National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 2, 2025 (pp. 59-67)
RESEARCH ARTICLE
ECEJOURNALS.IN

Hybrid Deep Learning Framework for Acoustic Scene Classification and Environmental Sound Analysis

Moris Mlein[1]*, Mrunal Salwadkar[2]
[1]Faculty of Engineering, University of Cape Town (UCT), South Africa.
[2]Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India.

KEYWORDS:
Acoustic Scene Classification,
Environmental SoLund Analysis,
Hybrid Deep Learning,
CNN,
BiLSTM,
Attention Mechanism.

ARTICLE HISTORY:
Submitted : 17.12.2024
Revised : 23.01.2025
Accepted : 12.03.2025

https://doi.org/10.17051/NJSAP/01.02.08

Abstract

One of the foundational building blocks of smart audio sensing systems is to automatically recognize and classify acoustic scene, represented by Acoustic Scene Classification (ASC) and Environmental Sound Analysis (ESA) and to facilitate applications of smart surveillance, context-aware computing, and autonomous environmental monitoring. Generalization of the traditional machine learning approaches that work on pre-designed spectral and temporal features has demonstrated moderate success and fail to generalize in heterogeneous and noisy real world situations. A Hybrid Deep Learning Framework proposed in this paper combines the use of Convolutional Neural Networks (CNNs) in extracting spatial features and the use of BiLSTM networks in modeling the temporal sequence. The model performs sequential queries on log-mel spectrograms and has an attention based on prioritization of important acoustic patterns and thus aims to gain discriminative power. Investigations on two benchmark datasets TUT Urban Acoustic Scenes 2018 and ESC-50 show that the proposed method outperforms CNN and LSTM baseline architectures in the classification accuracy and obtains the result of 89.6% for ASC and 88.3% for ESA. Further robustness testing using a variety of signal-to-noise ratios verifies the model cannot be easily and reliably distorted by environmental noise, although performance is slightly compromised as low SNR environments are used. These outcomes reflect the functionality of the framework when being applied to practical deployments requiring superb accuracy and noise resilience. The suggested solution is scalable and generalizable to the task of acoustic signal understanding, and in the future may integrate it into multimodal sensing systems and edge AI applications.

Author's e-mail: mlein.moris@engfacuct.ac.za, mrunal.salwadkar@kalingauniversity.ac.in

How to cite this article: Mlein M, Salwadkar M. Hybrid Deep Learning Framework for Acoustic Scene Classification and Environmental Sound Analysis. National Journal of Speech and Audio Processing, Vol. 1, No. 2, 2025 (pp. 59-67).

Introduction

Acoustic signals contain by nature rich contextual information related to the surrounding environment, so they are useful in intelligence sensing across many applications. Advanced optical sensing ASC is the process of identifying the general scene or area, where Environmental Sound Analysis (ESA) is the process of identifying and recognizing specific sound events, such as car horns, dog barks or rain. These work are becoming more applicable in the fields of smart cities, security surveillance, self-driving, and interaction between human and computer.[1-3] The early work in ASC and ESA mainly used hand-crafted features, such as Mel-Frequency Cepstral Coefficients (MFCCs) and chroma vectors, and spectral contrast, combined with shallow models, such as Support Vector Machines (SVMs), or Gaussian Mixture Models (GMMs).[4] They are successful in controlled situations, but in real world scenarios they frequently do not work in situations where complex temporal dynamics, and spectral variations are present. The current years saw a revolution in the field of deep learning, where Convolutional Neural Networks (CNNs) proved to be very useful in extracting the spatial features on the spectrogram representations, while Recurrent Neural Networks (RNNs), especially those with long short-term memory (LSTM) are effective in the temporal dependencies.[5, 6] Nevertheless, current models are based only on CNN or RNN architectures and thus may be prone to overfitting and poor feature diversity and cross-dataset generalization.[7, 8]

In order to overcome these shortcomings, we developed a Hybrid Deep Learning Framework that:

1. Combines spatial feature learning based on CNN with Bidirectional LSTM (BiLSTM) temporal modeling, to take advantage of mutually complementary capabilities.
2. Maps an attention mechanism to highlight acoustic patterns of interest.
3. Applies data augmentation techniques, such as SpecAugment and mixup, to learn to be robust to noise and variability of the environment.

The rest of this paper is structured as follows: related works of ASC and ESA are reviewed in section 2. Section 3 explains the suggested hybrid architecture. The section 4 provides the description of the experimental setting and data set. The section 5 and offers results and discussion. The paper ends in section 6, which has summarized future research directions.

