B.Tech project

ACOUSTIC GUARD :SECURITY CHECK SYSTEM USING CNN

Presented by Lovkash Garg
Mis No: 112215104 (2nd Year)

Abstract

There are many sounds all around us and our brain can easily and clearly identify them. Furthermore, our brain processes the received sound signals continuously and provides us with relevant environmental knowledge. Although not up to the level of accuracy of the brain, there are some smart devices which can extract necessary information from an audio signal with the help of different algorithms. In this presentation we converge audio classification with real-world applications. Several model like ANN, CNN, LSTM (long short term memory) are used . Through a comprehensive overview, we unravel the process of feature extraction, model training, and evaluation metrics, empowering you to discern birdcalls . Embark with us as we explore the fusion of spectrograms, MFCCs, and convolutional neural networks, illuminating the path towards robust audio classification systems.

Abstract

The novelty of our research lies in showing that the long-short term memory (LSTM) shows a better result in classification accuracy compared to CNN for many features used. Moreover, we have tested the accuracy of the models based on different techniques such as augmentation and stacking of different spectral features. In such a way it was possible with our LSTM model, to reach an accuracy of 98.81%, which is state-of-the-art performance on the UrbanSound8k dataset. Here We have upto 90 % accuracy .

Motivation of the Idea

Imagine a Cricket match without a Commentry, A Youtube video or a Reel. Sounds Boring ???
Media has been one of the Tool which is consumed by the most of the Individual. Researching about the topic selection of topic . I found a Interesting insight about how are sound recognizd by the computer and Machine . How does a glass break system work ??
It is Interesting How the Sound Travels . How sound wave can be represented in form of arrays.

How Sound Travels

Sound travels in waves by vibrating matter molecules through a medium (gases, liquids, and solids)
Sound can not travel through empty space, because there are no molecules to vibrate

Sound travels through a gas

Sound Travels through a liquid

Sound Travels through a solid

Introduction

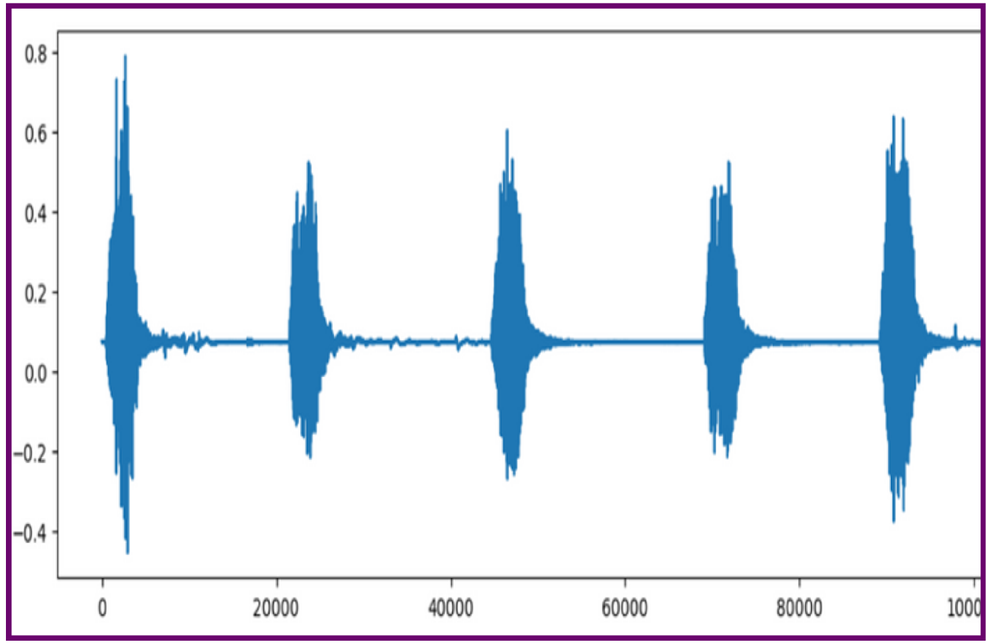
The primary goal of audio classification is to automate the process of identifying and organizing audio data, making it easier to search, analyze, and manage large volumes of audio content. This is particularly useful in scenarios where manual classification is impractical or time-consuming.

The process of audio classification typically involves extracting relevant features from audio signals and using machine learning algorithms to learn patterns and relationships within the data. These features may include spectral features such as frequency content, temporal features such as rhythm and timing, and higher-level features .

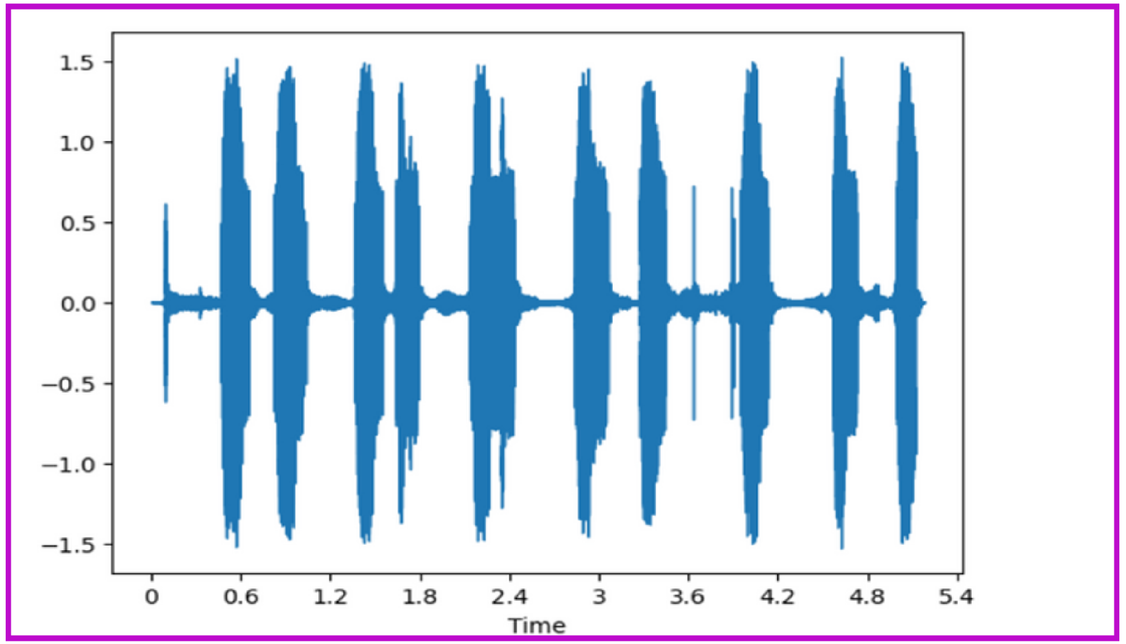
	Mono	Stereo
Introduction (from Wikipedia)	Monaural or monophonic sound reproduction is intended to be heard as if it were a single channel of sound perceived as coming from one position.	Stereophonic sound or, more commonly, stereo, is a method of sound reproduction that creates an illusion of multi-directional audible perspective.
Cost	Less expensive for recording and reproduction	More expensive for recording and reproduction
Recording	Easy to record, requires only basic equipment	Requires technical knowledge and skill to record, apart from equipment. It's important to know the relative position of the objects and events.
Key feature	Audio signals are routed through a single channel	Audio signals are routed through 2 or more channels to simulate depth/direction perception, like in the real world.
Channels	1	2

