IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.006, LIIR355
DOI: https://doi.org/10.52731/liir.v006.355

# Lightweight Convolutional Recurrent Neural Networks for Sound Event Classification

Nayana Agrahara Dattatri [*], Cheng Siong Chin [*],
Daniel Archambault [*], Caizhi Zhang [†]

## Abstract

Sound Event Classification (SEC) is essential for applications like urban noise monitoring and smart home automation, but modern models often struggle with efficiency and deployability. This study evaluated lightweight SEC architectures namely CNN, CRNN, and Transformer using the UrbanSound8K dataset, considering both accuracy and resource consumption. CRNN emerged as the top performer, achieving around 90% accuracy with only 175,754 parameters, surpassing the efficiency of CNNs and Transformers. These results underscore the CRNN's potential for scalable and cost-effective SEC solutions, making it ideal for smart city infrastructure and resource-limited IoT applications.

*Keywords:* Sound Event Classification, Convolutional Neural Network, Convolutional Recurrent Neural Network, Transformer.

## 1 Introduction

Sound Event Classification (SEC) is a key audio processing technology that identifies and classifies sounds along with their timing. It has practical applications in urban noise monitoring [1][2], security systems for detecting unusual sounds like gunshots, smart homes for alarms or voice commands, autonomous vehicles for critical audio cues, wildlife monitoring, and healthcare for real-time patient observation. Hybrid models, such as Convolutional-Recurrent Neural Networks (CRNNs), combine the spatial strengths of CNNs with the temporal modeling of RNNs, offering high performance with moderate computational demand [3]. Capsule Networks and Transformer-based models like BERT and 1D-DETR deliver accurate results but require significant resources [4][5][6]. Long Short-Term Memory (LSTM) networks are effective for long-term dependencies, while attention mechanisms in Transformers enable models to focus on key input sequences, improving accuracy [7]. However, these advancements face challenges like weakly labeled data, poor domain adaptation, and high computational costs, especially for real-time or resource-constrained applications [8]. With the rise of edge computing, lightweight SEC models have gained importance. On-device processing enhances energy efficiency, reduces hardware costs, and protects user privacy by eliminating reliance on cloud services. This study compares CNN, CRNN, and Transformer models for SEC, focusing on accuracy and efficiency. The CRNN emerged as the most effective, leveraging convolutional and recurrent layers to capture spatial and temporal features efficiently. The findings emphasize the importance of balancing performance and resource usage, paving the way for practical and scalable SEC systems.

[*] Newcastle University, Newcastle upon Tyne, United Kingdom
[†] Chongqing University, China

## 2 Dataset and Model

This study evaluates and compares three deep learning models—CNN, CRNN, and Transformer for Sound Event Classification (SEC) using the UrbanSound8K dataset. Each model was implemented, trained, and assessed based on its classification accuracy across diverse audio events. The methodology covers data preprocessing, model architecture, training procedures, evaluation metrics, and design rationale. UrbanSound8K, introduced by Salamon, Jacoby, and Bello in 2014, consists of 8,732 labeled audio clips (up to 4 seconds each) across ten urban sound classes: air-conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. Its standardized 10-fold cross-validation structure makes it a widely adopted benchmark for SEC research. This study utilized UrbanSound8K for its variety of sound events and robust cross-validation setup, ensuring generalizable evaluations. Before training, audio clips were resampled to 16 kHz using Librosa, a rate balancing efficiency and auditory detail. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to capture spectral features crucial for SEC. The extraction process began with a Short-Time Fourier Transform (STFT) using a 25ms Hann window and a 10ms hop length for high temporal resolution. A Mel spectrogram with 40 frequency bins was computed, optimized for detail and efficiency. Dynamic range compression and decorrelation were applied to align with human auditory perception and enhance noise resilience. Each clip was converted into a (40×T) feature matrix, where T represents time frames, capturing the audio's temporal structure.

### 2.1 Data Augmentation

To enhance model robustness and generalization, particularly with limited data, data augmentation techniques were implemented for the CNN model. **Time shifting:** Audio samples were shifted by 10 frames to simulate slight variations in event timing. This augmentation helps the model become more invariant to small temporal shifts in sound events, which is important for real-world SEC tasks where sound events may not occur at precisely the same time across different recordings. **Noise injection:** Gaussian noise with a mean of 0 and a standard deviation of 0.05 was added to the original samples. This technique simulates real-world audio conditions where background noise is present, thereby improving the model's ability to generalize to noisy environments. By introducing variations in the training data, these augmentations help the model learn to recognize sound events under different conditions, reducing the risk of overfitting.

