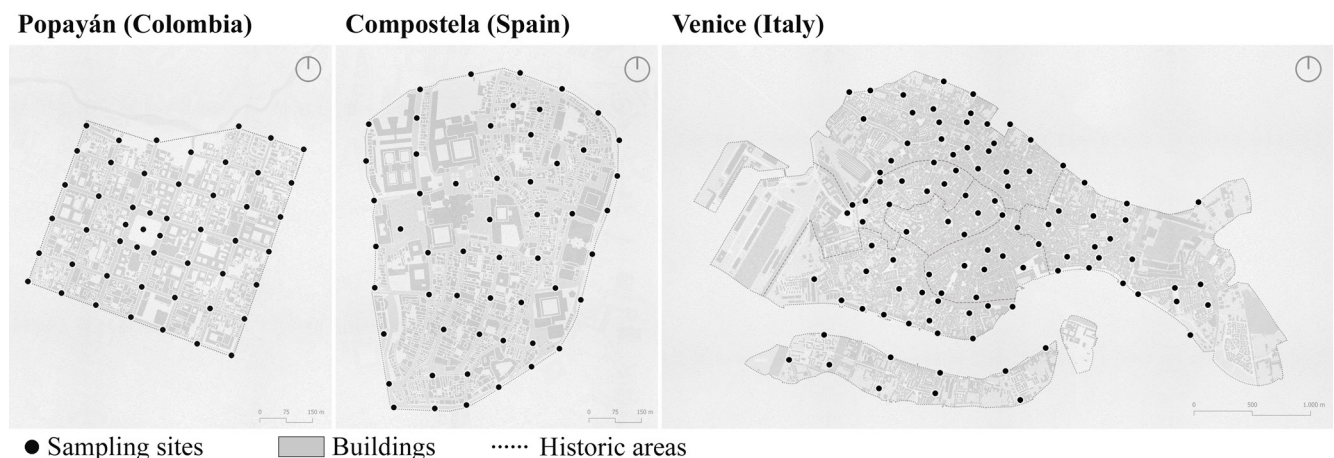


Figure 1. Sampling sites in Popayán, Santiago de Compostela, and Venice
Source: Authors

motorized transport (MT). The training of the CNNs began with the identification of basic features, such as the shape and texture of the spectrograms. Progressively, the model advanced towards extracting more complex features. During this learning process, the visual features of the extracted spectrograms were correlated with their corresponding known labels in order to classify each sound.

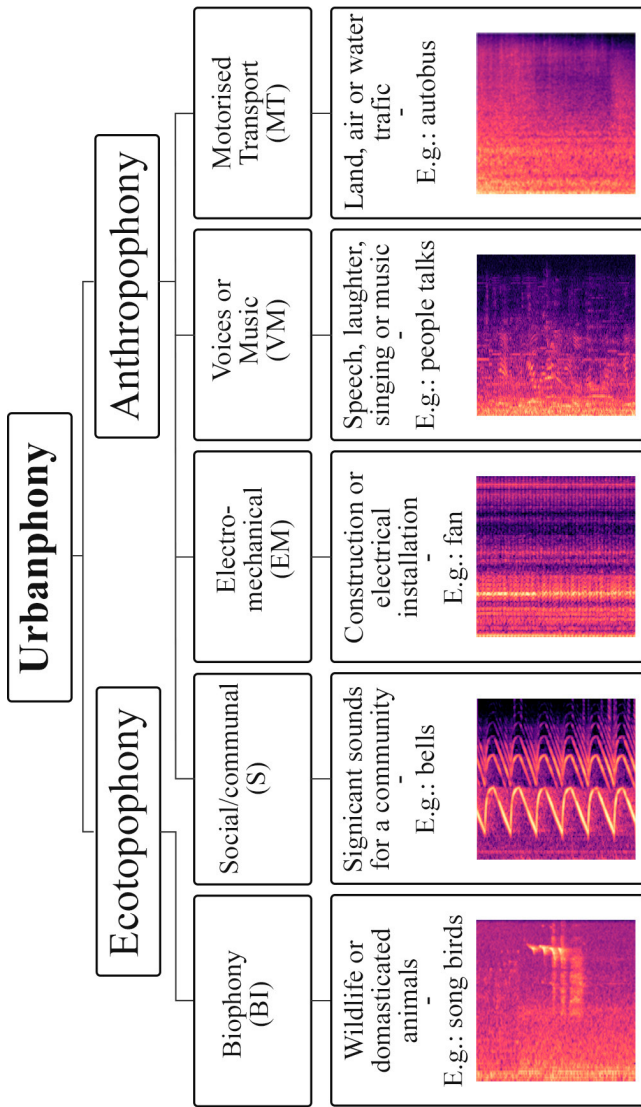


Figure 2. Urbanphony taxonomy adjusted for this study
Source: Elaborated by the authors based on [38]

Following [39], for model training, we used the well-established pre-trained architecture of EfficientNetB3, a prominent CNN that was trained on the extensive and diverse ImageNet dataset for transfer learning (<https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet>). This neural network allowed classifying complex visual patterns in spectrograms obtained from raw acoustic data.

Labeled datasets

To label UP3 according to the urbanphony taxonomy, each recording from the sampled cities was divided into

2 s segments (640 recordings). This allowed randomly selecting some segments from each recording until the dataset was completed, avoiding sequentiality in sample selection and providing representative data of each recording. Subsequently, the selected samples were labeled after being listened to and previewed in spectrograms with a range of 0-11 kHz and a window of 512 samples. This process resulted in a total of 6185 labels. To label the other datasets, the audio durations were adjusted to 2 s in order to match the length of the UP3 segments. The 5 s samples of the ESC50 dataset were split into 2 s segments, excluding the last second of each recording (for a total of 2279 segments). As for US8K, its 4 s samples were simply divided into two exact 2 s segments (14 605 segments). The predefined classes from both ESC50 and US8K were regrouped to align with the urbanphony taxonomy. Due to the quantitative imbalance of data for UP3 and ESC50, data augmentation techniques were applied, including random rotation, scaling, translation, and reflection [40]. Additionally, subsampling was applied to US8K, given its larger data volume. This approach ensured that all datasets were balanced in terms of both the data distribution across each class and the total data quantity. The test data not used during model training were set aside to evaluate model performance according to each dataset (Table I).

