Research Article

Deepank Verma*, Arnab Jana, and Krithi Ramamritham

Classification and mapping of sound sources in local urban streets through AudioSet data and Bayesian optimized Neural Networks

https://doi.org/10.1515/noise-2019-0005
Received Mar 23, 2019; accepted Jun 21, 2019

Abstract: Deep learning (DL) methods have provided several breakthroughs in conventional data analysis techniques, especially with image and audio datasets. Rapid assessment and large-scale quantification of environmental attributes have been possible through such models. This study focuses on the creation of Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) based models to classify sound sources from manually collected sound clips in local streets. A subset of an openly available AudioSet data is used to train and evaluate the model against the common sound classes present in the urban streets. The collection of audio data is done at random locations in the selected study area of 0.2 sq. km. The audio clips are further classified according to the extent of anthropogenic (mainly traffic), natural and human-based sounds present in particular locations. Rather than the manual tuning of model hyperparameters, the study utilizes Bayesian Optimization to obtain hyperparameter values of Neural Network models. The optimized models produce an overall accuracy of 89 percent and 60 percent on the evaluation set for three and fifteen-class model respectively. The model detections are mapped in the study area with the help of the Inverse Distance Weighted (IDW) spatial interpolation method.

1 Introduction

Sounds are dynamic and persistent property of the environment [65]. The assessment of environmental sound characteristics has been an essential part of research in urban planning and design [9, 32, 58], environmental monitoring [61, 64, 65], and eco-acoustics [79, 82]. In recent years, the research on evaluation of various acoustic landscapes (soundscapes) has been gaining momentum especially in the built environment and individual health contexts, mainly due to its relevance in rapidly changing urbanscapes [5, 40, 41].

1.1 Relevance of sound sources in soundscapes assessment

Majority of previous studies have evaluated the people’s perception of sounds in a limited context of noise. Further, the quality of the urban living environment has been consistently being assessed through noise annoyance and its prolonged exposure to the inhabitants [14, 17, 52]. Several studies have focused on the understanding of the interrelations between the noise-inducing sounds such as traffic and airplane and its effect on the individuals’ activity disturbance such as speech interference and sleep deprivation [47] leading to long-term mental and physical health deterioration [37, 77]. The use of Sound Pressure Level (SPL) as a quantitative metric to measure noise intensity has been widely accepted as a part of such studies and noise management policies such as the Environmental Noise Directive (END) [19]. However, in recent years, due to the awareness of people and local authorities and the desire to improve the sound environment, the focus of new researches has been shifted to a proper understanding of the auditory realm and the means to evaluate them. The recent studies in “Soundscapes,” [71] provides an alternative viewpoint in the evaluation of auditory landscape.

Soundscapes has been defined as an “acoustic environment as perceived or experienced and/or understood by a
person or people, in context” [36]. As Davies et al. [16] explains, soundscape evaluation is simply how a listener categorizes the sounds and the feelings and emotions which it stimulates. More specifically, the studies in soundscapes assessment comprise of physical and social dimensions defined by acoustical characteristics of sounds and the perception of sounds by humans, respectively [35]. In soundscape studies, the assessment of soundscapes is carried out with the help of responses gathered regarding perceptual attributes and sound sources from soundwalks. While, the sound characteristics such as the psychoacoustic properties are generally collected with the help of binaural audio recordings [3, 42]. Recent studies have analyzed soundscape as a combination of entities such as experienced, acoustic and Extra-acoustic environment which included audio, photo and video recordings for soundscape assessment [44, 45].

The accessibility of an individual to the pleasant sounds positively affect wellbeing and helps in attention restoration [4, 41, 83]. Further, listening to natural sounds such as birds, water streams, and rustling leaves have been found to contribute to overall well-being and in creation of health-promoting urban environments [25, 29]. The studies involving assessment of soundscapes have helped in understanding the thoughts and emotions of the individuals regarding the particular sonic scene [16]. More specifically, the individuals’ perceptions of the acoustic environment have been found to correlate with the composition of the scene with natural and human-made sound sources [6, 46]. Studies have explored the relationships between the perceptual attributes and the presence of specific sounds such as bird vocalization, sounds of vehicles, water, wind, construction noise, human-based sounds such as conversation and walking, and street music [8, 33, 80]. With the help of responses gathered from the people at large, these sounds can be categorized as positive and negative sounds [35]. Kawai et al. [43], conducted an experiment which included participants which grouped various sound sources into larger groups such as natural, transportation, and the household which was further utilized in the psychological evaluation of environmental sounds. Liu et al. [50] studied the spatiotemporal changes in the presence of different sources of sound in a multifunctional urban area. With the help of the raters, recorded sound clips were evaluated on the basis of anthropophony, biophony, and geophony and its variation over the selected area. Similarly, Hong et al. [34] investigated the variability of soundscapes in different urban built morphologies such as Business areas, low and high-density commercial, and residential. Kang et al. [42] proposed a model for soundscape mapping in smart cities and conducted soundwalks to capture the responses on the presence of common street sounds, and further, studied relationship between them with the auditory perceptual attributes.

The soundscape studies are mainly dependent on the manual labeling of sound sources collected from the environment with the help of survey respondents [34, 44, 50, 81]. These studies, therefore have been subjected to inherent human biases and constrained to few observation sites due to the limited respondents to participate in soundscape research.

### 1.2 Advances in the computer-based classification of sound sources

Automated extraction of sound sources has been mainly implemented as a part of eco-acoustics and environmental monitoring [1, 20]. The sounds have been collected with the help of microphones installed for extended durations to understand species abundance in natural habitats [21, 61]. While the studies have successfully implemented several classification models to classify biophonic sounds [1, 10, 78, 82], classification of common everyday sounds are comparatively complex even by state of the art computer algorithms. Ambient noise and overlapping sound sources in unstructured environments such as urban streets pose computational challenges in the implementation of well-established sound classification algorithms [11, 18] such as Gaussian Mixture Models (GMM) [7] and Hidden Markov Models (HMM) [30, 51].

In recent years, Deep Learning (DL) has been utilized to remodel almost every data analysis task. DL models are a part of broader family of Machine Learning (ML) methods, in which deep architectures are used to find representations from the large datasets required for classification tasks [48]. DL based visual analysis tasks such as object detection and semantic segmentation have shown near human accuracy in identifying the visual elements in the surroundings. However, this has not been entirely true for audio classification tasks — the primary reason being the unavailability of large annotated datasets corresponding to the common sound classes. Researchers have extensively utilized the annotated datasets such as UrbanSound8k [70], ESC-50 [63], TUT sound events database [55] for urban-sounds related research. Such datasets have been widely used in benchmarking the performance of different classification models [12, 62]. However, these datasets include fewer sound classes corresponding to common urban street sounds, therefore not adequate for the creation of a generalized street-level sound classification model.
which can be utilized for large-scale urban studies. Recently AudioSet [24], a large corpus of audio data collected from YouTube clips by Google is released which consists of 632 audio event classes comprising of a wide range of common everyday sounds present in indoors and outdoors such as human and animal sounds, music, sounds of instruments, traffic, and vehicles. A subset of AudioSet data is utilized in this study for model training.