### 2.2 Model Architectures

Three common model architectures were implemented and compared. The Convolutional Neural Network (CNN) uses three convolutional layers (32, 64, and 128 filters) with Batch Normalization, MaxPooling, and Dropout to prevent overfitting. It includes a fully connected layer (128 units) and employs ReLU and soft-max activation functions, totaling around 1.2 million parameters. The Convolutional Recurrent Neural Network (CRNN) combines CNNs and RNNs, featuring two convolutional layers (32 and 64 filters) and an LSTM layer (128 units) for temporal dynamics. It concludes with a fully connected layer (64 units), using ReLU and Softmax activations, and has around 175,754 parameters. The Transformer model processes audio as a sequence, using 8 attention heads and two feed-forward layers (512 units each). A final fully connected layer (256 units) integrates in-formation, with ReLU and Softmax activations. With around 2.5 million parameters, it excels at learning complex audio patterns

# 3 Experimental Setup

This section compares the performance of CNN, CRNN, and Transformer models to determine the most effective approach for Sound Event Classification (SEC). These models were selected for their unique strengths: CNNs excel in capturing localized spatial features, CRNNs blend spatial and temporal modeling through their hybrid architecture, and Transformers leverage self-attention for long-range dependencies. Among them, CRNN emerged as the top performer due to its ability to effectively capture both frequency components and temporal dynamics, making it ideal for SEC tasks. The experiments were conducted on a system with an Intel Core i7-13700HX processor, 16 GB RAM, and an NVIDIA RTX 4070 GPU (12 GB GDDR6X), ensuring sufficient computational capacity. All models used the Adam optimizer for efficient convergence. Key hyperparameters, such as a batch size of 16 and adaptive learning rates, were optimized through iterative testing, with CNNs employing early stopping (typically under 50 epochs) and CRNN/Transformer models following fixed schedules. The CRNN achieved optimal performance with approximately 175,754 parameters, compared to 1.2 million for CNNs and 2.5 million for Transformers. Not all hyperparameters are detailed here to maintain focus on critical comparisons, and because hyperparameters often require fine-tuning based on specific tasks and datasets. The results highlight the trade-offs between model complexity and performance, offering valuable insights into resource-constrained applications.

Table 1: Hyperparameter comparison

| **Hyperparameter** | **CNN** | **CRNN** | **Transformer** |
|---|---|---|---|
| Batch Size | 16 | 16 | 16 |
| Learning Rate | 0.001 (with scheduler) | 0.001 (fixed) | 0.001 (fixed) |
| Epochs | 50 | 50 | 10 |
| Optimizer | Adam | Adam | Adam |
| Regularization | L2 (0.001) | None | None |
| Dropout | 0.3 | 0.3 | None |
| Number of parameters | ~ 1.2M | ~ 175k | ~ 2.5 M |
| Activation Function | ReLU (Conv Layers) | ReLU (Conv, LSTM Layers) | ReLU (FF Layers) |

# 4 Results and Evaluation

This section presents the performance evaluation of three models: Convolutional Neural Network (CNN), Convolutional Recurrent Neural Network (CRNN), and Transformer on the UrbanSound8K dataset. Each model's performance is measured in terms of accuracy, computational efficiency, and class-wise recognition, with a focus on CRNN as the proposed solution.

## 4.1 Model Performance Assessment

To evaluate the performance of CNN, CRNN, and Transformer models, a combination of quantitative metrics and qualitative analysis was used. Classification accuracy was the primary metric, representing the proportion of correctly classified sound events. Accuracy trends were tracked and plotted over epochs to observe learning progress, while cross-entropy loss was used to measure prediction errors. For the Transformer model, training and validation losses were closely monitored to identify overfitting, with loss curves providing insights into convergence and learning stability. As shown in Fig. 1, CNN achieved a final validation accuracy of approximately 75%. Training progress was enhanced by using a learning rate scheduler and early stopping, which helped improve generalization on the limited UrbanSound8K dataset. The learning rate scheduler adjusted the rate dynamically for smoother training, while early stopping prevented overfitting by halting training at the optimal point. Despite these strategies, some overfitting persisted after fine-tuning, reflecting the challenges posed by the dataset's constraints. As shown in Fig. 2, the CRNN consistently outperformed the other models, reaching about 90% validation accuracy. Its strength comes from combining convolutional layers (capturing spatial features like frequency patterns) with recurrent layers (modeling temporal dependencies such as rhythm), allowing it to understand sound events deeply. Trained for a fixed 50-epoch, the CRNN showed steady improvement without overfitting, unlike the CNN, demonstrating strong generalization. Both accuracy and loss curves remained stable, highlighting its reliability and efficiency, making it a top choice for real-time audio tasks. In Fig. 3, the Transformer reached around 85% accuracy. Though promising, it didn't surpass CRNN or CNN. It was only trained for 10 epochs due to its heavy computational demands and long training times. This shorter training limited its convergence and overall performance, suggesting it needs more epochs, better tuning, or a larger dataset to fully shine in SEC tasks. The Transformer's complexity and resource needs make it less practical here, despite its theoretical power.

