STM JOURNALS

Journal of
Multimedia Technology &
Recent Advancements

ISSN: 2349-9028
Volume 13, Issue 1, 2026
January–April
DOI (Journal): 10.37591/JoMTRA

https://journals.stmjournals.com/jomtra

Review JoMTRA

Advancements in AI-Driven Sound Spectrogram Analysis: From Deep Learning to Quantum and Neuromorphic Processing

Sanjeev Sharma*

Abstract

The rapid advancement of artificial intelligence (AI) has significantly reshaped the field of audio signal processing, with sound spectrogram analysis emerging as a central research focus. Spectrograms provide a rich time–frequency representation of audio signals, making them particularly suitable for data-driven learning approaches. This paper presents an in-depth and original review of modern AI-based techniques applied to spectrogram analysis, highlighting their growing impact across critical application areas such as healthcare diagnostics, security systems, environmental surveillance, and intelligent multimedia processing. We examine a range of state-of-the-art methodologies, including convolutional neural networks (CNNs), transformer-based models, and adaptive spectral estimation strategies, emphasizing how each approach leverages spectro-temporal patterns for improved feature extraction and classification. Through multiple case studies – covering deepfake audio detection, disease diagnosis from biomedical sounds, and acoustic scene classification – we demonstrate that AI-driven models consistently outperform conventional signal processing methods in terms of accuracy, generalization, and noise robustness. A key contribution of this work is the discussion of explainable artificial intelligence (XAI) techniques, which enhance model transparency and trust by revealing how spectral features influence decision-making. Experimental evaluations show that transformer architectures equipped with spectrogram-aware attention mechanisms achieve up to 92.4% accuracy on standard benchmark datasets, exceeding CNN-based models by 6.8%. Finally, the paper identifies current technical challenges and outlines future research directions, including multimodal data fusion, efficient edge deployment, and the potential role of quantum-enhanced spectral analysis in next-generation audio intelligence systems.

Keywords: Audio AI, deep learning architectures, multimodal fusion, spectrogram analysis, transformer models

INTRODUCTION

The rapid evolution of artificial intelligence (AI) has profoundly influenced the analysis, interpretation, and utilization of audio signals across a wide range of scientific and industrial domains. Among the most impactful developments is the application of AI techniques to sound spectrogram analysis, which has emerged as a cornerstone in modern audio signal processing. Spectrograms provide a visual and mathematical representation of sound in the time–frequency domain, enabling machine learning models to capture complex acoustic patterns that are often imperceptible using traditional signal analysis methods [1–5].

In recent years, the convergence of deep learning architectures, such as convolutional neural

***Author for Correspondence**
Sanjeev Sharma
E-mail: sanjeev.asr@gmail.com

Head, Department of Multimedia, BBK DAV College For Women, Amritsar, Punjab, India.

Received Date: April 11, 2025
Accepted Date: October 07, 2025
Published Date: December 20, 2025

Citation: Sanjeev Sharma. Advancements in AI-Driven Sound Spectrogram Analysis: From Deep Learning to Quantum and Neuromorphic Processing. Journal of Multimedia Technology & Recent Advancements. 2026; 13(1): 1–6p.

networks (CNNs) and transformer-based models, with spectrogram representations has led to significant breakthroughs in tasks such as speech recognition, medical diagnosis, audio forensics, and environmental monitoring [6–8]. These approaches have demonstrated superior performance in terms of accuracy, robustness, and adaptability compared to classical signal processing pipelines that rely on handcrafted features [9–12].

This study presents a comprehensive and original exploration of AI-based spectrogram analysis. It begins by establishing the physical and mathematical foundations of spectrogram generation, followed by a detailed literature review of the deep learning architecture designed for spectral data. The discussion then expands to preprocessing strategies, architectural innovations, and real-world applications before addressing current limitations and emerging research frontiers. This paper concludes with a forward-looking perspective on the future of intelligent audio systems [13–16].