2 Overview of Deep Learning Models

Among the various Deep Learning (DL) models, Convolutional Neural Networks (CNN) have shown extraordinary performance in image and video classification tasks. CNNs learn the relevant features from the spatial representation of the input data to frame meaningful relations between dataset and the corresponding labels. The Mel-frequency cepstral coefficients (MFCC) based audio representation is commonly utilized in audio classification tasks to leverage the benefits of the CNN Based Classifiers [11, 68, 69]. Similar methodology has been followed while preprocessing sound samples in AudioSet data [24]. A 128-dimensional high-level embedding for each second of the segment obtained from custom VGG [72] CNN (vggish) architecture is provided as a part of AudioSet rather than the raw audio samples (More on [31]). These embeddings are the simplified features of complex sound samples which can be used as an input to other machine learning models.

In this study, the sound classification task is implemented with the help of Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN). ANNs are the simplest form of DL algorithms in which the nodes carrying input data in the form of a layer are fully connected with the nodes in hidden and output layers. Such connections have weights, which recursively update itself with the help of forward-passes, loss functions, activations, and backpropagation algorithm. These operations are used in DNNs with varying levels of complexity. Recurrent Neural Networks are types of DNNs which are efficient in training the sequential input data such as sound, text and time series [48]. RNNs learn the important features from the incoming data sequences and store it in memory cells to further predict the next time steps (More on [49]). One of the widely studied RNN models, Long Short-Term Memory (LSTM) is utilized in this study. LSTM fixes the loopholes in vanilla RNN models such as vanishing and exploding gradients [60]. The workings of ANN and LSTM models are discussed widely in literature such as [27]; therefore, in this study, we have only discussed the modifications made in the final prepared models.

2.1 Hyperparameter optimization

Manual hyperparameter tuning in Machine Learning Algorithms, especially DNNs requires prior experience, rules of thumb and expert knowledge [73]. While gathering the prior knowledge to set up a working model with the help of past literature is not difficult, however, to find the optimum parameters in the especially novel datasets is often challenging.

Typically, optimization problem includes three parts: (a) Objective function such as model loss or accuracy in machine learning algorithms, which is to be minimized or maximized. (b) Domain space, which includes the range of hyperparameter values to be evaluated. (c) The optimization algorithm, which consists of the method to choose the hyperparameter values, evaluate the objective function and iterate over the domain space. Common hyperparameter optimization methods such as Grid search and Random Search evaluate the entire search space, which for complex models and a large number of hyperparameters is often infeasible. These methods involve independent evaluations of the model for each set of parameters from a search space which requires a more extended run time and hence computationally expensive. Further, the choice of the next set of hyperparameters is completely unrelated to the past selection and the results from evaluations, thus not efficient to conduct optimizations for large models.

Recently, Bayesian Optimization (BO) methods have been utilized for complex Machine learning problems [13]. BO is a method to solve black-box optimization problems which involve computationally expensive calculations [23, 74]. On the contrary to the Grid and random based optimization methods, BO methods utilize the results from past evaluations in the search space to choose the next optimal values to evaluate the objective function. In this way, the model selects the set of hyperparameters based on an earlier set of inputs which have performed better at evaluation thus limiting poor hyperparameter choices.

BO starts with creating a simpler surrogate model $f^*$ of the unknown objective function $f$ and incorporates the prior belief in searching the hyperparameter values in domain space. It further employs an acquisition function which is computed from the $f^*$ and used for guiding the selection of the next evaluation point [22]. The objective function is modeled with the help of Gaussian Process (GP) algorithm. GPs are the generalization of a Gaussian distribu-
tion to distribution over functions instead of random variables (More on [67]). The acquisition functions address exploitation vs. exploration trade-off in which the function trades with an exploration of optimal areas with an exploration of the unexplored regions [23]. One of the examples of acquisition function, Expected Improvement (EI) is used in this study.

The hyperparameters of ANN and LSTM models are optimized using BO, the details including the selection of domain space and calculated optimized values are discussed in subsequent sections.

3 Objectives of the study

This objective of this paper is to create a methodology to classify street-level urban sounds with the help of DL algorithms. The step-by-step process includes: (1) Creation of subset of the large corpus of openly available and annotated AudioSet data, (2) Preparation of LSTM and ANN models and comparison of accuracy between the two, (3) Utilization of Bayesian Optimization to select hyperparameters for the model, (4) Collection of sound clips from various locations from local study area, (5) Manual assessment of the model predictions, and (6) the Preparation of sound maps for the study area. The flowchart showing the methodology is given in Figure 1.

![Figure 1: Flowchart showing the methodology of the study.](https://research.google.com/audioset/ontology/index.html)

4 Data preparation

The complete AudioSet data is available in three segments: (a) Evaluation and (b) Balanced train, which consists of approx. 20000 segments each including approx. 59 samples from 527 sound classes, and (c) Unbalanced train, which includes the remaining 2 million segments from all the available classes with a variable number of samples in each class. The length of each segment is 10 seconds and has been manually labeled using structured hierarchical ontology [24]. The dataset is curated from YouTube videos uploaded from all over the world, which ensures the mix quality of sounds and therefore, increases the generalization of the existence of a particular sound class in real life scenarios.

As AudioSet data comprises of a wide variety of sound classes, the model training including all the classes would result in increasing computational costs and the inability of the trained classifier to interpret correct street-level sound classes. Therefore, we selected 31 classes from the 632 audio event classes which represent the sounds that are typically present in the streets. Appendix 1 provides details regarding the selected AudioSet classes. The column “Sound Classes” provides the name of the sound classes in AudioSet data; the numbers preceding the classes are the original class numbers as present in the data. The audio classes in AudioSet are arranged in a hierarchical fashion with a maximum depth of six levels (e.g., Sounds of things > Vehicle > Motor Vehicle (Road) > Car > Vehicle horn > Toot). In this study, we classify the classes at the level-sixth node as the child classes and the classes at the higher hierarchy as the parent classes. The “Relation” column in Appendix 1 indicates this hierarchical setup; further, the numbers in brackets are provided for the parent classes which indicate the class number of the respective child classes. The quality assessment of each class in AudioSet has been already conducted by the experts and the “Data quality” rating indicating high, medium and low is supplied with the data (More details on https://research.google.com/audioset/download.html). Appendix 1 also provides the number of samples present in unbalanced train segment of AudioSet data for each of the 31 classes which can be used for model training purposes.