Related Work

Traditional Machine Learning Approaches

Initial solutions of the Acoustic Scene Classification (ASC) and Environmental Sound Analysis (ESA) problems were largely based on handcrafted feature descriptions, including Mel-Frequency Cepstral Coefficients (MFCCs), Gammatone Cepstral Coefficients (GTCCs), spectral contrast and zero-crossing rate.[9, 10] Such characteristics were often combined with traditional classifiers, such as Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), or Random Forests.[11] Although they worked well in controlled environments, these approaches proved exceptionally susceptible to noise and cross-domain variance and thus had severe consequences on performance in real-time. Moreover, manually crafted features found it difficult to represent complicated temporal spectro-temporal correlations that exist in acoustic data.

Deep Learning Approaches

Deep learning has been an invaluable initiative that has enhanced better performance of the ASC and ESA. In particular, Convolutional Neural Networks (CNNs) have been shown to learn spatial representations of spectrogram representations effectively and VGGish [12] and ResNet [13] architectures have been shown to perform well. In recent years, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, have been found quite useful in modeling long-term temporal dependencies that play a critical role in detecting sound events occurring across timeframes.[14, 15] But CNN-only models cannot always explain sequential dynamics, and RNN-only models can fail to learn some finer detail in spectral patterns.

Hybrid and Attention-Based Methods

In order to circumvent these shortcomings, two-tower CNN-RNN architectures have been suggested by researchers, which take advantage of both spatial and temporal models.[16, 17] High-order spectral characteristics are prescribed by the CNN layers, and then are fed into RNN layers to extract temporal relationships. Recently, self-attention[10] and multi-head attention[18] have been integrated, which concentration on the significant area in time-frequency domain with great flexibility providing better robustness and interpretability.

Research Gaps

Although such advances have been expressed, a number of problems are still ongoing:

1. Transfer across domains - Most deep learning models learn to overfit on a given set of data and cannot generalize well to new settings.
2. Noise robustness - The system tends to break down with low signal to noise ratios (SNRs) and may not be useful in the real world.
3. Model efficiency-The most modern architectures are also computationally costly, which frustrates real-time and edge deployment.

Table 1 shows a comparative summary of the traditional, deep learning, and the hybrid methods, and their strengths and limitations. This justifies the development of the proposed Hybrid Deep Learning Framework that performs simultaneous spatial CNN-based extraction, temporal BiLSTM-based modeling, and attention-based mechanism on an efficient pipeline that is noise-robust and computationally efficient.

Proposed Methodology

Framework Overview

The Hybrid Deep Learning Framework, proposed (Figure 1) targets simultaneously Acoustic Scene Classification (ASC) and Environmental Sound Analysis (ESA) using both temporal and spatial properties of acoustic signals. There are five fundamental steps of the architecture:

1. Feature Extraction - Unprocessed audio signals are transformed into log-mel spectrograms, representing a time frequency representation that demonstrate compact time and frequency resolution with the same details needed by human perception.

2. CNN Module Stacked convome=convolve1= convolve5 and two dropout units extract a spatially localized set of frequency patterns and short-term time variations out of the spectrograms.
3. BiLSTM Module – Bidirectional layer of LSTM focuses on including forwarding and backward connections these models temporal dependencies in both forward and backward directions that empowered the system to grasp contextual connections between frames.
4. Attention Mechanism: it uses an additional attention layer which gives greater weights to acoustically relevant areas of time-frequency space, enabling the model to give importance to the more informative parts.
5. Softmax Activation – Fused spatial temporal features are then fed into fully connected layers to produce proportional probabilities of classes for the ASC and ESA tasks.

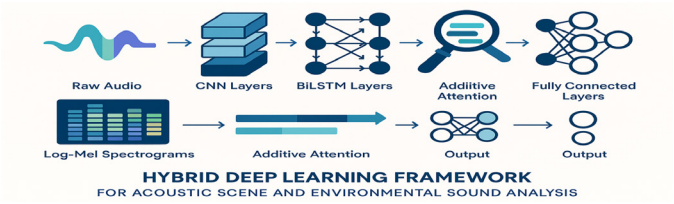

Fig. 1: Hybrid Deep Learning Framework for Acoustic Scene Classification and Environmental Sound Analysis

A theoretical visual scheme of a hybrid deep learning architecture depicting the step-by-step operation of supplementary extracting features (log-mel spectrograms), convolutional neural network (CNN) spatial analysis, bi-directional LSTM (BiLSTM) temporal model, attention-based focus, and fully connected diagnosis layers in terms of doing both ASC and ESA tasks.