librosa

Normal Sound Wave

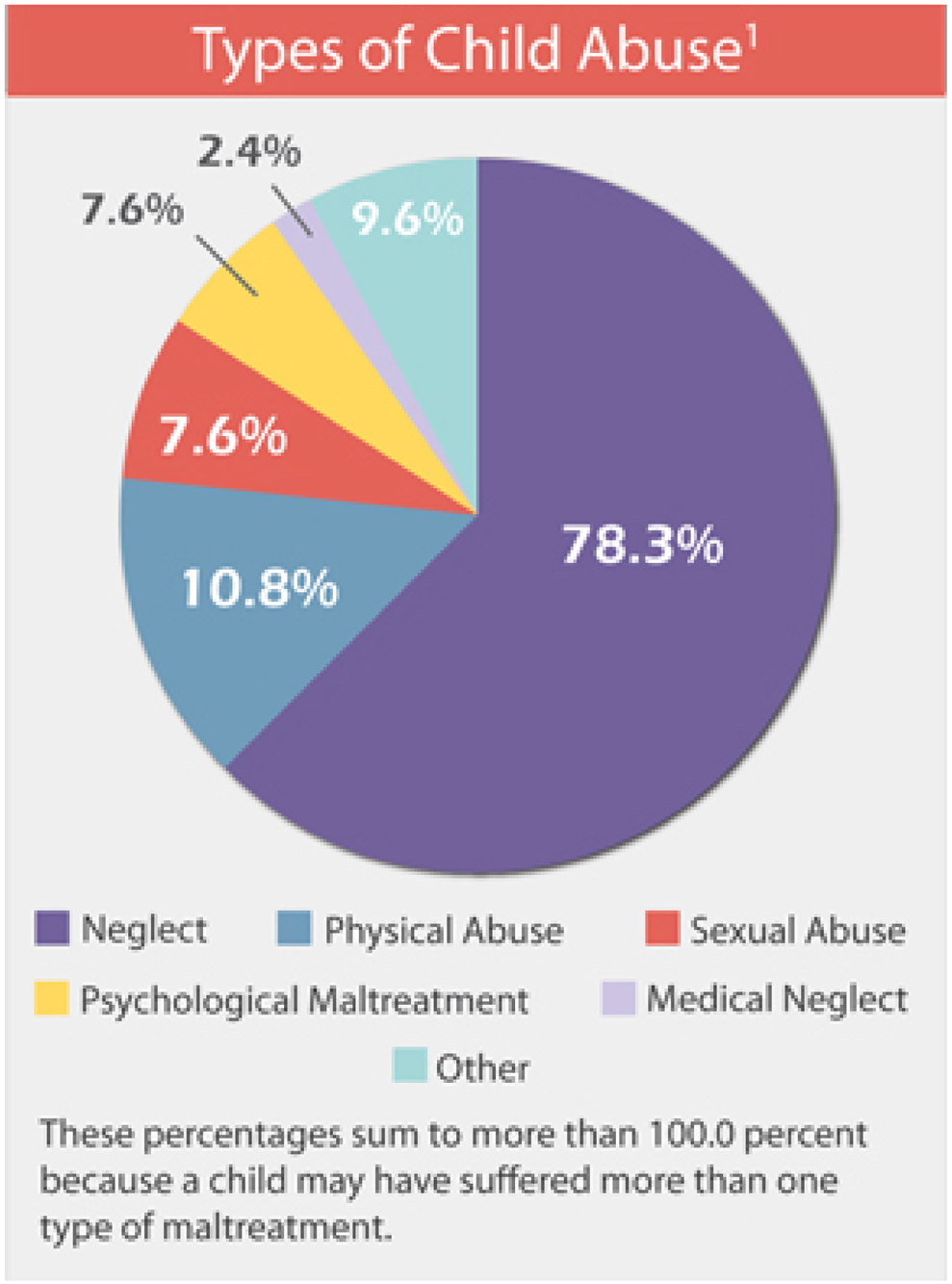


With Librosa



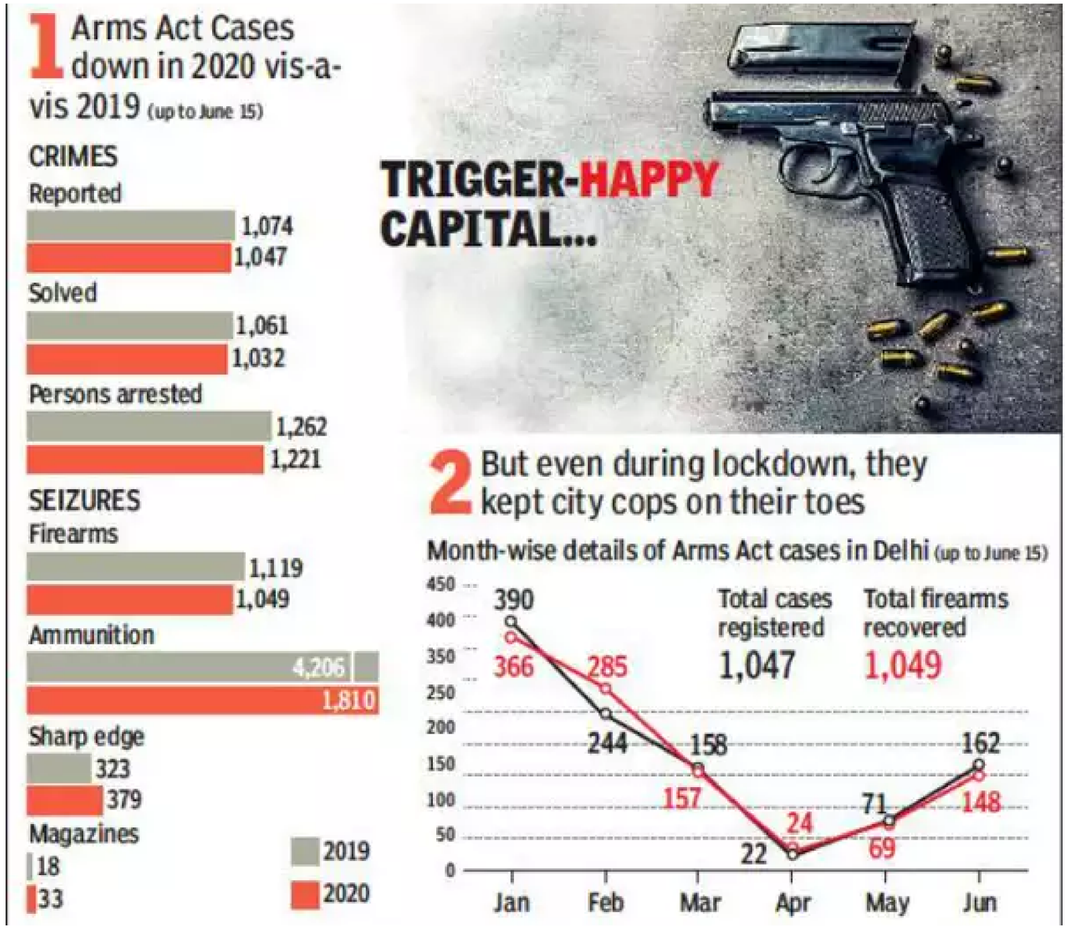
Applications of Audio Classification

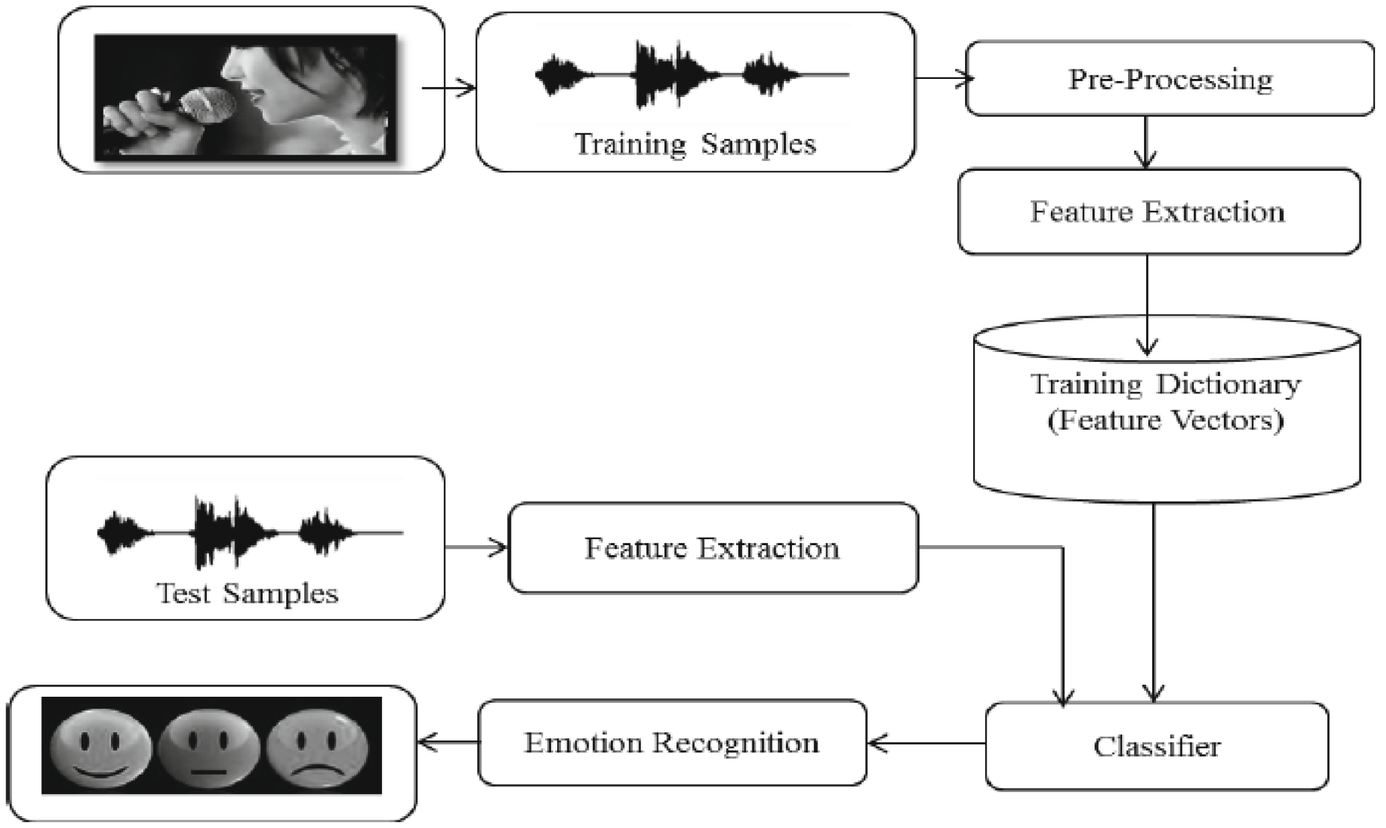
- Emotion Recognition: Analyzing audio data to detect emotions such as happiness, sadness, anger, etc., finds applications in customer service (e.g., analyzing call center conversations), mental health monitoring, and affective computing.
- Security Breaches and Protection:It is also used for predictive maintenance by detecting sound discrepancies in factory machinery, detecting security breaches, .
- Fauna Recognition:It is even used to differentiate animal calls for wildlife observation and preservation and also birds species recognition.
- Medical Diagnosis: Analyzing sounds from medical devices such as stethoscopes or ultrasound machines can aid in diagnosing conditions like heart murmurs, respiratory disorders, and fetal abnormalities.
- Content Moderation: Detecting and filtering inappropriate or harmful audio content in platforms such as social media, online gaming, and streaming services to ensure a safe and healthy online environment.



