A Transfer Learning Approach For Audio Identification And Classification

Dr. David Raj Micheal
Department of Mathematics
School of Advanced Sciences
Vellore Institute of Technology Chennai
Tamil Nadu – 600127
davidraj.micheal@vit.ac.in

Amal Thomas
Department of Mathematics
School of Advanced Sciences
Vellore Institute of Technology Chennai
Tamil Nadu – 600127
amal.thomas2023@vitstudent.ac.in

***Abstract*—This paper presents a novel audio classification system combining signal processing with deep learning. Audio files are resampled and segmented to produce mel-spectrograms, which are converted into log-scaled feature maps reflecting different audio components. The VGG16 model, initially trained on image data, is repurposed to extract features from these spectrograms. A custom deep neural network then classifies these features using multiple dense layers with dropout. The system's performance is validated with accuracy and loss metrics. Predictions are generated for new audio samples and compared to actual labels. Visualization of waveforms and feature maps further illustrates the model's effectiveness. This framework highlights an innovative use of transfer learning for accurate audio classification.**

***Index Terms*—Audio Classification, Signal Processing, Deep Learning, Mel-Spectrograms, VGG16 Model, Feature Extraction, Transfer Learning, Neural Networks, Performance Metrics, Data Visualization.**

I. Introduction

Audio classification has seen extensive research, especially in music and speech recognition domains. However, urban sound data remains a relatively underexplored area in this field. Today, Convolutional Neural Networks (CNNs) are commonly used in audio classification to detect features from audio datasets, with their application spanning fields like image processing, image stitching, and document analysis. Much work has been done to classify audio using convolutional networks, with datasets like ESC-10, ESC-50, and UrbanSound8k frequently used to provide a wide range of audio samples. Despite the notable progress achieved through these networks, there is a sense that alternative approaches could potentially yield even greater accuracy. This gap becomes particularly significant when considering pedestrian safety. A study from the United States reveals a roughly 200% increase in accidents attributed to reduced situational awareness. This startling statistic raises an important question: ”Could sound classification play a role in injury prevention?”

II. Objective

The goal of this work is to develop an efficient audio classification framework using signal processing and transfer learning. By extracting deep features from mel-spectrograms with a pre-trained VGG16 model, the system aims to enhance classification accuracy while reducing computational overhead. A custom neural network is implemented to classify audio data, and dropout layers are used to prevent overfitting. The framework is designed to handle diverse audio samples, providing accurate predictions and robust performance.

III. Literature Review

[1]. In the study by Zhang et al. (2021), a new attention-based convolutional recurrent neural network (ACRNN) is introduced for environmental sound classification. The method starts by extracting spectro-temporal features using log-gammatone spectrograms, which are then processed by a CNN to learn high-level feature representations, while bidirectional gated recurrent units (Bi-GRUs) capture the temporal relationships within the sound data. To improve focus on meaningful sound frames, a frame-level attention mechanism is applied, minimizing the influence of irrelevant or silent segments. This strategy significantly boosts the model's ability to distinguish between sound classes, leading to top-tier classification performance on both the ESC-10 and ESC-50 datasets. The integration of CNNs, RNNs, and attention mechanisms enables the model to manage the intricate temporal dynamics of environmental sounds efficiently while keeping the computational complexity low.

In the paper titled ”Deep Convolutional Neural Network with Mixup for Environmental Sound Classification”[2], the methodology for audio detection follows several essential steps. Initially, the audio data undergoes preprocessing, where it is resampled and segmented into overlapping windows to capture detailed information. Mel-spectrograms are then created from these segments, reflecting the frequency content of the audio across time. These mel-spectrograms are further processed into log-scaled feature maps, which emphasize both the full spectrum and harmonic/percussive elements of the sound. The VGG16 model, pre-trained on image datasets, is repurposed to extract meaningful audio features from these spectrograms. The extracted features are subsequently input into a custom deep neural network for classification. This network is composed of multiple dense layers with dropout applied to prevent overfitting, ensuring reliable performance.

The model is trained on the prepared data and evaluated using metrics such as accuracy and loss. Predictions are made on test audio samples, with results visualized through waveforms and feature maps, demonstrating the model's ability to distinguish between various audio classes effectively.

