Deep Learning-Based Background Noise Classification and Reduction for Audio Enhancement

August 15, 2024

1 Abstract

Background noise can deteriorate audio quality and interfere with the ability to concentrate on the intended sounds. The categorization and reduction of background noise are crucial in audio enhancement since they help to identify and diminish undesired noise. Implementing this technique improves audio recording intelligibility and quality, ensuring that the intended sound is dominant and free of background noise. The objective is to improve the user's auditory experience by minimizing ambient or unnecessary sounds in audio applications. In order to accurately categorize audio due to fluctuations in sound pitch and volume, a Deep Learning-Based Background Noise Classification and Reduction for Audio Enhancement is proposed in this research. By employing deep learning techniques, particularly Long Short-Term Memory (LSTM) networks, the methodology achieves an impressive accuracy rate of 88.35%. To train the proposed system, the Urbansound8 dataset is used, which has approximately 8,732 audio files in WAV format. The enhancement of background noise identification is achieved by employing an innovative sliding window technique that incorporates both audio wavelet characteristics and time information. In addition, a comparative analysis of Tiny Machine Learning (TinyML) models is used to construct a robust, efficient and effective background noise categorization and reduction model. Further research will develop an LSTM model using TinyML as a foundation to enhance and rectify noise in practical situations.

2 Introduction

Real-world audio recordings consist of a combination of various sound signals originating from diverse sources. Our everyday activities are outlined by sounds, which include the music we listen to, the discussions we have with others, and all the various sounds of our surroundings.

Background noise classification is the process of identifying and categorizing different types of background noise [25]. Selective background noise reduction is an essential procedure utilized in audio enhancement to improve the intelligibility and overall quality of audio content in a wide range of applications. For example, news information providers would like to label the huge amount of news audio data they collect every day in a reliable and easy way and video classification systems can use the audio information along with the video stream to achieve higher accuracy [5]. Through the process of isolating and eradicating extraneous noise, it guarantees that the primary audio maintains its clarity. In situations where speech intelligibility is critical, such as in recordings, film dialogues or conference calls, this is of the utmost importance. Within the domain of music production, this procedure aids in the enhancement of recordings by ensuring that the melodic components are prominently featured, devoid of any extraneous ambiance. Also, selective background noise reduction is very important in forensic analysis and the law because it makes it easier to get important information from audio evidence, which is very helpful for investigations and court cases. In its entirety, this methodology serves as an essential instrument for preserving the authenticity and excellence of auditory material across diverse domains, guaranteeing audiences and listeners a cohesive and engrossing auditory encounter. In recent years, the problem of classifying and identifying environmental sounds has received more attention in [3, 4, 16]. Particular circumstances may require the elimination of particular background noise as a result of the importance attributed to particular ambient sounds. For example, in situations pertaining to abductions, the presence of background noise may conceivably facilitate the pinpointing of the location.

While background noise has been acknowledged as a variety of audio in some prior studies pertaining to classification tasks, its impact on classification performance has not been extensively examined [25]. Only specific suspicious behaviors, such as those that include banging and screaming sounds, would be detectable by this type of surveillance system. This is because the supervised models would have been trained on sound classes that were obtained offline. But it's also important for the system to find strange events that it has never seen before [19]. In recent advancements in audio signal processing, deep neural networks have demonstrated significant progress, particularly in applications such as speech detection [13,28], audio enhancement [12,21], music information retrieval [6] and source separation [6,27]. A growing number of algorithms have been created to efficiently sort and analyze sound patterns. These algorithms use a variety of deep learning methods, such as CNN, RNN, and LSTM architectures [2,17]. The IEEE Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 dataset is used to compare different deep learning models for finding sounds in the surroundings [18].

This paper proposes a novel method for classifying and reducing background noise in audio improvement by utilizing deep learning techniques, particularly

Long Short-Term Memory (LSTM) networks. The methodology achieves a significant accuracy rate of 88.35% by using a large dataset consisting of 8,732 audio files in WAV format. The progress in identifying background noise is achieved through the use of a novel sliding window technique that combines audio wavelet properties with time information. In addition, a comparative analysis of Tiny Machine Learning (TinyML) models is used to create a strong, efficient and successful model for categorizing and reducing background noise. As for future study, we want to quickly and accurately sort background noise into different categories, so we plan to build an LSTM model based on TinyML's ideas.

