International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.131

Suspicious Activity Detection based on Audio Detecting Methodology using Deep Learning

Adarsh Shailendra
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, India
as8969@srmist.edu.in

Chirag Bengani
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, India
cb4849@srmist.edu.in

K. Shantha Kumari
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, India
shanthak@srmist.edu.in

P. Senthilraja
Dept. of CSE
K.S.Rangasamy College of Technology
Namakkal, India
visitsenthilraja@gmail.com

A Prithivi
Data Science and Business Systems SRM Institute of Science and Technology
Kattankulathur, India
ap5851@srmist.edu.in

Shruti Ramesh
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, India
sr9478@srmist.edu.in

Abstract*—Suspicious Activity refer to actions that appear unusual or questionable and may indicate the possibility of potentially illegal, harmful or illicit activities. These activities can be an early warning sign of criminal activity and their detection and prevention is necessary. This in turn helps in protection of assets, protection of individuals and prevention of the crime. The already increased rate of crime causes a lot of significant economic and personal damage. For detection, the counter is often standard human surveillance at all times. A smart surveillance system should be able to identify these activities through any means. In this paper, we propose a method using Deep Learning with Tensorflow to classify suspicious sounds like gun shots, jackhammer or glass breaking. A sophisticated microphone can be placed in Areas of interest like a Bank Locker or ATM and using the model, incoming sounds can be detected and classified. The Deep Learning Models that are trained, evaluated and tested are a Dense Model and an LSTM Model. The LSTM Model performs the best for identification of the sounds with an accuracy of 96.02%.

***Keywords*—Deep Learning, Suspicious Activity, Tensorflow, LSTM, Suspicious Sound, Sound Classification**

I. INTRODUCTION

In today's world, the amount of activity that takes place has exponentially risen, but this warrants that even unlawful or unwanted activities could have risen, which does not bode well for the society and economy in general. People have to show extra care and precaution today, especially in public places because of the fact that we are the most likely to come under scrutiny for any action that may look shady or incorrect. Apart from this factor, there also is an economic factor linked to any and all suspicious activities.

There has been a rise in crimes in the past few years, but we still have archaic methods of dealing with the consequences after. If we update our methods of dealing with the same, then we actually have a fighting chance to nip the illicit activities in the bud itself. That is achievable but often at a price. Our intention is to research and find ways that can make our life safe and secure using methods that are inexpensive to setup and also easy to update and fix.

Considering the presence of sophisticated hardware i.e. Sensitive Microphones in certain Areas of Interest which are difficult to monitor in Night Time or are usually isolated at night like ATMs, Bank Lockers or for personal surveillance.

The main purpose of the paper is to classify different types of suspicious sounds present in the dataset. The present sounds are Gun Shot, Jackhammer, Drilling, Dog Barking and Siren. This was done with the help of Deep Learning Models. The main aim of the paper is to develop a reliable and accurate model which can accurately classify sounds in their respective classes and distinguish them from other sounds. This will also in turn improve public safety by pairing it up with an alert system which can alert law enforcement or security personnel. Finally, the paper aims to compare the performance of multiple models to find the best fit model for the job.

To make the model able to classify different kinds of sounds, two Deep Learning models are constructed. The first model contains entirely of Dense layers provided by Tensorflow's Deep Learning API Keras , and dropout layers. The second model contains an LSTM Layer along with Dense and Dropout layers. The paper aims to classify the sounds based on their extracted features. The extracted features are in the form of Mel-frequency Cepstral Coefficients which collectively make up a Mel-frequency cepstrum. The models are trained, validated and evaluated on this data with appropriate training methods. The performance of the models are then compared and the model with the best performance is converted into an exportable format for further development.

In spite of having the accomplished crime departments who have been managing these issuesfor quite a while, why is there a need for DL, having said that answer is very straightforward, the experienced departments can check on 3-4 parameters or can cover and comb through footage manually only in batches at a time with the help of the human capability present on site at that moment but DL, on other hand, can cover the numerous batches in one go, once provided the sufficient computing resources. In our scenario,DL actually helps out by covering various batches that would otherwise not be covered because of the sheer amount of data that is present which would make it nearly impossible to actually reach the right conclusion. Our intentions here are to try to provide solutions that are scalable and not limited to only a particular use case.