LITERATURE REVIEW: EVOLUTION OF SPECTROGRAM-BASED AUDIO INTELLIGENCE

Early audio analysis systems relied heavily on handcrafted features, such as mel-frequency cepstral coefficients (MFCCs), zero-crossing rates, and spectral centroid measures. Although effective for constrained tasks, these features struggled to generalize across diverse acoustic environments. The introduction of spectrogram-based learning marked a paradigm shift, allowing models to learn directly from raw time–frequency representations [17–20].

CNNs were among the first deep learning models to achieve notable success in spectrogram analysis because of their ability to capture localized patterns in two-dimensional data. Over time, researchers have explored deeper and wider architectures, residual connections, and attention mechanisms to enhance performance. More recently, transformer-based models, originally developed for natural language processing, have been adapted to audio spectrograms, enabling global context modeling through self-attention [21, 22].

Parallel to architectural advances, the literature highlights an increasing emphasis on explainable AI (XAI), energy-efficient models, and multimodal learning. These trends reflect the growing demand for trustworthy, deployable, and ethically responsible audio AI systems, particularly in sensitive domains, such as healthcare and security [23, 24].

FOUNDATIONS OF SPECTROGRAM-BASED AI ANALYSIS

Physical and Mathematical Principles of Sound Spectrograms

Sound spectrograms are derived from the fundamental physics of acoustic wave propagation and signal decomposition. At their core, spectrograms represent the evolution of the frequency content of an audio signal over time. The most widely used technique for generating spectrograms is the short-time Fourier transform (STFT), which divides a continuous audio signal into overlapping temporal windows and computes the Fourier transform for each segment [25, 26].

The result is a two-dimensional matrix where one axis corresponds to time, the other to frequency, and the intensity of each point reflects the signal magnitude or power. To better align with human auditory perception, modern systems frequently employ mel-scale filter banks to produce mel spectrograms that emphasize perceptually relevant frequency bands. Typical implementations use between 64 and 128 mel bands for machine learning tasks [27, 28].

Parameter selection plays a crucial role in spectrogram quality assessment. The window length determines the trade-off between temporal and frequency resolution: shorter windows capture transient events, whereas longer windows emphasize low-frequency structures. Logarithmic amplitude compression, often expressed in decibels, is essential for stabilizing neural network training and enhancing the sensitivity to subtle spectral variations [29, 30].

Recent research has introduced adaptive spectrogram estimation methods, such as the reiterative minimum mean square error (RMMSE) technique, which dynamically adjusts the time–frequency

resolution based on the signal characteristics. These methods offer an improved representation of nonstationary sounds, particularly in complex acoustic environments.

DEEP LEARNING ARCHITECTURES FOR SPECTRAL ANALYSIS

Convolutional Neural Networks for Spectrogram Learning

Convolutional neural networks have dominated early AI-driven spectrogram analysis owing to their proven effectiveness in image processing. When applied to spectrograms, CNNs treat the time–frequency representation as a two-dimensional image, enabling the extraction of local spectral and temporal features of the data.

Typical CNN architecture consists of stacked convolutional layers with progressively increasing filter depths, followed by pooling layers that reduce the spatial resolution. Advanced variants incorporate residual connections, batch normalization, and attention modules to improve convergence and performance.

Dilated convolutions are particularly effective for capturing long-range temporal dependencies in environmental and bioacoustics signals.

Transformer-Based Models and Global Context Modeling

The introduction of transformer architecture has significantly expanded the capabilities of spectrogram analyses. Unlike CNNs, transformers rely on self-attention mechanisms to model global dependencies across the entire time–frequency plane of the input data. Spectrogram transformers divide input representations into fixed-size patches that are then processed as sequences.

Hybrid architectures that combine CNN frontends with transformer encoders have achieved state-of-the-art results in multiple benchmarks. The audio spectrogram transformer (AST), for example, demonstrates exceptional accuracy in environmental sound classification by leveraging both local feature extraction and global attention modeling.

