International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Neural Network Architectures for Extracting Meaningful Representations from Audio Data

[1]Jeevitha SL, [2]Khushi Jain, [3]Kushi Prasanna, [4]M Shubha
[1]-[5] Department of Computer Science and Engineering in Artificial Intelligence and Machine Learning, Vidyavardhaka College of Engineering, Mysuru, India
[1]jeevithasl93@gmail.com,[2]khushijainmootha@gmail.com, [3]Khushiprasanna05@gmail.com, [4]shubhamanjunath7873@gmail.com

Abstract*— Audio data carries rich information in the form of speech, music, and environmental sounds, but its raw waveform is often complex and high-dimensional, making direct analysis difficult. Neural network architectures have emerged as powerful tools for extracting meaningful representations from audio signals, enabling efficient analysis and interpretation. Recurrent Neural Networks (RNNs), and Transformer-based models— for learning robust and discriminative features from audio. By automatically capturing temporal, spectral, and contextual patterns, these architectures significantly improve performance in tasks such as speech recognition, speaker identification, music classification, and environmental sound detection. The findings highlight the potential of neural networks to replace traditional handcrafted features, thereby advancing the development of scalable, accurate, and realtime audio processing applications. The rapid growth of audio data across domains such as speech, music, healthcare, and environmental monitoring has created a strong need for effective methods to extract meaningful representations from complex audio signals. Traditional approaches rely on handcrafted features like MFCCs and spectrogram descriptors, which often fail to capture the full temporal and spectral dynamics present in raw audio.

Keywords: Neural Networks, Deep Learning, Audio Representation Learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformer Models, Feature Extraction, Speech Recognition, Speaker Identification, Sound Event Detection, Spectrogram Analysis, Audio Signal Processing.

I. INTRODUCTION

Audio signals are inherently complex, high-dimensional, and temporally dynamic, making the extraction of meaningful representations a crucial step for tasks such as speech recognition, music information retrieval, speaker identification, and environmental sound classification. Traditional approaches relied heavily on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram statistics, or filter banks. While effective to some extent, these features often lacked robustness and generalization across diverse audio domains. In recent years, neural network architectures have revolutionized audio representation learning by automatically discovering hierarchical features directly from raw waveforms or time-frequency representations. Convolutional Neural Networks (CNNs) capture local spectral-temporal patterns, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) units model long-range temporal dependencies, while attention-based architectures such as Transformers provide the ability to learn global contextual relationships. Moreover, self-supervised and contrastive learning paradigms have further advanced audio representation learning by leveraging large amounts of unlabeled audio data to learn transferable embeddings. The study of neural architectures for audio data is thus central to modern audio processing, bridging low-level acoustic features with high-level semantic representations. A literature review in this area highlights the progression from handcrafted descriptors to deep, end-to-end models and explores how different architectural choices impact the quality, interpretability, and applicability of learned audio representations across domains. Beyond their success in task-specific applications, neural network-based audio representations have become foundational for a wide range of downstream tasks, much like word embeddings in natural language processing. Pretrained audio models, such as wav2vec, YAMNet, and Audio Spectrogram Transformers, demonstrate how large-scale representation learning can capture universal acoustic patterns transferable across domains. These representations reduce the reliance on large labeled datasets, improve generalization to unseen conditions, and support cross-modal tasks like audio-visual learning. As a result, the focus of current research is not only on improving architectural efficiency and performance but also on enhancing interpretability, robustness to noise, and adaptability to resource-constrained environments.

II. RESEARCH PAPERS

[1] The paper *"Audio Spectrogram Representations for Convolutional Neural Networks"* explores how different audio representations, particularly spectrograms, can be used with CNNs for generative tasks like style transfer. The study compares raw audio, MFCCs, and spectrograms, highlighting spectrograms as a balanced choice between information retention and dimensionality reduction. A key strength is the demonstration that style transfer can work even with random weights, though networks trained on audio data yield better results. However, treating spectrograms as images is limited since pitch shifts and overlapping audio events are not well captured by 2D convolution. The proposed methodology experiments with both image-based and channel-stacked spectrogram inputs, showing differences in synthesis quality.

The main drawback is the need for deep networks with large channel counts, making computation heavy. The novelty lies in revealing that meaningful style-content integration in audio is possible without pretrained weights, suggesting a new direction for audio generative modeling.

