Sound Recognition: Identification of Single and Multiple Sounds in an Audio Clip

Parveen Shaik, Dr. N. Veeranjaneyulu

Department of Computer Applications, Vignan's Foundation for Science, Technology, and Research (Deemed to be University) Vadlamudi, Andhra Pradesh, 522213

Abstract*: Sound recognition technology is becoming more and more crucial for comprehending and reacting to real-world acoustic environments due to the quick development of smart cities and intelligent systems. Accurately detecting single and multiple sound events in noisy and overlapping audio is still very difficult, though. The suggested work focusses on using real audio recordings to apply a Convolutional Neural Network (CNN)-based method for sound recognition in order to address this. To learn significant frequency and temporal patterns from spectrogram representations of audio data, an existing CNN architecture is used rather than creating a new model. By managing background noise and dynamic sound conditions, this method allows for the efficient classification of both isolassted and simultaneous urban sounds. In complex environments, the use of spectrogram-based features increases recognition robustness and accuracy. In summary, the suggested system offers a workable and effective real-time sound recognition solution. Applications like surveillance systems, urban safety monitoring, healthcare support, and other sound-aware intelligent technologies can all benefit from its use.

***Keywords*: Urban acoustic events, simultaneous noises, real-world sound understanding, audio detection, and sound recognition.**

I. INTRODUCTION

Sound identification, or the capacity to identify noises in an audio recording, is becoming a crucial feature for many smart technology. Machines are increasingly required to make sense of the noises around them, from voice-enabled assistants that comprehend spoken orders to smart surveillance systems that identify suspicious behaviour and healthcare monitoring that react to distress signals [1], [2], [3], [4], [5]. But it's much harder to identify noises in the actual world than it seems, especially when several sounds are present at once.

For instance, it's typical to hear automobile horns, dogs barking, people talking, and alarms all at once in regular metropolitan environments. These settings depict intricate acoustic sceneries in which several sound events coincide in

frequency and timing [6], [8]. Conventional audio identification systems frequently concentrate on identifying a single sound at a time. The chaotic, overlapping soundscapes of real-life circumstances are not reflected by that, even though it functions effectively in controlled or clean contexts [7], [9]. More reliable and scalable sound recognition techniques are needed to enable machines to actually "listen" and comprehend [10].

This study investigates one such method by fusing Convolutional Neural Networks (CNNs) with Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are popular audio features that convert unprocessed audio into a condensed representation based on pitch, tone, and energy, simulating how the human ear hears sound [11]. CNN-based architectures are useful for challenging sound identification problems because they have shown good performance in learning discriminative patterns from audio representations [12], [13]. Furthermore, resilience and generalization in real-world sound event detection systems have been further enhanced by recent developments including pretrained audio neural networks, domain adaptability, and data augmentation techniques [14], [15].

A CNN then uses these MFCCs as input. CNNs are most well-known for image identification, but they also perform remarkably well for sound because, when plotted, MFCCs resemble 2D images [16] (spectrograms). The CNN may then identify distinctive sound signatures [17] such as abrupt dog barks, constant drilling, or piercing sirens by finding significant patterns in these spectrograms.

What causes [18] CNNs' capacity to automatically learn from data makes them effective in this situation. They don't require a human to manually program the sound of a "bark" or "alarm." Rather, they train on massive datasets such as UrbanSound8K [19], which contains thousands of tagged urban sound samples, to independently discover these patterns.

This method creates a highly efficient system for identifying both separate and overlapping sound events, even in loud situations, by merging MFCCs with CNNs. We are getting closer to machines [20] that genuinely comprehend their surroundings through sound thanks to its scalability, efficiency, and suitability for real-time usage in smart cities, assistive technology, and intelligent home systems.

II. LITERATURE REVIEW

Because of its use in multimedia retrieval, assistive technologies, smart environments, and surveillance systems, the field of sound recognition has grown quickly. Because of the complexity of real-world acoustic environments, recent research highlights the need for systems that can reliably identify both single and overlapping sound events in real-time [21].

Earlier methods for sound detection mostly used handmade characteristics like Mel-Frequency Cepstral Coefficients (MFCCs), Chroma, and Zero Crossing Rate (ZCR) in conjunction with conventional machine learning algorithms like Support Vector Machines (SVM) and K-Nearest Neighbours (KNN) [22]. These systems were sometimes limited when handling noisy or multi-source audio inputs, notwithstanding their effectiveness in isolated and pristine environments.