In the paper titled ”Emotion Recognition from Audio, Dimensional and Discrete Categorization using CNNs”[3], the methodology for recognizing emotions from audio involves a combination of discrete and dimensional classification using Convolutional Neural Networks (CNNs). Initially, Mel-frequency cepstral coefficients (MFCCs) are extracted from the audio as these features closely mimic the behavior of the human vocal tract. The extracted features are then fed into CNN models designed for both discrete and dimensional emotion categorization. The discrete model focuses on classifying audio into specific emotional categories like happy, sad, or angry, utilizing a traditional CNN architecture with dropout layers for regularization. In contrast, the dimensional model maps emotions onto a 2D plane using valence (positivity/negativity) and arousal (emotional intensity) and applies a 3D CNN to predict these coordinates. This enables a more detailed and continuous representation of emotions beyond categorical classification. For the dimensional model, three levels of valence and arousal (high, medium, and low) are encoded using one-hot encoding. The models are evaluated on the RAVDESS dataset, and the results demonstrate that the dimensional model, particularly in predicting emotions in specific quadrants of the valence-arousal space, provides enhanced accuracy compared to the discrete model.

This paper by Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier[4] provides an extensive review of datasets utilized for training audio-language models (ALMs), focusing on those that offer textual descriptions of sounds and events. The authors highlight the growing trend of leveraging large, varied datasets, such as those from Freesound and AudioSet, to enhance ALM capabilities. The survey not only categorizes these datasets but also examines issues such as data leakage and distributional shift, particularly in scenarios with limited audio data. By analyzing the origins, features, and potential biases of these datasets, the paper aims to help researchers effectively use audio-language data for tasks like audio captioning, audio-text retrieval, and audio generation.

The paper ”A Survey of Audio Classification Using Deep Learning”[5] presents an overview of various deep learning models applied to audio classification. It focuses on five key architectures: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders, Transformers, and hybrid models. CNNs are effective for tasks like speech and environmental sound classification, as they convert audio signals into spectrograms to capture time-frequency data. RNNs are ideal for handling sequential data, such as speech processing. Autoencoders are employed for unsupervised feature extraction, while Transformers utilize attention mechanisms for processing large-scale audio datasets. Hybrid models combine multiple architectures, like CNN-RNN, to improve performance by leveraging both spatial and temporal features. The paper evaluates the strengths and limitations of these models across different audio classification tasks.

The paper ”A Survey of Audio Classification Using Deep Learning”[6] reviews five main deep learning models for audio classification: CNNs, RNNs, Autoencoders, Transformers, and hybrid models. CNNs convert audio signals into spectrograms for tasks like speech recognition, while RNNs are suited for sequential data processing. Autoencoders aid in unsupervised learning and feature extraction. Transformers efficiently handle large audio datasets through attention mechanisms. Hybrid models combine different architectures, such as CNNs with RNNs, to manage both spatial and temporal data. The paper compares these methods, detailing their strengths and limitations in audio classification.

The method described in the paper investigates various Convolutional Neural Network (CNN) architectures for large-scale audio classification using a dataset of 70 million YouTube videos[7]. The audio is processed into log-mel spectrogram patches, which are then fed into different CNN architectures, including AlexNet, VGG, Inception, and ResNet. These architectures, originally designed for image classification, are adapted for audio by modifying input parameters such as stride and pooling layers. The study compares performance across models using metrics like area under the curve (AUC) and mean average precision (mAP) for classification accuracy, revealing that deeper networks like Inception and ResNet perform best. The models are also tested for Acoustic Event Detection (AED), demonstrating improved performance when using embeddings from CNN classifiers.

The method outlined in Hershey et al. (2017) employs CNN architectures, originally designed for image classification, for large-scale audio classification[8]. The audio data is transformed into log-mel spectrograms, which are processed as 2D image-like inputs. The study evaluates different CNN models such as AlexNet, VGG, Inception, and ResNet using a dataset of 70 million labeled YouTube videos, aiming to predict video-level tags based on audio alone. The results show that deeper networks like ResNet and Inception offer the best performance. Additionally, the study examines the effect of training set size and tests the models' ability to improve acoustic event detection using learned embeddings.

