SOUND CLASSIFICATION IN INDIAN CITIES USING MULTI-LABEL DATA AND TRANSFER LEARNING

Rishi Gupta, Vaibhav Kumar *
GeoAI4Cities Lab, Department of Data Science and Engineering
IISER Bhopal
Madhya Pradesh, India
{rishi20,vaibhav}@iiserb.ac.in

ABSTRACT

The existing research primarily focuses on single-label classifications of complex urban sounds, ignoring crucial factors like sound mixtures and duration. This work proposes a novel transfer learning approach for multi-label sound classification. We fine-tuned a pretrained VGGish model using a manually labeled audio dataset containing representative classes from diverse Indian cities, collected through various avenues. Our model achieves remarkable performance, demonstrating a significant 32% increase in F1-score compared to models trained on the AudioSet benchmark dataset.

1 INTRODUCTION

City areas are filled with lots of different sounds, like noises from human activities, transportation, and nature, all mixing together. Deciphering these diverse sounds provides valuable insights into the environment, enabling the creation of soundscapes that contribute to urban well-being and infrastructure planning(Berman et al., 2014). Classification of these varied sounds poses a formidable task for deep learning models, owing to factors such as data complexity, class imbalance (Anders et al., 2021), acoustic overlap (Cakir et al., 2015), and a shortage of labeled data (Mushtaq & Su, 2020). In the case of Indian environments, characterized by an exceptionally diverse urban composition, these challenges become even more pronounced (Verma et al., 2020). Recent efforts have focused on classifying outdoor sounds, leveraging datasets such as ESC-10, ESC-50 (Piczak, 2015), Urbansound, Urbansound8K (Salamon et al., 2014), AudioSet (Gemmeke et al., 2017), Urban-SED (Salamon et al., 2017), and SONYC-UST (Cartwright et al., 2019). However, there are two key assumptions associated with studies using these datasets. Firstly, the sounds within these datasets may not accurately represent the complexity of urban environments in developing nations. Secondly, many studies treat sounds as single-label classification problems (Lu et al., 2021; Mu et al., 2021), often overlooking essential factors such as the mix of classes, the duration of sounds, and the exclusion of classes like ambient sound, imposing limitations on the classification of urban sounds. To address these limitations, we propose a transfer learning approach. In this paper, we fine-tune a VGGish model pre-trained on AudioSet data using our manually labeled multi-label audio dataset. This dataset is carefully curated from diverse urban environments in Indian cities. This marks the first study to enhance the accuracy and relevance of urban sound classification in the Indian urban environment.

2 METHODOLOGY AND EXPERIMENTATION

We prepared a dataset comprising 1772 audio clips, each lasting four seconds to train the model. Teh clips were divided in to 4 second durion for better performance and representation of environment as suggested by Chu et al. (2009). To ensure diversity, we gathered audio samples from various sources, including manual collection in Bhopal, Madhya Pradesh, India, and online platforms like YouTube and Freesound (Fonseca et al., 2017) from various Indian cities. RODE VideoMicro microphone was employed to capture the data. Each clip could receive multiple labels: Human, Vehicle, Mammal,

*https://sites.google.com/view/vaibhavkumar1

Birds, Bicycle Bell, and Ambient Sound. The Ambient Sound label was assigned exclusively to clips devoid of other sounds. The Vehicle and Human classes boast 1015 and 740 positive samples, respectively, while the bell and mammal classes each contain only 88 and 61 samples. The Ambient Sound and Bird classes have 207 and 294 samples, respectively. We have detailed the annotation process in Appendix section A.2.

We fine-tuned the VGGish model, a pretrained CNN model by Google (Hershey et al., 2017). The model is pretrained on AudioSet data. To prepare the audio clips from our dataset for finetuning, we ensured they were all monaural and resampled them to a consistent sampling rate of 16 kHz, matching the input format. We then applied zero-padding to the audio clips and converted them into Log Mel spectrograms with 64 frequency bands spanning 125Hz to 7500Hz. The Mel scale in Log Mel spectrograms approximates how humans perceive frequencies, while the log scale mimics human loudness perception. Each spectrogram consisted of 96 frames, each lasting 10 msec without overlap. To focus on extracting higher-level features and refining the final output, we trained only the last three linear layers and the output layer, while keeping the convolutional layers frozen. The training process in detail is discussed in Appendix A.1.