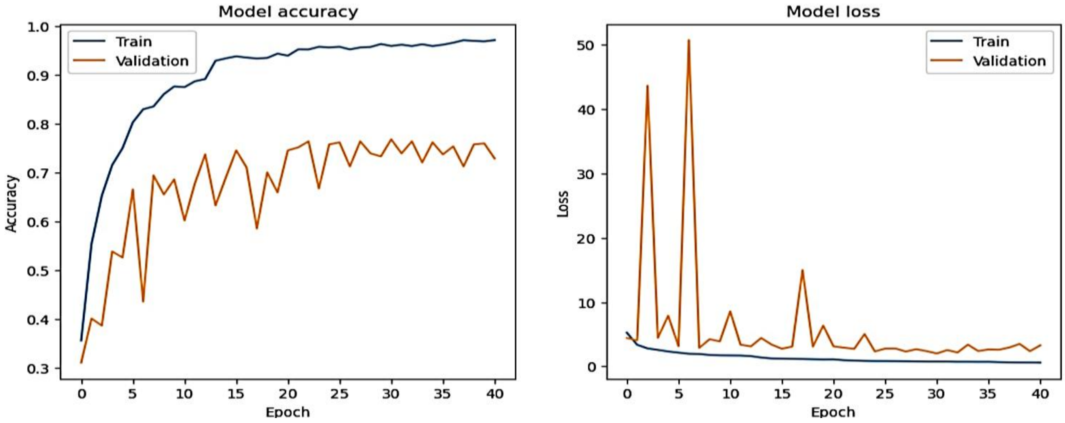

Figure 1: CNN model accuracy and loss curves during training

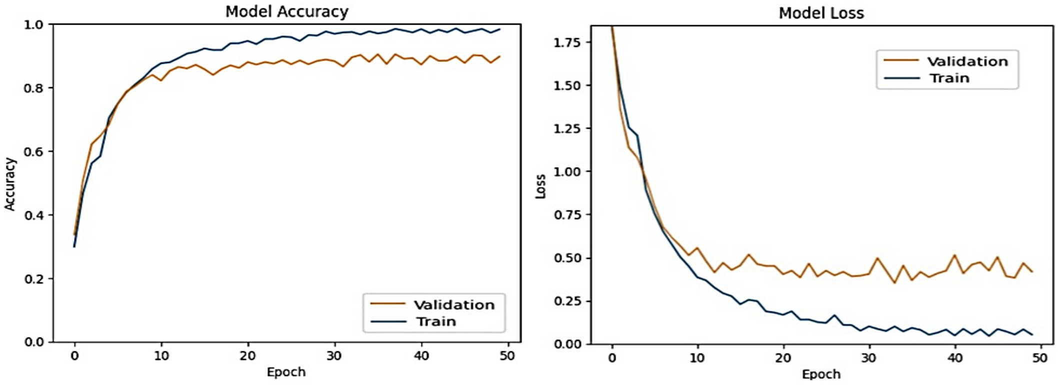

Figure 2: CRNN model accuracy and loss curves during training

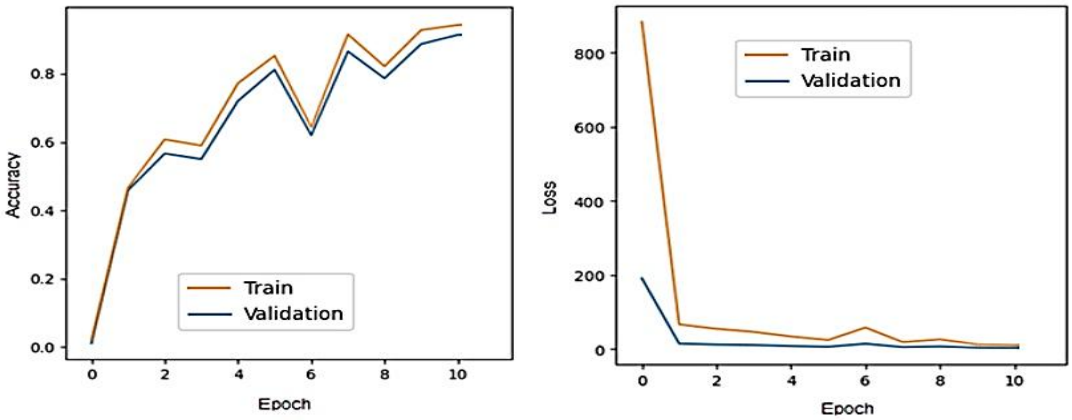

Figure 3: Transformer model training and validation loss, and validation accuracy curves

### 4.2 Comparative Analysis

A critical aspect of this research was to assess the computational efficiency of each model. As seen in Table 2, the analysis included training time, inference time and memory usage. The Transformer model required significantly more time to train per epoch compared to the CNN and CRNN models due to its complex architecture and higher parameter count. In contrast, the CNN model, with fewer parameters, trained much faster while still achieving reasonable performance. The inference time for the CRNN was slightly longer than that of the CNN but shorter than the Transformer. The CRNN's balance of accuracy and computational efficiency made it the most suitable for real-time applications on resource-constrained platforms. The memory footprint of the Transformer model was the largest, which posed challenges for deployment on devices with limited memory resources. The CNN and CRNN models, with their smaller memory requirements, are more practical for such environments. These efficiency metrics are crucial for determining the feasibility of deploying these models in real-world applications.

Table 2: Performance Comparison

| Model | Accuracy | Final Validation Loss | Parameters | Training Time | Epochs to Convergence |
|---|---|---|---|---|---|
| **CNN** | 75% | 0.35 | 1.2M | 7200s | 40 epochs |
| **CRNN** | 90% | 0.30 | 175k | 1320s | 50 epochs |
| **Transformer** | 85% | 0.45 | 2.5M | 3744s | 10 epochs |