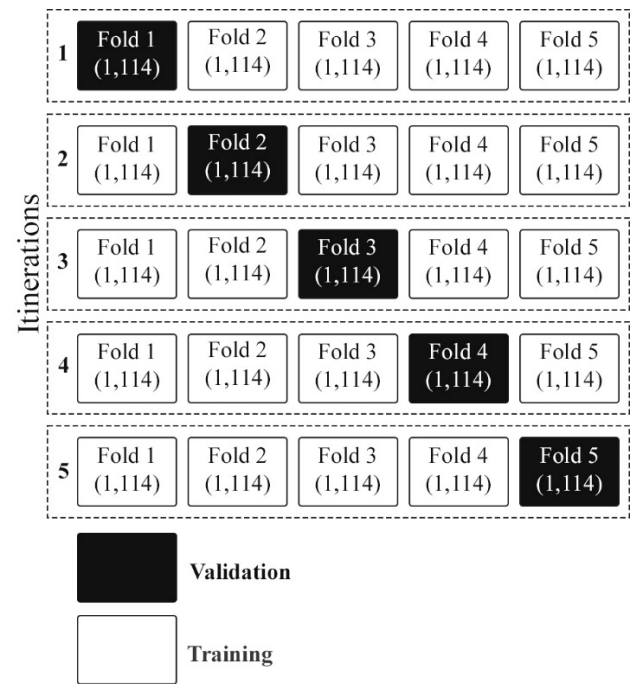


Figure 3. Cross-validation splits for each dataset in this study (UP3, ESC50, and US8K)
Source: Authors

Spectrogram generation and cross-validation

To optimally represent the labeled urban sounds in this study, the calculation of 2 s Mel spectrograms was chosen for model training. These spectrograms were obtained using

Table I. Number of sampled Mel labels in the UP3, ESC50, and US8K datasets

Datasets	Classes	2-s Mel labels	Data augmentation	Undersampling	Training set size	Testing set size	Total 2-s Mels
UP3	BI	324	972	0	1 666	130	1296
	EM	76	724	0	720	80	800
	S	150	1050	0	1080	120	1200
	TM	2945	0	1473	1325	148	1473
	VM	1902	0	1416	1283	143	1426
	Total	5 397	2746	2889	5574	621	6195
ESC50	BI	841	455	0	1 666	130	1296
	EM	478	322	0	720	80	800
	S	377	823	0	1080	120	1200
	TM	238	1 235	0	1325	148	1473
	VM	345	1 081	0	1283	143	1426
	Total	2279	3916	0	5574	621	6195
US8K	BI	1393	0	1296	1666	130	1296
	EM	5136	0	800	720	80	800
	S	1797	0	1200	1080	120	1200
	TM	2322	0	1473	1325	148	1473
	VM	3957	0	1476	1283	143	1426
	Total	14 605	0	6195	5574	621	6195

Source: Authors

version 0.10.1 of the *Librosa* library, operating under Python 3.9.7. These spectrograms covered the entire spectrum of frequencies audible to humans (0-11 KHz). The resulting images, originally sized at 600 x 600 pixels in the RGB space, were resized to 300 x 300 pixels for CNN training and testing, as this is the default input size for the pre-trained EfficientNetB3.

For modeling, a five-fold cross-validation approach was employed. To implement it, the training set of each dataset was divided into five comparable folds. After partitioning, the folds were trained and validated in five (K) iterations, using a different fold for each iteration as validation data, while the remaining folds served as training data (Fig. 4). By folding the datasets, we avoided model overfitting, as the training process in each iteration was validated with different data. After each iteration, the model's accuracy, loss, and F1 score were determined. The confusion matrix and the ROC curve were also calculated using the test data reserved for

each dataset, as well as its precision, recall, and F1 score metrics regarding each sound class.

CNN transfer learning

The CNNs developed for classification according to the urbanphony taxonomy were implemented with Python 3.9.7 and Keras 3.0., using transfer learning to address the relatively small datasets available for this study, as well as to achieve faster training times. The pre-trained EfficientNetB3 architecture was used to train different models with each dataset, employing a fine-tuning technique. To this effect, all convolutional blocks were frozen, and training was conducted only on the top layers. The fully trained CNNs produced a vector of five probabilities using the *Softmax* classifier (one value for each sound class). The best CNN models derived from cross-validation iterations for the UP3, ESC50, and US8K datasets were named as follows: UP3-CNN, ESC50-CNN, and US8K-CNN. For the hyperparameters applied, please refer to Table II.

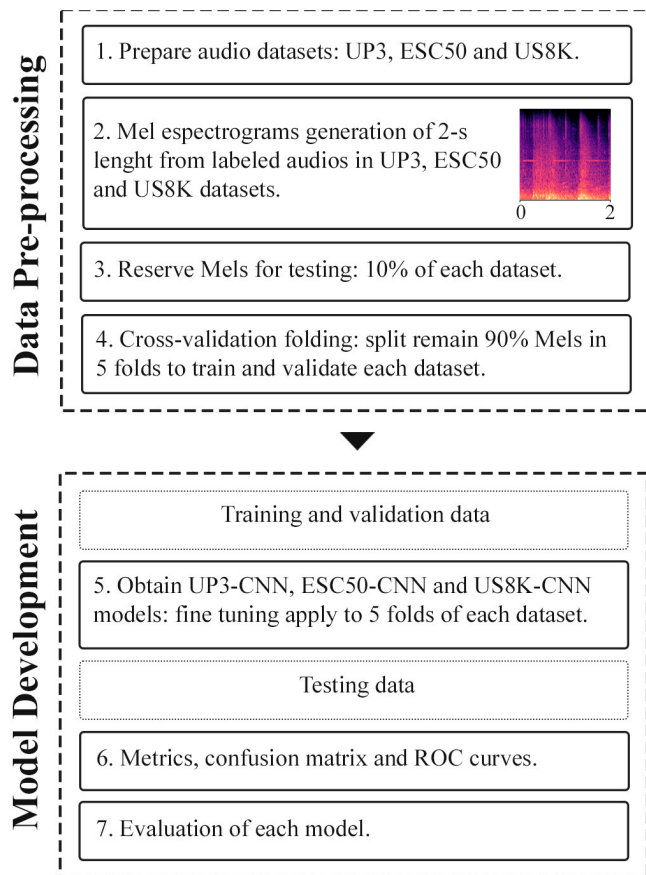


Figure 4. Workflow for data preprocessing (1-4) and model development (5-7)

Source: Authors

Table II. Hyperparameters used

Hyperparameter	Value
Learning rate	0.0001
Batch size	16
Number of epochs	100
Dropout rate	0,2
Activation function	Softmax
Optimization algorithm	Adam

Source: Authors

Results

Model performance

The performance of the models was determined from the results of cross-validation on UP3, ESC50, and US8K. Table III shows the overall results from the five-fold cross-validation in terms of the accuracy, loss, and F1-score of each dataset.