DOI: rp-9788770040723.131

LITERATURE REVIEW

A. Phishing attack detection using Machine Learning

This paper [1] acts as our initial dive into the realm of how machine learning is used in various applications like physical and network security. We aim to use this paper as a path on how we can try and implement the machine learning concepts in our current setup to enhance and derive results that provide us with a better overall knowledge and view.

B. A sound monitoring system for prevention of underground pipeline damage caused by construction

In this paper[2], the usage of sound detection is in a very specific use case linked to the construction industry, in which they want to use acoustic signals to reduce pipeline damage. The other common noises are collected as environment noises and the system is applied via 2 layers, one which detects the suspicious sounds and the other that fine-tunes the results of the 1st layer. The testing of the model leads to results of 95% of the noises being detected which is good for improvements later.

C. Audio IoT Analytics for Home Automation Safety

This paper [3] speaks about how audio analytics can help in the household environment to detect abuse, violence and other unnatural activities that might take place. The audio is recorded and split into chunks in the server for training the model and classifying different categories. This system enforces home safety and if any suspicious sound is generated, then emergency services are informed immediately to take action.

D. Audio analysis for surveillance applications

In a paper written by R. Radhakrishnan et al [4]. A novel approach of analyzing time series like data of audio classes for detecting crime which can happen in elevators is developed and proposed. The authors showcase the disadvantages of a completely supervised machine learning based audio classification system and propose a hybrid solution to overcome them. It consists of two parts, in which the first part does unsupervised audio analysis and the second part which uses an audio classification framework. The framework uses a Gaussian Mixture Model (GMM) to identify background sounds and updates the model. The results seemed to be highly promising.

E. Smart Speaker: Suspicious Event Detection with Reverse Mode Speakers

In another paper written by G.Kalmar [5], an exploration into using speakers in reverse mode for detection of suspicious events has been done. The paper studies how a loudspeaker in reverse mode can be used to detect suspicious events like Gun Shots or screaming. To study the impact of reverse mode's distorting effects on the classification of events, a full transformation of traditional audio event datasets into forms as if they were recorded by speakers is done. The results suggested that speakers in reverse mode can be used for event detection.

F. A New Approach to Real Time Impulsive Sound Detection for Surveillance Applications

In this paper [6], the authors discuss the disadvantages of primary video surveillance systems and propose how they should be accompanied by other sensors. The paper makes a review of impulsive sound detection algorithms such as gun shots, explosions, screaming et cetera. It makes a review of all sound detection algorithms and noise detection algorithms as well. An adaptation of algorithms for detection of impulsive noise is also done. The results showcase that WLP (Warped Linear Prediction) can be used for impulsive sound detection.

II. METHODOLOGY

A. Overview

The paper's main purpose is to create and compare the performance of Deep Learning Models to identify which model is better and can be used for further development in this domain. The Model's main aim is to classify the sounds in their respective classes. The models are only trained to identify Suspicious Sounds like Drilling, Jackhammer, Dog Barking, Gun shots et cetera. The whole process of training the model requires a very crucial step i.e. Feature Engineering. Deep Learning models cannot be directly trained on sounds. They can however be trained on the features of the sounds that's why Feature Engineering is a very significant step for this whole process. After the training of the models, their evaluation on completely unseen data i.e. testing data is very important. This helps identify which model can work well when facing completely unseen data and is not overfit for the training data. Figure 1 depicts the entire process in a workflow diagram.

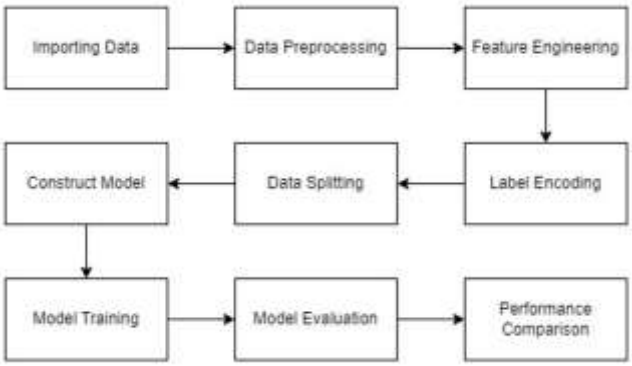

Fig. 1. Workflow Diagram

B. Environment

The entire project is implemented on Google Colaboratory, commonly known as Google Colab. It is a free, cloud-based Jupyter notebook environment that allows users to write Python code through the browser. Hosted by Google, Colab provides access to powerful computing resources such as Graphical Processing Unit (GPU) and Tensor Processing Unit (TPU), which allows for seamless work on the platform. Google Colab is thus particularly useful for researchers focused on data science and machine learning, who wish to collaborate with others on coding projects in a shared environment without hindrances. Unfortunately, Google Colab might be unable to process larger datasets as it is a limited platform, and does not support all Python libraries. However, for the purposes of this paper, Google Colab is a suitable work environment that is more than capable of handling the necessary datasets and

DOI: rp-9788770040723.131
libraries. In order to increase the scalability of this project, environments such as Kaggle can also be utilized.