The main contribution of this paper are as follows:

- A Deep Learning-Based Background Noise Classification and Reduction for Audio Enhancement was proposed to detect and reduce background noise using state-of-the-art deep learning techniques, particularly Long Short-Term Memory (LSTM) networks.
- The proposed methodology demonstrates a remarkable accuracy rate of 88.35% by using a large dataset consisting of 8,732 audio files in WAV format.
- A novel sliding window technique is introduced for background noise identification.
- The study conducts a comparative analysis of Tiny Machine Learning (TinyML) models. This analysis contributes to the construction of a background noise categorization and reduction model that is not only accurate but also efficient and resource-effective.

The rest of the paper is organized in the following manner: Section 3 represents the literature review. The proposed methodology and algorithm are described in Section 4. Section 5 illustrates the experimental results analysis. Performance evaluation is described in Section 6. Lastly, Section 7 concludes the paper.

3 Related Works

Wei Chu et al. [8] proposed a simpler FFT-based spectrum that is based on the self-normalization property of an early auditory model. This spectrum performs well in audio classification tasks even when there is noise. The research compares the proposed FFT-based spectrum to the auditory spectrum for speech, music, and noise categorization. For this job, a support vector machine (SVM) is used as the predictor. Salamon and Bello [24] introduced a shallow Convolutional Neural Network (CNN) for sound classification. They utilized the log-mel spectrogram as a feature in their approach. In addition, they demonstrated the advantages of employing data augmentation techniques for the UrbanSound dataset.

Nikhil Kandpal et al. [15] introduces a music enhancement model that sequentially enhances mel-spectrograms and synthesizes waveforms from them. The model is trained using high-quality public dataset samples and low-quality simulations of amateur recording artifacts. This model outperforms baselines in human MOS tests. Limitations in the study include the insufficiency of current objective measurements for audio augmentation, which fail to accurately capture human perception of music. The objective of Janvijay Singh et al.'s [25] study is to classify background noises in human speech-containing audio recordings via transfer learning based on convolutional neural networks. Two datasets are utilized in the paper: UrbanSound8K and YBSS-200. By employing data augmentation techniques, the classification outcomes are enhanced. The paper fails to present a comparative analysis with other contemporary models or methodologies for classifying background sounds.

The primary objective of Gupta [11], is to enhance speech signals by reducing noise. The Spectral Statistics Based on Minimum Statistics (SSBMS) method is introduced and assessed for estimating the power spectrum of non-standard noise signals using minimal statistics. The study evaluates the performance of SSBMS and improves its configurations for speech communication. The method is tested under various noise conditions and sound intensities using single-channel speech data in the study. This research admits the presence of little background noise in the system's output. Chu et al. [7], proposed a method to learn the initial background model using a semi-supervised learning approach. Their approach involved training classifiers to distinguish between foreground and background in an environment. They enhanced the background recognition procedure by self-training and classifying a significant amount of unlabeled audio data. They also developed an online K-means approximation algorithm to detect changes in the background. However, the model had limitations in classifying background sounds, resulting in a lower accuracy rate when unlabeled background noise merged with the foreground sound. Unfortunately, the source of the database used in the study is not specified.

A study conducted by Wei Chu et al. [9] introduced a simplified model for computing a new self-normalized spectrum based on the Fast Fourier Transform (FFT). The performance of this proposed FFT-based spectrum was evaluated through a three-class audio classification task, involving speech, music, and noise. In the classification task, a support vector machine (SVM) was used as the classifier. It achieved an error rate of 33.18%. Pablo Zinemanas et al. [29] introduced a new interpretable deep learning model named Audio Prototype Network (APNet) for automatic sound classification. The model utilizes a latent space consisting of learned prototypes to explain its predictions by measuring the similarity between the input and these prototypes. APNet consists of two main components: an autoencoder and a classifier. The suggested model underwent assessment across three sound classification tasks, encompassing speech, music, and environmental audio. The classifier component of APNet achieved an accuracy of 69.1% on the diverse datasets mentioned in the paper.