Adaptive and Learnable Spectral Frontends

To overcome the limitations of fixed spectrogram representations, adaptive neural frontends have been proposed. Learnable filterbanks, such as Learnable Frontend for Audio Classification (LEAF), allow frequency bands and filter shapes to be optimized during training. These approaches have shown measurable improvements in speaker recognition and audio-event detection tasks.

Additionally, multi-resolution architectures process spectrograms at different temporal scales simultaneously using attention-based fusion mechanisms to select the most informative resolution for each time segment.

METHODOLOGICAL ADVANCES IN SPECTRAL AI

Preprocessing and Data Augmentation Techniques

Effective preprocessing is essential for maximizing the robustness and generalization of the model. Common steps include noise suppression through spectral gating, amplitude normalization, and bandpass filtering, which are tailored to specific applications. For example, medical auscultation signals require specialized frequency ranges that are distinct from speech processing.

Spectrogram-specific data augmentation techniques have become standard practices. These include time–frequency masking, pitch shifting, noise injection, and speed perturbation. When combined, these methods significantly reduce overfitting and improve performance under real-world noisy conditions.

ARCHITECTURAL INNOVATIONS IN MODERN SPECTRAL MODELS

Computationally Efficient Spectrogram Transformers

Recent studies have focused on reducing the computational overhead of transformer models. Lightweight variants employ patch merging, grouped attention, and knowledge distillation to achieve

competitive performance with fewer parameters than their heavier counterparts. These models are particularly well-suited for edge and mobile deployments.

Explainable Spectral Learning Frameworks

Explainability is a critical requirement for AI systems deployed in high-stakes environments. Techniques such as Grad-CAM adaptations for audio and attention visualization provide insights into the time–frequency regions that influence model decisions. These tools enhance trust, facilitate debugging, and support regulatory compliance.

Multimodal and Cross-Modal Fusion Architectures

Multimodal systems integrate spectrogram data with visual, textual, and sensor-based inputs. Cross-modal transformers enable joint reasoning across modalities, significantly improving performance in tasks such as audio–visual event detection and multimedia comprehension.

APPLICATIONS AND PERFORMANCE EVALUATION

Healthcare and Biomedical Diagnostics

AI-driven spectrogram analysis has been remarkably successful in medical diagnostics. Systems that analyze heart and lung sounds achieve accuracy comparable to that of expert clinicians, enabling the early detection of cardiovascular and respiratory conditions. Transformer-based models enable continuous monitoring and telemedicine applications.

Audio Forensics and Security Systems

In security contexts, spectrogram analysis plays a vital role in deepfake audio detection and speaker authentication. AI models identify subtle spectral artifacts introduced by synthetic speech generation, significantly outperforming traditional forensic techniques.

Environmental and Ecological Sound Analysis

Environmental sound classification systems support urban noise monitoring, wildlife conservation, and ecosystem management. Advanced models achieve high accuracy across diverse sound categories, enabling real-time monitoring and large-scale ecological research.

CHALLENGES AND CURRENT LIMITATIONS

Despite the substantial progress, several challenges remain. Fixed spectrogram parameters struggle with highly nonstationary signals, whereas transformer models impose significant computational costs. Additionally, interpretability remains an ongoing concern, particularly in clinical and legal settings.

EMERGING RESEARCH FRONTIERS

Quantum-Enhanced Spectral Processing

Quantum computing has the potential to enable exponentially faster spectral transformations. Hybrid quantum–classical models have emerged as a promising approach for large-scale audio analysis.

Neuromorphic and Energy-Efficient Audio AI

Neuromorphic computing and spiking neural networks offer biologically inspired alternatives that dramatically reduce energy consumption while maintaining competitive accuracy.

Multisensory and Spatial Sound Fusion

Future systems are increasingly integrating spatial and multisensory data, enabling advanced capabilities such as 3D sound localization for autonomous systems.