[2] The paper *"Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals"* introduces AudioMNIST, a dataset of 30,000 spoken digit recordings, and applies deep neural networks to classify digits and speaker gender. Two models are compared: AlexNet trained on spectrograms and AudioNet trained on raw waveforms. To interpret decisions, the authors use Layer-wise Relevance Propagation (LRP), which reveals which input features most influence classification. Strengths of the work include introducing a benchmark dataset, achieving high classification accuracy (≈96% on spectrograms), and demonstrating that LRP can identify meaningful audio features like pitch and harmonics for gender recognition. Weaknesses include limited interpretability for higher-order features (e.g., phonemes) and dataset imbalance (more male than female speakers). The proposed methodology shows that manipulating LRP-identified features leads to large performance drops, proving the models' reliance on them. The novelty lies in extending interpretability research from vision to audio, providing both a dataset and a framework for understanding deep audio classifiers beyond accuracy.

[3] The paper *"Fundamental Technologies in Modern Speech Recognition"* reviews the shift from GMM-HMM to DNN-HMM models in acoustic modeling. It presents methods such as RBM-based pretraining, Deep Belief Networks, and discriminative fine-tuning that achieved significant accuracy improvements on benchmarks like TIMIT, Switchboard, and Google Voice Input. Strengths include the ability to model nonlinear manifolds, use larger acoustic contexts, and deliver major error rate reductions. Weaknesses include high computational cost, overfitting risks, and reliance on large labeled datasets. The methodology integrates generative pretraining with discriminative fine-tuning and sequence training methods like MMI. The novelty lies in proving that deep pretraining followed by fine-tuning allows DNN-HMM hybrids to surpass state-of-the-art GMM-HMM systems, marking a turning point in speech recognition.

[4] The paper *"auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks"* presents a Python toolkit for unsupervised representation learning from acoustic data. It employs sequence-to-sequence autoencoders with RNNs (LSTM/GRU) to learn fixed-length features from variable-length spectrograms. Strengths include its unsupervised nature, ability to capture temporal dynamics, and modular design with both API and CLI support. Weaknesses are reliance on spectrogram preprocessing and sometimes lower performance compared to large-scale pre-trained models like SoundNet. The methodology involves extracting spectrograms, training autoencoders, and exporting learned features for classifiers like MLP or LibLINEAR. The novelty lies in providing the first open-source toolkit for deep unsupervised audio representation learning, achieving competitive results across multiple benchmarks without external pretraining.

[5] The paper *"Monaural Audio Source Separation Using Deep Convolutional Neural Networks"* proposes a low-latency framework for separating vocals, drums, bass, and other instruments from music mixtures. Strengths include reduced parameters, faster processing (≈4x faster than MLPs), and competitive results in MIREX and SiSEC evaluations. Weaknesses involve difficulty handling the "other instruments" class due to its variability and reliance on spectrogram-based inputs. The methodology introduces time-frequency masking and optimized loss functions to improve separation quality. The novelty lies in demonstrating that CNNs tailored to audio spectrograms achieve real-time source separation with efficiency comparable to state-of-the-art systems.

[6] The paper *"Improved Music Feature Learning with Deep Neural Networks"* uses three main technologies: Rectified Linear Units (ReLUs), Dropout, and Hessian-Free (HF) optimization. The methodology directly trains deep networks with SGD (for ReLUs) and HF (for sigmoids), extracting hidden layer activations as features and employing Random Forest classifiers for genre prediction. Drawbacks include overfitting in large sigmoid architectures and inability to jointly use dropout regularization with HF optimization. Novelty lies in showing competitive results without pretraining, leveraging ReLUs, Dropout, and HF on randomly initialized weights, and validating feature robustness across datasets.

[7] The paper *"Exploring Convolutional, Recurrent and Hybrid Deep Neural Networks for Speech and Music Detection in a Large Audio Dataset"* uses CNN, LSTM, hybrid CNN-LSTM, and fully connected networks with Keras-TensorFlow. Strengths are high accuracy (~85%) and strong time-frequency modeling through hybrid architectures. Weaknesses are reliance on weakly labeled AudioSet data and errors from distractor events like whispering or solo flute. The methodology involves training separate vs. joint neural networks on mel-spectrogram features for speech and music detection. Drawbacks include misclassifications, noisy labels, and heavy computational requirements. Novelty lies in the hybrid CNN-LSTM design, focus on speech/music classes, and detailed distractor analysis.