C. Importing Data

Importing the data from its source can be a hassle. Specially when the data is large. We used the UrbanSounds8k Dataset [7] which contains 8732 sound samples of urban sounds like engine idling, children playing, street music et cetera. To import this data into our Google Collaboratory workspace, we use an external library called opendatasets [8] which uses Kaggle's official API to download datasets at high speeds.

D. Data Preprocessing

Data preprocessing is a crucial step in any machine learning project. It is a technique that involves transforming raw data into well-formed datasets and cleaning the raw data in order to check whether or not it is fit for analysis. More often than not, raw data is incomplete and inconsistent. Therefore, preparing the data has a direct impact on the final results post-analysis.

The main steps that are involved in Data Preprocessing include data cleaning, integrating the data. Transforming the data so that it can be used for different purposes and reducing the dimensions of the data. The dataset contains sound samples of different classes. Out of these classes, we only need five classes which can be categorized into Suspicious Sounds. We separate these classes and make a new dataset with only the metadata of sound samples of these classes. To depict the difference in these sound samples and a normal sound sample, we plot the waveform of the sounds.

The waveforms of the sound samples clearly represent the difference between them. This was made possible with the help of librosa [9]. librosa is a python package used for audio analysis. It makes the building blocks of audio processing systems. Figure 2 represents the waveforms of the class 'Dog Bark' and figure 3 indicates howthe waveform of the class 'Gun Shot' looks

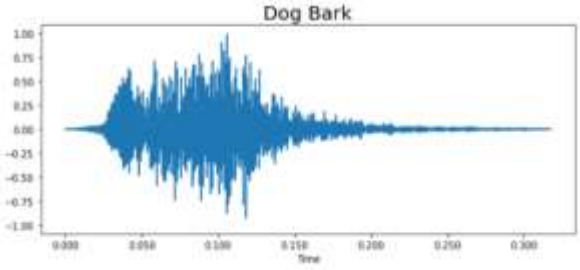

Fig. 2. Waveform of a Dog Bark

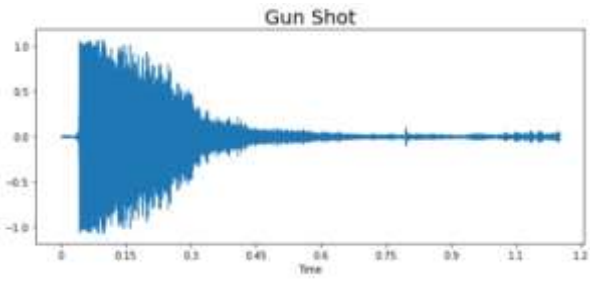

Fig. 3. Waveform of a Gun Shot

The same difference can be represented in the form of a spectrogram [10]. Figure 4 depicts the spectrogram for the class 'Dog Bark' and figure 5 depicts the spectrogram for class 'Gun Shot'

