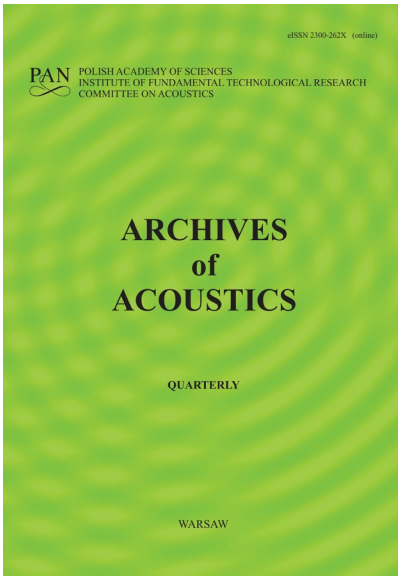
JOURNAL PRE-PROOF

This is an early version of the article, published prior to copyediting, typesetting, and editorial correction. The manuscript has been accepted for publication and is now available online to ensure early dissemination, author visibility, and citation tracking prior to the formal issue publication.

It has not undergone final language verification, formatting, or technical editing by the journal's editorial team. Content is subject to change in the final Version of Record.

To differentiate this version, it is marked as "PRE-PROOF PUBLICATION" and should be cited with the provided DOI. A visible watermark on each page indicates its preliminary status.

The final version will appear in a regular issue of *Archives of Acoustics*, with final metadata, layout, and pagination.



Title: Audio Strips Network (ASNet) and Amalgamation Audio Features (A2F): A Synergistic Approach for Audio Source Separation

Author(s): S.P Sakthidevi, C. Divya

DOI: https://doi.org/10.24423/archacoust.2026.4251

Journal: *Archives of Acoustics*

ISSN: 0137-5075, e-ISSN: 2300-262X

Publication status: In press

Received: 2025-05-29

Revised: 2025-12-10

Accepted: 2026-04-25

Published pre-proof: 2026-04-29

Please cite this article as:

Sakthidevi S.P., Divya C. (2026), Audio Strips Network (ASNet) and Amalgamation Audio Features (A2F): A Synergistic Approach for Audio Source Separation, *Archives of Acoustics*, https://doi.org/10.24423/archacoust.2026.4251

Audio Strips Network (ASNet) and Amalgamation Audio Features (A2F): A Synergistic Approach for Audio Source Separation

S.P Sakthidevi [(1)]*, C. Divya [(2)]

[(1)] https://orcid.org/0009-0004-8500-7611, [(2)] https://orcid.org/0000-0001-7364-0269

Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

*Corresponding Author e-mail: spsakthidevi2000@gmail.com

Abstract

Audio Source Separation refers to the procedure of decomposing a mixed audio signal into its constituent components. This technique enables numerous applications, including creative music production, educational tools, karaoke, transcription, and music analysis. Despite the recent success of deep learning-based source separation techniques, these techniques often do not perform very accurately and do not provide high-quality separation of sources when many contain complex combinations in their mixtures. Source separation techniques generally rely on temporal or spectral features for analysis, which does not fully capture the complex dynamics of audio signals. To address these limitations, proposed the Amalgamation Audio Features (A2F), a hybrid representation combining temporal and spectral features. Then, Proposed the Audio Strips Network (ASNet), a novel framework designed to achieve clean and precise separation of individual audio sources with enhanced performance. ASNet utilized A2F, to separate sources more effectively. The model is trained and evaluated on the MUSDB, DSD100 and MUSDB18-HQ dataset, a benchmark for music source separation, and its standard measures like the Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR) are used to examine performance. ASNet achieves enhanced separation performance with SDR values of drums 12.63, vocal 11.42, bass 12.01 and other 11.14, and SIR values of drums 9.57, vocal 9.61, bass 9.66 and other 9.67. This advancement benefits musicians through high-quality remixing and creativity while aiding researchers in improving Deep Learning and hybrid audio processing models.

Keywords: Feature Extraction; A2F; Source Extraction; ASNet.