Regunathan Radhakrishnan et al. [23] propose a hybrid approach comprising two components: unsupervised audio analysis and audio classification using an offline-trained framework. Their approach uses a Gaussian Mixture Model (GMM) to represent background noise and it updates the model gradually as new audio data come in. However, the paper does not mention a specific accuracy rate, suggesting that the work may be incomplete or lacks this particular detail. Achyut Mani Tripathi et al. [26] presents a new deep learning model that incorporates attention mechanisms to identify semantically important frames in a signal's spectrogram. This attention-guided model effectively captures spatiotemporal relationships within the spectrogram data. The researchers evaluate the proposed method using two popular datasets for Environmental Sound Classification: ESC-10 and DCASE 2019 Task-1(A). The results show that the proposed model achieves an accuracy of 92% and 82% for the respective datasets.

Felix Gontier et al. [10] proposed a two-stage approach in their paper. Initially, they introduced a self-supervised stage where they developed a pretext task called Audio2Vec skip-gram inpainting. This task was applied to unlabeled spectrograms obtained from an acoustic sensor network. In the next stage, a supervised stage, they formulated a multilabel urban sound classification task using synthetic scenes. The system they developed achieved an average accuracy of 73.6%. When they removed the self-supervised learning stage, the accuracy slightly decreased to 72.9%. Additionally, reducing the synthesis of polyphonic training sets had a significant impact, resulting in an accuracy of only 49.6%.

4 Proposed Methodology and Algorithm

4.1 Proposed Methodology

The proposed system aims to reduce noise from audio files in a multi-phase approach. Initially, audio files from various sound sources are inputted. Subsequently, it employs wavelet decomposition techniques to separate the audio signals into background and foreground wavelets. The background wavelet is continuously analyzed and processed in a sequential and overlapping manner using a sliding window technique. This algorithm permits a structured analysis of consecutive parts of the original signal, facilitating the identification of important attributes and the assessment of the accompanying background interference. It plays a crucial role in the overall system workflow by facilitating ongoing evaluation and contributing to subsequent stages of noise recognition and removal. To categorize the noise sounds present in the audio files, LSTM

Table 1: Overview of Related Works.

Research	**Year**	**Dataset**	**Modeling**	**Accuracy**	**Classification type**
Janvijay et al. [25]	2019	UrbanSound, ESC50, and AUDIOSET	CNN, SVM, and FNN	72%	Different types of sounds
Chu et al. [7]	2009	Manual	Gaussian Mixture model technique	75.9%	Coffee room Courtyard Subway platform audio sound classification
Michel Olvera et al. [20]	2020	(DESED)	CNN	N/A	Foreground and background sound separation
Awotunde et al. [1]	2020	TIMIT	CNN	N/A	Speech segregation in background noise
Pablo Zinemanas et al. [29]	2021	UrbanSound8k	APNet, SB-CNN	69.1%	Sound Classification
Achyut Mani Tripathi et al. [26]	2021	ESC-10 dataset	SVM, GMM	82%	Environmental sound classification
Felix Gontier et al. [10]	2021	CENSE-2k	CNN	73.6%	Urban sound classification

is employed. Finally, the system combines the foreground audio sound with the noise-free background wavelet to produce the noise-free audio output shown in Fig 1.

4.2 Sliding Window

In Fig 2 the sliding window algorithm divides the sound wave into smaller frames, where the window size determines the duration of each frame. The LSTM model is utilized to perform sound categorization for each frame, specifically tailored to capture temporal dependencies in sequential data. At first, the LSTM model takes the current frame as input and predicts the sound class associated with it. Afterwards, the predicted sound class for each frame is stored, resulting in a list of predicted sound classes for the entire signal. By sliding the

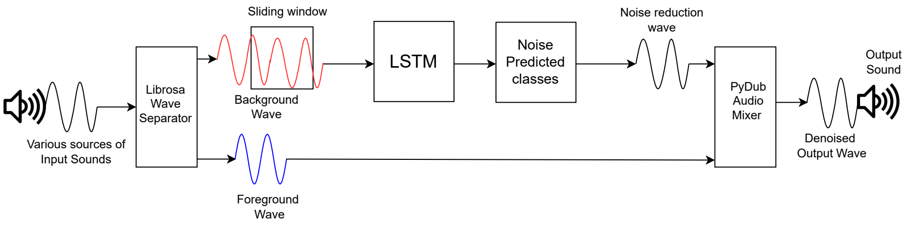


Figure 1: Proposed LSTM And Sliding Model

window with a certain hop size, the algorithm covers the entire sound signal, analyzing it frame by frame and capturing temporal information. This process continues until all frames have been analyzed. Through this approach, the algorithm effectively segments the sound signal, applies LSTM-based classification to each segment and produces predictions for the corresponding sound classes, enabling the classification of different sounds based on the analyzed frames.