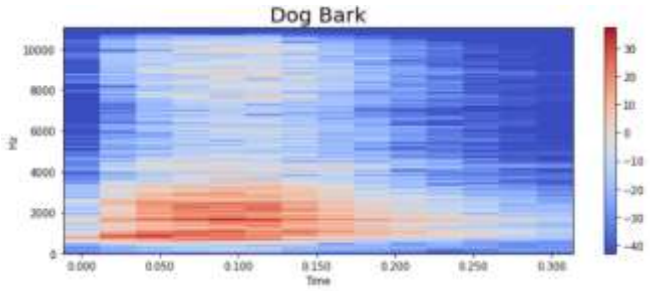

Fig. 4. Spectrogram of a Dog Bark

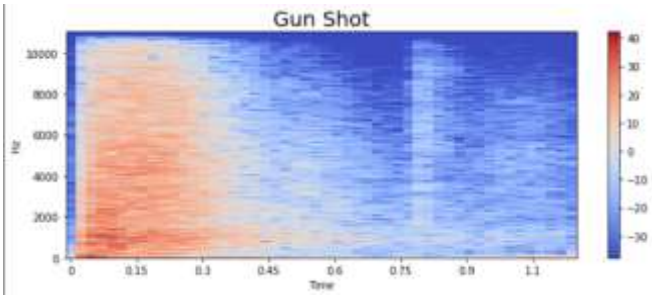

Fig. 5. Spectrogram of a Gun Shot

E. Feature Engineering

Deep Learning models are not able to identify sound samples in their raw forms. Providing a raw sound sample to a Deep Learning model for training yields nothing. For a Deep Learning model to identify a sound sample, its features can be supplied to it for training. For training our models, we extract the features of the sound samples. Since their are multiple types of features which can be extracted from a sound sample, we work with Mel-frequency Cepstral Coefficients [11]. The MFCC collectively make up the MFC or the Mel Frequency Cepstsrum which represents the change in the acoustic power of a sound based on a cosine transform of the power spectrum on a mel scale. This is done with the help of librosa. After extraction of features of every sound sample, a new dataset with the extracted features over fifty steps and its class labels is produced. The class labels are the encoded in the form of a distinct integer. This is called label encoding and is performed with the help of scikit-learn which provides efficient tools for data analysis and a premium library for Machine Learning in a Python Ecosystem.

F. Data Splitting

Data splitting enables us to split the whole data into bits which will be used for different purposes. In our case, we split our data into three subsets. The training data which is 80% of the original data, the validation data which is 10% and the testing data which is 10% as well. The training data will be used by the model for training, while it makes a pass over it and backtracks to adjust its wrights accordingly, and the validation data is used for checking how the model performs on relatively unseen data. Validation data is not used for training at all. The testing data in the end will be

DOI: rp-9788770040723.131

used for evaluating the model's performance finally on a completely unseen subset of data.

G. Model Construction

We For the betterment of results, we construct two relatively different models. The first model has four hidden layers. All these layers are Dense layers from the Keras API [12]. All these layers use 'relu' [13] as their activation function. The final output layer is a Dense layer with 'softmax' activation function [14] which is used in Multi-class classification. The model uses Categorical Crossentropy as its loss function along with Adam as its optimizer. This model is referenced as Dense Model throughout the paper.

The second model, referenced as the LSTM Model throughout the paper consists of three hidden layers. The first layer of this model is a LSTM layer [15]which returns only the final output. LSTMs are characterized by their ability to selectively remember or forget information from previous time steps, making them well-suited for tasks involving sequences of input data. Figure 6 depicts the architecture of a LSTM cell.

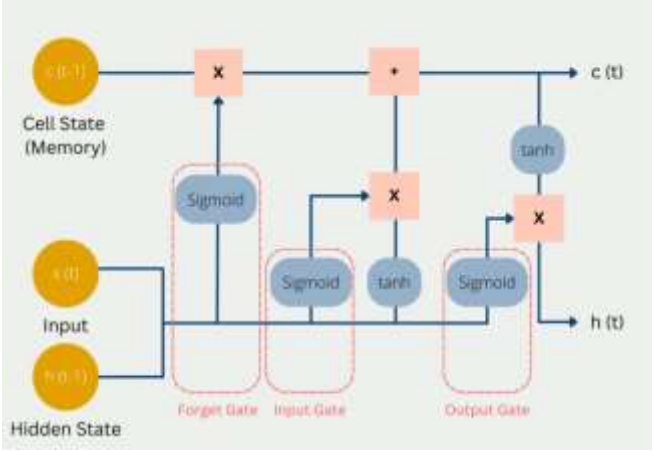

Fig. 6. Architecture of LSTM Cell

All the dense layers use 'relu' activation function. The output layer of the model uses 'softmax' activation function for multi-class classification. The model uses Categorical Crossentropy as its loss function along with Adam as its optimizer.

H. Model Training and Evaluation

Model training adjusts the weights of the models according to the training data provided. For our paper, we set both the model to be trained for 200 epochs with a batch size of 32. Along with this, an optional callback of Early Stopping is used. Early stopping observes a prescribed metric of the model while training. If the said metric does not increase or decrease or remains constant for a specific patience value i.e. number of epochs, Early Stopping stops the training of the model and reserves the best weights till the last epoch.