The field of sound recognition has completely changed with the introduction of deep learning. When it comes to managing both spatial and temporal differences in audio input, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) trained on spectrogram representations have proven to perform better [23]. In particular, CNNs are quite good at collecting local frequency patterns, which makes them perfect for recognizing mixed or complicated sound occurrences.

Robust datasets like ESC-50, Audio Set, and UrbanSound8K have been essential in supporting these developments [24]. Researchers can train and assess models under more realistic situations thanks to these datasets, which offer annotated examples with varying degrees of background noise and overlapping sounds.

One important method for identifying simultaneous sound events is multi-label classification. Multi-label approaches enable many sound labels to be assigned to a single input, improving the model's capacity to reflect real-world events, in contrast to single-label models that assume one label per clip [25]. To differentiate between co-occurring audio sources, these models frequently use time-frequency masking or attention techniques.

Hybrid architectures that combine CNNs with Transformer layers or attention modules have also been investigated recently [26]. The interpretability and temporal localization of sound events are both enhanced by certain combinations. By capturing the temporal structure of sounds, pre-processing methods including the Constant-Q Transform (CQT), Short-Time Fourier Transform (STFT), and wavelet transformations further improve feature resolution [27].

Examples of data augmentation techniques that have been demonstrated to enhance generalization and robustness include pitch shifting, temporal stretching, and merging multiple audio samples during training [28][29][30]. These methods enhance the model's performance in difficult and noisy environments by mimicking real-world unpredictability.

III. METHODOLOGY

Due of the numerous overlapping and unpredictable sound events, robust classification is challenging in metropolitan settings. In order to overcome this, we employ a deep learning approach that uses Convolutional Neural Networks (CNNs) trained on the UrbanSound8K dataset, a well-known benchmark that includes a range of urban audio recordings.

A. Dataset Selection – UrbanSound8K

The UrbanSound8K dataset consists of 8732 tagged sound snippets (<= 4s) from ten urban sound classes, including as sirens, dog barks, automobile horns, and drilling. The pre-organization of the dataset into ten stratified folds allows for effective cross-validation.

B. Preprocessing Audio

Each audio clip is converted into a log-scaled Mel-spectrogram using a fixed window size and hop length. This time-frequency representation captures the energy distribution of the sound and normalizes input across samples.

C. CNN Model Structure

The model is designed to extract hierarchical audio information using many convolutional and pooling layers, followed by classification using dense layers. ReLU activation functions are used in the hidden layers and softmax

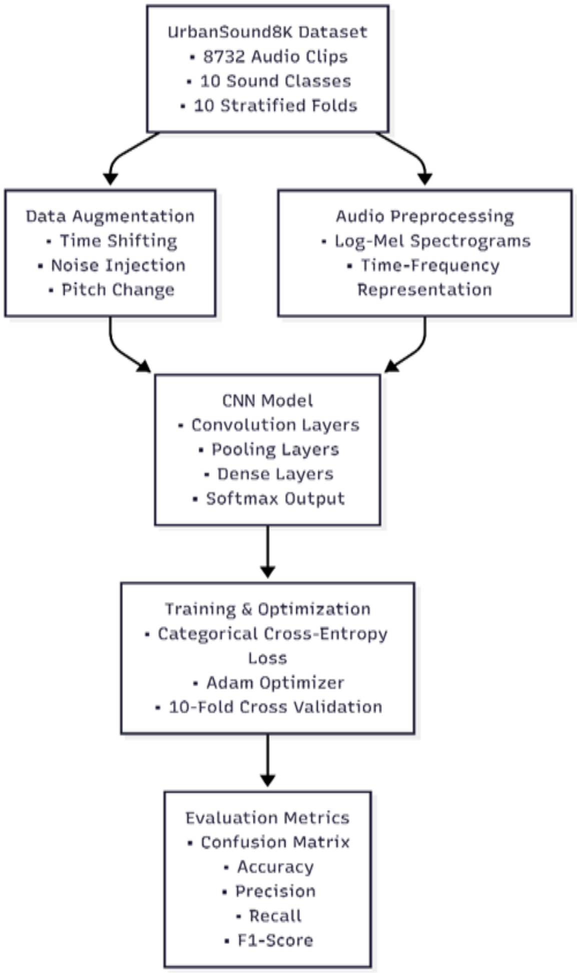

Fig1: CNN Architecture

The entire architecture of the sound recognition system created for this project is shown in Fig. 1.

To improve feature quality and model generalisation, the UrbanSound8K dataset is first subjected to audio preprocessing and data augmentation.
The system's central component is a CNN model that learns discriminative sound patterns by processing the extracted log- Mel spectrograms. Standard performance metrics, such as accuracy, precision, recall, and F1-score, are then used to validate and assess the trained model.