The paper ”Unsupervised Feature Learning for Audio Classification Using Convolutional Deep Belief Networks” investigates the use of convolutional deep belief networks (CDBNs) for learning audio features without labeled data. By applying CDBNs to tasks like speaker identification, phone classification, and music genre classification, the authors show that the learned hierarchical features outperform conventional methods like spectrograms and MFCCs, particularly in settings with limited labeled data. The study demonstrates that stacking layers of convolutional restricted Boltzmann machines (CRBMs) enables the model to capture more complex audio features, leading to higher classification accuracy[9]. The results highlight the potential of deep learning techniques in enhancing audio classification performance.

The model proposed in the paper introduces a Temporal-Frequency Attention based Convolutional Neural Network (TFCNN) for environmental sound classification. This model addresses the complex and unstructured nature of environmental sounds by leveraging two novel attention mechanisms—temporal and frequency attention[10]. These mechanisms focus on key frequency bands and critical time frames within a log-Mel spectrogram, effectively reducing noise and irrelevant data. The model combines these mechanisms to enhance feature extraction, improving classification accuracy. Tested on public datasets like UrbanSound8K and ESC-50, the TFCNN achieved high accuracy rates, demonstrating its effectiveness in capturing essential time-frequency patterns in audio data.

This paper proposes a novel approach for environmental sound classification using a Temporal-Frequency Attention based Convolutional Neural Network (TFCNN)[11]. The TFCNN model utilizes a dual-attention mechanism to selectively focus on crucial time segments and frequency bands within log-Mel spectrograms, thereby mitigating the impact of noise and irrelevant data on feature extraction and classification accuracy. Experimental results on the UrbanSound8K and ESC-50 datasets demonstrate the TFCNN's superior performance compared to existing state-of-the-art methods, showcasing its ability to effectively capture salient time-frequency patterns in audio data. A thorough review of relevant literature is also provided, covering various audio classification techniques, including spectrograms, MFCCs, and deep learning approaches such as CNNs and RNNs. The significance of attention mechanisms in audio classification tasks is discussed, and the potential of the TFCNN model for real-world applications is highlighted.

The paper ”Audio Based Bird Species Identification using Deep Learning Techniques” introduces an innovative method for classifying bird species using deep learning, specifically employing a convolutional neural network (CNN)[12]. Unlike traditional approaches that utilize template matching or decision trees, this study incorporates techniques from speech recognition alongside modern advancements in deep learning. The CNN model, consisting of five convolutional layers with max-pooling layers, processes spectrograms derived from bird audio recordings. To boost model accuracy, the researchers applied unique preprocessing techniques to separate bird sounds from background noise and employed data augmentation strategies, such as time shifts and audio mixing. Trained on the largest publicly available bird sound dataset, the model outperformed existing methods and won the BirdCLEF 2016 Recognition Challenge. The CNN's success in distinguishing both primary and background bird species demonstrates its effectiveness. The paper also addresses challenges like background noise, multi-label classification, and varying recording lengths, offering solutions through advanced data augmentation and feature extraction techniques.

The paper ”A Scalogram-Based CNN Approach for Audio Classification in Construction Sites” proposes an innovative technique for classifying sounds on construction sites using audio data processed by convolutional neural networks (CNNs)[13]. Unlike conventional methods that use spectrograms, this research applies scalograms, generated from the Continuous Wavelet Transform (CWT), which provide enhanced time-frequency resolution. The authors adapt the AlexNet CNN model, tailoring it to handle scalogram inputs. By leveraging the detailed information in scalograms, the approach significantly improves the classification accuracy of construction site sounds, including machinery noise. The method was tested on audio data from five types of equipment, and the results indicate that this scalogram-based CNN outperforms traditional techniques, making it a highly effective solution for real-time automatic monitoring of construction activities.