The parameters for UP3 indicate a notable performance, with a training accuracy of 72.679%, a training loss of 0.7412, and an F1-score of 0.7582. In contrast, ESC50 exhibits a training accuracy of 84.55%, a training loss of 0.4623, and an F1-score of 0.8451, while US8K reports a training accuracy of 83.30%, a training loss of 0.4680, and an F1-score of 0.8150. The acceptable performance demonstrated across all three datasets suggests that the models are effective in classifying urban sounds, although the performances of ESC50 and US8K are considerably better than that of UP3.

Table III. Overall performance of cross-validation across the datasets utilized in this study

Parameters	UP3	ESC50	US8K
Accuracy training	72.679%	84.553%	83.303%
Accuracy validation	75.179%	85.009%	82.062%
Loss training	0.7412	0.4923	0.4680
Loss validation	0.6788	0.4530	0.5109
F1-score	0.7582	0.8451	0.8150

Source: Authors

The convergence graphs for UP3-CNN, ESC50-CNN, and US8K-CNN are presented in Figs. 5, 6, and 7, respectively. In all three figures, note that the accuracy training and validation, as well as loss training and validation graphs, are very close to each other. This proves that the models do not overfit during training. In addition, US8K-CNN converges faster than other datasets.

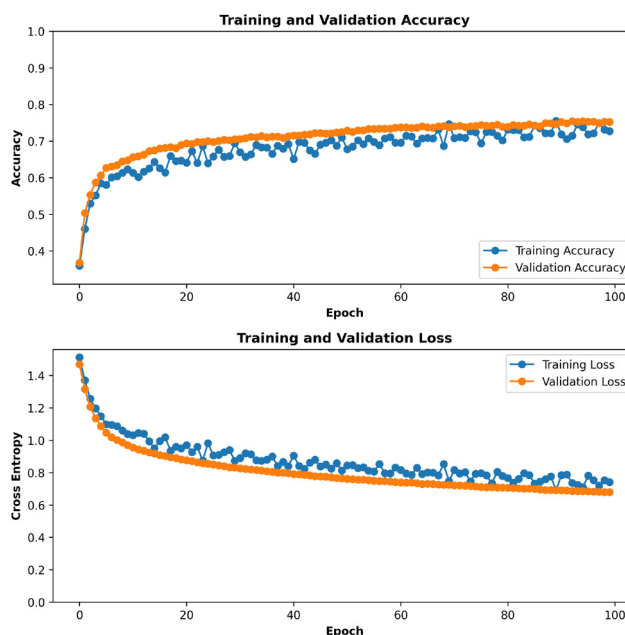


Figure 5. Convergence graphs of UP3-CNN (iteration 5)
Source: Authors

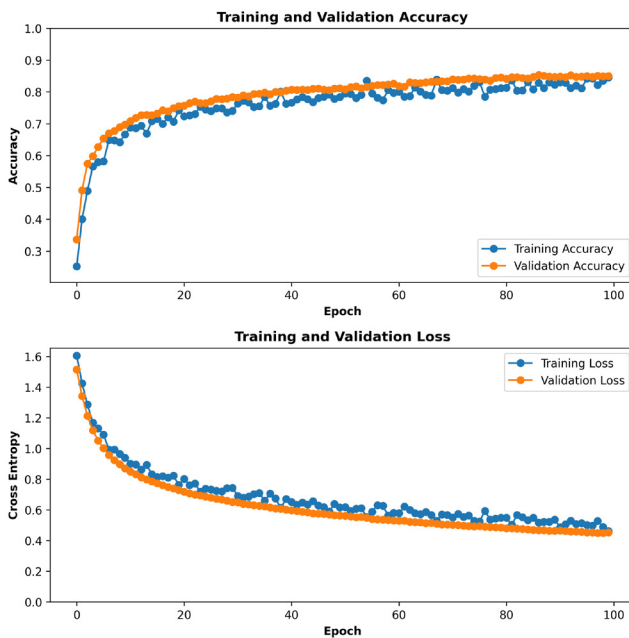


Figure 6. Convergence graphs of ESC50-CNN (iteration 4)
Source: Authors

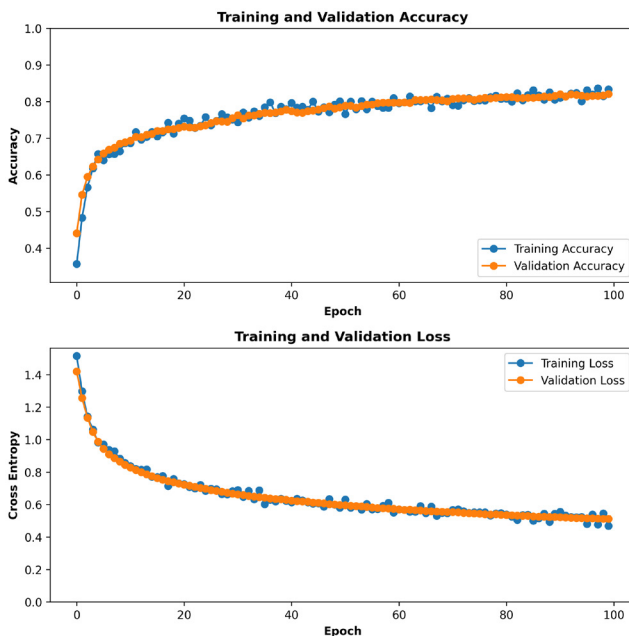


Figure 7. Convergence graphs of US8K-CNN (iteration 3)
Source: Authors

Confusion matrix and ROC curve

The confusion matrix derived from the UP3-CNN model during the test phase is shown in Fig. 8 (iteration 5). The model labeled MT with the highest accuracy (90.54%), as well as the electromechanical (EM) and biphony (BI) classes with the lowest accuracy (59.23 and 58.75%, respectively). The high accuracy regarding MT can be attributed to the distinct and prominent visual patterns present in the spectrograms,

which facilitate its recognition. In contrast, EM and BI exhibit patterns that may overlap with the dominant sound of MT. Furthermore, EM was particularly impacted by its low data availability during training, which may have affected its classification.