To further assess the variability and uniformity among the audio samples in 31 selected classes, we manually selected 10 random samples from each of the classes and listened to the corresponding audio from AudioSet website (https://research.google.com/audioset/ontology/index.html). The consistency among the samples of the
Table 1: Table showing domain space and optimized values for ANN and LSTM classification models.

| Hyperparameter                | Domain space | Optimized value | Hyperparameter                | Domain space | Optimized value |
|-------------------------------|--------------|-----------------|-------------------------------|--------------|-----------------|
| 1. Learning Rate              | $10^{-6} - 10^{-2}$ | $2.3 \times 10^{-5}$ | 1. Learning Rate              | $10^{-6} - 10^{-2}$ | $1.5 \times 10^{-4}$ |
| 2. Dropout                    | 0.1 - 0.9    | 0.4             | 2. Dropout                    | 0.1 - 0.9    | 0.7             |
| 3. Activation Function        | Relu, Sigmoid| Sigmoid         | 3. Activation Function        | Relu, Sigmoid| Relu            |
| (at hidden layers)            |              |                 | 4. No. of LSTM layers         | 1 - 3        | 1               |
| 4. No. of hidden (dense) layers | 1-3         | 3               | 5. No. of units in each LSTM cell | 4 - 1024 | 512             |
| 5. No. of nodes in hidden (dense) layers | 32 – 2048 | 1394           | 6. No. of nodes in dense layers | 15 – 1024 | 15              |

classes are judged by (a) the extent of inclusion of sound events related to other similar classes, (b) the sounds of classes being relevant to the street-based sounds. The parent classes are avoided where the children classes are not predominantly street-based, and (c) data quality of the class as indicated in the AudioSet. Finally, a total of 15 classes from Appendix 1 is selected to create a subset from the larger AudioSet data. The reasons for eliminating particular classes are discussed under the “Remarks” column in Appendix 1.

We selected the unbalanced train to create the training set as it consists of significantly more samples than the balanced train set. In machine learning problems, general practice is to divide the dataset into three parts train, validation and test set prior to model training and evaluation. The train set is the actual dataset used to build the model, while the validation set is used to evaluate the performance of the model while tuning the model hyperparameters. The test set is unseen by the model and is utilized to provide an unbiased evaluation of the trained model. As a rule-of-thumb, researchers divide the data into train, test, and validation in ratios 6:2:2 or 5:2.5:2.5 depending upon the number of training samples. With the increase in training samples, the share of test and validation sets becomes smaller [57].

In this particular case, AudioSet consists of evaluation segment, exclusively for the purpose of testing the trained model. Hence, the total no. of samples (52845) from the unbalanced train for 15 selected classes is divided into train and validation set in the ratio of 7:3. The assessment of the performance of the final trained model is conducted with the help of the evaluation segment of the AudioSet data as a test set.

5 Model Preparation and Evaluation

The models are prepared with the help of Python-based Keras [15] library with Tensorflow [53] backend. The optimization is achieved with the help of Skopt (https://scikit-optimize.github.io/) library. Each model is run 50 times to obtain optimum values for each of the selected hyperparameters (Table 1). ANN and LSTM models are programmed to execute 50 and 10 training epochs respectively at each run. The complete hyperparameter optimization process took approximately four days on a machine with 16 core Xeon processor and 32 GB memory with Nvidia K2000 2GB Graphics. During the model training process, the neural network with best hyperparameters is saved to the disk, which is further utilized for inferences and model evaluation.

The input layer in the ANN model takes 10-seconds input with dimension $10 \times 128$. The input is flattened to produce 1280 nodes which are fully connected to 3 sigmoid activated hidden layers with 1394 nodes each as determined by BO. The dropout function with a rate of 0.4 is added after each hidden layer. The nodes in the third hidden layer are connected to the softmax activated output layer with 15 nodes. The sparse categorical cross-entropy is used as a loss function with Adam as a gradient descent optimizer. Similarly, in the LSTM model, the input is passed to the LSTM layer with 512 units which is fully connected with a 15-node dense layer having a Relu activation function. The dropout function with a rate of 0.7 is added after the LSTM layer. The implementation of the output layer, loss function, and the optimizer is same as in the ANN model.
The softmax function applied in the output layer provides probabilities for each class considered in the model. The class with the highest probability is considered as a model prediction for the corresponding sample.

The Accuracy is the commonly used performance measure in classification tasks which is given as the ratio of correctly classified samples to the total samples. However, the accuracy measure does not consider number of correct labels from other classes and is sensitive to imbalanced datasets [75]. It is therefore suggested to evaluate the model performance through Precision and Recall metrics. Precision is defined by the ratio between correctly classified positive samples and the number of samples labeled by the classifier as positive. Whereas Precision refers to the ratio between the number of correctly classified positive samples and the number of true positive samples in the dataset [76]. Precision and Recall metrics is further studied in unison as F-measure which is the harmonic mean of Precision and Recall. Kappa statistic [54] is often used to compare the performance of various classifiers. The Kappa metric compares the observed accuracy with the random chance, also called as expected accuracy. For example, a Kappa value of 0.56 suggests that the performance of the classifier is 56% better than the assignment of the classes by random chance. In this study, F1-score is used to evaluate per class performance while overall accuracy and Kappa statistic is used to compare the performance between two classifiers.

The optimized models are evaluated against the test set (evaluation set in AudioSet data), and the results are shown in Figure 2 and Table 2. Surprisingly, the performance of the LSTM model over ANN does not show a significant difference in overall accuracy (0.60 vs. 0.56) or kappa metrics (0.56 vs. 0.52) (Table 2). However, the capability of the LSTM model to take the variable size of input shows dominance over ANNs in sound classification tasks. The sound classification with variable size input hints more use cases such as live audio monitoring and sound mapping.

The confusion matrix (Figure 2) shows that the majority of misrepresented samples belong to common traffic and human-based sounds. Among all the classes corresponding to common traffic sounds, classes such as Light Engine and Engine Starting show poor performance in both the models. Further, except Airhorn, all classes show poor classification with F1-score less than 0.55 (Table 2). Among the human-based sounds, most samples of children shouting are misinterpreted by both the classifiers as the crowd. Upon manual inspection of the test samples, it is observed that several samples are indistinguishable from the crowd even to human ears. Further, samples of crowd class in the evaluation set comprise of groups of people conversing and celebrating with each other which has similarity to the sounds of children shouting. Similarly, the Hubbub class corresponds to the common speech sounds from different sources, which is difficult to separate from the crowd. On the other hand, the conversation class includes
Table 2: Class-wise model performance of 15-class ANN and LSTM models.