The paper titled ”Design and Implementation of Environmental Warning Systems for Drivers with Hearing Impairments” outlines the creation of a system that assists drivers with hearing disabilities by transforming important environmental sounds, such as fire truck and ambulance sirens, into visual notifications. The system is built using a Recurrent Neural Network (RNN) combined with Long Short-Term Memory (LSTM) to analyze sound data. Implemented on a Raspberry Pi 4, along with a microphone, LCD screen, and LEDs, the system categorizes sounds into three types: traffic noise, ambulance sirens, and fire truck sirens[14]. These sounds are then converted into visual cues, including text displayed on the screen and flashing LED indicators, to warn the driver. Testing the system across different sound recognition intervals—2, 3, and 4 seconds—produced accuracy rates of 78

The paper titled ”Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models” examines the application of deep learning for categorizing both musical instruments and environmental sounds. It utilizes Mel Frequency Cepstral Coefficients (MFCC) to extract features from audio, closely simulating the human auditory system. To enhance model accuracy, the study integrates data augmentation methods, such as pitch shifting, time stretching, and noise injection. Two models are employed: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). CNNs effectively capture spatial patterns in the audio, while RNNs, particularly with Long Short-Term Memory (LSTM) units, are well-suited for processing sequential data[15]. Despite the use of relatively small datasets, both models performed well, demonstrating the potential of deep learning in audio classification tasks like speech recognition and music analysis.

The paper investigates the application of Convolutional Neural Networks (CNNs) for classifying environmental sounds, focusing on their ability to process complex and varied audio events. The proposed model includes two convolutional layers with max-pooling, followed by two fully connected layers. It is trained using segmented spectrograms of audio data and incorporates delta features for enhanced performance[16]. The model's performance was tested on three public datasets: ESC-50, ESC-10, and UrbanSound8K. Results showed that the CNN outperformed traditional models based on handengineered features like mel-frequency cepstral coefficients (MFCCs) and achieved performance comparable to leading methods. The study concludes that CNNs are a promising solution for environmental sound classification, even with limited training data, particularly when using data augmentation techniques like time-stretching and pitch-shifting.

The paper introduces a comprehensive method for classifying environmental sounds using a 1D Convolutional Neural Network (CNN), which learns directly from raw audio data without relying on handcrafted features or 2D representations such as spectrograms. The model comprises several convolutional layers to extract features from time-domain audio signals, followed by fully connected layers for classification. This approach was evaluated using the UrbanSound8k dataset, where it achieved an average accuracy of 89, surpassing many contemporary models that depend on 2D CNNs or manually designed features[17]. The network's performance is further enhanced by initializing the first convolutional layer with Gammatone filters, which simulate human auditory processing. This design optimizes accuracy while keeping the model efficient, making it ideal for applications with limited data or processing power. The research emphasizes the effectiveness of 1D CNNs in sound classification and suggests future improvements through the combination of 1D and 2D representations.

The paper investigates how people naturally categorize everyday sounds, emphasizing the spontaneous classification of sounds based on their environment. It reviews psychological theories, experimental data, and proposed classification systems for understanding how sounds are organized[18]. The main approach involves studying how participants group sounds according to attributes like the source of the sound, the action causing it, and the context in which it occurs. Techniques such as sorting tasks, similarity assessments, and multidimensional scaling are used. The study shows the role of both bottom-up (stimulus-driven) and top-down (knowledge-driven) processes in sound categorization. It reveals that sound classification is influenced not only by acoustic features but also by contextual and semantic factors, providing insights for improving computational sound recognition systems.

The paper explains a speech recognition system named ”IDENTIFY,” which is tailored for performing speech recognition, identification, and categorization experiments on PCs. This system enables researchers to present up to 100 different speech waveform files as stimuli, with subjects providing their responses using a Koala Pad digitizing tablet[19]. Developed in Turbo Pascal, the program can manage various stimulus and response configurations. It can generate a confusion matrix when the number of stimuli matches the number of responses, and it is also capable of running categorization experiments that utilize a binary response format. This functionality is particularly valuable in psychoacoustic research. The program is compatible with platforms like ILS and C-Speech, and it requires specific hardware, including a Data Translation DT2801-A D/A output board.