### 4.3 Computational Efficiency

The CRNN model strikes a good balance between accuracy and computational efficiency, making it a standout choice for real-time Sound Event Classification tasks. The CRNN model excels in real-time Sound Event Classification by balancing accuracy and efficiency. With only 175,754 parameters, it outperforms the 2.5M-parameter Transformer model while remaining lightweight. Its hybrid architecture, combining convolutional and recurrent layers, effectively captures spatial and temporal patterns, enabling strong generalization with minimal computational cost. Unlike the Transformer, whose self-attention mechanism adds size and complexity without significant accuracy gains, the CRNN's design ensures faster training, lower memory usage, and suitability for resource-constrained platforms like mobile devices. This makes the CRNN an ideal choice for practical, real-time sound classification systems.

### 4.4 Proposed CRNN Model

To highlight the CRNN model's strengths, we evaluated its performance across diverse sound classes. The confusion matrix in Fig. 4 demonstrates its high accuracy, excelling in challenging categories like air-conditioner, jackhammer, and siren. This reflects the model's robustness in handling complex and overlapping sound patterns typical in real-world environments. The Precision-Recall (PR) curve in Fig. 5 confirms the CRNN's ability to balance precision and recall, essential for real-time Sound Event Classification (SEC) tasks where minimizing both false positives and false negatives is critical. Additionally, the ROC curve in Fig. 6, with AUC values nearing 1.0 for many classes, showcases its exceptional discriminatory power and reliability. Comparative evaluations with CNNs and Transformer-based architectures were excluded as these models did not meet the task's constraints. CNNs struggled with temporal dynamics crucial for SEC, while Transformers, despite their popularity, faced challenges due to high computational demands and limited temporal coherence on smaller audio datasets. CRNN emerged as the optimal solution, offering a practical, high-performing architecture suited for real-world applications.