| S. no | Class                        | Evaluation samples | F1-score | ANN | LSTM |
|-------|------------------------------|--------------------|----------|-----|------|
| 1     | Conversation                 | 60                 | 0.77     | 0.8 |      |
| 2     | Children shouting            | 60                 | 0.06     | 0.33|      |
| 3     | Walk-footsteps               | 60                 | 0.67     | 0.73|      |
| 4     | Crowd                        | 80                 | 0.62     | 0.68|      |
| 5     | Hubbub                       | 57                 | 0.55     | 0.59|      |
| 6     | Children playing             | 60                 | 0.81     | 0.87|      |
| 7     | Bird vocalization            | 169                | 0.78     | 0.83|      |
| 8     | Air-truck horn               | 60                 | 0.76     | 0.81|      |
| 9     | Motorcycle                   | 60                 | 0.49     | 0.45|      |
| 10    | Traffic-noise                | 64                 | 0.54     | 0.47|      |
| 11    | Light-engine                 | 59                 | 0        | 0.15|      |
| 12    | Medium Engine                | 54                 | 0.32     | 0.34|      |
| 13    | Engine Starting              | 44                 | 0.04     | 0.18|      |
| 14    | Idling                       | 64                 | 0.32     | 0.34|      |
| 15    | Silence                      | 58                 | 0.68     | 0.67|      |
|       | Overall Accuracy             |                    | 0.563    | 0.601| 0.527 | 0.567 |

The process of dataset preparation, training, and evaluation of the ANN and LSTM models are repeated with the revised number of classes while optimizing the hyperparameters (Table 3). The architecture of the new models is similar to the 15-class models with the difference in the output layer, where 15-node output layer is replaced by a 3-node layer corresponding to the revised number of classes. The domain space for Bayesian Optimization process is modified accordingly (Table 3).

The optimized 3-class ANN and LSTM model shows an overall accuracy of 85 and 89 percent respectively (Table 4). Apart from a few classification errors, both the classifiers show robust classification performance in all the three classes (Figure 3). We utilize LSTM based model to further classify manually-collected sounds from the local streets. The inference model is designed to take variable lengths of input up to 10 seconds, in which, the sound clip is first converted to the required vggish embeddings format before passing it to the trained 3-class LSTM model.

6 Street level sound collection and model inferences

To demonstrate the applicability of the discussed method in real life scenarios, an area of approximately 0.2 sq. Km is chosen to collect sound clips in the city of Mumbai, India (Figure 5a). The area is composed of predominantly residential buildings with varying built typology (Figure 4). The street-level sounds were recorded at 23 random locations within the area with the help of high-resolution handheld sound recorder (Zoom H4N Pro) with binaural microphones (Roland CS 10-EM, frequency range 20 Hz – 20 kHz). One of the authors wore the microphone and recorder setup and traveled along the street while capturing the sounds at specific locations. Audio clips of 30 seconds length are recorded at each location from 1730-1830 hrs.

Further, each 30-sec clip is split into 5-sec segments which are passed to the inference model to get a set of 6 inferences from 3-class LSTM classifier. As an example, for survey location no. 2, The recorded 30-second clip is further split into 5-second smaller clips. The model inferences obtained from each of the six clips are (1) Anthropogenic, (2) Biotic, (3) Anthropogenic, (4) Biotic, (5) Biotic, and (6) Human-based. Similarly, the inferences for all the loca-
Table 3: Table showing domain space and optimized values for 3-class ANN and LSTM classification models.

| Hyperparameter          | Domain space     | Optimized value | Hyperparameter          | Domain space     | Optimized value |
|-------------------------|------------------|-----------------|-------------------------|------------------|-----------------|
| 1. Learning Rate        | \(10^{-6} \text{ - } 10^{-2}\) | 0.7 \times 10^{-3} | 1. Learning Rate        | \(10^{-6} \text{ - } 10^{-2}\) | 0.3 \times 10^{-4} |
| 2. Dropout              | 0.1 – 0.9        | 0.6             | 2. Dropout              | 0.1 – 0.9        | 0.1             |
| 3. Activation Function  | \(\text{Relu, Sigmoid}\) | \(\text{Sigmoid}\) | 3. Activation Function  | \(\text{Relu, Sigmoid}\) | \(\text{Sigmoid}\) |
| (at hidden layers)      |                  |                 | (at hidden layers)      |                  |                 |
| 4. No. of hidden        | 1-3              | 1               | 4. No. of LSTM layers   | 1-3              | 1               |
| (dense) layers          |                  |                 |                         |                  |                 |
| 5. No. of nodes in      | 32 – 2048        | 248             | 5. No. of units in each | 4 – 1024         | 969             |
| hidden (dense) layers   |                  |                 | LSTM cell               |                  |                 |
|                         |                  |                 | 6. No. of nodes in      | 3 – 1024         | 726             |
|                         |                  |                 | dense layers            |                  |                 |

Figure 3: Confusion matrix showing the classification results of 3-class (a) ANN and (b) LSTM models

Table 4: Class-wise model performance of 3-class classifiers.

| Class         | Evaluation samples | F1-score | F1-score |
|---------------|--------------------|----------|----------|
|               |                    | ANN      | LSTM     |
| Human-based   | 377                | 0.85     | 0.89     |
| Biotic        | 227                | 0.78     | 0.83     |
| Anthropogenic | 405                | 0.9      | 0.93     |
| O.A           |                    | 0.85     | 0.89     |
| Kappa         |                    | 0.77     | 0.83     |

The experts are Ph.D. students at the university working on various urban-related problems and are well versed with the locations from which the sound clips are collected. The experts were given an overall understanding of the study and were asked to browse through the AudioSet ontology and the samples. Further, a total of four 30-sec sound clips from manually collected audio clips are provided before the start of the process to familiarize them with the data. Finally, the experts were given 25 random 5-sec sound clips, where they were asked to classify the sound sources as human-based, biotic and anthropogenic (Appendix 2). The responses given by each expert is compared with the LSTM model predictions to determine consistency or agreement between the two through kappa metrics. The results show kappa values ranging from 0.5-0.75 indicating moderate to substantial agreement [54] with the model predictions (Appendix 3).
Figure 4: Photographs taken at some of the survey locations.
Table 5: Table showing the existence of different sound sources in each location.

| Survey Location | Instances of Human-based sounds | Instances of Biotic Sounds | Instances of Anthropogenic sounds | Total Instances |
|-----------------|---------------------------------|---------------------------|-----------------------------------|-----------------|
| 1               | 0                               | 0                         | 6                                 | 6               |
| 2               | 1                               | 3                         | 2                                 | 6               |
| 3               | 1                               | 2                         | 3                                 | 6               |
| 4               | 2                               | 3                         | 1                                 | 6               |
| 5               | 2                               | 2                         | 2                                 | 6               |
| 6               | 0                               | 0                         | 6                                 | 6               |
| 7               | 1                               | 0                         | 5                                 | 6               |
| 8               | 0                               | 0                         | 6                                 | 6               |
| 9               | 1                               | 3                         | 2                                 | 6               |
| 10              | 5                               | 0                         | 1                                 | 6               |
| 11              | 5                               | 0                         | 1                                 | 6               |
| 12              | 6                               | 0                         | 0                                 | 6               |
| 13              | 6                               | 0                         | 0                                 | 6               |
| 14              | 2                               | 1                         | 3                                 | 6               |
| 15              | 3                               | 0                         | 3                                 | 6               |
| 16              | 2                               | 0                         | 4                                 | 6               |
| 17              | 4                               | 0                         | 2                                 | 6               |
| 18              | 1                               | 0                         | 5                                 | 6               |
| 19              | 6                               | 0                         | 0                                 | 6               |
| 20              | 5                               | 0                         | 1                                 | 6               |
| 21              | 4                               | 0                         | 2                                 | 6               |
| 22              | 0                               | 0                         | 6                                 | 6               |
| 23              | 0                               | 0                         | 6                                 | 6               |

6.1 Sound Mapping using GIS

The model predictions show robust performance in detecting major sound sources in local streets, the results of which are comparable to that of human experts. The study area, therefore, can be translated to sound maps with the help of GIS and spatial interpolation techniques. Sound maps can provide detail insights into the exposure of different types of sound and may help in increasing public awareness regarding the same. Further, similar to noise maps and its significance in framing management policies; DL-based local sound classification and GIS-based mapping may find relevance in designing future sound identification and management guidelines.