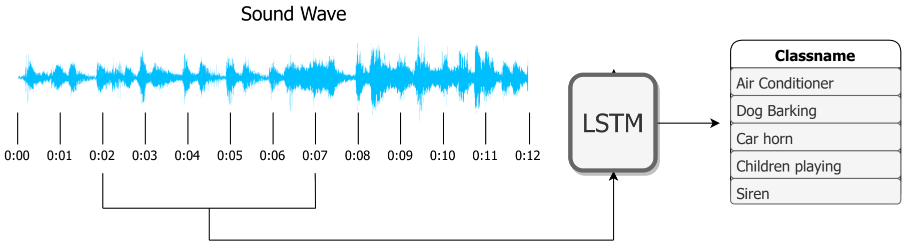


Figure 2: Sliding Window

4.3 Algorithm

4.4 Explanation of The Algorithm

Sound Classification Sliding Window with LSTM: Here's an expanded explanation of the sound classification sliding window algorithm using LSTM with a window size of 5 seconds:

- **Input:** The algorithm takes a sound signal, denoted as S, as input, along with a window size of W.
- **Parameters:** The sampling rate of the sound signal is denoted as SR.
- **Output:** The algorithm outputs the predicted sound classes.
- **Frame Calculation:** The frame size is calculated as the product of the window size W and the sampling rate SR. It determines the number of samples contained in each frame.

```
Algorithm 1 Noise Classification
Input: Sound signal S, Window size W
Output: Predicted sound classes
Parameters: Sampling rate SR
Frame size ← W × SR;
Hop size ← Frame size ÷ 2;
model ← load_model(”path_to_model.h5”)
Predicted sound classes ← empty list;
for start position = 0 to (length of sound signal - Frame size):
    Current frame ← sound signal[start position: start position + Frame size];
    Predicted class ← model.predict(np.expand dims(Current frame, axis=0))[0];
    Predicted sound classes.append(Predicted class);
endfor
```

- **Hop Size:** The hop size is computed as half the frame size, resulting in a 50% overlap between consecutive frames. It determines the number of samples to move forward when sliding the window.

- **Initialize LSTM Model:** An LSTM model is initialized with appropriate parameters. The specific architecture may vary but typically includes an LSTM layer followed by one or more dense layers for classification.

- **Initialize Predicted Sound Classes:** An empty list is initialized to store the predicted sound classes for each frame.

- **Sliding Window Process:** The algorithm iterates over the sound signal using a sliding window approach. Starting from the beginning of the signal, the window is moved by the hop size until the end of the signal is reached.

- **Frame Extraction:** At each position, a current frame is extracted from the sound signal. The frame consists of a segment of the sound signal corresponding to the frame size.

- **Classification:** The current frame is passed to the LSTM model for classification. The LSTM model predicts the sound class associated with the frame.

- **Storing Predicted Sound Class:** The predicted sound class is appended to the list of predicted sound classes.

- **Repeating the Process:** The sliding window process continues until all frames in the sound signal have been processed.

- **Output:** The algorithm returns the list of predicted sound classes as the output. Each predicted sound class corresponds to a particular frame in the sound signal, providing a classification for that segment.

5 Experiment and Result Analysis

5.1 Dataset

UrbanSound8K [22] is an openly accessible dataset of urban sounds. This dataset consists of ten different classes of 8,732 audio samples. Each sample has an average duration of approximately 4 seconds, resulting in a total of 9 hours of audio. The dataset is considered state-of-the-art and includes spectrograms of the audio. The provided spectrograms represent the easier [14] samples from the dataset, while the dataset itself contains numerous challenging samples where detecting distant and faint sounds of interest is difficult. The categories within the dataset encompass a wide range of sounds, including children playing, dog barks, street music, jackhammers, engine idling, air conditioners, drilling, sirens, car horns, and gunshots shown in Table 2 and the pie chart in Fig 3.