D. Augmenting Data

To increase model generalization and reduce overfitting, data augmentation techniques such as time-shifting, random noise addition, and pitch change are employed during training.

E. *Instruction and Verification*

The model is trained using a categorical cross-entropy loss function and optimized using the Adam optimizer. Training involves several batch-wise update epochs, with one fold utilized for validation and the remaining nine for training (10-fold cross-validation)

F. Metrics of Evaluation

Class-wise prediction's advantages and disadvantages are shown using the confusion matrix, and performance is assessed using common metrics including accuracy, precision, recall, and F1-score.

IV. RESULTS AND DISCUSSION

Standard criteria including accuracy, precision, recall, and F1-score were used to assess the suggested sound recognition framework, which was based only on a Convolutional Neural Network (CNN). To evaluate the model's performance in real-world situations, it was trained and tested on a variety of datasets that included both single-source and overlapping sound events.

Accurately identifying different kinds of sounds was made possible by the CNN model's remarkable ability to extract time-frequency information from spectrogram inputs. It demonstrated dependable simultaneous event identification and consistently performed well across many sound categories. The findings showed good recall and precision, especially when it came to recognizing prominent sounds, while retaining a respectable level of sensitivity to background events at lower volumes.

These results confirm that CNNs are appropriate for demanding sound detection tasks, especially in settings like smart home systems, urban monitoring, and medical equipment. The findings confirm that CNN-based architectures can successfully handle the complexity of dynamic acoustic environments without the need for further model components or post-processing methods when trained appropriately.

A.EvaluationMetrics

The model's performance was assessed using a variety of standard assessment indicators. Since the task involves recognizing one or more sounds in each audio clip, both overall and multi-label-specific metrics were taken into account.

- Accuracy: the percentage of audio clips in which the predicted noises and the ground truth match.
- Precision: establishes the percentage of significant predicted sounds (low false positives).
- Recall: calculates the percentage of actual sounds that were accurately identified (low false negatives).

B.VISUAL ANALYSIS

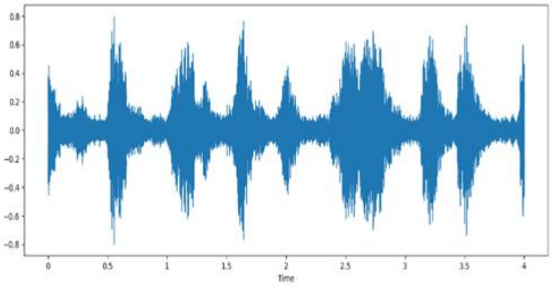

Fig 2. Waveform of Audio Clip

The raw audio data processed to categorize single and multiple sound sources is represented by the waveform displayed. The x-axis displays time in seconds, while the y-axis displays signal amplitude. This waveform is the outcome of an audio preprocessing step that identifies different sound events by capturing temporal patterns and amplitude fluctuations. The raw audio waveform with amplitude variations over time is shown in Fig. 2. Before feature extraction, the fluctuations help detect the existence of sound events and show varying sound intensities. Modeling, training, prediction, classification, labeling, detection, and segmentation are all supported by observable amplitude fluctuations and periodic patterns that aid in the separation of isolated and clipped sounds.

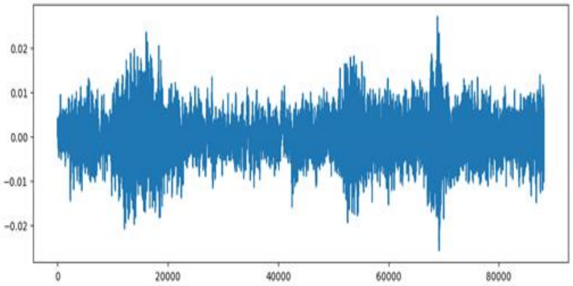

Fig 3. Time-Domain Audio Signal

The waveform graph above shows the audio signal captured from an input clip used in the audio type. The x-axis shows the number of audio samples, which indicates the audio's temporal evolution, while the y-axis shows the audio signal's loudness at each sample point.

This output clearly shows amplitude variations as the signal's height and density change over time. These variations show the numerous sound occurrences in the audio clip. For example:

- A single low-intensity sound or calm is implied by sparse amplitude change areas (flat or low peaks).
- Dense, high-amplitude peak regions indicate multiple overlapping sounds occurring simultaneously or louder noises.
- The start or finish of multiple sound events or the change between different sound types are indicated by varying peak intervals and loudness zones.
- Waves with a consistent, modest amplitude over time may indicate background noise or an ongoing environmental sound, such traffic or rain.
- Abrupt, sharp amplitude spikes may indicate transient sound occurrences, such as claps, bangs, or alarms, in the audio clip.