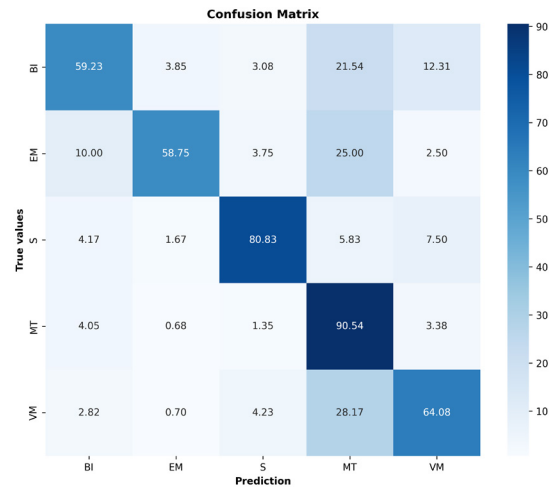


Figure 8. Confusion matrix for the accuracy value of the UP3-CNN model (iteration 5)
Source: Authors

The confusion matrix obtained for the ESC50-CNN model during the test phase is presented in Fig. 9 (iteration 4). Note that the model labels most classes in a varied manner. MT achieves the highest accuracy (98.65%), while the BI and the voices or music (VM) categories are recognized with the lowest accuracy (57.69 and 69.93%, respectively).

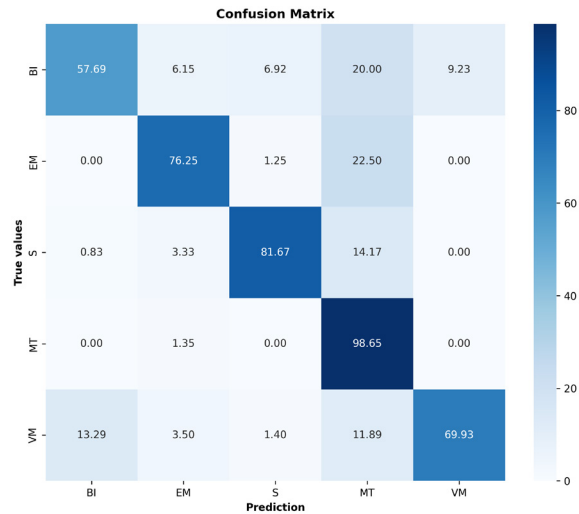


Figure 9. Confusion matrix for the accuracy value of the ESC50-CNN model (iteration 4)
Source: Authors

The confusion matrix for the US8K-CNN model during the test phase is given in Fig. 10 (iteration 3). The model identifies most of the classes similarly, with accuracy values ranging from 79.02 (VM) to 90.00% (S) for all classes. This

result is consistent with the outstanding fit demonstrated by US8K during training. The classification behavior observed in the confusion matrices of the three datasets is further corroborated in Table IV.

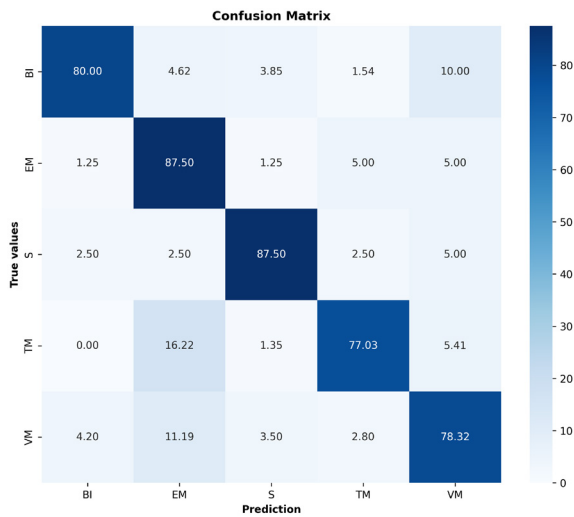


Figure 10. Confusion matrix for the accuracy value of the US8K-CNN model (iteration 3)

Source: Authors

Table IV. Metrics for the urbanphony taxonomy in each model tested

Model	Class	Precision	Recall	F1-score	
US8K-CNN	BI	0.7700	0.5923	0.6696	
	EM	0.8393	0.5875	0.6912	
	UP3-CNN	S	0.8661	0.8083	0.8362
	MT	0.5852	0.9054	0.7109	
	VM	0.7398	0.6408	0.6868	
ESC50-CNN	BI	0.7895	0.5769	0.6667	
	EM	0.7625	0.7625	0.7625	
	US8K-CNN	S	0.8909	0.8167	0.8522
	MT	0.6518	0.9865	0.7849	
	VM	0.8929	0.6993	0.7843	
UP3-CNN	BI	0.9123	0.8000	0.8525	
	EM	0.6000	0.8250	0.6947	
	US8K-CNN	S	0.8710	0.9000	0.8852
	MT	0.8864	0.7905	0.8357	
	VM	0.8014	0.7902	0.7958	

Source: Authors

The area under the ROC curve (AUC) is a significant measure of overall model performance; a higher AUC value indicates a better classification performance. To provide further insights in this regard, the ROC curves for UP3-CNN, ESC50-CNN, and US8K-CNN are presented in Figs. 11, 12, and 13, respectively. Note that the ROC curve for each model approaches the upper left corner, demonstrating a desirable performance across all datasets, albeit with slight differences. The curves show that, for UP3-CNN and ESC50-CNN, the increase in the true positive rates (TPR) at lower false positive rates (FPR) is somewhat lower, indicating a relatively reduced recall rate at low thresholds. In contrast, the ROC curve for US8K-CNN is closer to the upper left corner, reflecting a higher TPR at the same FPR, suggesting that US8K has higher sensitivity and better overall classification capabilities.