We utilized the Inverse Distance Weighted (IDW) spatial interpolation technique to estimate the presence of anthropogenic, biotic and human-based sounds in the selected area. Other interpolation techniques such as Kriging and Spline methods have been implemented by different studies [33, 42, 50] in the creation of sound maps. We selected the IDW interpolation technique over other methods as it is easier to define the parameters and to understand the results while also being comparable to the performance of other interpolation methods [26, 28].

Spatial interpolation techniques, in general, estimates the values of locations with unknown values with the help of known ones. IDW uses local interpolation which gives more weight to nearby points with known values as opposed to all the available points. Therefore, the estimated value of an unknown point is directly influenced by the more nearby points and diminishes with the increase in distance. The IDW function is implemented in ArcMap 10.1. The results are shown in Figure 5b, 5c, and 5d. IDW tool takes input regarding the search radius as distance or no. of points and power. Power controls the significance each nearby point has on the interpolated value; higher power suggests the low significance from the distant points. We took the default values for variable search radius in ArcMap, i.e. the no. of neighbor points and power at 12 and 2 respectively.

The sound maps estimate the presence of sound sources in different parts of the study area. The western part of the map is dominated by the human-based sounds, mainly due to the presence of street shops and vegetable
market (Figure 4d, 4e, 4f) (Figure 5b). Generally, low traffic is seen in this area as streets are occupied by pedestrians. The northern part of the study area consists of a major road, due to which, the survey points near the road such as 1, 23, and 22 show the dominance of traffic sounds over others. Similarly, streets which lead to the major commercial center in the southern part of the area show similar characteristics (Figure 4c) (Figure 5d). The western part of the study area, which comprises low-rise built typology with narrower streets shows the prevalence of silence and bird sound events and the absence of vehicles (Figure 4a, 4b) (Figure 5c).

7 Discussion

This study provides a DL based methodology to classify local street-level sounds from openly available AudioSet data. While many studies have utilized similar methods to automate the identification of sound sources and events, the scope of the studies have been limited to fewer sound classes or have been conducted in relatively homogeneous environments. Large-scale environmental sound classification studies have utilized algorithms to classify sound sources mainly in natural environments [61, 66, 82]. Re-
cent studies which have experimented with DL-based methods in sound classification are based on selected sound classes or environments [10, 12, 69]. In this study, however, we showed how crowdsourced audio dataset (AudioSet) could be utilized to identify major sound sources in unstructured environments such as urban streets with regular DL algorithms.

As evident from erstwhile studies, the individual perception regarding soundscapes relates well with the presence of particular sounds. Apart from the analysis of psychoacoustic parameters such as loudness, roughness, and sharpness captured from the collected audio, the discussed method would enable the researchers to identify sound-sources which will help enhance the prospects and scale of soundscape studies. This study considers a smaller area to demonstrate the applicability of the proposed method in classifying street-level sounds. However, with means to rapidly collect such sounds at different locations all over the city will assist in creation of sound maps at the city or sub-city level. Previous city-wide sound mapping studies have utilized proxy indicators such as sound tags and comments collected from online sound archives such as Freesound to categorize the sound sources [2]. With the classification model discussed in this study, rather than text-based input [2], sound clips from such archives can be used directly to obtain predictions on the presence of different sound sources. Further, the availability of low-cost acoustic sensors can be used to record and monitor urban sounds at large spatial and temporal scale [56, 59], which can help prepare soundscape maps and changes in real-time. Overall, the sound classification task without manual intervention holds promise in Smart city development [42], where automated analysis of sound sources may help in city-wide soundscape mapping.

Earlier urban perception studies have put major emphasis in visual evaluation of urban areas through street view imagery to cover large urban contexts; expanding such studies to include sound collection and analysis may help in multi-modal urban assessment. The identification of street-level sounds along with visual elements can further assist in redefining place locale [84]. Future studies may discuss the variability in sound sources at different timings and can relate to the overlying land use and function [50]. Apart from evaluating the presence of noise as a sole criterion of a nuisance, urban management authorities can make use of such tools to determine sound sources. Further, such a tool might provide insights to urban designers and planners into the potential design alternatives based on the sounds which are audible to the users [38, 39].

While the discussed method provides an easier alternative to the tedious process of manual collection and annotation of data samples, the shortcomings of the proposed approach cannot be ignored. The AudioSet data, due to inherent diversity in the sound quality and the context, is prone to misclassification on training with a large number of classes. Further, the 10-second samples in AudioSet data usually include unwanted noise elements or the sound events which show similarity with other classes, the inclusion of which is subjective and unavoidable in most cases. The discussed method may not perform well enough in all the use cases due to the large variations within the samples of each class and the devices used to collect these samples as YouTube videos. However, such issues can be solved if samples collected from the local areas are annotated and merged with the AudioSet before model training.

8 Conclusions

The study utilizes openly available AudioSet data and developed the LSTM model to detect sound classes from audio data collected from the streets. It utilizes Bayesian optimization to obtain hyperparameters for the selected models which provide an overall accuracy of 60 and 89 percent for 15 and 3 classes, respectively. The model predictions are further checked by human subjects to determine accuracy on the performance of the model on locally collected street-level audio clips. Further, the results of the sound classification are mapped in the survey area with the help of GIS-based interpolation tools. The stepwise methodology used in the study can be reiterated to be utilized in new soundscape studies. The DL models can be integrated into IoT devices connected with acoustic sensors for real-time updates on urban soundscapes. Further, the setup can be deployed in environments other than urban streets such as urban parks, forests by varying the selection of sound classes in AudioSet data used in this study.

The study discussed how new technological advancements could help increase the premise of well-established soundscape studies. We believe that such studies will pave the path for exciting new applications in noise management which will further benefit the urban population at large.

Acknowledgement: The authors would like to thank anonymous reviewers for their comments. The authors are also thankful to the Ministry of Human Resource Development (MHRD), India and Industrial Research and Consul-
tancy Centre (IRCC), IIT Bombay for funding this study under the grant titled Frontier Areas of Science and Technology (FAST), Centre of Excellence in Urban Science and Engineering (Grant Number 14MHRD005).