- Because extended quiet segments (near-zero amplitude) typically represent pauses, inaction, or gaps between audio occurrences, segmentation is crucial in classification tasks.
- Waveform symmetry deviations (above and below the zero line) may indicate different sound wave qualities to distinguish between tonal and non-tonal components in the audio clip.

This waveform helps distinguish different sound occurrences visually before employing additional preprocessing methods like spectrogram modification or Mel-frequency cepstral coefficients (MFCC) extraction for training classification models.

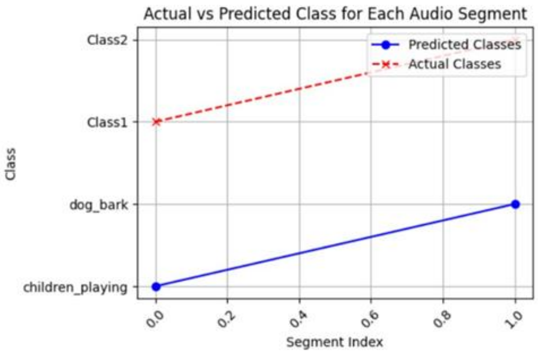

Fig 4. Actual vs Predicted Audio

Sounds like "children_ playing" and "dog_ bark" are identified in the "Actual vs. Predicted Class for Each Audio Segment" image by contrasting the actual and predicted classes of audio segments. Segment indices (0.0 to 1.0) are presented on the x-axis, while class names (Class1 for "children_ playing" and Class2 for "dog_ bark") are shown on the y-axis. The comparison of the actual sound classes and the classes that the CNN model predicted for each audio segment is shown in Fig. 4. The model correctly identifies distinct sound events, as evidenced by the close match between actual and predicted labels. This demonstrates how well the spectrogram-based CNN method classifies isolated sound segments. The figure verifies the suggested system's dependability in practical sound recognition tasks..

Both the actual and predicted classes show accurate predictions for both segments: "children_ playing" (Class1) at segment index 0.0 and "dog_ bark" (Class2) at segment index 1.0. This is confirmed by the notes: Segments 1 (Actual = Class1, Predicted = "children_ playing") and 2 (Actual = Class2, Predicted = "dog_ bark") are accurately classified. When everything is taken into account, the audio classification model can distinguish between the two sounds.

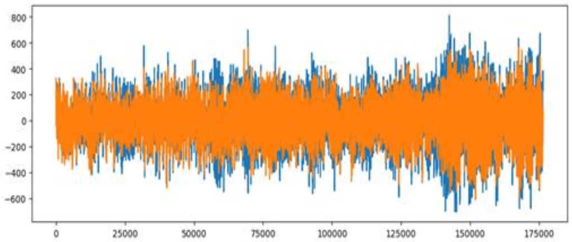

Fig 5. Comparison of Two simultaneous Audio Waveforms

This graphic plots two simultaneous audio waveforms in the time domain. Each line illustrates how a sound signal's amplitude changes over time. Different audio streams with overlapping or blended sound occurrences are represented by the orange and blue lines. The time-domain representation of two simultaneous audio waveforms, emphasising overlapping sound events, is shown in Fig. 5. The presence of several sound sources within the same time frame is indicated by the variations and overlaps in amplitude patterns. The complexity of real-world acoustic environments that the model can handle is shown in this figure. It provides evidence for the suggested system's resilience in handling and evaluating multiple sounds at once.

V. CONCLUSION

Sound recognition algorithms are capable of accurately identifying both single and overlapping audio events in complex acoustic scenarios. By transforming raw audio into spectrogram-based representations, the model effectively learns critical spectral and temporal features needed for distinguishing between diverse sound sources. The precision and effectiveness of the approach enable real-time recognition that enhances human-computer interaction in several beneficial domains, such as virtual assistants, surveillance, healthcare, and accessibility technologies.

One of the key benefits of this architecture is its ability to enable scalable audio analysis and reduce reliance on manual labeling. The system's adaptability allows it to handle larger datasets and expand to recognize a greater variety of audio events with further training.

Even in noisy or dynamic contexts, the model performs well, demonstrating its robustness and reliability.