Finally, after training the model, a separate performance evaluation of the models is done on the Test Data. All the metrics of performance of the models on the Training, Validation and Testing data are used for evaluation.

III. Results And Discussuions

The primary aim of this study was to classify suspicious sounds into their respective classes based on their extracted features. This aim was achieved with the help of Deep Learning concepts in which our approach is to take two models and compare their performance to evaluate which model works better and can be used in a real-time audio stream. One point to be noted is that both models have not been trained for the same number of epochs, due to the usage of 'EarlyStopping'. Early Stopping is a method in Deep Learning that helps you to decide an arbitrary number of epochs at which the training of the model stops due to no improvement of results on the validation dataset.

A. Model I- Dense Model

The first model overall performed well on the dataset. It trained for 57 epochs and achieved a maximum training accuracy of 85.96% and a maximum validation accuracy of 88.63%. The training and validation losses of the model were 0.3889 and 0.3779 respectively. Fig 7 gives us a graphical representation of the training and validation accuracy. Fig 8 gives us the loss values for the training and validation loss.

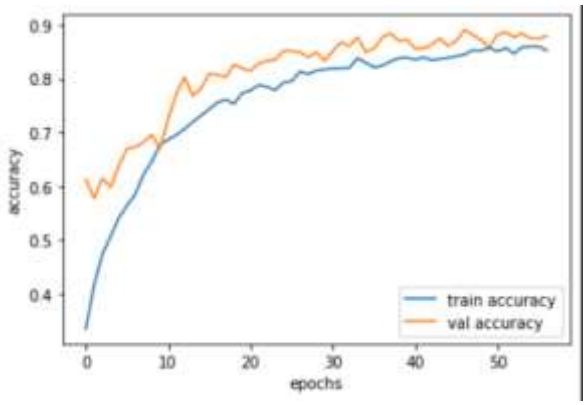

Fig. 7. Training and Validation Accuracy (Model I)

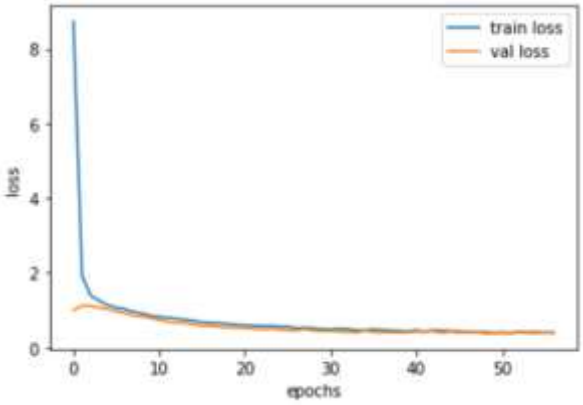

Fig. 8. Training and Validation Loss (Model I)

B. Model II- LSTM Model

This model performed better when compared to the previous model. It trained for 37 epochs and The model reached a training accuracy of 96.02% maximum and a validation accuracy of 90.49% maximum. The training and validation losses of the model are 0.1036 and 0.3184

respectively. Fig 9 gives us a look at the training and validation accuracy curves. Fig 10 gives us the loss values for the training and validation loss.

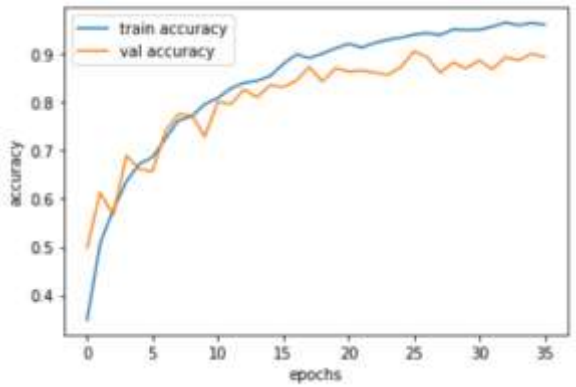

Fig. 9. Training and Validation Accuracy (Model II)