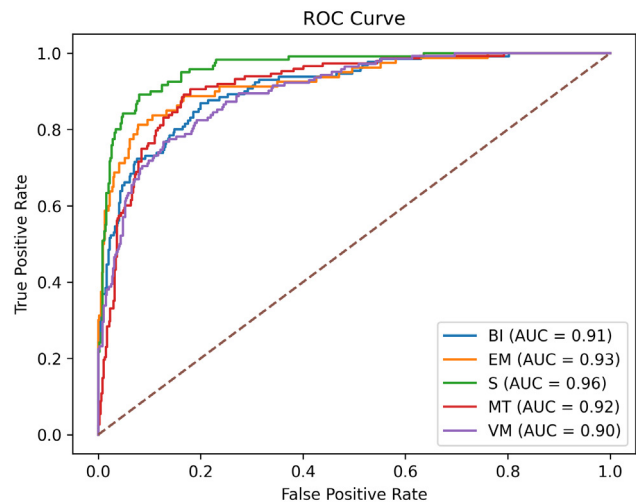


Figure 11. ROC curve for the UP3-CNN model

Source: Authors

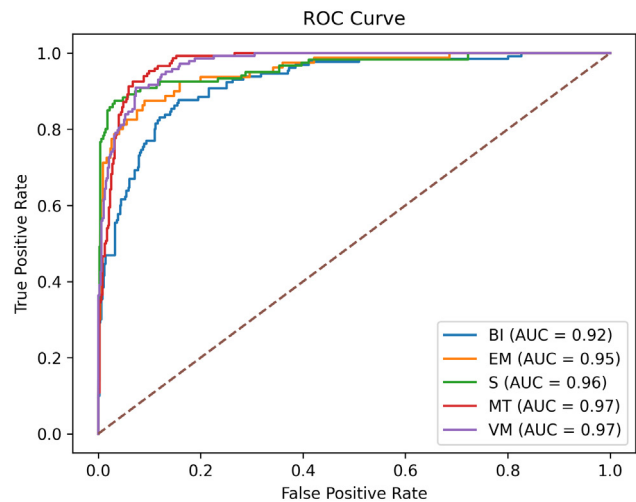


Figure 12. ROC curve for the ESC50-CNN model

Source: Authors

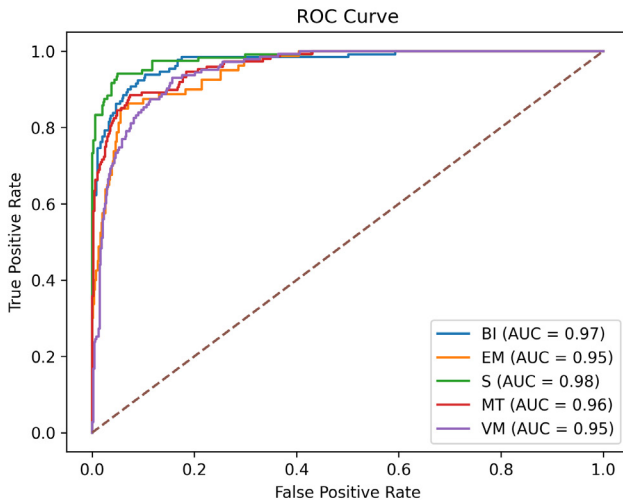


Figure 13. ROC curve for the US8K model
Source: Authors

Comparison with other studies

Table V presents a comparison between the proposed models and other deep learning solutions applied to urban sound classification. The proposed models, enhanced through transfer learning and validated by means of cross-validation, achieve competitive accuracy, including the UP3-CNN model, which utilizes real acoustic environments from historic cities with sound overlap. However, in comparison with the other models analyzed, the performance of the proposed models tends to be relatively lower. This gap can be attributed to the fact that the compared models implement dedicated data augmentation techniques [30], transformer-based architectures [40], attention mechanisms [42], and auxiliary pre-training in related audio domains [43]. Thus, it is noted that our models need to adopt similar strategies to ultimately achieve the best possible results.

Table V. Comparison of the accuracy values obtained with the proposed method against other techniques (%)

Study and datasets	Method	Accuracy
Jahangir <i>et al.</i> (2023) [40] (UrbanSound8K, ESC50, ESC10)	Transformer + data augmentation	US8K: 98%; ESC-50: 94%; ESC-10: 97%
Mushtaq and Su (2020) [41] (UrbanSound8K, ESC50, ESC10)	CNN pre-trained (DenseNet-161) + data augmentation	US8K: 97%; ESC50: 98%; ESC10: 99%
Quinn <i>et al.</i> (2022) [30] (self-collected in Sonoma County, USA + Freesound)	CNN pre-trained (MobileNetV2) + transfer learning + threshold optimization	90%
Mu <i>et al.</i> (2021) [42] (UrbanSound8K, ESC50)	CNN 1D with time-frequency attention mechanisms	US8K: 93%; ESC-50: 84%

Proposed UP3-CNN	CNN pre-trained (EfficientNetB3) + transfer learning + cross-validation	72%
Proposed ESC50-CNN	CNN pre-trained (EfficientNetB3) + transfer learning + cross-validation	84%
Proposed US8K-CNN	CNN pre-trained (EfficientNetB3) + transfer learning + cross-validation	83%

Source: Authors

Discussion

In this study, urbanphony taxonomy prediction models based on CNNs were evaluated. The cross-validation results indicate that CNN models achieve an acceptable accuracy across different datasets in predicting five classes of urban sounds represented through Mel spectrograms. This performance demonstrates that image-based approaches hold significant promise in soundscape assessment for urban planning and design, as previously suggested by [16]. The primary advantage of this approach lies in the ability to automatically identify multiple sounds from the vast amounts of acoustic information generated in cities. This classification task allowed automatically revealing the unknown and diverse sound experiences within urban environments, characterized by specific dominant sounds, including the annoying anthropophonic sounds of motorized transport or the pleasant ecotopophonic sounds coming from the biophonies of bird songs. Therefore, CNNs have the potential to uncover a broad spectrum of urban soundscapes that deserve to be preserved and promoted, given the ecological, social, and affective values of each kind of sound in them [44].