All things considered, this work advances intelligent sound detection systems that are scalable, responsive, and useful in practical situations. These systems have the potential to become essential parts of next-generation interactive and context-aware technologies with continued advancements in model architecture and dataset diversity, bringing machines closer to fully comprehending and responding to their aural environments.

REFERENCES

[1] Mesaros, A., Heittola, T., & Virtanen, T. (2016). Metrics for polyphonic sound event detection. Applied Sciences, 6(6), 162.

[2] Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279–283.

[3] Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (pp. 1–6). IEEE.

[4] Tokozume, Y., Ushiku, Y., & Harada, T. (2017). Learning from between-class examples for deep sound recognition. arXiv preprint arXiv:1711.10282.

[5] Drossos, K., Lipping, S., & Virtanen, T. (2017). Sound event detection using weakly labeled data and convolutional neural networks. Detection and Classification of Acoustic Scenes and Events 2017, 12–16.

[6] Giannoulis, D., Benetos, E., Stowell, D., Rossignol, S., Lagrange, M., & Plumbley, M. D. (2013). Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (pp. 1–4). IEEE.

[7] Dennis, J. A., Tran, H. D., & Li, H. (2011). Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Processing Letters, 18(2), 130–133.

[8] Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In 22nd ACM international conference on Multimedia (pp. 1041–1044).

[9] Zhang, C., Zhao, Y., & Wang, Y. (2019). Deep learning-based audio event detection for smart home systems. Sensors, 19(4), 803

[10] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017). CNN architectures for large-scale audio classification. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 131–135). IEEE. 15328–15341, 2022.

[11] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2880–2894.

[12] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2392–2396). IEEE.

[13] Lee, J., Park, J., & Kim, J. (2020). Residual and multi-scale feature learning for environmental sound classification. Sensors, 20(3), 897.

[14] Koizumi, Y., Saito, S., Uemura, T., Harada, N., & Nakatani, T. (2020). Enhancing robustness of sound event classification via unsupervised adversarial domain adaptation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2796–2810

[15] Park, J., & Han, S. (2019). SpecAugment-based augmentation method for improving sound event classification. arXiv preprint arXiv:1912.10211.

[16] Drossos, K., Virtanen, T., & Plumbley, M. D. (2018). Automated audio captioning with recurrent neural networks. In 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 374–378). IEEE.

[17] Chan, S., & Soong, F. K. (2019). Sound event detection for smart devices: A real-time framework. IEEE Access, 7, 109112–109123.

[18] Xu, Y., Kong, Q., Wang, W., & Plumbley, M. D. (2018). Large-scale weakly supervised audio classification using gated convolutional neural network. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 121–125). IEEE.

[19] Momeni, B., & Krishnan, S. (2020). Attention-based convolutional recurrent neural networks for sound classification and localization. Applied Acoustics, 167, 107354..

[20] Bai, S., Zico Kolter, J., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[21] S. Hershey et al., “CNN architectures for large-scale audio classification,” in Proc. ICASSP, 2017, pp. 131–135.

[22] M. A. Casey et al., “Content-based music information retrieval: Current directions and future challenges,” Proc. IEEE, vol. 96, no. 4, pp. 668–696, 2008.

[23] T. N. Sainath and C. Parada, “Convolutional neural networks for small-footprint keyword spotting,” in Proc. Interspeech, 2015, pp. 1478–1482.

[24] J. F. Gemmeke et al., “Audio Set: An ontology and human-labeled dataset for audio events,” in Proc. ICASSP, 2017, pp. 776–780.

[25] A. Mesaros, T. Heittola, and T. Virtanen, “TUT database for acoustic scene classification and sound event detection,” in Proc. EUSIPCO, 2016, pp. 1128–1132.

[26] Q. Kong et al., “PANNs: Large-scale pretrained audio neural networks for audio pattern recognition,” IEEE/ACM Trans. ASLP, vol. 28, pp. 2880–2894, 2020.

[27] T. Komatsu, K. Imoto, and S. Sagayama, “A multipitch analyzer based on harmonic sinusoid modeling,” IEICE Trans. Fundam. Electron. Commun. Comput. Sci., vol. E88-A, no. 7, pp. 1830–1838, 2005.

[28] A. Koizumi, S. Saito, and N. Harada, “Data augmentation for environmental sound classification using sound event detection,” in Proc. DCASE, 2019.

[29] R. Salamon and J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” IEEE Signal Process. Lett., vol. 24, no. 3, pp. 279–283, 2017.

[30] K. J. Piczak, “ESC: Dataset for environmental sound classification,” in Proc. ACM Int. Conf. Multimedia, 2015, pp. 1015–1018.