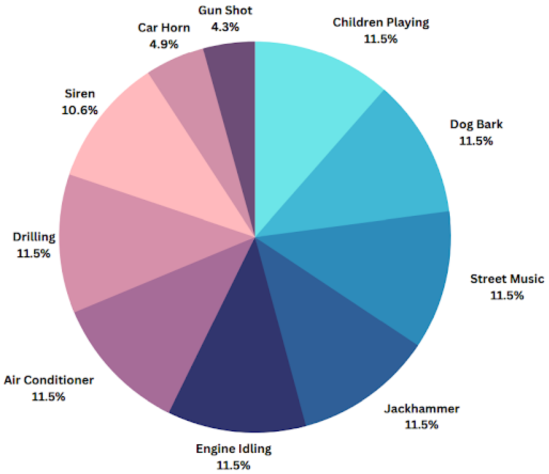


Figure 3: Pie Chart Of The Dataset

5.2 Experiment Setup

The implementation of the models utilized the free version of Google Colab, a cloud-based platform that provides access to GPU and TPU resources for training machine learning models without any cost. The models were developed using well-known machine learning frameworks like TensorFlow, Keras and Scikit-learn and the primary programming language used was Python.

6 Performance Evaluation

A dataset of over 8,732 audio samples, split into ten classes, was used to test how well two different federated models could recognize different types of audio

Table 2: Class of sound

Type	**Set**
Children Playing	1000
Dog Bark	1000
Street Music	1000
Jackhammer	1000
Engine Idling	1000
Air Conditioner	1000
Drilling	1000
Siren	929
Car horn	429
Gun Shot	374

Table 3: Evaluation Matrices

LSTM Model	Test Data	Train Data
Accuracy	88.35	99.81
Precision	88.47	99.82
Recall	88.27	99.81
F1-Score	88.27	99.81

noise. The models under scrutiny were LSTM-based.

Table 3 presents the assessment metrics for all the models developed and tested in this study. The LSTM models that were evaluated achieved the highest test accuracy, reaching an impressive 88.35%. Regarding the precision metric in the test data, our models performed well with values of 88.47%. Conversely, the recall metric showed more variability among the models, with LSTM achieving the value of 88.27%. The F1-score metric, which considers both precision and recall, showed that LSTM achieved the value of 88.27%. In the train data, the evaluated LSTM models achieved the highest train accuracy of 99.81%. Regarding the precision metric in the train data, our models performed well with values of 99.82%. On the other hand, the recall metric showed more variability among the models, with LSTM achieving the value of 99.81%. The F1-score metric, which considers both precision and recall, showed that LSTM achieved the value of 99.81%.

Loss and accuracy graphs are important tools for evaluating model performance in machine learning research. Loss graphs track the error between predicted and actual values during training, while accuracy graphs measure the proportion of correctly predicted values. Analyzing these graphs helps to assess