Audio recordings are relevant elements in training CNNs for automatic soundscape classification, as they are the primary resource for distinguishing the sounds that constitute a soundscape. The CNN models employed in this study classify urban sounds through Mel spectrograms derived from a self-collected UP3 audio dataset (including recordings from Popayán, Santiago de Compostela, and Venice) and the well-known ESC50 and US8K datasets. This result confirms CNNs’ ability to identify the diverse sounds produced within urban soundscapes by means of spectrograms obtained from audio recordings [45].

Remarkably, a successful classification was achieved with the self-collected UP3 dataset (UP3-CNN model), despite the fact that the audios were sourced directly from the real-world acoustic environments of urban settings and their typically chaotic and overlapping character. Thus, it is possible to move beyond the limitations of previous studies that rely solely on established audio datasets, where sounds are often recorded in quiet or controlled

environments with minimal sound overlap [46]–[50]. With this in mind, our work posits that the central challenge of urban soundscape classification with CNNs lies in ensuring that their development is grounded in real-world scenarios. Only then can these types of models be useful and deployed in related applications, as is the case with monitoring urban soundscape quality through intelligent sensor networks or similar technologies [12], [51].

It should also be noted that the relatively lower performance of UP3-CNN compared to ESC50-CNN and US8K-CNN suggests that CNNs are less efficient at classifying overlapping urban sounds, especially with regard to sound classes such as BI and EM. Therefore, the automatic classification of urban soundscapes through spectrogram representations depends on strategically managing the overlap of sounds that tend to cause greater confusion during classification [52], [53].

Furthermore, this study corroborates that CNNs can classify the diverse sounds that compose anthropophony. This finding was verified through an analysis of confusion matrices, obtaining differentiated performances across diverse anthropophonic sound classes (EM, VM, social/communal, and MT). This underscores the fact that CNNs can transcend the conventional classification of anthropophony as just a single class, which has been the norm in previous studies [54]. Overall, CNN models are flexible enough to be adapted for discriminating anthropophony into a much greater number of distinct sound classes, a capability that is essential in urban contexts, where human-generated sounds are the predominant component of the soundscape and significantly influence urban sound perceptions.

Conclusions

This study leverages the efficiency and performance of the proposed CNN models in soundscape classification in acoustic environments with and without sound overlap in urban areas. The best models for the UP3, ESC50, and US8K datasets were UP3-CNN, ESC50-CNN, and US8K-CNN, with accuracy rates of 75.6, 84.55, and 83.30%, respectively. Indeed, the models revealed their ability to address the automatic broad taxonomic classification of urban soundscapes, using Mel spectrograms to differentiate urban sounds in data from real acoustic environments (with sound overlap) and other well-known datasets (without sound overlap).

The limitations of this study lie in its small number of labeled samples from real-world urban soundscapes and in the mismatch between predefined classes in public datasets and standard urban soundscape taxonomy. Future studies should address this limitation by employing more dedicated data augmentation techniques or similar methods. Another important future direction is to improve the classification of urban sounds that frequently overlap, such as S and MT, by adopting advanced labeling strategies, such as multi-label annotations. Likewise, expanding the urban soundscape

taxonomy for automatic classification is another path to follow. For instance, it would be beneficial to include different kinds of vocal sounds, depending on whether they originate from people talking, laughing, or singing. Apart from the recognition of sound classes, other aspects of urban soundscapes, such as emotional responses to sound or cultural sound preferences across different urban contexts, should also be considered.

Finally, we believe it to be of particular importance that we have designed and used CNN models to automatically detect sounds based on an urban soundscape taxonomy while including mixed sounds from the environment. This opens an opportunity to think about the adaptation of other AI models for managing the classification of massive amounts of acoustic information in urban settings. By developing such adaptable models, we are confident that a better understanding of urban soundscapes can be achieved, thereby informing strategies that enhance human health, livability, and well-being in the cities of tomorrow.

Acknowledgements

This work was supported by a scientific research strategy on soundscapes known as *Mantra*, along with its main associated project, *Urbanphony*, which was funded by the Ministry of Science of Colombia (Contract 2022-0650) and implemented by Institución Universitaria Colegio Mayor del Cauca. Special gratitude is extended to the Vice-Principal of this university, Paola Umaña Aedo, for her profound emotional and reasoned commitment to *Mantra* and *Urbanphony*. Deep appreciation is also conveyed to Carlos Realpe, Pablo Garzón, Hermin Castro, Sebastián Agredo, Santiago Prieto, and Jineth Córdoba for their invaluable assistance in data labeling for training our CNN models. In the same manner, acknowledgment is given to the student group of Universidad del Cauca's Aerospace Electronic System Society for their voluntary involvement during the analysis and discuss of the results of this research. Finally, we thank the peer reviewers for their contributions to improving the quality of this paper and the *Ingeniería e Investigación* journal for the opportunity granted.

Data and materials availability

The source code for *Urbansonic-CNNs*, including all scripts required to reproduce the results, is available at <https://zenodo.org/records/15523650>, and the corresponding dataset can be accessed via <https://zenodo.org/uploads/15499672>

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have had any bearing on the work reported in this paper.

CRedit author statement

Carlos Durán: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing (original draft).

Julián Grijalba: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, supervision, visualization, writing (review and editing).

Sebastián Cajas: methodology, data curation, formal analysis, investigation, software, writing (review and editing).

Camilo Ferreira: supervision, formal analysis, investigation, writing (review and editing).