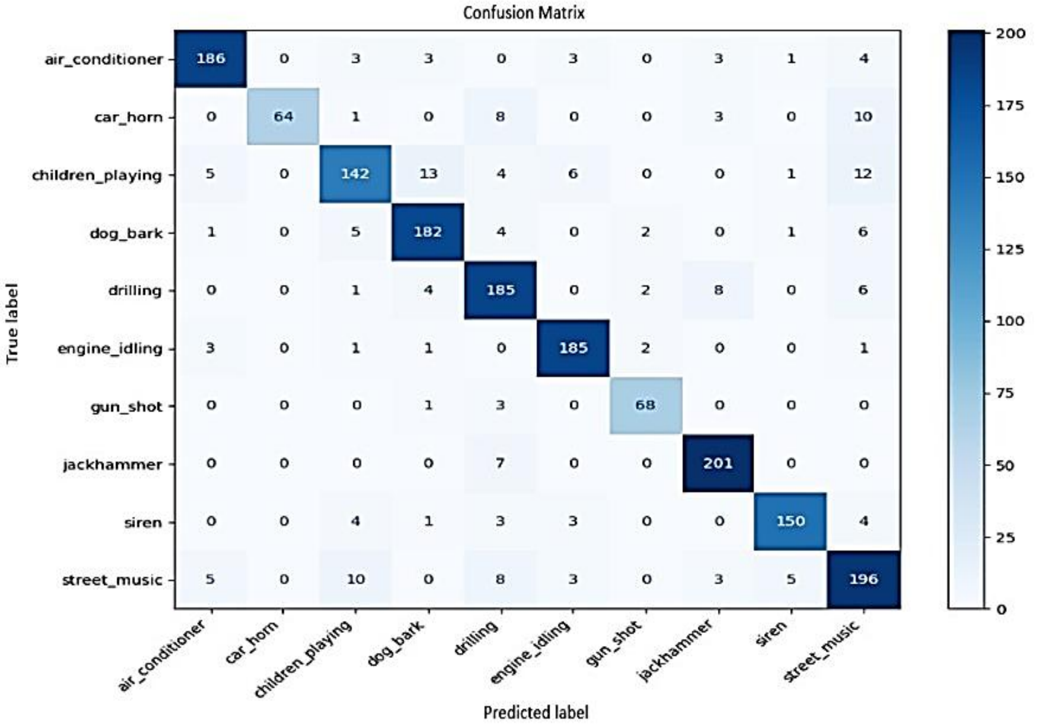

Figure 4: Confusion Matrix of CRNN model on the UrbanSound8K dataset

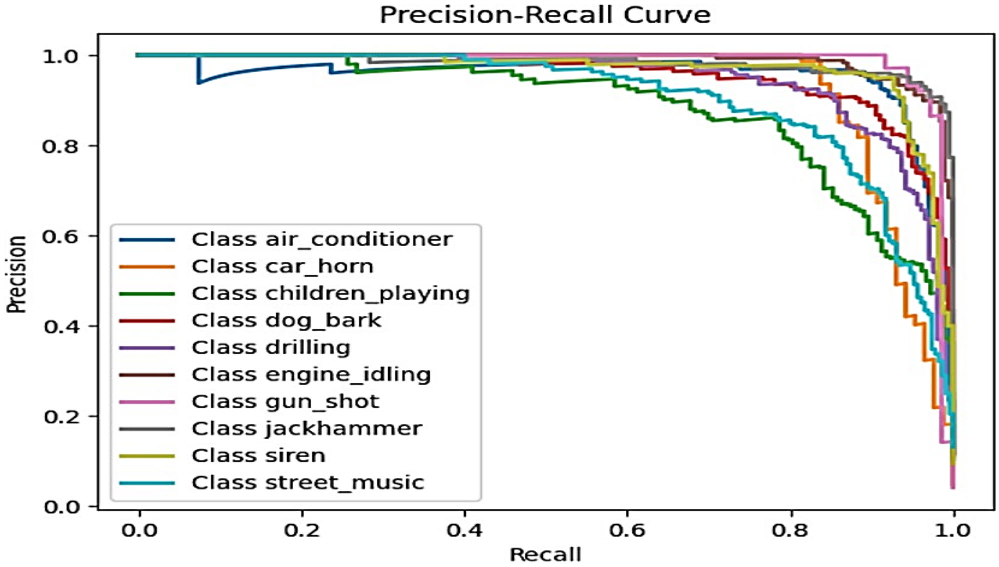

Figure 5: Precision-Recall curves for different sound classes using CRNN Model

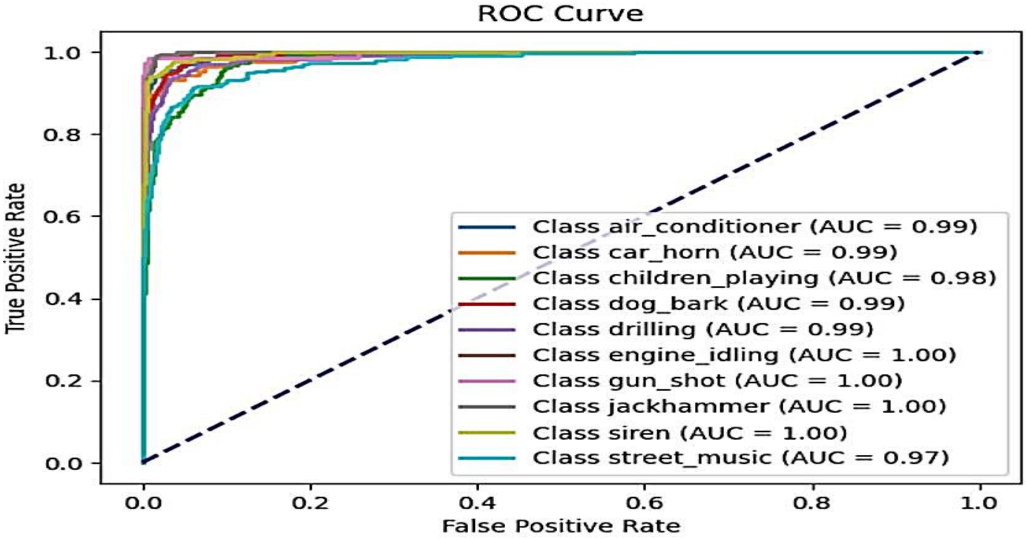

Figure 6: ROC curves and AUC values for different sound classes Using CRNN model

## 5 Conclusion

This study evaluated three deep learning architectures namely CNN, CRNN, and Transformer for Sound Event Classification (SEC) using the UrbanSound8K dataset. The CRNN outperformed its counterparts, achieving around 90% validation accuracy, compared to 75% for the CNN and 85% for the Transformer. The CRNN's ability to integrate spatial and temporal audio features makes it particularly effective for SEC tasks. With just 175,754 parameters, it balances complexity and performance, meeting the demand for lightweight, real-time solutions. While the Transformer excelled at capturing complex temporal relationships, its high computational cost and extended training times limit its practicality for resource-constrained applications. Conversely, the CNN, though computationally efficient, lacked the temporal modeling needed for accurate classification. Future work could focus on optimizing the CRNN for low-resource environments and integrating multimodal inputs, such as visual data, to enhance robustness and adaptability in diverse settings.

## Acknowledgement

We extend our gratitude to Newcastle University for their invaluable support and assistance throughout this project.