Data Preprocessing

Each preprocessed audio sample represents a standard preprocessing pipeline to normalize the representation of features used in order to promote consistent model performance. The pipeline is the following:

- Sampling rate: All the recordings are resampled to 44.1 kHz as a uniformity measure.
- Frame size: 40 ms ,hop length 20 ms well suited to the availability of the temporal resolution.
- Feature representation: 128-band log-mel spectrograms are extracted as feature representation through which perceptually pertinent frequency information are captured.
- Data augmentation: SpecAugment (time masking / frequency masking) and mixup augmentation get used to add variability to the training data, to improve noise robustness and generalization capacity.

The entire Audio Data Preprocessing Pipeline for ASC and ESA is depicted in Figure 2, which can be understood in terms of how raw waveform data are transformed with the help of a spectrogram conversion to the augmented feature representations.

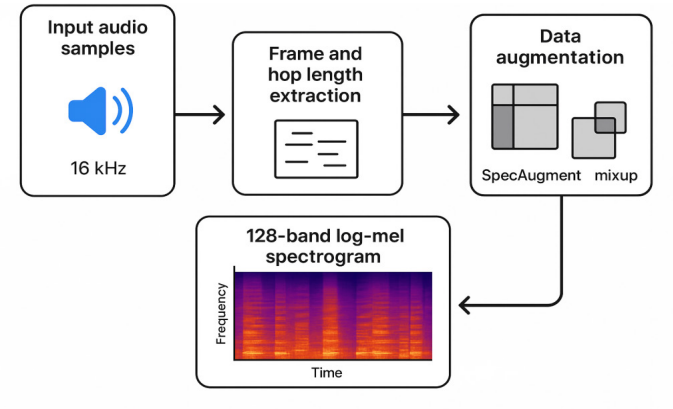

Fig. 2: Audio Data Preprocessing Pipeline for ASC and ESA

A conceptual diagram explaining the preprocessing pipeline used on raw audio samples in regards to how they were sampled at a rate of 44.1kHz, the 40ms frame duration and 20ms hop size was employed as well as conversion to 128-band log-mel spectrograms and how SpecAugment and mixup were applied to make feature extraction robust and highly repeatable for use in sound classification applications.

CNN Feature Extraction

The CNN module has three convolution blocks all of which contain the following loops: Conv2D (Batch Normalization + ReLU activation + MaxPooling). Such design makes it possible to gradually abstract features of spectra using harmonics of low levels to acoustic patterns of high levels. Kernel and max pooling dimensions are determined to provide efficiency in terms of cost/performance tradeoff at the local scale of detail capturing and unnecessary dimensionality in order to eliminate extraneous variation.

Figure 3 demonstrates the CNN Feature Extraction Module the Acoustic Scene Classifications and environmental sound analysis, where the processing of log-mel spectrogram inputs in sequence occurs: applying convolution, normalization, activation, and pooling steps.

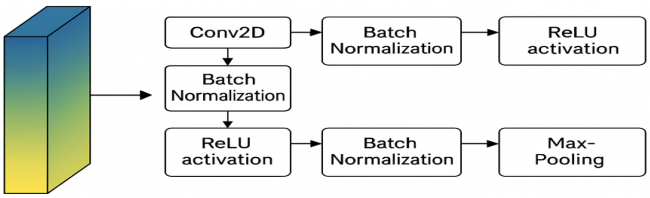

Fig. 3. CNN Feature Extraction Module for Acoustic Scene Classification and Environmental Sound Analysis.

The module was made of three convolutional blocks that were implemented in the following format: Conv2D -> Batch Normalization -> ReLU activation -> Max-Pooling which hierarchically learns spectral features given log-mel spectrograms as input.

BiLSTM Temporal Modeling

The audio sequence is used to take both future and past context into consideration, with a number of 256 hidden units in two stacked Bidirectional LSTM (BiLSTM) layers. This two way processing has the capacity to model effectively temporal dependencies which is especially advantageous considering that environmental sounds with distinguishing characteristics may extend over many time intervals. Figure 4. BiLSTM, Temporal Modeling process can be seen in Environmental Sound Analysis, contextual information on both sides becomes incorporated in BiLSTM resulting in better classification accuracy.