Vulgarity on Internet!!!

Emotion Recognition using audio Classification





Scikit Learn

Scikit-learn is one of the most popular open-source machine learning libraries for Python. Here are some of its most important points:

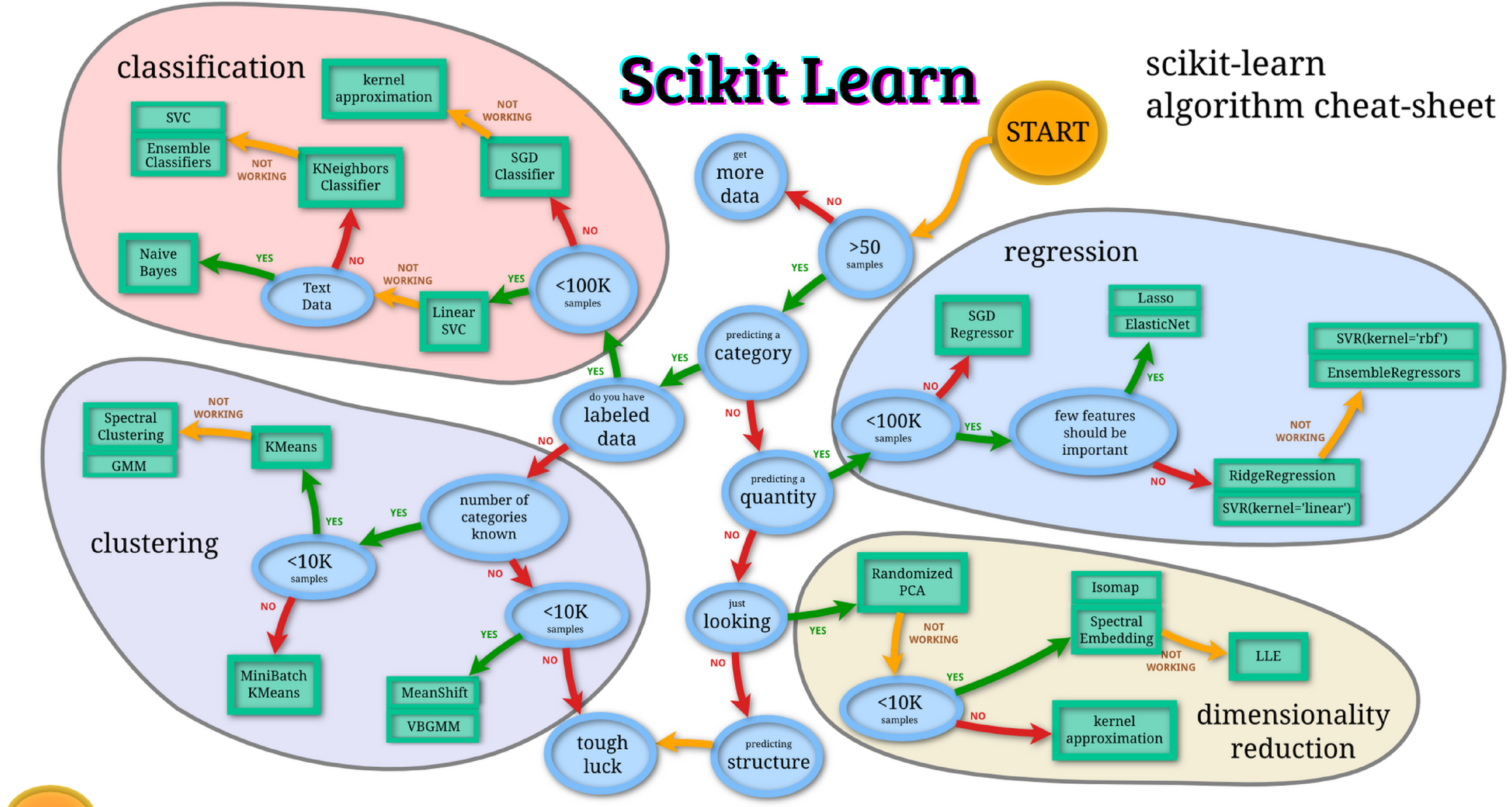
- Ease of Use: Scikit-learn provides a simple and efficient interface for implementing various machine learning algorithms. Its makes it easy for users to quickly prototype and deploy machine learning models.

- Wide Range of Algorithms: It offers a rich set of algorithms for various machine learning tasks, including classification, regression, clustering, dimensionality reduction, and more. These algorithms are implemented in a modular fashion, making it easy to experiment with different techniques.

- Integration with NumPy and SciPy: Scikit-learn seamlessly integrates with other Python libraries such as NumPy and SciPy, making it easy to manipulate data and perform advanced mathematical operations.