3 Results and Discussion

We use multi-label confusion matrices introduced by Heydarian et al. (2022) to evaluate class-specific performance. We obtained the following normalized confusion matrix, on the pretrained model (left) and the finetuned model (right).

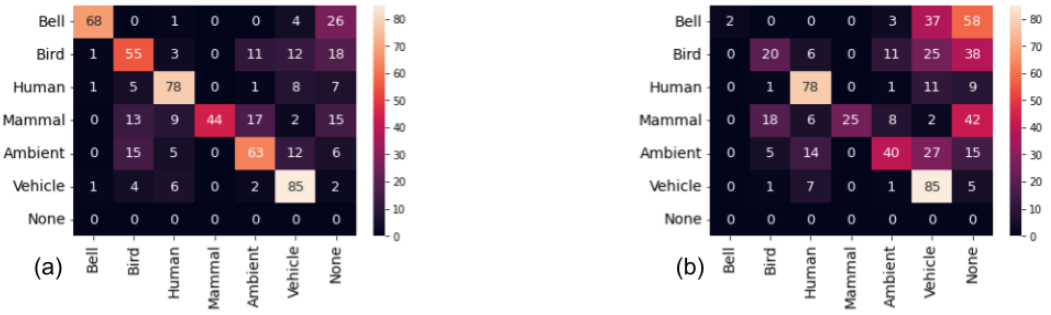

Figure 1: Normalized confusion matrix: (a) VGGish pretrained on AudioSet (b) VGGish fine-tuned on our data

Table 1: Performance of Finetuned Model vs Pretrained Model at Threshold 0.4

Training Mode	F1 Score	Jaccard Index
Pretrained	0.507699	0.392629
Finetuned	**0.826125**	**0.622463**

Macro averaged F1 score and Jaccard Index were employed to evaluate our models, due to their suitability for multi-label classification tasks. Macro averaging ensures equitable consideration of all classes. Our proposed approach significantly improves finetuned model performance compared to the benchmark pretrained model, with a remarkable 32% increase in F1 score and a 23% increase in Jaccard Index (Table 1). These significant results highlight the effectiveness of our approach. Furthermore, our findings strongly suggest that the existing AudioSet benchmark dataset inadequately represents the soundscape of the Indian urban environment. Our data, with its inherent diversity, effectively captures the unique soundscape of the city, leading to superior feature learning. This is further substantiated by the confusion matrices (Figure 1), where our model excels at classifying both 'Bicycle Bell' and 'Ambient Sound' classes, which the pretrained VGGish model fails to do entirely. This emphasizes the insufficient representation of these classes in AudioSet. Our work opens avenues for further research and validation across diverse cities. By generating soundscape maps, it can provide crucial data to inform various planning agencies.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

DATASET

The dataset is available here: `https://data.mendeley.com/datasets/4553jk9dvz/1`

REFERENCES

Franz Anders, Ammie K Kalan, Hjalmar S Kühl, and Mirco Fuchs. Compensating class imbalance for acoustic chimpanzee detection with convolutional recurrent neural networks. *Ecological Informatics*, 65:101423, 2021.

Marc G Berman, Michael C Hout, Omid Kardan, MaryCarol R Hunter, Grigori Yourganov, John M Henderson, Taylor Hanayik, Hossein Karimi, and John Jonides. The perception of naturalness correlates with low-level visual features of environmental scenes. *PloS one*, 9(12):e114572, 2014.

Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Polyphonic sound event detection using multi label deep neural networks. In *2015 international joint conference on neural networks (IJCNN)*, pp. 1–7. IEEE, 2015.

Mark Cartwright, Ana Elisa Mendez Mendez, Aurora Cramer, Vincent Lostanlen, Graham Dove, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Bello. Sonyc urban sound tagging (sonyc-ust): A multilabel dataset from an urban acoustic sensor network. 2019.

Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6): 1142–1158, 2009.

Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 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, 2017.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Mohammadreza Heydarian, Thomas E Doyle, and Reza Samavi. Mlcm: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095, 2022.

Meng Liu, Ke Liang, Dayu Hu, Hao Yu, Yue Liu, Lingyuan Meng, Wenxuan Tu, Sihang Zhou, and Xinwang Liu. Tmac: Temporal multi-modal graph learning for acoustic event classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3365–3374, 2023.