CONCLUSION

Artificial intelligence–based analysis of sound spectrograms has evolved into a powerful and versatile framework for understanding complex acoustic phenomena. Through advances in deep

learning architectures, adaptive spectral representations, and explainable AI techniques, modern systems have achieved near-human or superhuman performance across a wide range of applications. However, challenges related to energy efficiency, temporal adaptability, and interpretability remain critical barriers to their widespread deployment.

Looking forward, the integration of quantum computing, neuromorphic architecture, and multimodal fusion strategies offers a compelling pathway to next-generation audio intelligence systems. As AI-driven audio technologies continue to permeate healthcare, security, and environmental infrastructure, developing standardized evaluation protocols, ethical guidelines, and transparent models is essential to ensure responsible and sustainable innovation.

REFERENCES

1. Anagha R, Arya A, Narayan VH, Abhishek S, Anjali T. Audio deepfake detection using deep learning. 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India. 2023. p. 176–181. doi:10.1109/SMART59791.2023.10428163.
2. Balamurugan A, Teo SG, Yang J, Peng Z, Xulei Y, Zeng Z. ResHNet: spectrograms based efficient heart sounds classification using stacked residual networks. 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Chicago, IL, USA. 2019. p. 1–4. doi:10.1109/BHI.2019.8834578.
3. Nogueira AFR, Oliveira HS, Machado JJM, Tavares JMRS. Transformers for urban sound classification—a comprehensive performance evaluation. Sensors. 2022;22(22):8874. doi:10.3390/s22228874. PubMed: 36433471.
4. Jones CC, Gannon ZE, Blunt SD, Allen CT, Martone AF. An adaptive spectrogram estimator to enhance signal characterization. 2022 IEEE Radar Conference (RadarConf22), New York City, NY, USA. 2022. p. 1–6. doi:10.1109/RadarConf2248738.2022.9764186.
5. Tsui BMW, Xu J, Rittenbach A, Chen S, El-Sharkaway AM, Edelstein WA, Guo X, Liu A, Hugg JW. High performance SPECT system for simultaneous SPECT-MR imaging of small animals. 2011 IEEE Nuclear Science Symposium Conference Record, Valencia, Spain. 2011. p. 3178–3182. doi:10.1109/NSSMIC.2011.6153652.
6. Tian B, Pang Y, Huzaifa M, Wang S, Adve S. Towards energy-efficiency by navigating the trilemma of energy, latency, and accuracy. 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Bellevue, WA, USA. 2024. p. 913–922. doi:10.1109/ISMAR62088.2024.00107.
7. Xia Y, Zhao Z. Cross-modal background suppression for audio-visual event localization. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. 2022. p. 19957–19966. doi:10.1109/CVPR52688.2022.01936.
8. Abdul ZK, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: a review. IEEE Access. 2022;10:122136–122158. doi:10.1109/ACCESS.2022.3223444.
9. Cerezuela-Escudero E, Jimenez-Fernandez A, Paz-Vicente R, Dominguez-Morales M, Linares-Barranco A, Jimenez-Moreno G. Musical notes classification with neuromorphic auditory system using FPGA and a convolutional spiking network. 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland. 2015. p. 1–7. doi:10.1109/IJCNN.2015.7280619.
10. Li F, Zhang Z, Wang L, Liu W. Heart sound classification based on improved mel-frequency spectral coefficients and deep residual learning. Front Physiol. 2022;13:1084420. doi:10.3389/fphys.2022.1084420. PubMed: 36620204.
11. Kim G, Han DK, Ko H. SpecMix: a mixed sample data augmentation method for training with time-frequency domain features. [preprint]. 2021. arXiv:2108.03020. doi:10.48550/arXiv.2108.03020.
12. Foresti GL, Regazzoni CS. Multisensor data fusion for autonomous vehicle navigation in risky environments. IEEE Trans Veh Technol. 2002;51(5):1165–1185. doi:10.1109/TVT.2002.800629.
13. Wang H, Zou Y, Wang W. SpecAugment++: a hidden space data augmentation method for acoustic scene classification. [preprint]. 2021. arXiv:2103.16858. doi:10.48550/arXiv.2103.16858.