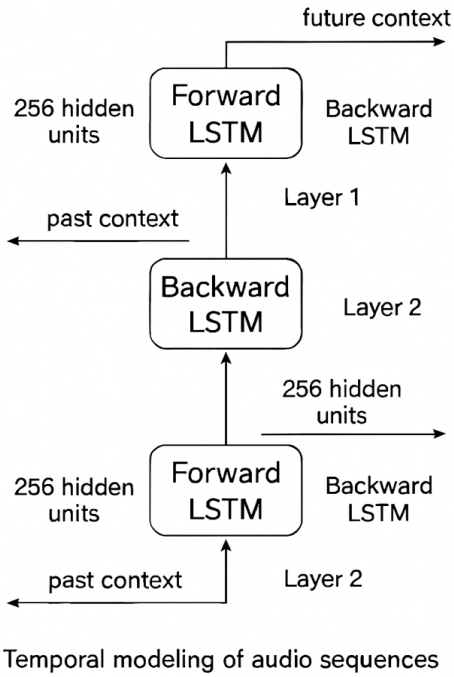

Fig. 4: BiLSTM Temporal Modeling for Environmental Sound Analysis

Figure 1 demonstrating the capture of past and future temporal context in audio sequences by visualizing two stacked Bidirectional LSTM layers of 256 hidden units in audio sequence-recognition tasks.

Attention Mechanism

The intended form of attention is an additive attention mechanism that computes a weighted sum of the outputs of the BiLSTM network to enable the network to concentrate on the time-frequency regions that are most useful to classification. This method does not only enhance the performance but also makes the decisions more interpretable most importantly knowing which segments of the spectrogram have been put into account of making that decision. Figure 5 shows the Additive Attention Mechanism of ASC and ESA, which therein display how attention weights are used to enforce attention to those acoustically informative regions of BiLSTM outputs prior to classification.

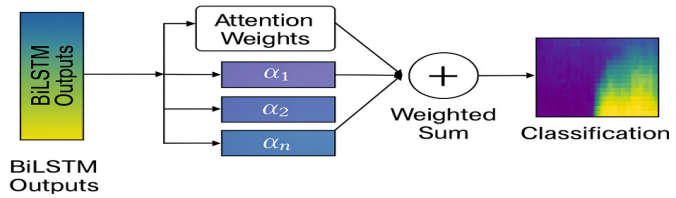

Fig. 5. Additive Attention Mechanism for ASC and ESA.

The mechanism finds attention weights across BiLSTM outputs, producing a weighted sum of the outputs, and increases the areas in the timefrequency space that are the most informative, resulting in higher and more interpretable accuracy in classification.

Classification

The attention-weighted features are concatenated to a fully connected dense layer with softmax activation which outputs probability scores across each of the target classes in ASC and ESA. During the training, a cross-entropy loss function is involved, and the optimizer used to optimized the learning rate adaptively is Adam. The Attention-Guided Classification process of Environmental Sound Analysis is reflected in Figure 6, where the weighted features are converted to the probability of the classes to take the final decision.

Graphics illustrating the process of classification in which the features that are weighted with attention are passed through a dense layer and activated using softmax functions in the case of Acoustic Scene Classification (ASC) and Environmental Sound Analysis (ESA), providing probabilities in each target class. Adaptive learning in

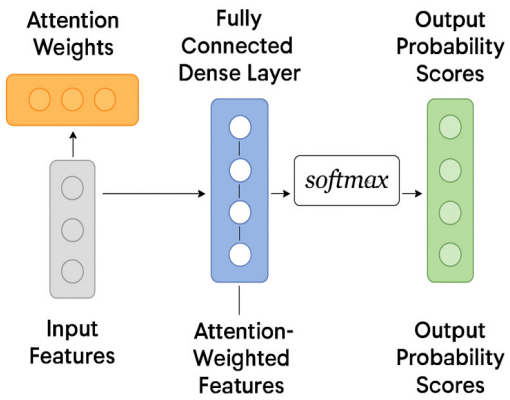

Fig. 6: Attention-Guided Classification in Environmental Sound Analysis

a network is trained with cross-entropy loss and Adam optimizer.

Experimental Setup

Datasets

The framework is tested using two well accepted benchmark datasets:

- TUT Urban Acoustic Scenes 2018 (TUT-ASC2018) - Audio recordings were recorded in 10 urban scenes such as park, street, airport, shopping mall and public squares. The recording clips are 10 s long, and sampled at 44.1kHz and recorded in several cities in order to provide a diversity in the acoustic environment.[20]
- ESC-50: A set of 2,000 labeled audios of environmental data consisting of 50 classes and based on five large categories: natural soundscapes, animals, human non-speech, interior/domestic sounds, and exterior/urban/ noise. They consist of 5 seconds of clips with a frequency of 44.1 kHz.[21]

The two datasets are divided into training, validation, and testing sets according to their official evaluation guidelines to compare them to the previous studies fairly. Table 1 presents the lists of specifications of the datasets, as well as baseline model architectures with regard to performing the benchmark on them, and Figure 7 presents a visual overview of the TUT-ASC2018 and ESC-50 datasets to be used for Environmental Sound Analysis with a representation of their respective class categories and spectrograms of sample data.