References

[1] Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., & Aide, T. M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics, 4(4), 206–214. https://doi.org/10.1016/j.ecoinf.2009.06.005

[2] Aiello, L. M., Schifanella, R., Quercia, D., & Aletta, F. (2016). Chatty maps: constructing sound maps of urban areas from social media data. Royal Society Open Science, 3(3), 150690. https://doi.org/10.1098/rsos.150690

[3] Aletta, F., Kang, J., & Axelsson, Ö. (2016). Soundscape descriptors and a conceptual framework for developing predictive soundscape models. Landscape and Urban Planning, 149, 65–74. https://doi.org/10.1016/j.landurbplan.2016.02.001

[4] Aletta, F., Oberman, T., & Kang, J. (2018). Associations between positive health-related effects and soundscape perceptual constructs: A systematic review. International Journal of Environmental Research and Public Health, 15(11), 1–15. https://doi.org/10.3390/ijerph15112392

[5] Aletta, F., & Xiao, J. (2018). What are the Current Priorities and Challenges for (Urban) Soundscape Research? Challenges, 9(1), 16. https://doi.org/10.3390/challe9010016

[6] Axelsson, Ö., Nilsson, M. E., & Berglund, B. (2010). A principal components model of soundscape perception. The Journal of the Acoustical Society of America, 128(5), 2836–2846. https://doi.org/10.1121/1.3493436

[7] Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic Scene Classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine, 32(3), 16–34. https://doi.org/10.1109/MSP.2014.2326181

[8] Berglund, B., & Nilsson, M. E. (2006). On a tool for measuring soundscape quality in urban residential areas. Acta Acustica United with Acustica, 92(6), 938–944.

[9] Brown, A. L., Kang, J., & Gjestland, T. (2011). Towards standardization in soundscape preference assessment. Applied Acoustics, 72(6), 387–392. https://doi.org/10.1016/j.apacoust.2011.01.001

[10] Cakir, E., Adavanne, S., Parascandolo, G., Drossos, K., & Virtanen, T. (2017). Convolutional recurrent neural networks for bird audio detection. In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 1744–1748). IEEE. https://doi.org/10.23919/EUSIPCO.2017.8081508

[11] Cakir, E., Parascandolo, G., Heltola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12), 1291–1303. https://doi.org/10.1109/TASLP.2017.2690575

[12] Chen, Yan, Guo, Q., Liang, X., Wang, J., & Qian, Y. (2019). Environmental sound classification with dilated convolutions. Applied Acoustics, 148, 123–132. https://doi.org/10.1016/j.apacoust.2018.12.019

[13] Chen, Yutian, Huang, A., Wang, Z., Antonoglou, I., Schrittwieser, J., Silver, D., & de Freitas, N. (2018). Bayesian Optimization in AlphaGo, (1). https://doi.org/arXiv:1812.06855v1

[14] Chew, Y. R., & Wu, B. S. (2016). A soundscape approach to analyze traffic noise in the city of Taipei, Taiwan. Computers, Environment and Urban Systems, 59, 78–85. https://doi.org/10.1016/j.compenvurbsys.2016.05.002

[15] Chollet, F., & others. (2015). Keras.

[16] Davies, W. J., Adams, M. D., Bruce, N. S., Cain, R., Carlyle, A., Cusack, P., . . . Poxon, J. (2013). Perception of soundscapes: An interdisciplinary approach. Applied Acoustics, 74(2), 224–231. https://doi.org/10.1016/j.apacoust.2012.05.010

[17] De Coensel, B., & Botteldooren, D. (2007). The Rhythm of the Urban Soundscape. Noise & Vibration Worldwide, 38(9), 11–17. https://doi.org/10.26957/095745607782689827

[18] Dennis, J. W. (2014). Sound Event Recognition in Unstructured Environments using Spectrogram Image Processing. Thesis. Nanyang Technological University. Retrieved from http://www.ntu.edu.sg/home/aseschng/Thesis/JohnDennis_PhDThesis2014.pdf

[19] EC. (2002). Environmental Noise Directive. Retrieved from http://ec.europa.eu/environment/noise/directive_en.htm

[20] Eichinski, P., & Roe, P. (2017). Clustering and Visualization of Long-Duration Audio Recordings for Rapid Exploration Avian Surveys. In 2017 IEEE 13th International Conference on e-Science (e-Science) (pp. 168–176). IEEE. https://doi.org/10.1109/eScience.2017.29

[21] Farina, A., Pleretti, N., & Piccioli, L. (2011). The soundscape methodology for long-term bird monitoring: A Mediterranean Europe case-study. Ecological Informatics, 6(6), 354–363. https://doi.org/10.1016/j.ecoinf.2011.07.004

[22] Frazier, P. I. (2018). A Tutorial on Bayesian Optimization. Retrieved from http://arxiv.org/abs/1807.02811

[23] Gelbart, M. A., Snoek, J., & Adams, R. P. (2014). Bayesian Optimization with Unknown Constraints, 1–14. https://doi.org/10.1017/CBO9781107415324.004

[24] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., . . . Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 776–780). IEEE. https://doi.org/10.1109/ICASSP.2017.7952261

[25] Gidlöf-Gunnarsson, A., & Öhrström, E. (2007). Noise and well-being in urban residential environments: The potential role of perceived availability to nearby green areas. Landscape and Urban Planning, 83(2–3), 115–126. https://doi.org/10.1016/j.landurbplan.2007.03.003

[26] Gong, G., Matteveda, S., & O’Bryant, S. E. (2014). Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas. Environmental Research, 130, 59–69. https://doi.org/10.1016/j.envres.2013.12.005

[27] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[28] Harman, B. I., Koseoglu, H., & Yigit, C. O. (2016). Performance evaluation of IDW, Kriging and multiquadratic interpolation methods in producing noise mapping: A case study at the city of Isparta, Turkey. Applied Acoustics, 112, 147–157. https://doi.org/10.1016/j.apacoust.2016.05.024
Kang, J., Aletta, F., Margaritis, E., & Yang, M. (2018). A model for implementing soundscape maps in smart cities. *Noise Mapping*, 5(1), 46–59. https://doi.org/10.1515/noise-2018-0004

Kawai, K., Kojima, T., Hirate, K., & Yasuoka, M. (2004). Personal evaluation structure of environmental sounds: Experiments of subjective evaluation using subjects’ own terms. *Journal of Sound and Vibration*, 277(3), 523–533. https://doi.org/10.1016/j.jsv.2004.03.013

Kogel, P., Arenas, J. P., Bermejo, F., Hinalaf, M., & Turra, B. (2018). A Green Soundscape Index (GSI): The potential of assessing the perceived balance between natural sound and traffic noise. *Science of the Total Environment*, 642, 463–472. https://doi.org/10.1016/j.scitotenv.2018.06.023

Kogel, P., Turra, B., Arenas, J. P., & Hinalaf, M. (2017). A comprehensive methodology for the multidimensional and synchronic data collecting in soundscape. *Science of the Total Environment*, 580, 1068–1077. https://doi.org/10.1016/j.scitotenv.2016.12.061