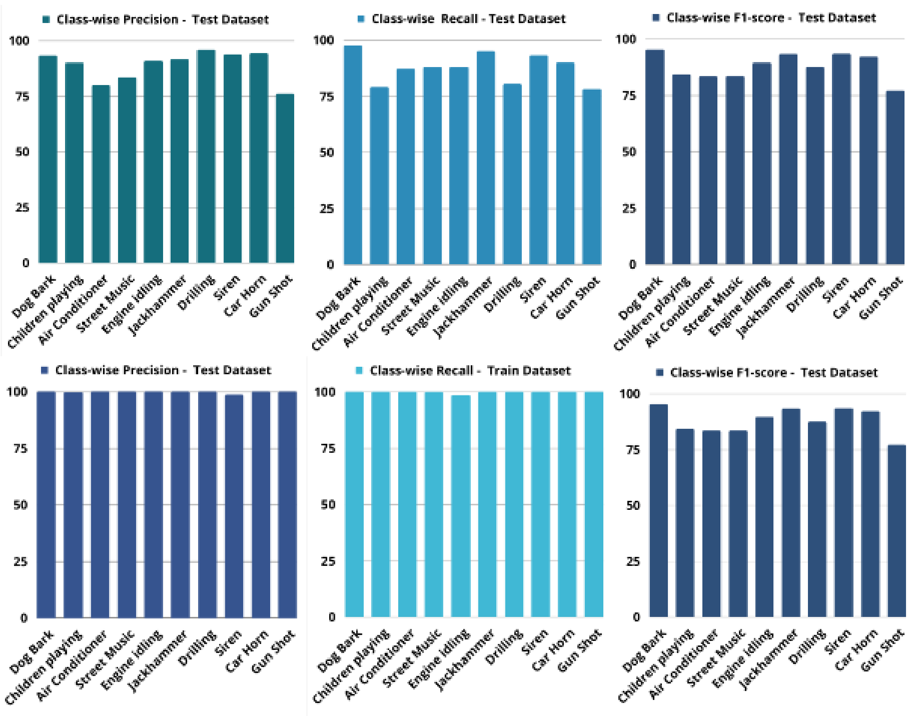


Figure 4: Train and Test data

model learning progress, identify over fitting or under fitting issues and understand the trade-off between model complexity and generalization. Comparison of loss and accuracy graphs can guide model selection, hyper-parameter tuning and model improvement strategies. Graphs are essential tools for evaluating and enhancing the performance of machine learning models in research scenarios.

In this context, Fig 5 displays the accuracy values of both the test and train data for various machine learning models across multiple epochs during testing and training. The x-axis represents the epochs, ranging from 1 to 10, while the y-axis represents the accuracy values, which range from 0 to 2. The lines depicted on the graph illustrate the accuracy trends observed for each individual model. Overall, the graph shows that the models LSTM consistently perform better in terms of accuracy.

Similarly, figure 4 shows the loss of test and train data are different machine learning models over epochs during testing and training. The x-axis denotes the epochs, ranging from 1 to 10, while the y-axis represents the corresponding loss values, ranging from 0 to 2.

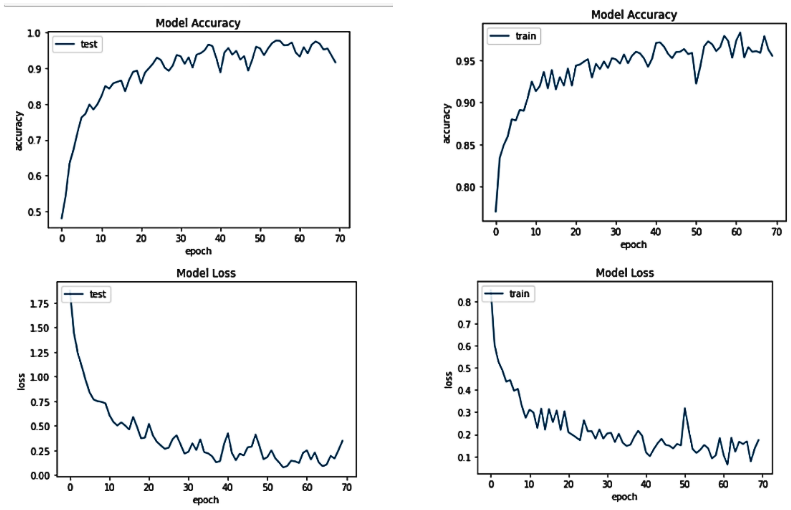
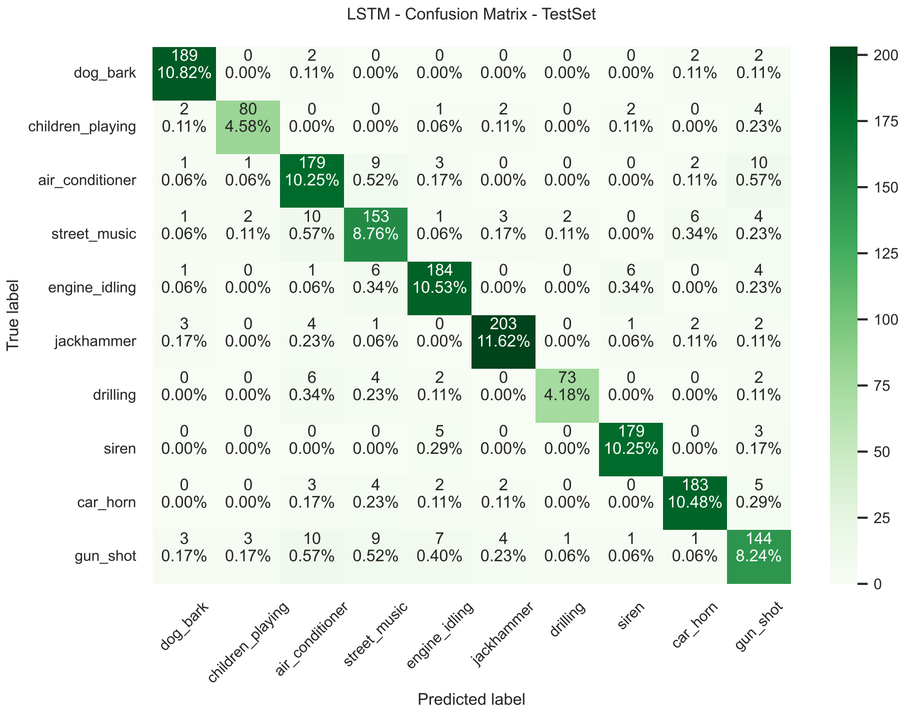