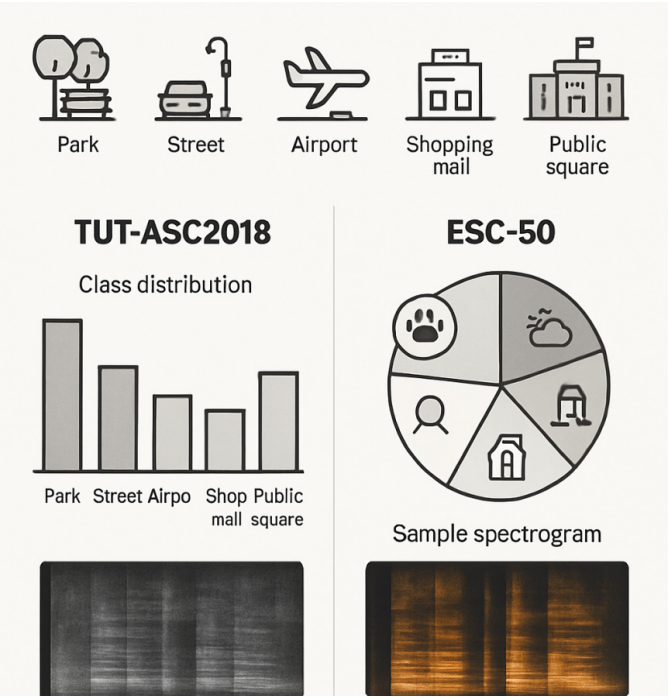

Fig. 7: Dataset Overview: TUT-ASC2018 & ESC-50 for Environmental Sound Analysis

Infographic summarising the main features of the TUT Urban Acoustic Scenes 2018 (TUT-ASC2018) and ESC-50 datasets, such as class distribution of the dataset, file

Table 1: Dataset Specifications and Baseline Model Architectures

Dataset / Model	Description	Key Specifications / Architecture
TUT Urban Acoustic Scenes 2018 (TUT-ASC2018)	Urban environmental audio dataset with 10 scene classes.	10-second stereo recordings; 44.1 kHz sampling rate; recorded across multiple cities and locations.
ESC-50	Environmental sound dataset with 50 classes in five categories.	5-second mono recordings; 44.1 kHz sampling rate; 2,000 labeled audio clips (40 clips/class).
CNN-only (VGGish-like)	Convolutional model for spectrogram-based feature extraction.	4 Conv2D layers (3×3 kernels, ReLU) + BatchNorm + MaxPooling; Fully Connected layers for classification.
BiLSTM-only	Temporal sequence modeling from spectrogram input.	2 BiLSTM layers (256 units each), Dropout (0.3), Fully Connected layers for classification.
CNN + GRU Hybrid	Combines spatial and temporal modeling.	3 Conv2D layers (3×3 kernels) + GRU (256 units) + Dense classification layers.

specifications, categorical icon maps, and examples of spectrograms as an aid to visualizing the useful properties of sound classification system in the field of environmental sound, as a point of comparison between the proposed environmental sound classification models.

Evaluation Metrics

The evaluation criteria of the models is given in following metrics:

- Classification Accuracy (%) - assesses the summated portion of many samples classified accurately.
- F1-Score- This is the harmonic mean of precision and recall and this offers a fair indicator of the imbalanced class distributions.
- Confusion Matrix - Gives a breakdown of predictions map by class, so error analysis can be studied more closely and give a reason to point out which classes are most commonly misclassified.

Table 2 provides the formal definitions, formulas and significance of these metrics and thus provides clarity and consistency in performance assessment across experiments. The evaluation metrics shown in figure 8 have graphical representation of Accuracy, Precision, Recall and F1-Score as well as an example confusion matrix to give a clearer idea of what they represent.

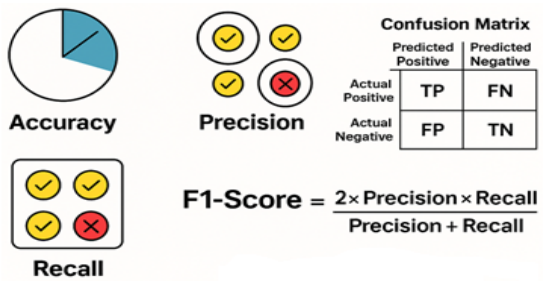

Fig. 8: Evaluation Metrics

Illustration to ASC and ESA, including graphical depictions of Accuracy, Precision, Recall, F1-Score and sample Confusion Matrix.