Jianrui Lu, Ruofei Ma, Gongliang Liu, and Zhiliang Qin. Deep convolutional neural network with transfer learning for environmental sound classification. In *2021 International Conference on Computer, Control and Robotics (ICCCR)*, pp. 242–245. IEEE, 2021.

Wenjie Mu, Bo Yin, Xianqing Huang, Jiali Xu, and Zehua Du. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11(1): 21552, 2021.

Zohaib Mushtaq and Shun-Feng Su. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167:107389, 2020.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 344–348. IEEE, 2017.

Amir Shirian, Krishna Somandepalli, Victor Sanchez, and Tanaya Guha. Visually-aware acoustic event detection using heterogeneous graphs. *arXiv preprint arXiv:2207.07935*, 2022.

Eleni Tsalera, Andreas Papadakis, and Maria Samarakou. Comparison of pre-trained cnns for audio classification using transfer learning. *Journal of Sensor and Actuator Networks*, 10(4):72, 2021.

Deepank Verma, Arnab Jana, and Krithi Ramamritham. Predicting human perception of the urban environment in a spatiotemporal urban setting using locally acquired street view images and audio clips. *Building and Environment*, 186:107340, 2020.

Khalid Zaman, Melike Sah, Cem Direkoglu, and Masashi Unoki. A survey of audio classification using deep learning. *IEEE Access*, 2023.

A Appendix

A.1 VGGish Model

Among other technologies used for audio classification, such as transformer-based architectures (Gong et al., 2021), and graph neural networks (Shirian et al., 2022; Liu et al., 2023), CNNs have proven very effective for environmental sound classification (Zaman et al., 2023), due to their ability to extract spatial features from audio signals. The consistent success shown by the VGGish model in prior audio classification studies, particularly its impressive performance in environmental sound recognition as reported by Tsalera et al. (2021), motivated us to choose it for our work.

VGGish is a variant of the VGG model, pretrained on Google's AudioSet dataset (Hershey et al., 2017). Our selection of the VGGish model was motivated by its demonstrated effectiveness in various audio tasks, including music genre classification and sound event detection. Notably, VGGish embeddings have also shown promising results in conjunction with Wav2Vec2 embeddings for environmental sound classification.

The VGGish architecture relies on 3x3 convolutional kernels with a sequence of operations. The model takes a 3-channel audio input and applies several convolutional layers: 64 filters followed by a max pooling layer, 128 filters and another max pooling layer, two additional convolutional layers with 256 filters each followed by another max pooling, and finally two layers with 512 filters each before the final max pooling. The network then utilizes three fully connected layers with sizes 4096, 4096, and 128, respectively. The ReLU activation function is used for all hidden layers to introduce non-linearity, while the final embedding layer employs sigmoid activation. This process yields a 128-dimensional embedding that serves as input for downstream tasks such as classification.

While fine-tuning the model, we freeze the convolutional layers and only train the final three linear layers and the output layer responsible for higher-level features. We use a learning rate of 3e-5 for the pre-trained model and 9e-5 for fine-tuning the model. We use the binary cross entropy loss and Adam optimizer with weight decay of 1e-5. The training batch size is 32. We applied a threshold of 0.4 to accept a value as positive. The pre-trained model was fully frozen, except for the final (output) layer. Training was done on a NVIDIA A100 GPU with 80GB memory.

A.2 CURATION OF AUDIO CLIPS

We collected the audio clips through various sources, manual collection, and web sources. For the manual collection, we used a RODE VideoMicro microphone to record audio clips across various locations in Bhopal, Madhya Pradesh, India. Our recordings encompassed the IISER Bhopal campus, Bairagarh, Lalghati Square, and Karond, each offering distinct acoustic environments representative of typical Indian cities.