## References

[1] Alsina-Pagès, RM., Benocci, R., Brambilla, G., Zambon, G.: Methods for Noise Event Detection and Assessment of the Sonic Environment by the Harmonica Index, Appl. Sci.11(17), 8031 (2021).

[2] Diez, I., Saratxaga, I., Salegi, U., Navas, E., Hernaez, I.: NoisenSECB: An Urban Sound Event Database to Develop Neural Classification Systems for Noise-Monitoring Applications, Applied Sciences, 13(16), 9358 (2023).

[3] Çakır, E., Parascandolo, G., Heittola, T., Huttunen H., Virtanen, T.: Convolutional Recur-rent Neural Networks for Polyphonic Sound Event Classification, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(6), 1291-1303 (2017).

[4] Sabour, S., Frosst, H., Hinton, G. E.: Dynamic routing between capsules, In NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Sys-tems, pp. 3859 – 3869, Long Beach California, USA (2017).

[5] Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: Pre-Training of Deep Bidirec-tional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-guage Technologies, pp. 4171-4186, Minneapolis, Minnesota (2019).

[6] Ye, Z., Wang X., Liu, H., Qian, Y., Tao, R., Yan, L., Ouchi, K.: Sound Event Classifica-tion Transformer: An Event-baSEC End-to-End Model for Sound Event Classification, arXiv:2110.02011, https://arxiv.org/abs/2110.02011, last accesSEC 2025/1/31.

[7] Khan, MS., Shah, M., Khan, A., Aldweesh, A., Ali, M., Eldin, ET., Ishaq, W., Hussain, L.: Improved Multi-Model Classification Technique for Sound Event Classification in Ur-ban Environments, Applied Sciences, 12(19), 9907 (2022).

[8] Filippov, SN., Heinosaari, T., Leppäjärvi, L.: A necessary condition for incompatibility of observables in general probabilistic theories, Phys. Rev. A 95, 032127 (2017).