Krzypicka, P., & Byrka, K. (2017). Restorative qualities of and preference for natural and urban soundscapes. *Frontiers in Psychology*, 8(OC1), 1–13. https://doi.org/10.3389/fpsyg.2017.01705

Laszlo, H. E., McRobie, E. S., Stansfeld, S. A., & Hansell, A. L. (2012). Annoyance and other reaction measures to changes in noise exposure - A review. *Science of the Total Environment*, 435–436, 551–562. https://doi.org/10.1016/j.scitotenv.2012.06.112

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning, 1–38. https://doi.org/10.1145/2647868.2654889

Liu, J., Kang, J., Luo, T., Behm, H., & Coppack, T. (2013). Spatiotemporal variability of soundscapes in a multiple functional urban area. *Landscape and Urban Planning*, 115, 1–9. https://doi.org/10.1016/j.landurbplan.2013.03.008

Marchegiani, L., & Posner, I. (2017). Leveraging the urban soundscape: Auditory perception for smart vehicles. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6547–6554). IEEE. https://doi.org/10.1109/ICRA.2017.7989774

Marquis-Favre, C., Premat, E., Aubree, D., & Vallet, M. (2005). Noise and its Effects A Review on Qualitative Aspects of Sound. Part I: Notions and Acoustic Ratings. *Acta Acustica United with Acustica*, 91(4), 613–625. Retrieved from http://www.ingentaconnect.com/content/avaa/aaau/2005/00000001/00000004/art00001

Martin, A., Ashish, A., Paul, B., Eugene, B., Zhifeng, C., Craig, C., … Xiaojiang, Z. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved from https://www.tensorflow.org/

McHugh, M. L. (2012). Lessons in biostatistics Interrater reliability: the kappa statistic. *Biochimia Medica*, 22(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mesaros, A., Heittola, T., & Virtanen, T. (2016). TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)* (Vol. 2016-Novem, pp. 1128–1132). IEEE. https://doi.org/10.1109/EUSIPCO.2016.7760424

Mydlarz, C., Salamon, J., & Bello, J. P. (2017). The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics*, 117, 207–218. https://doi.org/10.1016/j.apacoust.2016.06.010

Ng, A. (2019). Train / Dev / Test sets - Coursera. Retrieved May 10, 2019, from https://www.coursera.org/lecture/deep-neural-network/train-dev-test-sets-cxG1s
### Appendix 1: Common Street-level sound classes sampled from AudioSet data.

| S.no | Sound Classes                                      | Relation       | Data Quality | No. of samples in Unbalanced train. | Remarks                                                                 | Status  |
|------|----------------------------------------------------|----------------|--------------|-------------------------------------|-------------------------------------------------------------------------|---------|
| 1.   | 4. Conversation                                    | child          | high         | 2144                                |                                                                          | Accept  |
| 2.   | 13. Children shouting                              | child          | high         | 553                                 |                                                                          | Accept  |
| 3.   | 53. Walk, footsteps                                | child          | high         | 1563                                |                                                                          | Accept  |
| 4.   | 69. Crowd                                          | child          | medium       | 10116                               |                                                                          | Accept  |
| 5.   | 70. Hubbub, speech noise, speech babble            | child          | high         | 1360                                |                                                                          | Accept  |
| 6.   | 71. Children playing                               | child          | high         | 692                                 |                                                                          | Accept  |
| 7.   | 112. Bird vocalization, bird call, bird song       | Parent         | high         | 6075                                |                                                                          | Accept  |
| 8.   | 283. Wind                                          | Parent [284, 285] | high       | 6454                                | Includes samples of wind sounds collected in open areas. Not relevant to street level wind sounds | Reject  |
| 9.   | 284. Rustling leaves                               | child          | low          | 1608                                | Low data quality. A mix of classes such as wind, rain, speech, bird vocalization. | Reject  |
| 10.  | 285. Wind noise (microphone)                       | child          | high         | 6172                                | Similar to samples in [283].                                           | Reject  |
| 11.  | 300. Vehicle                                       | Parent [306]   | high         | 12641                               | Samples include different types of vehicles (road, rail, and water).     | Reject  |
| 12.  | 306. Motor vehicle (road)                          | Parent [307, 316, 321, 326, 327] | high       | 8891                                | Irregularities among the samples of child classes. Further, not all of the child classes are street-based. | Reject  |
| 13.  | 307. Car                                           | Parent [308, 314] | high       | 40978                               | Includes sounds of the car alarm, power windows, skidding, tire squeal which are irrelevant in this case. | Reject  |
| 14.  | 308. Vehicle horn, car horn, honking               | Parent         | medium       | 3587                                | The samples of Vehicle horn are mixed with other vehicle sounds such as a passing vehicle, car racing, train horn, etc. | Reject  |
| 15.  | 314. Car passing by                                | child          | medium       | 3684                                | Include sounds mostly from car racing.                                   | Reject  |
| 16.  | 316. Truck                                         | Parent [318]   | high         | 10870                               | Includes sounds of reversing beeps and ice cream truck which is irrelevant | Reject  |
| 17.  | 318. Air horn, truck hom                           | child          | high         | 691                                 |                                                                          | Accept  |
### Appendix 1: ... continued

| S.no | Sound Classes | Relation | Data quality | No. of samples in Unbalanced train. | Remarks | Status |
|------|---------------|----------|--------------|------------------------------------|---------|--------|
| 18. | 321. Bus | child | high | 5171 | Samples include sounds collected from the interior and exterior of the bus. | Reject |
| 19. | 326. Motorcycle | child | high | 7140 | Accept |
| 20. | 327. Traffic noise, roadway noise | child | high | 1566 | Accept |
| 21. | 343. Engine | Parent [344, 348, 349, 351, 352, 353] | high | 15870 | Includes samples of all kinds of engine sounds such as the Jet engine and Heavy engine which are not typically present in the streets. | Reject |
| 22. | 344. Light engine (high frequency) | Parent | high | 313 | Accept |
| 23. | 348. Medium engine (mid frequency) | child | high | 6012 | Accept |
| 24. | 349. Heavy engine (low frequency) | child | low | 4133 | Low-quality dataset. Also includes engine sounds from heavy automobiles, steam and rocket engines. | Reject |
| 25. | 351. Engine starting | child | high | 988 | Accept |
| 26. | 352. Idling | child | high | 6089 | Accept |
| 27. | 353. Accelerating, revving, vroom | child | high | 11827 | Samples include a mix of sounds from jet engines, racing cars, motorcycle, etc. | Reject |
| 28. | 500. Silence | child | high | 7543 | Includes sounds recorded in the outside environment, consists of a mix of vehicle sounds, crowd, people talking, etc. Samples do not show a distinction from samples of classes such as [327], [500]. | Accept |
| 29. | 509. Outside, urban or manmade | child | high | 33861 | Includes mix sounds samples of animals, vehicles, water bodies, etc. Similar to [509], [510]. | Reject |
| 30. | 510. Outside, rural or natural | child | high | 35291 | Includes sounds recorded in natural environments, however similar to [509]. | Reject |
| 31. | 514. Environmental noise | child | medium | 1120 | | Reje
Appendix 2: Flyer distributed to the experts as a part of collection of responses regarding the presence of various street level sounds.