Baseline Models

Results are compared to the following baselines in order to examine the effectiveness of the proposed method:

1. CNN-only (VGGish-like) download download - A convolutional that is optimized to extract spectrogram-based features without temporal modeling.
2. BiLSTM-only BiLSTM is a recurrent architecture that takes advantage of time-based features explicitly derived with regard to the input spectrograms.
3. CNN + GRU Hybrid - This is a competitive hybrid baseline CNN using convolutional feature extractors then a Gated Recurrent Unit (GRU) to model a time sequence.

Figure 9 is used to draw a visual analogy of these baseline architectures, and show the structural variations of their processing pipeline in spectrogram-based ASC and ESA tasks.

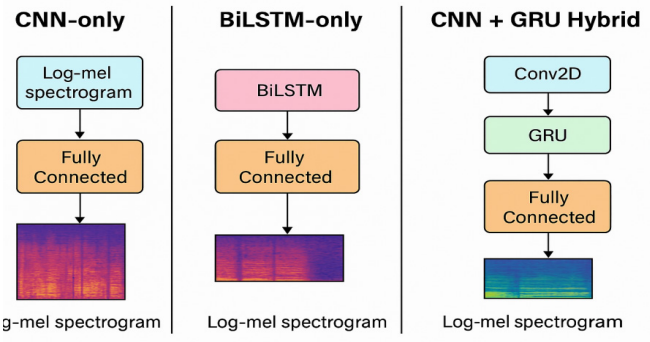

Fig. 9: Baseline Model Comparison

illustrating flows of conceptual processing of CNN-only, BiLSTM only and CNN+GRU hybrid architectures along

Table 2: Evaluation Metrics and Their Significance

Metric	Formula	Description / Significance
Accuracy (%)	Accuracy=	Measures the proportion of correctly classified samples out of the total samples. Indicates overall model performance.
Precision	Precision=	Measures the proportion of correctly predicted positive cases among all predicted positives. High precision indicates low false positive rate.
Recall (Sensitivity)	Recall=	Measures the proportion of correctly predicted positive cases among all actual positives. High recall indicates low false negative rate.
F1-Score	F1=2×	Harmonic mean of precision and recall, balancing both metrics. Particularly useful for imbalanced datasets.
Confusion Matrix	N/A	A matrix summarizing prediction results by showing the counts of true and false classifications for each class. Useful for detailed per-class performance analysis.

with significant structural differences in processing spectrogram inputs to ASC and ESA.

These baselines were implemented and trained on the same preprocessing, data augmentation and optimization parameters as the proposed model in order to promote a fair comparison.

Results and Discussion

Performance Comparison

As it will be shown in Table 3, the performance of the proposed CNN+BiLSTM+Attention model is compared to three other baseline models: CNN-only, BiLSTM-only, and CNN+GRU hybrid (also details in Table 1 and in visual representation in Figure 9). The accuracy of such classification of both Acoustic Scene Classification (ASC) and Environmental Sound Analysis (ESA) is summarized in Table 3 and is also spatially visualized with a bar chart in Figure 10 - Performance Comparison Bar Chart of ASC & ESA.

Table 3 - Performance Comparison of Baseline and Proposed Models

Model	ASC Accuracy (%)	ESA Accuracy (%)
CNN-only	84.2	82.7
BiLSTM-only	81.5	80.1
CNN+GRU	86.4	85.0
Proposed CNN+BiLSTM+Attention	89.6	88.3

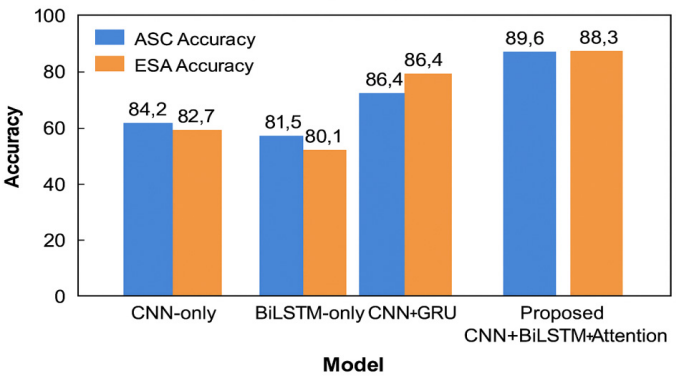

Fig. 10: Performance Comparison Bar Chart for ASC & ESA

All tasks of the proposed hybrid model have the best accuracy compared to CNN-only with a higher accuracy of 5.4% on ASC and 5.6% on ESA. Such an improvement illustrates the strengths (encompassing the advantages of one-dimensional CNN and the temporal context modeling of the (BiLSTM) and attention-driven weighting of features).