Scikit Learn

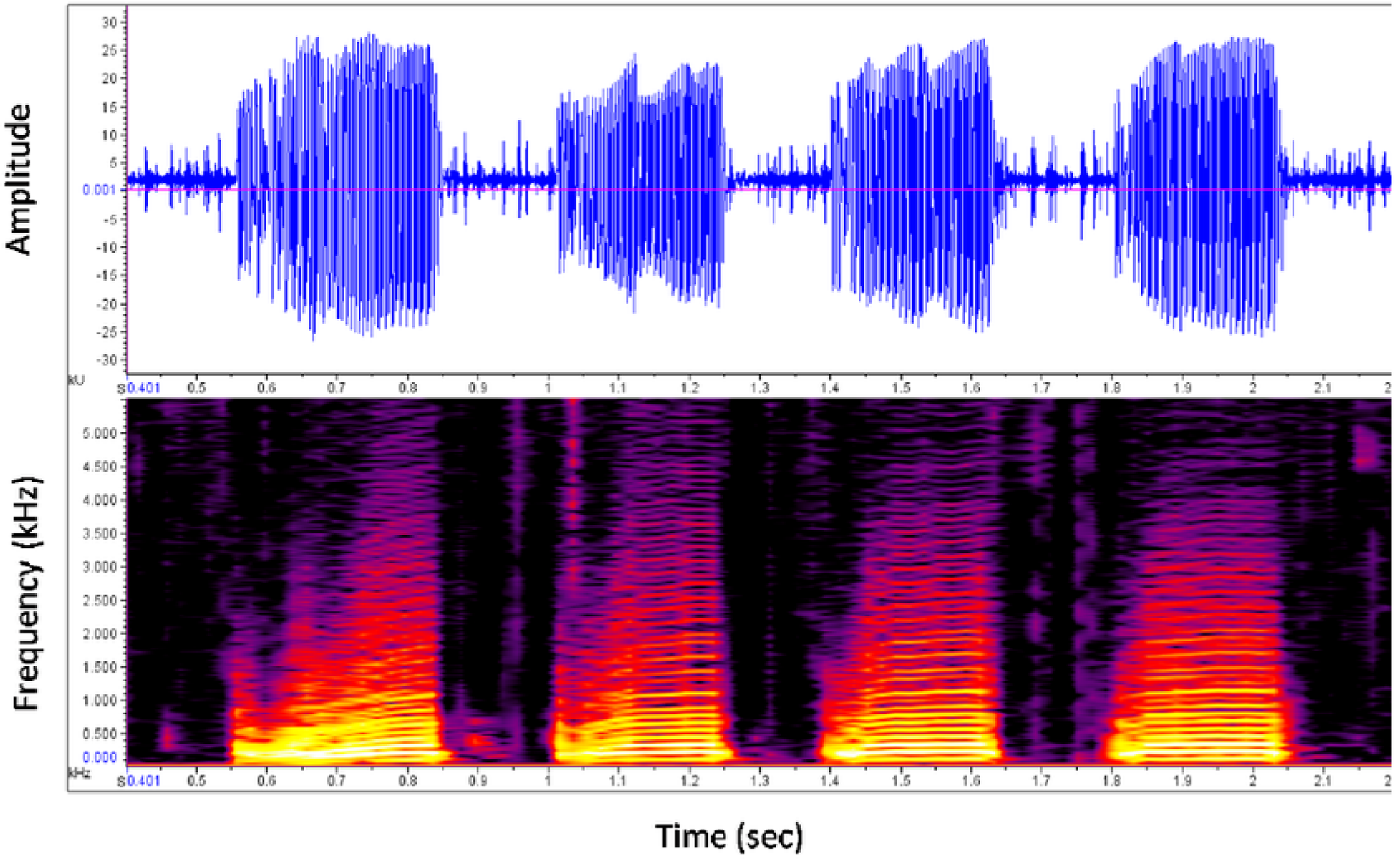
scikit-learn
algorithm cheat-sheet



Back

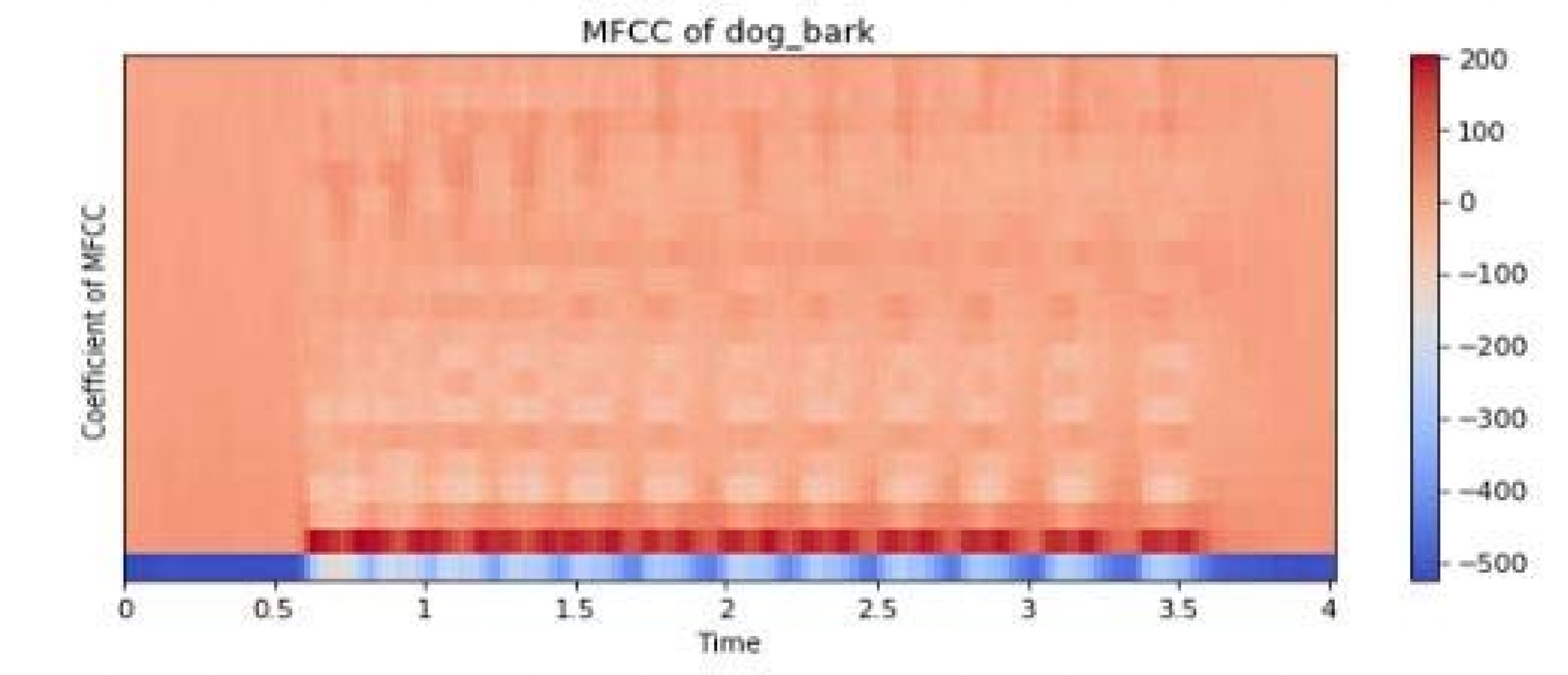
scikit learn

Spectogram



MFCC

Mel frequency cepstral coefficients in figure are compact representations of the spectrum are typically used to automatically identify speech and it is also used as a primary feature in many research areas that include audio signals