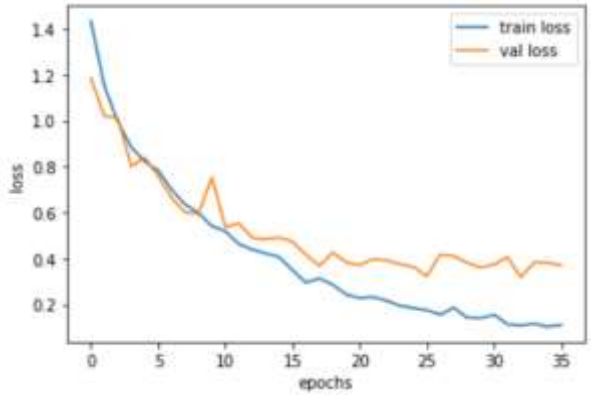

Fig. 10. Training and Validation Loss (Model II)

To compare the performance of the models in a more structured and elaborative manner, two tables are made to depict their performance. These tables reference to the performance of both the models collectively on their training data, validation data and testing data respectively. Table I gives us an overview of the training and validation loss of the model.Table II gives us an insight on the training, validation and testing loss of both the models..

TABLE I. TRAINING AND VALIDATION ACCURACY OF THE MODEL

Model	Training Accuracy	Validation Accuracy	Test Accuracy
Model I	85.96%	88.63%	88.86%
Model 2	96.02%	90.49%	90.25%

TABLE II. TRAINING AND VALIDATION LOSS OF THE MODEL

Model	Training Loss	Validation Loss	Test Loss
Model 1	0.3889	0.3779	0.3488
Model 2	0.1036	0.3184	0.3471

After analysing and comparing the results for both models, it can be said that Model 2 which is the LSTM model can be further developed and used to detect suspicious sounds in real-time continuous audio streams.

IV. CONCLUSION

The result of our paper, as proven by the Results and Detection Section shows us that the LSTM model performed better than the Dense Model. This in turn helps us prove that interpreting the features as a time-series array proves beneficial for sound classification models. There can be addition of more classes of suspicious sounds like footsteps or glass breaking, which will allow the model to detect a broader spectrum of classes, which also makes the smart surveillance system even smarter. We can also improve it to detect a live audio feed and also to alert personnel when needed.

REFERENCES

[1] Sundara Pandiyan S, Prabha Selvaraj, Vijay Kumar Burugari, Julian Benadit P, and Kanmani P, "Phishing attack detection using Machine Learning", Measurement: Sensors, vol. 24, pp. 2665-9174, 2022,ISSN,https://doi.org/10.1016/j.measen.2022.100476. (https://www.sciencedirect.com/science/article/pii/S2665917422001106)

[2] Dhanabalan, S. S., Sitharthan, R., Madurakavi, K., Thirumurugan, A., Rajesh, M., Avaninathan, S. R., & Carrasco, M. F. (2022). Flexible compact system for wearable health monitoring applications.Computers and Electrical Engineering, 102, 108130.

[3] S. K. Shah, Z. Tariq, and Y. Lee, "Audio IoT Analytics for Home Automation Safety," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 5181-5186, doi: 10.1109/BigData.2018.8622587.

[4] R. Radhakrishnan, A. Divakaran and A. Smaragdis, "Audio analysis for surveillance applications," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005., New Paltz, NY, USA, 2005, pp. 158-161, doi: 10.1109/ASPAA.2005.1540194.

[5] G. Kalmar, "Smart Speaker: Suspicious Event Detection with Reverse Mode Speakers," 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, 2019, pp. 509-512, doi: 10.1109/TSP.2019.8769024.

[6] Arslan, and Yuksel,"A New Approach to Real Time Impulsive Sound Detection for Surveillance Applications," DOI.org (Datacite), 2019. https://doi.org/10.48550/ARXIV.1906.06586.

[7] 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 (MM '14)," Association for Computing Machinery, New York, NY, USA, pp. 1041–1044, 2014. https://doi.org/10.1145/2647868.2655045

[8] https://github.com/JovianHQ/opendatasets

[9] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[10] \https://en.wikipedia.org/wiki/Spectrogram

[11] https://en.wikipedia.org/wiki/Mel-frequency_cepstrum#:~:text=Mel-frequency cepstral coefficients (MFCCs,-a-spectrum").

[12] https://keras.io/api/layers/core_layers/dense/

[13] https://en.wikipedia.org/wiki/Rectifier_(neural_networks)

[14] https://en.wikipedia.org/wiki/Softmax_function#:~:text=The softmax function is often,based on Luce's choice axiom

[15] https://keras.io/api/layers/recurrent_layers/lstm/