[8] The paper *"Deep Neural Networks-based Classification Methodologies of Speech, Audio and Music, and its Integration for Audio Metadata Tagging"* presents an automatic system for generating audio metadata. The proposed methodology involves segmenting audio and applying ASR, AEC, ASC, and music detection modules to automatically produce metadata tags. Drawbacks include high computational complexity and limited robustness under overlapping sounds. The novelty is its integrated framework combining speech, music, and scene understanding for automatic video storytelling and metadata tagging.

[9] The paper *"Neural Network Architecture for Extracting Meaningful Representations from Raw Audio Data"* explores deep learning methods for audio analysis. Strengths include the ability to automatically learn robust features from raw audio and achieve high accuracy in distinguishing speech, music, and noise. Weaknesses are high computational cost and performance drops in noisy or overlapping environments. The methodology converts raw

audio to spectrograms, processes them through CNN and LSTM layers, and classifies the outputs into target categories. Drawbacks include dependency on large labeled datasets and limited generalization to unseen conditions. Novelty lies in combining CNN and LSTM to jointly capture time–frequency and temporal dependencies, enabling richer and more meaningful audio representations.

[10] The paper *"Improved Music Feature Learning with Deep Neural Networks"* by Siddharth Sigtia and Simon Dixon explores better feature learning approaches for Music Information Retrieval. The methodology involves training ReLU networks with SGD and Dropout, as well as sigmoid networks with HF optimization, and comparing their learned audio features on benchmark datasets. Drawbacks include dataset limitations (e.g., GTZAN flaws) and heavy computational requirements. The novelty is combining ReLU, Dropout, and HF optimization to significantly improve training efficiency and audio feature quality over traditional methods.

[11] The paper *"Graph-Based Audio Classification Using Pre-Trained Models and Graph Neural Networks"* presents a method to classify audio by representing it as graphs and applying GCN, GraphSAGE, and GAT with features from VGGish, YAMNet, and PANNs. The approach effectively captures structural dependencies, with PANNs and GAT giving the best accuracy. Strengths include improved classification performance and applicability to ecoacoustics, while weaknesses are high computational cost and graph construction complexity. The methodology extracts embeddings from pretrained models, builds graphs using k-NN, and classifies nodes with GNNs. Drawbacks are scalability and long training times. The novelty lies in combining pre-trained models with GNNs to outperform CNN-based methods in audio classification.

[12] The paper *"Deep Learning for Audio Signal Processing"* reviews state-of-the-art deep learning methods for speech, music, and environmental sound analysis. It discusses techniques like CNNs, RNNs (LSTM, GRU), autoencoders, GANs, and spectrogram- or waveform-based models. Strengths include the ability to learn powerful hierarchical features and improve across recognition and synthesis tasks, while weaknesses involve high computational cost, large data requirements, and difficulties in interpretability. The methodology emphasizes transfer learning, data augmentation, and end-to-end deep architectures for recognition and generation. Drawbacks lie in scalability, generalization to unseen domains, and evaluation challenges. The novelty is in unifying speech, music, and environmental sound processing under deep learning, highlighting cross-domain similarities and opportunities for shared advances.

[13] The paper *"A Review of Deep Learning Techniques for Speech Processing"* surveys how deep learning has revolutionized speech tasks such as recognition, synthesis, speaker identification, and translation. The methodology is a structured review comparing architectures, datasets, and tasks while highlighting transfer learning and multimodal approaches. Drawbacks include scalability challenges and lack of parameter-efficient, explainable models. The novelty lies in unifying diverse speech-processing tasks under deep learning, tracing their evolution, and mapping future research directions.

[14] The paper *"Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions"* provides a comprehensive survey of deep learning techniques and applications. It reviews supervised, semi-supervised, unsupervised, and reinforcement learning, with a strong focus on CNN architectures like AlexNet, VGG, ResNet, and HRNet. Strengths include automation of feature extraction and applicability across domains like NLP, computer vision, and healthcare. Weaknesses include overfitting, vanishing gradients, and data imbalance. The proposed methodology provides an integrated review of architectures, challenges, computational tools (CPU, GPU, FPGA), and applications to guide researchers. Drawbacks include lack of experimental validation and limited coverage of newer models. The novelty lies in offering a holistic survey combining concepts, architectures, and applications in a single work.