Figure 5: Model Accuracy and Loss



	dog_bark	children_playing	air_conditioner	street_music	engine_idling	jackhammer	drilling	siren	car_horn	gun_shot
dog_bark	189 10.82%	0 0.00%	2 0.11%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	2 0.11%	2 0.11%
children_playing	2 0.11%	80 4.58%	0 0.00%	0 0.00%	1 0.06%	2 0.11%	0 0.00%	2 0.11%	0 0.00%	4 0.23%
air_conditioner	1 0.06%	1 0.06%	179 10.25%	9 0.52%	3 0.17%	0 0.00%	0 0.00%	0 0.00%	2 0.11%	10 0.57%
street_music	1 0.06%	2 0.11%	10 0.57%	153 8.76%	1 0.06%	3 0.17%	2 0.11%	0 0.00%	6 0.34%	4 0.23%
engine_idling	1 0.06%	0 0.00%	1 0.06%	6 0.34%	184 10.53%	0 0.00%	0 0.00%	6 0.34%	0 0.00%	4 0.23%
jackhammer	3 0.17%	0 0.00%	4 0.23%	1 0.06%	0 0.00%	203 11.62%	0 0.00%	1 0.06%	2 0.11%	2 0.11%
drilling	0 0.00%	0 0.00%	6 0.34%	4 0.23%	2 0.11%	0 0.00%	73 4.18%	0 0.00%	0 0.00%	2 0.11%
siren	0 0.00%	0 0.00%	0 0.00%	0 0.00%	5 0.29%	0 0.00%	0 0.00%	179 10.25%	0 0.00%	3 0.17%
car_horn	0 0.00%	0 0.00%	3 0.17%	4 0.23%	2 0.11%	2 0.11%	0 0.00%	0 0.00%	183 10.48%	5 0.29%
gun_shot	3 0.17%	3 0.17%	10 0.57%	9 0.52%	7 0.40%	4 0.23%	1 0.06%	1 0.06%	1 0.06%	144 8.24%

Figure 6: Confusion Matrix

Figure 6 illustrates the confusion matrix, which offers valuable information about how different models perform on the test data. The matrix provides insights into the accuracy of class predictions across various categories. The results indicate that the majority of classes demonstrate favorable scores, with the exception of children playing and drilling. Despite these lower scores, the overall performance of the model is promising, as it successfully classifies the majority of samples accurately. This indicates that the model is generally successful in accurately recognizing the categories within the dataset.

7 Conclusion

The research contributes to the field of audio signal processing by offering an innovative solution that overcomes some of the limitations of existing techniques. The LSTM-based Tiny deep learning model exhibits improved accuracy and robustness in background sound categorization, enabling better identification and classification of different types of sounds in real-world scenarios. Furthermore, the denoising aspect of the technique enhances the quality of audio signals by effectively reducing unwanted background noise. This has significant implications in various domains such as speech recognition, audio recording, and acoustic analysis, where high-quality audio signals are essential for accurate interpretation and analysis.

References

[1] Joseph Bamidele Awotunde, Roseline Oluwaseun Ogundokun, Femi Emmanuel Ayo, and Opeyemi Emmanuel Matiluko. Speech segregation in background noise based on deep learning. *IEEE Access*, 8:169568–169575, 2020.

[2] Soo Hyun Bae, In Kyu Choi, and Nam Soo Kim. Acoustic scene classification using parallel combination of lstm and cnn. In *DCASE*, pages 11–15, 2016.