The document provides a comprehensive overview of audio event detection (AED) techniques, tracing the process from feature extraction to classification[20]. It explores various algorithms employed in AED, focusing on preprocessing, feature extraction, and classification stages. The paper covers different models, including supervised, unsupervised, and semi-supervised learning methods such as support vector machines (SVMs), Gaussian mixture models (GMMs), and hidden Markov models (HMMs). It highlights the crucial role of feature and classifier selection in enhancing AED accuracy, particularly for applications in areas like security, speech recognition, and health monitoring. A range of classifiers, including neural networks, ensemble techniques, and rule-based systems, are evaluated, discussing their strengths and challenges in different AED use cases.

IV. Dataset

The model is applied to the “UrbanSound8K” dataset, which consists of urban sounds categorized into 10 distinct classes, totaling 8,732 labeled audio excerpts of up to 4 seconds in duration. The classes include:

- Air conditioner
- Car horn
- Children playing
- Dog bark
- Drilling
- Engine idling
- Gunshot
- Jackhammer
- Siren
- Street music

All classes are derived from the urban sound taxonomy. The dataset is organized into 10 folders, labeled fold1 to fold10, which facilitate the automated classification process described earlier. Additionally, a CSV file is provided that contains metadata for each audio excerpt.

V. Methodology

A. CNN

Convolutional Neural Networks (CNNs) have gained significant traction in deep learning due to their unique ability to learn diverse filters tailored to specific problems, such as sound classification, image recognition, and image processing tasks. Although primarily developed for two-dimensional images, CNNs can also handle one-dimensional and three-dimensional data. CNNs utilize small filters, often referred to as kernels, that focus on extracting specific features from the input. These filters are applied across sections of the data through inner product operations, capturing essential patterns that contribute to the model's understanding of complex input data.

B. Convolution Layer

The convolutional layer is fundamental to the functionality of CNNs. Within this layer, arrays representing sound data are multiplied by a two-dimensional array of weights, called a filter, resulting in a scalar product, which makes convolution a linear operation. CNNs typically use multiple convolutional layers in succession to extract meaningful features from input data, which are then organized into feature maps. As filters pass over the sound data, they detect specific patterns. When a filter identifies a matching pattern, it returns a positive value; otherwise, it outputs a minimal value.

C. C. Pooling Layer

The pooling layer produces a set of pooled feature maps, with each feature map processed independently. Unlike other operations, pooling does not involve learned parameters; rather, it requires predefined functions. Some commonly used pooling functions include:

- **Average Pooling**: Computes the mean values over regions of the feature map, providing a smoothed version of the data.
- **Maximum Pooling (Max Pooling)**: Selects the highest value within each region of the feature map, capturing prominent features.

Typically, max pooling is favored over average pooling as it better preserves key features. The pooling layer is applied to each convolution layer, creating a more condensed version of the features and reducing the computational complexity in a CNN.

D. Dropout Layer

In CNNs, a regularization technique called dropout is frequently used to prevent overfitting. This involves training several neural networks with different architectures in parallel by temporarily omitting—or “dropping out”—certain nodes or layers. This omission alters the layer’s structure, effectively creating varying configurations of nodes and connections with the previous layer. During training, each update is applied to a layer with a slightly different “perspective,” enhancing the network’s robustness.

E. Softmax

The Softmax function converts a vector of K real numbers into a vector of K values that sum to 1. Each input value, regardless of being positive, negative, zero, or exceeding one, is transformed into a number between 0 and 1. This transformation allows the values to be interpreted as probabilities, making it especially useful in classification tasks.

F. MFCC

Mel-frequency cepstrum (MFC) represents the power spectrum of sound over specific time intervals. It is calculated by applying a linear cosine transform to a logarithmic power spectrum, performed on a nonlinear mel frequency scale. The Mel-frequency cepstral coefficients (MFCCs) are the features derived from MFC, widely used in various speech recognition applications. The main distinction between MFC and cepstrum lies in the spacing of frequency bands; MFC features are spaced evenly on the mel scale, allowing for a more precise representation of human auditory perception compared to the linearly spaced frequency bands of cepstrum. This method essentially involves frequency warping, enhancing the representation of sound signals.

MFCCs are typically derived through the following steps:

- Compute the Discrete Fourier Transform (DFT) of the input signal.
- Apply the mel frequency wrapping to the magnitude of the DFT.
- Take the logarithm of each mel frequency.
- Perform the Inverse Discrete Fourier Transform (IDFT) at each frequency.
- The resulting amplitudes from the spectrum constitute the MFCCs.