References

- [1] M. Basner *et al.*, "Auditory and nonauditory effects of noise on health," *The Lancet*, vol. 383, no. 9925, pp. 1325–1332, 2014. [https://doi.org/10.1016/S0140-6736\(13\)61613-X](https://doi.org/10.1016/S0140-6736(13)61613-X)
- [2] G. Licitra, L. Fredianelli, D. Petri, and M. A. Vigotti, "Annoyance evaluation due to overall railway noise and vibration in Pisa urban areas," *Sci. Total Environ.*, vol. 568, pp. 1315–1325, 2016. <https://doi.org/10.1016/j.scitotenv.2015.11.071>
- [3] L. C. Erickson and R. S. Newman, "Influences of background noise on infants and children," *Curr. Dir. Psychol. Sci.*, vol. 26, no. 5, pp. 451–457, 2017. <https://doi.org/10.1177/0963721417709087>
- [4] C. M. Tiesler *et al.*, "Exposure to road traffic noise and children's behavioural problems and sleep disturbance: Results from the GINIplus and LISApplus studies," *Environ. Res.*, vol. 123, pp. 1–8, 2013. <https://doi.org/10.1007/s40572-015-0044-1>
- [5] N. Bilenko *et al.*, "Traffic-related air pollution and noise and children's blood pressure: Results from the PIAMA birth cohort study," *Eur. J. Prevent. Card.*, vol. 22, no. 1, pp. 4–12, 2015. <https://doi.org/10.1177/2047487313505821>
- [6] N. ElBardisy, "An analytical investigation of environmental awareness about noise and visual pollution inside the Egyptian context," *Discover Cities*, vol. 2, art. 16, 2025. <https://doi.org/10.1007/s44327-025-00060-8>
- [7] H. Wang *et al.*, "Urban network noise control based on road grade optimization considering comprehensive traffic environment benefit," *J. Environ. Manage.*, vol. 364, art. 121451, 2024. <https://doi.org/10.1016/j.jenvman.2023.121451>
- [8] L. Fredianelli *et al.*, "Traffic flow detection using camera images and machine learning methods in ITS for noise map and action plan optimization," *Sensors*, vol. 22, no. 5, art. 1929, 2022. <https://doi.org/10.3390/s22051929>
- [9] A. Can, A. L'Hostis, P. Aumond, D. Botteldooren, M. C. Coelho, C. Guarnaccia, and J. Kang, "The future of urban sound environments: Impacting mobility trends and insights for noise assessment and mitigation," *App. Acous.*, vol. 170, art. 107518, 2020. <https://doi.org/10.1016/j.apacoust.2020.107518>
- [10] R. M. Schafer, *The soundscape: Our sonic environment and the tuning of the world*. Rochester, VT, USA: Destiny Books, 1993.
- [11] J. Wang, C. Li, Y. Lin, C. Weng, and Y. Jiao, "Smart soundscape sensing: A low-cost and integrated sensing system for urban soundscape ecology research," *Environ. Tech. Innov.*, vol. 29, art. 102965, 2023. <https://doi.org/10.1016/j.eti.2022.102965>
- [12] J. B. López Giler and C. S. Casquete Baidal, "¿Cómo plantear un proyecto de urbanismo que disminuya el impacto ambiental y ofrezca calidad?," *E-IDEA J. Eng. Sci.*, vol. 3, no. 6, pp. 17–32, 2021. <https://doi.org/10.53734/esci.vol3.id177>
- [13] Y. Xiang, Q. Meng, X. Zhang, M. Li, D. Yang, and Y. Wu, "Soundscape diversity: Evaluation indices of the sound environment in urban green spaces – effectiveness, role, and interpretation," *Ecol. Ind.*, vol. 154, art. 110725, 2023. <https://doi.org/10.1016/j.ecolind.2023.110725>
- [14] J. Y. Hong and J. Y. Jeon, "Relationship between spatiotemporal variability of soundscape and urban morphology in a multifunctional urban area: A case study in Seoul, Korea," *Build. Environ.*, vol. 126, pp. 382–395, 2017. <https://doi.org/10.1016/j.buildenv.2017.10.021>
- [15] H. I. Jo and J. Y. Jeon, "Urban soundscape categorization based on individual recognition, perception, and assessment of sound environments," *Landsc. Urban Plan.*, vol. 216, p. 104241, 2021. <https://doi.org/10.1016/j.landurbplan.2021.104241>
- [16] X. Fang, Y. Qi, M. Hedblom, T. Gao, and L. Qiu, "Do soundscape perceptions vary over length of stay within urban parks?," *J. Outdoor Rec. Tourism*, vol. 45, p. 100728, 2024. <https://doi.org/10.1016/j.jort.2023.100728>
- [17] Y. Zhang, C. Wang, and Z. Sun, "Soundscape diversity: Evaluation indices of the sound environment in urban green spaces – Effectiveness, role, and interpretation," *Ecol. Ind.*, vol. 147, art. 109063, 2023. <https://doi.org/10.1016/j.ecolind.2023.110725>
- [18] B. C. Pijanowski *et al.*, "Soundscape ecology: The science of sound in the landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, 2011. <https://doi.org/10.1525/bio.2011.61.3.6>
- [19] Y. Jia, H. Ma, J. Kang, and C. Wang, "The preservation value of urban soundscape and its determinant factors," *App. Acous.*, vol. 168, art. 107430, 2020. <https://doi.org/10.1016/j.apacoust.2020.107430>
- [20] H. I. Jo and J. Y. Jeon, "Effect of the appropriateness of sound environment on urban soundscape assessment," *Build. Environ.*, vol. 179, art. 106975, 2020. <https://doi.org/10.1016/j.buildenv.2020.106975>
- [21] S. Korpilo *et al.*, "Landscape and soundscape quality promote stress recovery in nearby urban nature: A multisensory field experiment," *Urban For. Urban Green.*, art. 128286, 2024. <https://doi.org/10.1016/j.ufug.2024.128286>
- [22] T. Ozseven, "Investigation of the effectiveness of time-frequency domain images and acoustic features in urban sound classification," *App. Acous.*, vol. 211, art. 109564, 2023. <https://doi.org/10.1016/j.apacoust.2023.109564>
- [23] A. Arnault, B. Hanssens, and N. Riche, "Urban sound classification: Striving towards a fair comparison," 2020. [Online]. Available: <https://doi.org/10.48550/arXiv.2010.11805>
- [24] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Net.*, vol. 61, pp. 85–117, 2015. <https://doi.org/10.1016/j.neunet.2014.09.003>