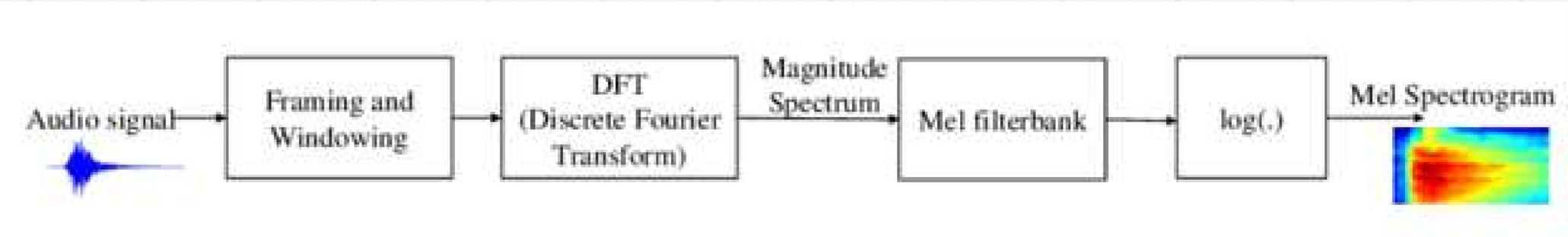


What is MFCCs ???

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum. This frequency warping can allow for better representation of sound, for example, in audio compression that might potentially reduce the transmission bandwidth and the storage requirements of audio signals.

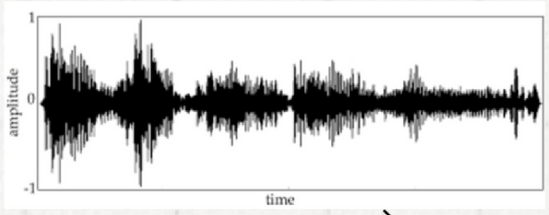
MFCCs are commonly derived as follows:

- Take the Fourier transform of (a windowed excerpt of) a signal.
- Map the powers of the spectrum obtained above onto the mel scale, using triangularoverlapping windows or alternatively, cosine overlapping windows.
- Take the logs of the powers at each of the mel frequencies.
- Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.



Audio wav File

Math plot



Librosa

```
librosa_data # from librosa
# Have range different from -1 to 1
```

```
array([-1.0687611e-14,  1.9862175e-13,  4.1020312e-13, ...,
        1.2874905e-05,  2.4051074e-06,  0.0000000e+00], dtype=float32)
```

CSV

Pandas

Metadata

MFCC feature classification

MFCC Scaled feature

Predicted Sound Category

Input File

Model.predict

Extracted Features

Sklearn training Model

Model.fit with fixed epochs and Batches

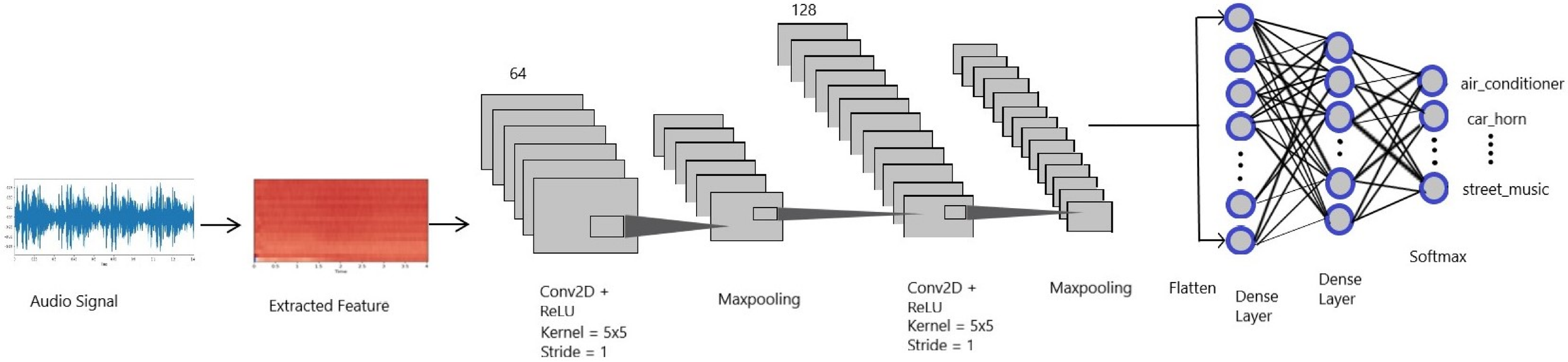
Sequential Keras Tensorflow Model

Saved Model as '.keras'