VI. Experimentation Setup

The following configurations and processes are implemented for the audio classification model:

- The sample rate is set to 22.05 kHz for subsequent processing.
- The normalized data encompasses bit depth values ranging from -1 to 1.
- Audio signals are converted to mono, ensuring a constant channel count of 1.
- Mel-frequency cepstral coefficients (MFCCs) are generated from the time series audio data.
- The primary function of the convolutional layers is to perform feature detection. This involves placing a filter window over the input data, followed by matrix multiplication, with the results stored in a feature map, a process referred to as Convolution.
- Filter parameters define the number of nodes in each layer, which increase in multiples of 16 across successive layers (i.e., 16, 32, 64, and 128). The kernel size is set to 2, resulting in a 2x2 filter matrix.
- The input state of (40, 174, 1) is passed to the first layer, where 40 denotes the number of MFCCs, 174 represents the total frames including padding, and the value 1 indicates a mono sound.
- ReLU is used as the activation function for the convolutional layers, and a dropout rate of 20
- Each convolutional layer is followed by a MaxPooling2D layer, while the final layer connects to a GlobalAveragePooling2D layer. These layers reduce model dimensionality by lowering boundaries and computational requirements, leading to shorter training times and decreased overfitting. The Max Pooling takes the maximum value from each window, while the Global Average Pooling computes the average.
- The output layer consists of 10 nodes corresponding to the number of possible classifications. The SoftMax activation function is applied to this layer, ensuring that the outputs sum to 1, which can be interpreted as probabilities for predictions made by the model.

VII. result

This section outlines the results of classifying urban sounds using CNNs. The model evaluation was carried out using 10-

fold cross-validation (UrbanSound8K), where one fold was set aside during training to serve as a validation set. From the results obtained, it is evident that as the number of epochs increases, the testing accuracy improves, ultimately reaching 91.57% on the test data. Additionally, there is a clear negative correlation between the testing loss and the number of epochs, with the loss reducing to around 30%. The model also achieves a high training accuracy of 95%, indicating minimal risk of overfitting, and the training loss decreases to a low 12%. These findings collectively demonstrate the effectiveness and improved efficiency of the model.

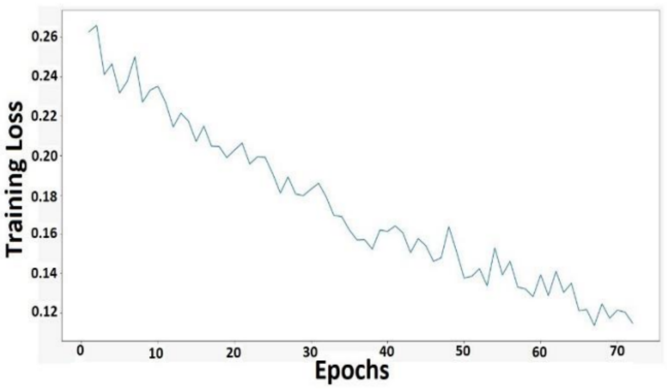


Fig. 1.

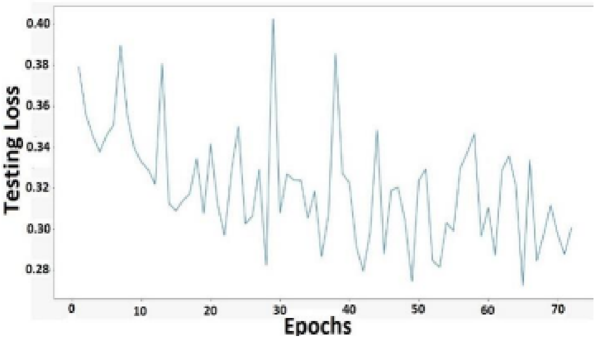
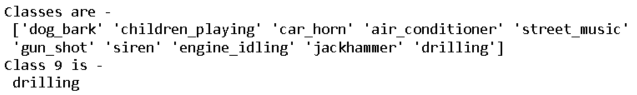


Fig. 2.



```
Classes are -
 ['dog_bark' 'children_playing' 'car_horn' 'air_conditioner' 'street_music'
 'gun_shot' 'siren' 'engine_idling' 'jackhammer' 'drilling']
Class 9 is -
 drilling
```

Fig. 3.