- [25] Ö. İnik, "CNN hyperparameter optimization for environmental sound classification," *App. Acous.*, vol. 202, art. 109168, Jan. 2023. <https://doi.org/10.1016/j.apacoust.2022.109168>
- [26] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. 2015 IEEE 25th Int. Work. Machine Learn. Signal Process. (MLSP)*, Boston, MA, USA, Sep. 2015, pp. 1–6. <https://doi.org/10.1109/MLSP.2015.7324337>
- [27] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Proc. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017. <https://doi.org/10.1109/LSP.2017.2657381>
- [28] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Sci. Rep.*, vol. 11, art. 21552, Nov. 2021. <https://doi.org/10.1038/s41598-021-01045-4>
- [29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015. <https://doi.org/10.1038/nature14539>
- [30] C. A. Quinn *et al.*, "Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data," *Ecol. Ind.*, vol. 138, art. 108831, 2022. <https://doi.org/10.1016/j.ecolind.2022.108831>
- [31] D. V. Devalraju and P. Rajan, "Multiview embeddings for soundscape classification," *IEEE/ACM Trans. Audio Speech Lang Proc.*, vol. 30, pp. 1197–1206, 2022. <https://doi.org/10.1109/TASLP.2022.3153272>
- [32] B. Mishachandar and S. Vairamuthu, "Diverse ocean noise classification using deep learning," *Applied Acoustics*, vol. 181, art. 108141, 2021. <https://doi.org/10.1016/j.apacoust.2021.108141>
- [33] R. M. Rehan, "The phonic identity of the city urban soundscape for sustainable spaces," *HBRC J.*, vol. 12, no. 3, pp. 337–349, 2016. <https://doi.org/10.1016/j.hbrj.2014.12.005>
- [34] A. Karapostoli and N. E. Votsi, "Urban soundscapes in the historic centre of Thessaloniki: Sonic architecture and sonic identity," *Sound Stud.*, vol. 4, no. 2, pp. 162–177, 2018. <https://doi.org/10.1080/20551940.2019.1582744>
- [35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Net.*, vol. 61, pp. 85–117, 2015. <https://doi.org/10.1016/j.neunet.2014.09.003>
- [36] A. China Manrique de Lara, "On the theory of deep learning: A theoretical physics perspective (Part I)," *Phys. A Stat. Mech. App.*, vol. 632, art. 129308, 2023. <https://doi.org/10.1016/j.physa.2023.129308>
- [37] J. A. Grijalba-Obando and V. Paül-Carril, "La influencia del paisaje sonoro en la calidad del entorno urbano. Un estudio en la ciudad de Popayán (Colombia)," *Urbano*, vol. 21, no. 38, pp. 70–83, 2018. <https://doi.org/10.22320/07183607.2018.21.38.06>
- [38] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. on Machine Learning (ICML)*, PMLR, vol. 97, pp. 6105–6114, 2019. <https://proceedings.mlr.press/v97/tan19a.html>
- [39] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecol. Infor.*, vol. 57, art. 101084, May 2020. <https://doi.org/10.1016/j.ecoinf.2020.101084>
- [40] R. Jahangir, M. A. Nauman, R. Alroobaea, J. Almotiri, M. M. Malik, and S. M. Alzahrani, "Deep learning-based environmental sound classification using feature fusion and data enhancement," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1069–1091, 2023. <https://doi.org/10.32604/cmc.2023.032719>
- [41] Z. Mushtaq and S.-F. Su, "Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images," *Symmetry*, vol. 12, no. 11, art. 1822, 2020. <https://doi.org/10.3390/sym12111822>
- [42] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Scientific Reports*, vol. 11, art. 21552, 2021. <https://doi.org/10.1038/s41598-021-01045-4>
- [43] M. Towsey, J. Wimmer, I. Williamson, and P. Roe, "The use of acoustic indices to determine avian species richness in audio-recordings of the environment," *Ecological Informatics*, vol. 21, pp. 110–119, 2014. <https://doi.org/10.1016/j.ecoinf.2013.11.007>
- [44] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1041–1044. <https://doi.org/10.1145/2647868.2655045>
- [45] B. Bahmei, E. Birmingham, and S. Arzanpour, "CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification," *IEEE Signal Proc. Lett.*, vol. 29, pp. 682–686, 2022. <https://doi.org/10.1109/LSP.2022.3150258>
- [46] M. Bubashait and N. Hewahi, "Urban sound classification using DNN, CNN & LSTM a Comparative Approach," in *2021 Int. Conf. Innov. Intel. Infor., Comp, Tech. (3ICT)*, Zallaq, Bahrain, 2021, pp. 46–50. <https://doi.org/10.1109/31CT53449.2021.9581339>
- [47] Z. Mushtaq, S. F. Su, and Q. V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," *Applied Acoustics*, vol. 172, p. 107581, 2021. <https://doi.org/10.1016/j.apacoust.2020.107581>
- [48] A. A. Rahman and J. Angel Arul Jothi, "Classification of UrbanSound8k: A Study Using Convolutional Neural Network and Multiple Data Augmentation Techniques," in *Soft Computing and its Engineering Applications*, K. K. Patel, D. Garg, A. Patel, and P. Lingras, Eds., *Communications in Computer and Information Science*, vol. 1374. Singapore: Springer, 2021. https://doi.org/10.1007/978-981-16-0708-0_5
- [49] B. Z. J. L. S. Thornton, "Audio recognition using Mel spectrograms and convolution neural networks," unpublished manuscript, 2019.
- [50] C. Mydlarz, M. Sharma, Y. Lockerman, B. Steers, C. Silva, and J. P. Bello, "The life of a New York City noise sensor network," *Sensors*, vol. 19, art. 1415, 2019. <https://doi.org/10.3390/s19061415>

- [51] S. Park, D. K. Han, and M. Elhilali, "Cross-referencing self-training network for sound event detection in audio mixtures," *IEEE Trans. Multimedia*, vol. 25, no. 10, pp. 4573–4585, Oct. 2023. <https://doi.org/10.1109/TMM.2023.3197123>
- [52] S. Zhang, Y. Zhang, Y. Liao, K. Pang, Z. Wan, and S. Zhou, "Polyphonic sound event localization and detection based on multiple attention fusion ResNet," *Mathematical Biosci. Eng.*, vol. 21, no. 2, pp. 2004–2023, 2024. <https://doi.org/10.3934/mbe.2024089>
- [53] A. Lie *et al.*, "Occupational noise exposure and hearing: A systematic review," *Int. Arch. Occup. Environ. Health*, vol. 89, pp. 351–372, 2016. <https://doi.org/10.1007/s00420-015-1083-5>
- [54] *Acoustics — Soundscape — Part 1: Definition and conceptual framework*, ISO 12913-1:2014, ISO, Geneva, Switzerland, 2014.