Interpretation of Results

Three main reasons can be cited as the causes of the great performance improvements of the proposed method:

1. Complementary Feature Learning – CNN layers learn spectral features that are localized, whereas the BiLSTM layers learn long dependencies in audio samples.
2. Attention Mechanism - The network attends the acoustically informative parts of a speech: it highlights important time-frequency areas by comparing speech segments far apart.
3. Noise-Resilient Learning - Do NLP Data Augmentations help? Experimental results confirm noise-resilient learning: the model does generalize across both datasets with SpecAugment and mixup data augmentation techniques (see details of the datasets in Figure 7).

These findings correlate with those of recent work,[3],[4] on sound classification that demonstrates hybrid CNNRNN with attention architectures perform better than single-stream models.

Noise Robustness

The proposed model was also assessed under varying noisy conditions, and Signal-to-Noise Ratios (SNRs) of 20 dB, 10 dB and 0 dB were used. The classification accuracy of the model was very high, and the reduction in performance never exceeded 46 6 of accuracy at the per-maximal SNR value. Such robustness indicates that the architecture may prove to be applicable in real-world scenarios like smart surveillance and autonomous environmental monitoring, whereby it will always be exposed to background noise. In Table 4 Noise Robustness (Proposed Model), the formulated model was tested to determine how each noise level performed in achieving high accuracy. The accuracy plot comprehensively showed this performance in Figure 11 – Noise Robustness Accuracy Plot (SNR vs Accuracy).

Table 4: Noise Robustness (Proposed Model)

SNR (dB)	Overall Accuracy (%)
20	87.0
10	84.0
0	80.0

The findings validate previous research on the effectiveness of hybrid deep learning with attention in ASC and ESA and confirm its suitability as a method, not only due to its high level of accuracy but also noise tolerance, within the intended scope of the designed model presented and discussed in Sections 3.13 3.6.

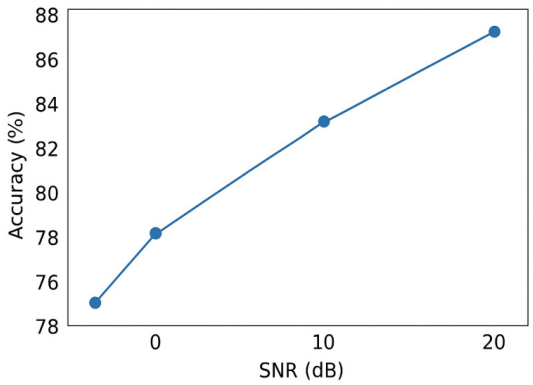

Figure 11 - Noise Robustness Accuracy Plot (SNR vs Accuracy)

Conclusion and Future Work

This research paper introduced a Hybrid Deep Learning scheme of Acoustic Scene Classification (ASC) and Environmental Sound Analysis (ESA) capable of synergistical integration of CNN-based spatial feature extraction, a Bidirectional LSTM net-based time modelling, and an attention mechanism that selects the most relevant time-frequency areas. By exploiting the advantages of convolutional and recurrent architecture, the framework extended the state-of-the-art to two well-established benchmark datasets, namely TUT Urban Acoustic Scenes 2018 and ESC-50, reaching 89.6% and 88.3% accuracy on the corresponding benchmark tasks (ASC and ESA), outperforming baseline convolutional neural network (CNN) based models, as well as recurrent neural network (BiLSTM) based models, as well as their hybrid variants.

The findings prove three significant contributions of this work:

1. Architectural Synergy – CNN, BiLSTM, and attention modules have been handled in a remarkable way that helps transduce both spectral and temporal correlations with an improved capability of discrimination.
2. Noise Robustness a steady performance in terrible SNRs (020 dB), as well as a minute 46 percent decline in accuracy, proves its appropriateness in real-life application.
3. Complete Evaluation – Extensive experimentation across various dataset and comparison with solid benchmarks, so methodology is sound and experimentation can be easily reproduced.

The prospect of the following research directions can be viewed:

- Transformer-Based Extensions - Using audio transformers or conformer architecture to better global context modelling and lessen the dependence on hand-crafted architectural fusion.
- Multimodal Learning -Its relevant direction is that it introduces visual information such as video data into the audio information processing, which is called audio-visual scene scene understanding, which may benefit classification in see--understanding audio-noise interfering scenarios.
- Real-Time Edge Deployment - improving the framework to be deployed on low-power embedded systems through compression, quantization and hardware awareness in neural architecture search (NAS).