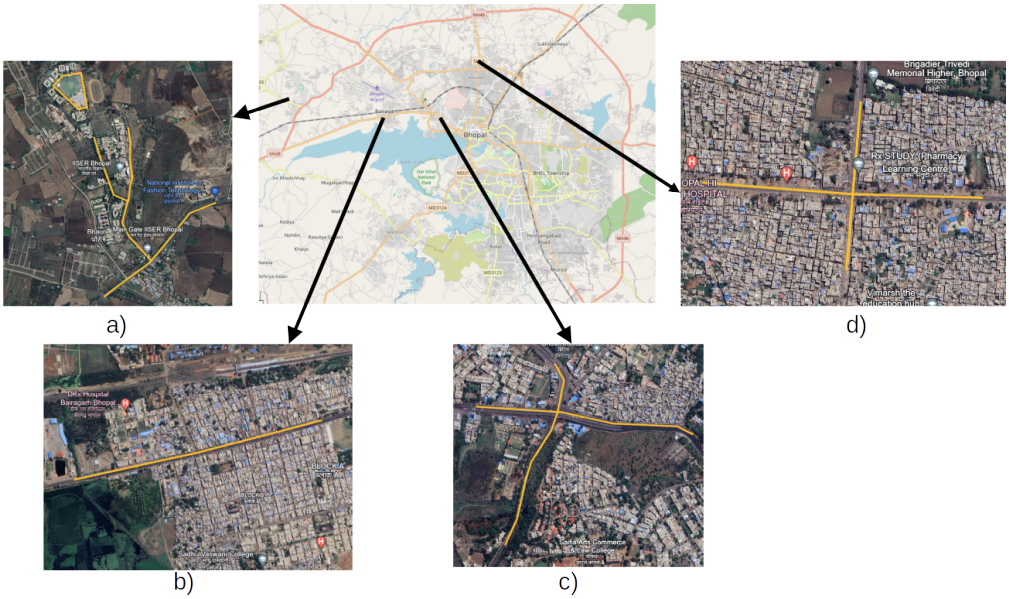

Figure 2: Various regions from the city of Bhopal where data was collected manually using a microphone, (a) IISER Bhopal, (b) Bairagarh, (c) Lalghati square, (d) Karond

We also extracted audio from various YouTube videos specific to regions of our interest. Keywords such as Indian street sounds, Indian city sounds, Indian traffic sounds, Indian market sounds, etc. were used to search the videos. We downloaded the selected YouTube videos and extracted audio from the videos at a sampling rate of 44.1 kHz. Our model uses a lower sampling rate of 16 kHz. To further increase the diversity in the dataset, we sourced audio clips from the Freesound database. Each audio clip was selected to represent the Indian environment. Featuring diverse durations ranging from a minute to over half an hour, our audio files were segmented into non-overlapping 4-second clips for efficient processing and subsequent labeling. Finally, we extracted and saved the trimmed clips as wav files in the pcm_s16le codec, since it is a lossless audio format.

Furthermore, to increase the variety in the dataset we sourced audio clips from the Freesound database. It was necessary that each audio clip selected accurately represents the Indian environment. For example, in case of the bicycle bell class, most of the audio clips found on Freesound did not correspond to the types of bells found on bicycle bells used in India. Similar to the YouTube clips, we then split the clips into 4 second long clips with FFmpeg, and saved them as wav files in the pcm_s16le codec.

We implemented a two-step annotation process. First, two annotators independently listened to each audio sample twice before assigning labels. In the second step, a script was utilized to test whether both annotators had assigned the same label to the audio. In instances of inter-annotator disagreement, a re-annotation process was initiated involving both annotators reviewing the audio sample until their classifications converged. This double-blind methodology aimed to minimize the influence of bias and guarantee the robustness of the annotation dataset.

The 'Human' class encompasses human speech, crowd chatter, and the occasional sounds of children playing. The 'Vehicle' class focuses on the sounds of horns and vehicles passing by. The 'Mammal' class is limited to sounds of mammals commonly found in Indian urban environments, such as dogs and cows. The 'Bird' class features birds chirping. The 'Bicycle Bell' class consists of sounds from

bicycle bells used on Indian bicycles. The 'Ambient Sound' class encompasses samples containing only background noise devoid of any of the above-mentioned sounds.

The selection of class labels for our dataset was driven by the composition of sounds found in the collected data. The resulting class balance reflects the natural proportion of each label encountered in the actual environment.

A.3 VISUALIZING SPECTROGRAMS

Figure 3 displays spectrograms illustrating the distinct characteristics of the various classes within our dataset. The x-axis signifies time, the y-axis denotes the frequency bins, and color signifies frequency amplitude at a specific time. The bicycle bell exhibits peaks at higher frequencies with a repetitive pattern. Bird chirps are repetitive, with the frequencies concentrated in a small area. Human speech displays inconsistency and brief periods of silence, featuring a more distributed frequency range. Dog barks peak at lower frequencies compared to other classes. The ambient sound class contains background noise but lacks a discernible pattern. The vehicle class lacks repetitive patterns but demonstrates distributed frequencies, exhibiting high values across all frequency ranges.