Respondent Name:
Objective: To manually classify the presence of Anthropogenic, Biotic and Human-based sound sources from sound clips collected from local urban streets.

Task 1: Visit Audio Set website. See the ontology and listen to few samples. (Use Headphones) https://research.google.com/audioset/ontology/index.html

Task 2: See the definitions of the classes.

| S.no | Sound Classes                          | Definitions                                                                 | Aggregated classes |
|------|----------------------------------------|-----------------------------------------------------------------------------|--------------------|
| 1    | Conversation                           | Interactive, spontaneous spoken communication between two or more people.    | Human-Based        |
| 2    | Children shouting                      | The boisterous vocalizations of a group of children, for instance in a playground. |                    |
| 3    | Walk, footsteps                         | The sound of feet or shoes contacting the ground in conventional human locomotion. |                    |
| 4    | Crowd                                  | The sound of a large group of people gathered together.                     |                    |
| 5    | Hubbub, speech noise, speech babble    | Loud, disordered, unintelligible speech noise from many sources.            |                    |
| 6    | Children playing                       | The sound of children at play, including any of vocalization or the sounds of their activities. | Biotic             |
| 7    | Bird vocalization, bird call, bird song| Bird communication calls, often considered melodious to the human ear.      |                    |
| 8    | Silence                                | The absence of audible sound or presence of sounds of very low intensity.    |                    |
| 9    | Air horn, truck horn                   | The sound of a pneumatic device mounted on large vehicles designed to create an extremely loud noise for signalling purposes. | Anthropogenic      |
| 10   | Motorcycle                             | Sounds of a small motor vehicle, usually with only two wheels. Motorcycles typically lack an external shell and seat riders astride the engine. |                    |
| 11   | Traffic noise, roadway noise           | The combined sounds of many motor vehicles traveling on roads.             |                    |
| 12   | Light engine (high frequency)          | The sound of a small engine such as a toy car, sewing machine, or moped.    |                    |
| 13   | Medium engine (mid frequency)          | The sound of a moderately-sized engine such as that which powers a motorcycle, sedan, or small truck. |                    |
| 14   | Engine starting                        | The sound of an engine starting from rest, which may involve a specific starter mechanism (as in a standard car). |                    |
| 15   | Idling                                 | The sound of an engine (typically in an automobile) running without any load and at minimal RPM. |                    |

Task 3: Familiarize yourself with the local street sounds. Listen to the 4 audio clips (30-sec each) collected from the streets. The clips are present in the folder 30_sec.

Task 4: Provide responses for 25 audio clips of 5-sec each [Human-based: 1, Biotic: 2, Anthropogenic: 3]. The clips are present in folder 5_sec.
| S.no | Samples   | Responses |
|------|-----------|-----------|
| 1    | 1-5.wav  |           |
| 2    | 2-4.wav  |           |
| 3    | 3-0.wav  |           |
| 4    | 3-4.wav  |           |
| 5    | 4-1.wav  |           |
| 6    | 4-2.wav  |           |
| 7    | 4-5.wav  |           |
| 8    | 5-4.wav  |           |
| 9    | 6-1.wav  |           |
| 10   | 6-5.wav  |           |
| 11   | 8-0.wav  |           |
| 12   | 10-3.wav |           |
| 13   | 12-3.wav |           |
| 14   | 13-0.wav |           |
| 15   | 13-4.wav |           |
| 16   | 14-3.wav |           |
| 17   | 15-0.wav |           |
| 18   | 16-1.wav |           |
| 19   | 17-0.wav |           |
| 20   | 18-1.wav |           |
| 21   | 19-4.wav |           |
| 22   | 21-2.wav |           |
| 23   | 21-4.wav |           |
| 24   | 22-4.wav |           |
| 25   | 23-0.wav |           |
### Appendix 3: Analysis of collected responses from experts and calculation of evaluation metrics.

| Samples       | Model Predictions | Responses_ah | Responses_an | Responses_de | Responses_ma | Responses_ni |
|---------------|-------------------|--------------|--------------|--------------|--------------|--------------|
| 1-5.wav       | 3                 | 3            | 3            | 3            | 3            | 3            |
| 2-4.wav       | 1                 | 2            | 1            | 1            | 3            | 1            |
| 3-0.wav       | 2                 | 2            | 2            | 1            | 1            | 2            |
| 3-4.wav       | 3                 | 3            | 3            | 3            | 3            | 3            |
| 4-1.wav       | 2                 | 2            | 2            | 2            | 2            | 3            |
| 4-2.wav       | 2                 | 2            | 2            | 2            | 2            | 2            |
| 4-5.wav       | 3                 | 3            | 3            | 3            | 3            | 3            |
| 5-4.wav       | 2                 | 2            | 2            | 2            | 2            | 2            |
| 6-1.wav       | 3                 | 3            | 3            | 3            | 3            | 3            |
| 6-5.wav       | 3                 | 3            | 3            | 3            | 3            | 3            |
| 8-0.wav       | 3                 | 1            | 3            | 3            | 3            | 3            |
| 10-3.wav      | 1                 | 1            | 1            | 1            | 1            | 1            |
| 12-3.wav      | 1                 | 1            | 1            | 1            | 1            | 1            |
| 13-0.wav      | 1                 | 1            | 1            | 1            | 1            | 1            |
| 13-4.wav      | 1                 | 3            | 3            | 3            | 1            | 3            |
| 14-3.wav      | 2                 | 2            | 2            | 1            | 2            | 2            |
| 15-0.wav      | 1                 | 3            | 1            | 1            | 3            | 3            |
| 16-1.wav      | 3                 | 3            | 1            | 3            | 3            | 3            |
| 17-0.wav      | 1                 | 1            | 1            | 1            | 3            | 2            |
| 18-1.wav      | 2                 | 2            | 2            | 2            | 3            | 2            |
| 19-4.wav      | 3                 | 1            | 1            | 1            | 1            | 1            |
| 21-2.wav      | 1                 | 1            | 2            | 1            | 2            | 2            |
| 21-4.wav      | 1                 | 1            | 1            | 3            | 2            | 2            |
| 22-4.wav      | 3                 | 2            | 3            | 3            | 3            | 3            |
| 23-0.wav      | 3                 | 3            | 3            | 2            | 3            | 3            |

| O.A | 0.76 | 0.84 | 0.76 | 0.68 | 0.72 |
| Kappa | 0.63 | 0.75 | 0.62 | 0.5  | 0.57 |

[Note: 1: Human-based, 2: Biotic, 3: Anthropogenic]