[15] This paper by Sören Becker et al. explores interpretability in deep neural networks for audio classification by introducing the Audio MNIST dataset and training two models—AlexNet on spectrograms and AudioNet on raw waveforms—both achieving high classification accuracy, particularly for gender and spoken digit recognition. Using layer-wise relevance propagation (LRP), the study reveals which input features are crucial for network decisions. It finds that gender classification relies heavily on lower frequency components and that models using raw waveforms depend on a small fraction of highly relevant samples. Results highlight the value of interpretability methods for understanding and improving model architectures in audio tasks and suggest future work applying these techniques to more complex datasets and comparing deep learning strategies to traditional audio features.

[16] The paper *"auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks"* by Michael Freitag et al. introduces an open-source Python toolkit for unsupervised representation learning from audio using recurrent sequence-to-sequence autoencoders. Fully unsupervised training eliminates the dependency on labeled datasets. The primary weakness is that state-of-the-art models pretrained on massive external corpora (such as SoundNet and L3) can outperform auDeep for some datasets. Some limits in feature-level fusion and configuration details may restrict reproducibility and generalization to non-audio sequential data when compared to broader multimodal approaches.

[17] This study by Muhammad Huzaifah investigates the impact of different time-frequency representations on environmental sound classification performance using CNNs. It compares STFT (linear and Mel scales), CQT, CWT, and MFCCs to evaluate their effectiveness as input features on ESC-50 and UrbanSound8K datasets. The paper emphasizes the importance of time-frequency representation choice, window size, and CNN architecture in improving classification accuracy.

[18] This article discusses *TA-AVN*, a system that predicts people's emotions by analyzing faces and voices. It mimics human multisensory integration and uses simple neural networks to quickly process asynchronous video and audio. TA-AVN handles challenges in emotion recognition, such as

differing video/audio speeds and limited labeled training examples. A random sampling strategy enhances training robustness. Testing across multiple datasets shows the system outperforms others in accuracy without requiring heavy computational resources.

[19] The research examines music genre classification using a 1D CNN on a dataset of 1,000 Nigerian traditional songs spanning seven genres. It details feature extraction, preprocessing, and CNN classifier training. The system achieved an overall accuracy of 92.5%, precision of 92.7%, recall of 92.5%, and F1 score of 92.5%. The study emphasizes the use of a culturally unique dataset and demonstrates the feasibility of 1D CNNs for real-time automated music genre classification.

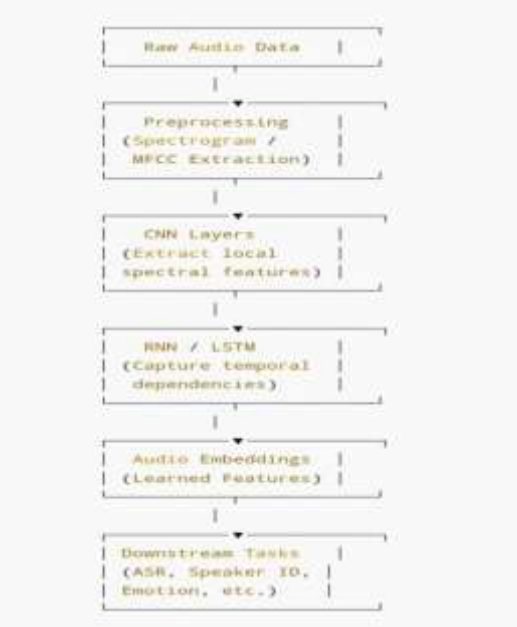
[20] Nilesh M. Patil and Milind U. Nemade present a system that uses machine learning and neural networks to organize and retrieve audio efficiently. Their approach addresses the challenge of large online audio datasets. They developed a Fuzzy Probabilistic Neural Network (FPNN), which combines fuzzy logic with probabilistic neural networks. This system outperforms traditional classifiers like SVM, k-NN, and standard PNNs.