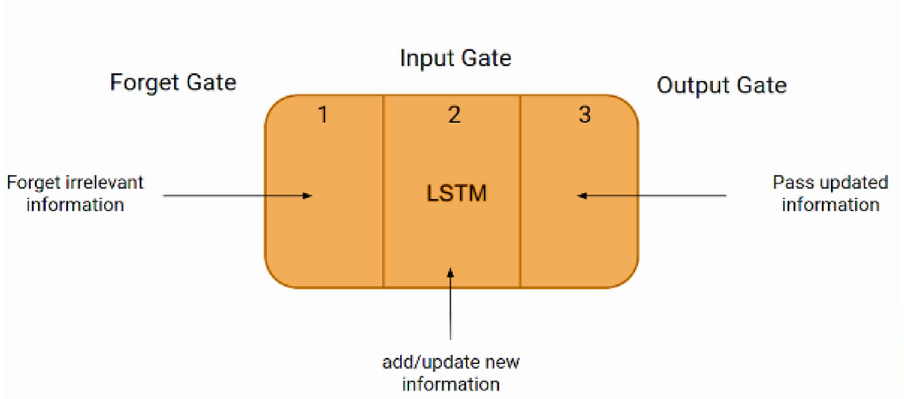
Formation of Layers

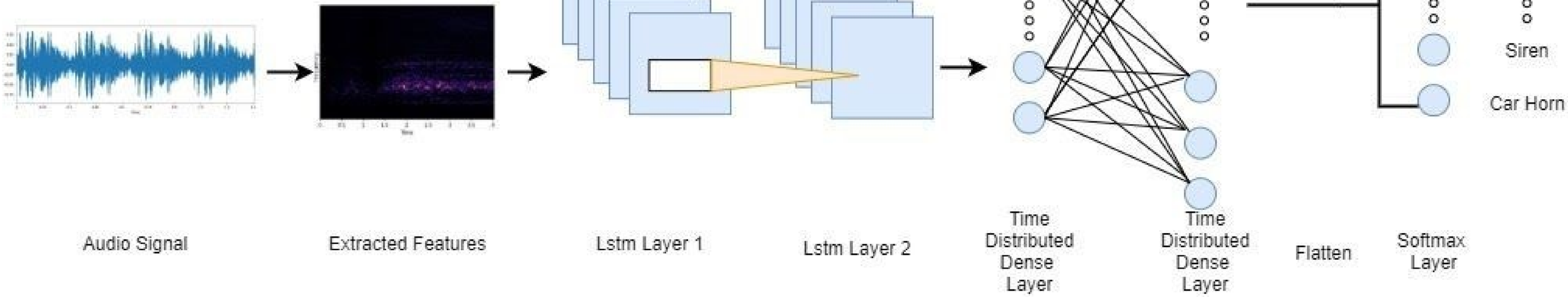
Scikit Learn Trained set

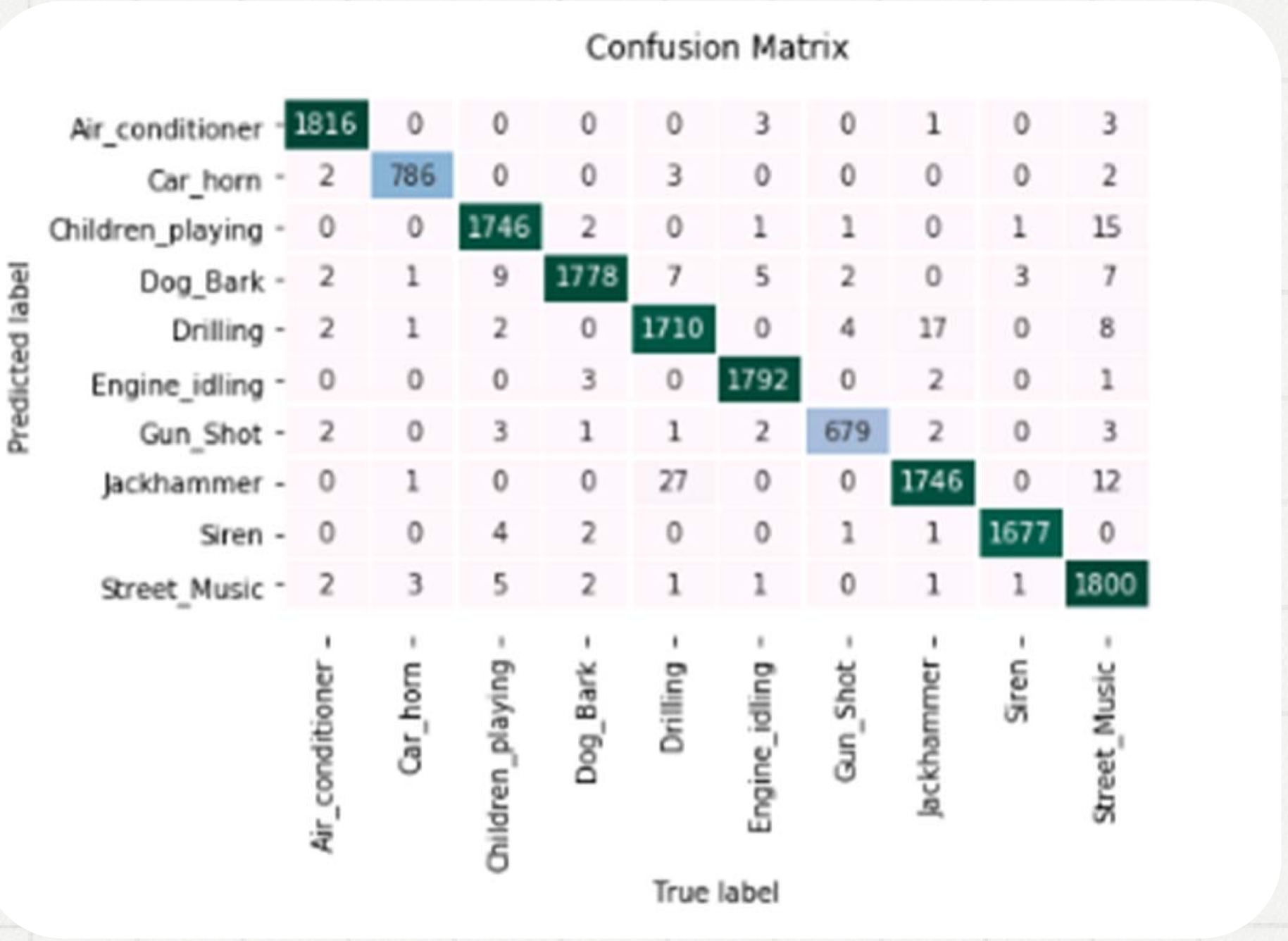
How a CNN Model Works ??