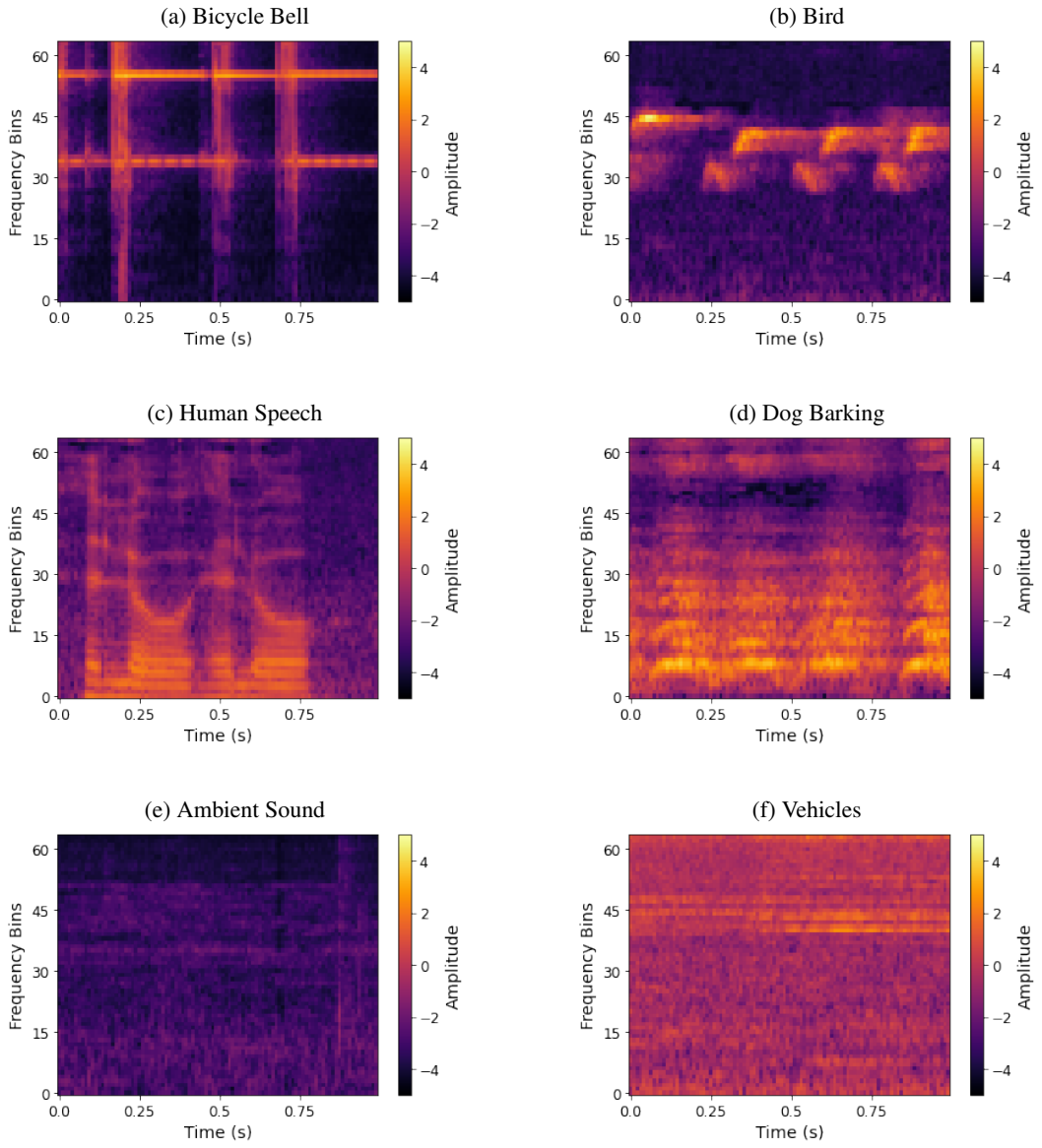

Figure 3: Spectrograms of various classes