III. RESEARCH GAPS

Neural network architectures provide powerful strengths for extracting meaningful representations from audio data. They can automatically learn features from raw waveforms or spectrograms, reducing the need for manual feature engineering. These models achieve high accuracy in tasks like speech recognition, music classification, and sound event detection, while also showing robustness against noise and variability. Their scalability makes them suitable for handling large datasets, and their versatility allows applications in multiple domains such as healthcare, entertainment, and environmental monitoring. Moreover, they can capture both low-level spectral details and high-level contextual information through hierarchical feature learning. On the other hand, neural networks also present certain weaknesses. They require high computational resources such as GPUs and large memory, and they depend heavily on large labeled datasets for effective training. Their black-box nature makes the learned features difficult to interpret, and they are prone to overfitting when trained on limited data. Additionally, complex architectures like Transformers may cause latency in real-time systems and demand significant energy consumption, making them challenging to deploy in resource-constrained environments.

IV. PROPOSED METHODOLOGY

The primary objective of this study is to review and compare existing neural network architectures for audio feature learning, including Convolutional Neural Networks (CNNs), Convolutional Recurrent Neural Networks (CRNNs), Transformers, and contrastive learning models, to understand their strengths, limitations, and applicability in learning meaningful audio representations across various domains. The project aims to implement a robust CRNN model capable of learning temporal-spectral patterns from spectrograms and to explore Transformer-based architectures such as Audio Spectrogram Transformer (AST) and Conformer for attention-based feature extraction, enabling better modeling of global contextual relationships in audio data. The models will be evaluated on multiple benchmark datasets including ESC-50, UrbanSound8K, and LibriSpeech to assess their performance in different audio classification and recognition tasks. Additionally, the learned representations will be visualized and analyzed using dimensionality reduction techniques such as t-SNE to interpret the model's internal feature spaces. The study will also benchmark the generalization of learned embeddings for downstream tasks like classification, retrieval, and anomaly detection. The dataset used in this project is AudioSet, which contains approximately 2 million human-labeled 10-second audio clips collected from YouTube, covering 527 sound classes such as speech, music, environmental sounds, and instruments, and is widely recognized for pretraining and benchmarking audio representation models.



The methodology begins with data collection and preprocessing, where raw audio is gathered, cleaned, and converted into spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) for easier analysis. After preprocessing, different neural network architectures are applied—Convolutional Neural Networks (CNNs) to capture local time-frequency features, Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to model temporal patterns, and Transformers to extract global contextual information. These models are trained on labeled datasets to learn meaningful feature embeddings, and their performance is evaluated using metrics such as accuracy, precision, recall, and F1-score across tasks like speech recognition and sound classification. The objective is to compare the strengths of each architecture and demonstrate how deep learning improves audio representation learning over traditional handcrafted methods. The proposed methodology focuses on extracting meaningful representations from raw audio data using a hybrid neural network architecture. First, raw audio waveforms are transformed into mel-spectrograms to capture both frequency content and temporal variations. CNNs are then applied to these spectrograms to automatically learn local patterns such as pitch, harmonics, and timbre, while pooling layers reduce

dimensionality and emphasize robust features. To model sequential dependencies in audio, LSTM layers are integrated after the CNN stage, enabling the system to capture rhythm, phoneme transitions, and long-term temporal context.

V. CONCLUSION

Neural network architectures have proven to be highly effective in extracting meaningful representations from complex audio data. Unlike traditional approaches that rely on handcrafted features, deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers can automatically learn hierarchical patterns, capturing both spectral and temporal information with greater accuracy. This capability significantly enhances performance in tasks such as speech recognition, speaker identification, music classification, and environmental sound detection. The proposed framework includes a comparative study implementing both classical transformer-based CNN/CRNN architectures and advanced models for audio representation learning. It also introduces a hybrid CRNN combined with self-supervised contrastive learning to enable the extraction of transferable and generalizable features. Furthermore, the study explores model performance across multiple datasets and tasks, including environmental sounds, speech, and music, while visualizing learned audio embeddings to analyze model behavior and class separability. Finally, the work proposes a modular architecture for audio understanding that is adaptable to low-resource edge devices, ensuring scalability, efficiency, and practical deployment in real-world applications.

REFERENCES

[1] Zhao, H., Chen, Y., Wang, R., & Malik, H. (2016). Audio Splicing Detection and Localization Using Environmental Signature. *Multimedia Tools and Applications (2023).*

[2] Leonzio, D. U., Cuccovillo, L., Bestagini, P., Marcon, M., Aichroth, P., & Tubaro, S**.** Audio Splicing Detection and Localization Based on Acquisition Device Traces. *IEEE Transactions on Information Forensics and Security, 18,* 4157–4172.