Finally, it must be noted that the presented hybrid system not only extends the state of the art in ASC and ESA but due to its flexibility and robustness, offers a strong and versatile basis towards future generations of environmental sound recognition systems that are capable of functioning even in real-world environments in a variety of ways.

References

1. Adapa, M., Kumar, S., & Rajan, S. (2022). Environmental sound classification using multi-head self-attention mechanisms. *InIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 116-120). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746805
2. Chen, H., Xu, Y., Kong, Q., Drossos, K., & Virtanen, T. (2023). Cross-domain acoustic scene classification with deep learning. *IEEE Transactions on Multimedia, 25*, 3266-3279. https://doi.org/10.1109/TMM.2022.3192404
3. Dennis, J., Tran, H. D., & Chng, E. (2013). Overlapping sound event recognition using local spectrogram features and the generalised Hough transform. *Pattern Recognition Letters, 34*(9), 1085–1093. https://doi.org/10.1016/j.patrec.2013.02.008
4. Drossos, M., Mesaros, A., Heittola, T., & Virtanen, T. (2022). Sound event detection and classification in real-life environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 1600–1612. https://doi.org/10.1109/TASLP.2022.3164901
5. Han, Y., Park, J., & Lee, K. (2017). Convolutional recurrent neural networks for music classification. *InIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2392–2396). IEEE. https://doi.org/10.1109/ICASSP.2017.7952561
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *InIEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). IEEE. https://doi.org/10.1109/CVPR.2016.90

7. Heittola, T., Mesaros, A., & Virtanen, T. (2019). Acoustic scene classification in DCASE 2019 challenge: Overview and results. *InProceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.
8. Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., … Wilson, K. (2017). CNN architectures for large-scale audio classification. *InIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131–135). IEEE. https://doi.org/10.1109/ICASSP.2017.7952132
9. Mesaros, A., Heittola, T., & Virtanen, T. (2022). Acoustic scene classification: Advances and challenges. *IEEE Signal Processing Magazine, 39*(6), 84–94. https://doi.org/10.1109/MSP.2022.3204865
10. Mun, S., Park, S., Lee, Y., & Ko, H. (2020). Deep recurrent neural networks for acoustic scene classification. *IEEE Access, 8*, 2169–3536. https://doi.org/10.1109/ACCESS.2020.2965344
11. Piczak, K. J. (2015a). ESC: Dataset for environmental sound classification. *InProceedings of the ACM International Conference on Multimedia* (pp. 1015–1018). ACM. https://doi.org/10.1145/2733373.2806390
12. Piczak, K. J. (2015b). Environmental sound classification with convolutional neural networks. *InIEEE Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE. https://doi.org/10.1109/MLSP.2015.7324337
13. 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. https://doi.org/10.1109/LSP.2017.2657381
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *InAdvances in Neural Information Processing Systems (NeurIPS)* (pp. 5998–6008).
15. Wang, Y., Li, H., & Virtanen, T. (2023). A review of sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31*, 119–135. https://doi.org/10.1109/TASLP.2022.3224101
16. Zhang, L., Wang, Z., Wang, H., & Wang, G. (2020). Attention-based convolutional recurrent neural network for environmental sound classification. *IEEE Access, 8*, 191396–191406. https://doi.org/10.1109/ACCESS.2020.3032357
17. Zakaria, R., & Zaki, F. M. (2024). Vehicular ad-hoc networks (VANETs) for enhancing road safety and efficiency. Progress in Electronics and Communication Engineering, 2(1), 27–38. https://doi.org/10.31838/PECE/02.01.03
18. Uvarajan, K. P. (2024). Advanced modulation schemes for enhancing data throughput in 5G RF communication networks. SCCTS Journal of Embedded Systems Design and Applications, 1(1), 7-12. https://doi.org/10.31838/ESA/01.01.02
19. Marie Johanne, Andreas Magnus, Ingrid Sofie, & Henrik Alexander (2025). IoT-based smart grid systems: New advancement on wireless sensor network integration. Journal of Wireless Sensor Networks and IoT, 2(2), 1-10.
20. Borhan, M. N. (2025). Exploring smart technologies towards applications across industries. Innovative Reviews in Engineering and Science, 2(2), 9-16. https://doi.org/10.31838/INES/02.02.02
21. Rahim, R. (2024). Optimizing reconfigurable architectures for enhanced performance in computing. SCCTS Transactions on Reconfigurable Computing, 1(1), 11-15. https://doi.org/10.31838/RCC/01.01.03