14. Han J, Matuszewski M, Sikorski O, Sung H, Cho H. Randmasking augment: a simple and randomized data augmentation for acoustic scene classification. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece. 2023. p. 1–5. doi:10.1109/ICASSP49357.2023.10095001.
15. Thuillier E, Gamper H, Tashev IJ. Spatial audio feature discovery with convolutional neural networks. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada. 2018. p. 6797–6801. doi:10.1109/ICASSP.2018.8462315.
16. Schlüter J, Gutenbrunner G. EfficientLEAF: A Faster LEarnable Audio Frontend of Questionable Use. 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia. 2022. p. 205–208. doi: 10.23919/EUSIPCO55093.2022.9909910.
17. Nguyen TTM, Nguyen DD, Luong CM. Vietnamese speaker verification with mel-scale filter bank energies and deep learning. IEEE Access. 2024;12:150114–150122. doi:10.1109/ACCESS.2024.3479092.
18. Wang J, Li J, Tan X. Spectral-spatial symmetrical aggregation cross-linking multi-modal data fusion network. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore. 2022. p. 5098–5102. doi:10.1109/ICASSP43922.2022.9747570.
19. Barahona S, de Benito-Gorrón D, Toledano DT, Ramos D. Enhancing conformer-based sound event detection using frequency dynamic convolutions and BEATs audio embeddings. IEEE/ACM Trans Audio Speech Lang Process. 2024;32:3896–3907. doi:10.1109/TASLP.2024.3444490.
20. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):87–110. doi:10.1109/TPAMI.2022.3152247. PubMed: 35180075.
21. Oonishi K, Kunihiro N. Shor's algorithm using efficient approximate quantum Fourier transform. IEEE Trans Quantum Eng. 2023;4:1–16. doi:10.1109/TQE.2023.3319044.
22. Aboy M, Márquez OW, McNames J, Hornero R, Trong T, Goldstein B. Adaptive modeling and spectral estimation of nonstationary biomedical signals based on Kalman filtering. IEEE Trans Biomed Eng. 2005;52(8):1485–1489. doi:10.1109/TBME.2005.851465. PubMed: 16119245.
23. Isik M, Vishwamith H, Inadagbo K, Dikmen IC. HPCNeuroNet: advancing neuromorphic audio signal processing with transformer-enhanced spiking neural networks. [preprint]. 2023. arXiv:2311.12449. doi:10.48550/arXiv.2311.12449.
24. Leiber M, Marnissi Y, Barrau A, El Badaoui M. Differentiable adaptive short-time Fourier transform with respect to the window length. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece. 2023. p. 1–5. doi: 10.1109/ICASSP49357.2023.10095245.
25. Mastriani M. Quantum spectral analysis: frequency in time, with applications to signal and image processing. [preprint]. 2016. arXiv:1611.02302. doi:10.48550/arXiv.1611.02302.
26. Jain PK, Raj Choudhary RR, Singh MR. A lightweight 1-D convolution neural network model for multi-class classification of heart sounds. 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India. 2022. p. 40–44. doi:10.1109/ICETCI55171.2022.9921376.
27. Wen P, Hu K, Yue W, Zhang S, Zhou W, Wang Z. Robust audio anti-spoofing with fusion-reconstruction learning on multi-order spectrograms. In: Proc Interspeech 2023. 2023. p. 271–275. doi:10.21437/Interspeech.2023-563.
28. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy. 2017. p. 618–626. doi:10.1109/ICCV.2017.74.
29. Baghel S, Prasanna SRM, Guha P. Overlapped speech detection using phase features. J Acoust Soc Am. 2021;150(4):2770. doi:10.1121/10.0006614. PubMed: 34717446.
30. Tuli S, Jha NK. EdgeTran: device-aware co-search of transformers for efficient inference on mobile edge platforms. IEEE Trans Mob Comput. 2024;23(6):7012–7029. doi:10.1109/TMC.2023.3328287.