[3] Zhang, Y., Dai, S., Song, W., Zhang, L., & Li, D. (2019). Exposing Speech Resampling Manipulation by Local Texture Analysis on Spectrogram Images. *Electronics, 9(1),* 23.

[4] Zhou, X., Zhang, Y., Wang, Y., Tian, J., & Xu, S. (2024). Pyramid Feature Attention Network for Speech Resampling Detection. *Applied Sciences, 14(11),* 4803.

[5] Zeng, C., Yang, Y., Wang, Z., Kong, S., Feng, S., & Zhao, N. (2022). Audio Tampering Forensics Based on Representation Learning of ENF Phase Sequence. *International Journal of Digital Crime and Forensics, 14(1),* 1–19.

[6] Wang, Z., Yang, Y., Zeng, C., Kong, S., & Feng, S. (2022). Shallow and Deep Feature Fusion for Digital Audio Tampering Detection. *EURASIP Journal on Advances in Signal Processing, 2022(1),* 69.

[7] Zeng, C., Kong, S., Wang, Z., Li, K., & Zhao, Y. (2023). Digital Audio Tampering Detection Based on Deep Temporal-Spatial Features of Electrical Network Frequency. *Information, 14(5),* 253.

[8] Zeng, C., Kong, S., Wang, Z., Wan, X., & Chen, Y. (2022). Digital Audio Tampering Detection Based on ENF Spatiotemporal Features Representation Learning. *CoRR, abs/2208.11920.*

[9] Ustübioğlu, B., Tahaoğlu, G., Ulutaş, G., Üstübioğlu, A., & Kılıç, M. (2024). Audio Forgery Detection and Localization with Super-Resolution Spectrogram and Keypoint-Based Clustering Approach. *Journal of Supercomputing, 80(1),* 486–518.

[10] Liu, T., & Yan, D. **(2021).** Identification of Fake Stereo Audio. *Information, 12(7),* 263.

[11] Xue, J., Fan, C., Lv, Z., Tao, J., Yi, J., Zheng, C., Wen, Z., Yuan, M., & Shao, S. (2022, October). Audio Deepfake Detection Based on a Combination of F₀ Information and Real Plus Imaginary Spectrogram Features. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Lisbon, Portugal.*

[12] Pianese, A., Cozzolino, D., Poggi, G., & Verdoliva, L. (2022, December). Deepfake Audio Detection by Speaker Verification. In *IEEE International Workshop on Information Forensics and Security (WIFS).*

[13] Kanwal, T., Mahum, R., AlSalman, A. M., Sharaf, M., & Hassan, H. (2024). Fake Speech Detection Using VGGish with Attention Block. *EURASIP Journal on Audio, Speech, and Music Processing, 2024(1),* 35.

[14] Pham, L., Lam, P., Nguyen, T., Nguyễn, H., & Schindler, A. (2024). Deepfake Audio Detection Using Spectrogram-Based Features and an Ensemble of Deep Learning Models. In *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2).*

[15] Zeng, C., Kong, S., Wang, Z., Wan, X., & Chen, Y. (2022). Digital Audio Tampering Detection Based on ENF Spatiotemporal Features Representation Learning. *CoRR, abs/2208.11920.*

[16] Huzaifah, M. (2017). Comparison of Time-Frequency Representations for Environmental Sound Classification Using Convolutional Neural Networks. *IEEE Signal Processing Letters.*

[17] Salamon, J., & Bello, J. P. **(2017).** Environmental Sound Classification Using Convolutional Neural Networks and Data Augmentation. *IEEE Signal Processing Letters.*

[18] Salamon, J., & Bello, J. P. **(2017).** Environmental Sound Classification Using Convolutional Neural Networks and Data Augmentation. *IEEE Signal Processing Letters.*

[19] Dal Rí, F. A., Ciardi, F. C., & Conci, N. (2023**).** Speech Emotion Recognition and Deep Learning: An Extensive Validation Using CNNs. *IEEE Access.*

[20] Patil, N. M., & Nemade, M. U. (2019). Content-Based Audio Classification and Retrieval Using Segmentation, Feature Extraction, and Neural Network Approach. In *Advances in Intelligent Systems and Computing (pp. 263–276).* Springer.