[3] Anam Bansal and Naresh Kumar Garg. Environmental sound classification: A descriptive review of the literature. *Intelligent Systems with Applications*, page 200115, 2022.

[4] Sachin Chachada and C-C Jay Kuo. Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, 3:e14, 2014.

[5] Lei Chen, Sule Gunduz, and M Tamer Ozsu. Mixed type audio classification with support vector machine. In *2006 IEEE International Conference on Multimedia and Expo*, pages 781–784. IEEE, 2006.

[6] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. A tutorial on deep learning for music information retrieval. *arXiv preprint arXiv:1709.04396*, 2017.

[7] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. A semi-supervised learning approach to online audio background detection. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1629–1632. IEEE, 2009.

[8] Wei Chu and Benoît Champagne. A noise-robust fft-based spectrum for audio classification. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[9] Wei Chu and Benoît Champagne. A noise-robust fft-based spectrum for audio classification. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[10] Félix Gontier, Vincent Lostanlen, Mathieu Lagrange, Nicolas Fortin, Catherine Lavandier, and Jean-François Petiot. Polyphonic training set synthesis improves self-supervised urban sound classification. *The Journal of the Acoustical Society of America*, 149(6):4309–4326, 2021.

[11] Monika Gupta, RK Singh, and Sachin Singh. Analysis of optimized spectral subtraction method for single channel speech enhancement. *Wireless Personal Communications*, 128(3):2203–2215, 2023.

[12] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

[13] Wenping Hu, Yao Qian, Frank K Soong, and Yong Wang. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166, 2015.

[14] jonnor.com. audio-classification-with-machine-learning-europython-2019. https://www.jonnor.com/2021/12/audio-classification-with-machine-learning-europython-2019/ , 2023, February 14. documentation.

[15] Nikhil Kandpal, Oriol Nieto, and Zeyu Jin. Music enhancement via image translation and vocoding. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3124–3128, 2022.

[16] Baljinder Kaur and Jaskirat Singh. Audio classification: Environmental sounds classification. 2021.

[17] Anurag Kumar, Maksim Khadkevich, and Christian Fügen. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 326–330. IEEE, 2018.

[18] Juncheng Li, Wei Dai, Florian Metze, Shuhui Qu, and Samarjit Das. A comparison of deep learning methods for environmental sound detection. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 126–130. IEEE, 2017.

[19] Fulufhelo V Nelwamondo, Tshilidzi Marwala, and Unathi Mahola. Early classifications of bearing faults using hidden markov models, gaussian mixture models, mel-frequency cepstral coefficients and fractals. *International Journal of Innovative Computing, Information and Control*, 2(6):1281–1299, 2006.

[20] Michel Olvera, Emmanuel Vincent, Romain Serizel, and Gilles Gasso. Foreground-background ambient sound scene separation. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 281–285. IEEE, 2021.

[21] Ashutosh Pandey and DeLiang Wang. A new framework for cnn-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7):1179–1188, 2019.

[22] paperswithcode.com. Dataset. https://paperswithcode.com/dataset/urban sound8k-1 , 2023, February 14. dataset documentation.

[23] Regunathan Radhakrishnan, Ajay Divakaran, and A Smaragdis. Audio analysis for surveillance applications. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pages 158–161. IEEE, 2005.

[24] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283, 2017.

[25] Raviraj Singh, Chuand Joshi. Background sound classification in speech audio segments. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE, 2019.

[26] Achyut Mani Tripathi and Aakansha Mishra. Environment sound classification using an attention-based residual neural network. *Neurocomputing*, 460:409–423, 2021.

[27] Alexander M von Benda-Beckmann, Paul J Wensveen, Filipa IP Samarra, S Peter Beerens, and Patrick JO Miller. Separating underwater ambient noise from flow noise recorded on stereo acoustic tags attached to marine mammals. *Journal of Experimental Biology*, 219(15):2271–2275, 2016.

[28] Xiao-Lei Zhang and DeLiang Wang. Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):252–264, 2015.

[29] Pablo Zinemanas, Martín Rocamora, Marius Miron, Frederic Font, and Xavier Serra. An interpretable deep learning model for automatic sound classification. *Electronics*, 10(7):850, 2021.