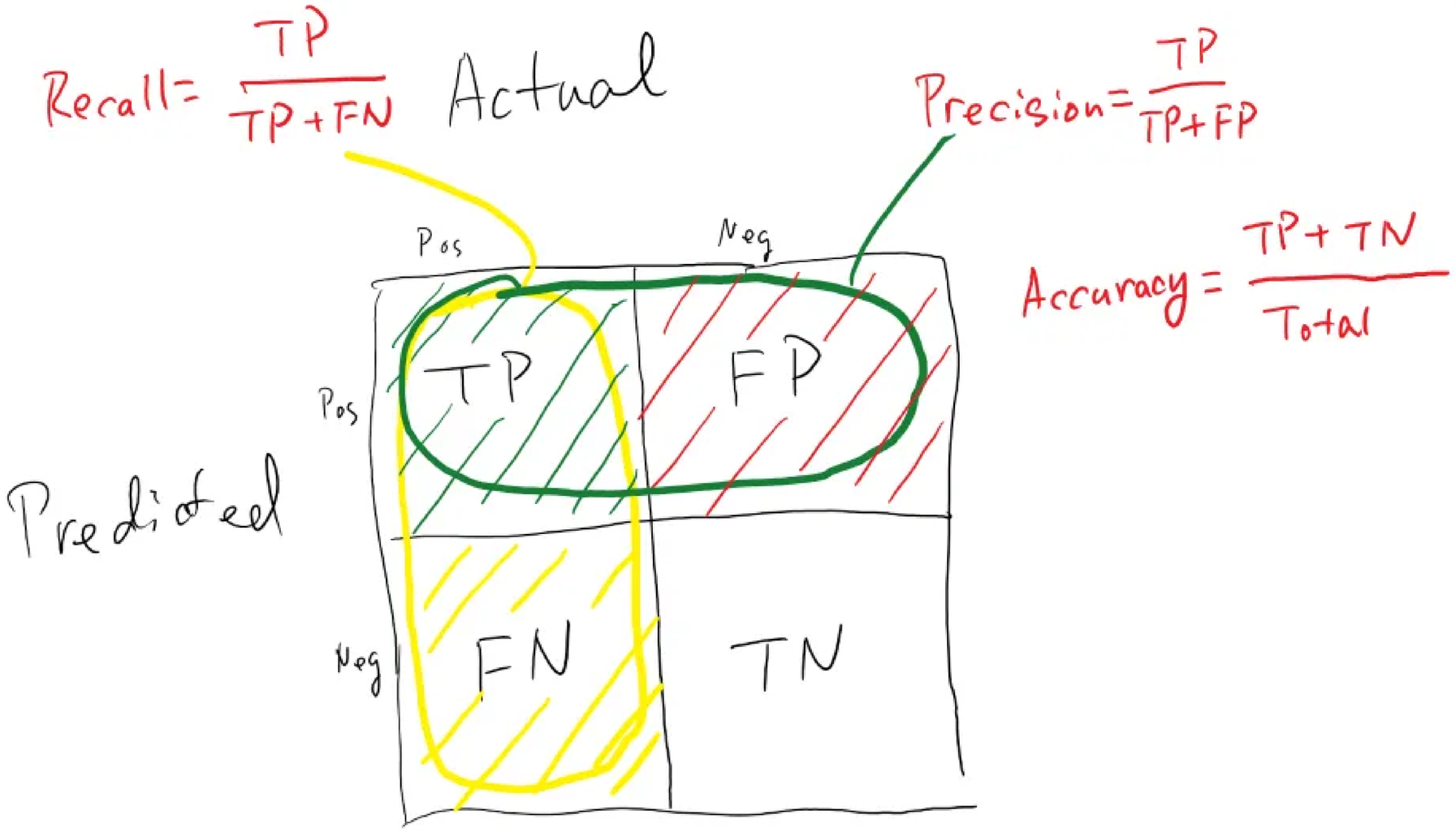
How a LSTM Model Work ??






Confusion Matrix
Predicted label
Air_conditioner 1816 0 0 0 0 3 0 1 0 3
Car_horn 2 786 0 0 3 0 0 0 0 2
Children_playing 0 0 1746 2 0 1 1 0 1 15
Dog_Bark 2 1 9 1778 7 5 2 0 3 7
Drilling 2 1 2 0 1710 0 4 17 0 8
Engine_idling 0 0 0 3 0 1792 0 2 0 1
Gun_Shot 2 0 3 1 1 2 679 2 0 3
Jackhammer 0 1 0 0 27 0 0 1746 0 12
Siren 0 0 4 2 0 0 1 1 1677 0
Street_Music 2 3 5 2 1 1 0 1 1 1800
Air_conditioner
Car_horn
Children_playing
Dog_Bark
Drilling
Engine_idling
Gun_Shot
Jackhammer
Siren
Street_Music
True label

Accuracy , Precision and Recall using Confusion Matrix



Conclusion

I have presented an approach to sound classification, which consists of multiple features stacking and two different neural network models which are CNN and LSTM .I have to conclusion that LSTM is best for sequential data such as time series, text, audio, or any data where the order of the elements matters. LSTM networks process sequences of data one element at a time, maintaining an internal state that captures temporal dependencies.
CNN is Primarily best for grid-like data such as images or spatial data. CNNs operate on fixed-size input grids (e.g., images), performing convolution operations across the spatial dimensions of the input to capture local patterns and hierarchies of features. The models have been trained and tested with original UrbanSound8K and its augmented dataset. Among these, using the MFCC which is used as a feature for the model and in this way we were able to achieve a state-of-the-art result.

Conclusion

LSTM performed better compared to CNN when it comes to audio classification. This is simply because the network of cells in LSTM take input from the previous state ht-1 and present input xt which is used together to predict the next output. Since spectrograms like MFCC have approximately the same pattern through the time domain, it works better with such features. Some of very few researches have also shown that LSTM is very good model for data related to time series such as speech recognition, natural language recognition etc Furthermore, we would like to proceed with this in my future work where we have planned to use novel unsupervised learning techniques that can be adopted to train, test the models, and check their accuracy.

References

[1] Chachada, S., & Kuo, C. (2014). Environmental sound recognition: A survey. APSIPA Transactions on Signal and InformatioProcessing, 3, E14. doi:10.1017/ATSIP.2014.12

[2] Sharma, Jivitesh & Granmo, Ole-Christoffer & Goodwin, Morten. (2019). Environment Sound Classification using Multiple Feature Channels and Deep Convolutional Neural Networks.

[3] Esmaeilpour, M., Cardinal, P., & Koerich, A. L. (2019). Unsupervised feature learning for environmental sound classification using cycle consistent generative adversarial network. arXiv preprint arXiv:1904.04221.

References

[4] Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., ... & Jaitly, N. (2018, April). State-of-the-art speech recognition with sequence-to-sequence models. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4774-4778). IEEE.

[5] Tzinis, E., Wisdom, S., Hershey, J. R., Jansen, A., & Ellis, D. P. (2020, May). Improving universal sound separation using sound classification. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 96-100). IEEE. averages. IEEE transactions on information theory, 58(9), 6093-6100

[6] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 confusion Matrix

[7]https://www.diffen.com/difference/Mono_vs_Stereo

Thank You !!

Btech Project

"Computers are able to see, hear and learn".