VIII. Conclusion

This paper aimed to evaluate the effectiveness of using convolutional neural networks (CNNs) for accurately and efficiently classifying urban sounds, making them suitable for a range of applications. A deep learning model with four convolutional layers, combined with max-pooling and GlobalAveragePooling2D, was developed and demonstrated to be robust enough for further use in sound classification tasks. Model accuracy was tested on the UrbanSound8k dataset, which contains over 8,700 audio samples, showing promising results.

An area for future exploration is to assess whether this model can generalize effectively to more diverse datasets and handle real-time applications.

References

[1] Zhang, Z., Xu, S., Zhang, S., Qiao, T., & Cao, S. (2021). Attention based convolutional recurrent neural network for environmental sound classification. Neurocomputing, 453, 896-903.

[2] Zhang, Z., Xu, S., Cao, S., & Zhang, S. (2018, November). Deep convolutional neural network with mixup for environmental sound classification. In Chinese conference on pattern recognition and computer vision (prcv) (pp. 356-367). Cham: Springer International Publishing.

[3] Rajak, R., & Mall, R. (2019, October). Emotion recognition from audio, dimensional and discrete categorization using CNNs. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 301-305). IEEE.

[4] Wijngaard, G., Formisano, E., Esposito, M., & Dumontier, M. (2024). Audio-Language Datasets of Scenes and Events: A Survey. arXiv preprint arXiv:2407.06947.

[5] Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. IEEE Access.

[6] Chandrakala, S., & Jayalakshmi, S. L. (2019). Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. ACM Computing Surveys (CSUR), 52(3), 1-34.

[7] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 131-135). IEEE.

[8] McLoughlin, I., Zhang, H., Xie, Z., Song, Y., & Xiao, W. (2015). Robust sound event classification using deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(3), 540-552.

[9] Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in Neural Information Processing Systems, 22.

[10] Mu, W., Yin, B., Huang, X., Xu, J., & Du, Z. (2021). Environmental sound classification using temporal-frequency attention based convolutional neural network. Scientific Reports, 11(1), 21552.

[11] Zhang, H., McLoughlin, I., & Song, Y. (2015, April). Robust sound event recognition using convolutional neural networks. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 559-563). IEEE.

[12] Sprengel, E., Jaggi, M., Kilcher, Y., & Hofmann, T. (2016). Audio based bird species identification using deep learning techniques. LifeCLEF 2016, 547-559.

[13] Scarpiniti, M., Parisi, R., & Lee, Y. C. (2023). A Scalogram-based CNN approach for audio classification in construction sites. Applied Sciences, 14(1), 90.

[14] Tasripan, T., Kusuma, H., & Huzaini, A. N. R. (2024). Design and Implementation of Environmental Warning Systems for Drivers with Hearing Impairments. European Journal of Electrical Engineering and Computer Science, 8(3), 7-13.

[15] Rahman, A., Shan-A-Khuda, M., Sulaiman, R. B., Shaikh, M. S. I., Abrar, M., Hamim, F. M., ... & Asha, U. F. T. Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models.

[16] Piczak, K. J. (2015, September). Environmental sound classification with convolutional neural networks. In 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP) (pp. 1-6). IEEE.

[17] Abdoli, S., Cardinal, P., & Koerich, A. L. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*, 136, 252-263.

[18] Guastavino, C. (2018). Everyday sound categorization. *Computational analysis of sound scenes and events*, 183-213.

[19] Shannon, R. V., Palumbo, R. L., & Grandgenett, M. (1988). IDENTIFY: A program for speech recognition, identification, and categorization experiments on PC compatible computers. The Journal of the Acoustical Society of America, 83(S1), S16-S16.

[20] Babaee, E., Anuar, N. B., Abdul Wahab, A. W., Shamshirband, S., & Chronopoulos, A. T. (2017). An overview of audio event detection methods from feature extraction to classification. *Applied Artificial Intelligence, 31*(9-10), 661-714.