1. Introduction

Audio signal processing [Zölzer, Udo., 2022] has expanded greatly in recent years, particularly in the area of source isolation: extracting different sources of a combined auditory signal from one another. Source isolation is crucial in many areas, including music production [Huber., et

al., 2023], spoken word transcription [O'Connell., et al., 2022], audio enhancement [Chuang., et al., 2022] and improvement, hearing aids [Sanders., et al., 2021], and forensic audio analysis [Renukadevi P., et al., 2024]. Even though advancements have been made, creating realistic source separation remains difficult due to how audio signals are often inherently complex and include overlapping frequencies, transient items, harmonics, and environmental noise and reverberation, all of which can impair model performance.

Advances in the field of Deep Learning (DL) [Issa., et al., 2021], have transformed the source isolation landscape by introducing powerful, data-driven approaches capable of learning complex patterns directly from raw or transformed audio inputs. Various techniques have achieved performance in extraction of sources tasks at the cutting edge, including Convolutional Neural Network (CNN) like U-Net [Deng, B., et al., 2024], Conv-TasNet [Défossez, A., et al., 2019], SCNet [Tong, W., et al., 2024], DTTNet [Chen, J., et al., 2024], and DenseNet [Takahashi, N., et al., 2020] have shown great success in modeling spectral and spatial correlations, as well as Recurrent Neural Network (RNN) [Sun, Chao, et al, 2021; Luo, Yi, et al., 2023] like Gated Linear Unit (GLU) [Tong, W., et al., 2024; Défossez, A., et al., 2019] and Bidirectional Long Short Term Memory (Bi-LSTM) [Chen, J., et al., 2024; Deng, B., et al., 2024; Défossez, A., et al., 2019] have demonstrated efficacy in capturing long-range temporal dependencies within audio streams. CNN do a good job of capturing spatial and spectral characteristics, while RNN are capable of capturing the temporal dependency of the audio segments while giving increased weight to the relevant parts of the audio signal. Nevertheless, both CNN and RNN face limitations, particularly when audio signals are highly overlapping or contaminated with noise.

In addition, there are many existing approaches that are computationally intensive, requiring great computational resources for training and inference and limiting their use cases in resource-constrained scenarios. Noise and reverberation further complicate separation, reducing the effectiveness of current approaches. To address these challenges, a novel methodology is proposed that combines the strengths of DL with advanced feature extraction techniques. This methodology introduces hybrid features that integrate temporal and spectral cues, enabling a more accurate representation of the complex dynamics in musical mixtures. Built on this feature set, the proposed Hybrid DL model achieves higher SDR and SIR scores across benchmark datasets, demonstrating superior separation quality and robustness compared to existing models. Its relevance is further reflected in practical applications such as music production, remixing, and research, where high-quality and efficient separation is essential.

2. Background Analysis

This section reviews existing methodologies of DL and Feature Extraction techniques, while addressing the challenges associated with these approaches. The gaps in current methods are identified to establish the foundation for the proposed framework

2.1. Analysis of Feature Extraction

Mel Frequency Cepstral Coefficients (MFCC) [Hamza, Ameer, et al., 2022] mimic human hearing by utilising the Fast Fourier Transform (FFT) [Rezaul, K. Mohammed, et al., 2024], pre-emphasis, framing, hamming window, Mel filterbanks, log-energy, and Discrete Cosine Transform (DCT) to extract speech and features of audio. However, they face challenges with efficiency and noise robustness. While effective, MFCC struggle with real-world noise, edge-device computation, and dynamic speech variations. [Rezaul, K. M., et al., 2024]. Crystalis (128-point FFT) and Spectral Feature Extraction (SFE) (512-point FFT + interpolation) cochlear implant strategies are compared, with SFE improving spectral resolution but lacking validation in noise robustness, efficiency, and generalizability. Future work should explore DL integration and lightweight SFE variants for clinical scalability, while long-term acclimatization and real-world performance remain unassessed [Zhang, Y., et al., 2025]. The 1-Transistor-1-Memristor (1T1R) neuromorphic system mimics biological neurons for audio feature extraction, achieving 85.38–95.91% accuracy on Spiking Heidelberg Digits (SHD) but lacks real-world noise robustness testing. Future work should improve Complementary Metal-Oxide-Semiconductor (CMOS) compatibility, scalability beyond memristors, and generalization to diverse acoustic environments [Wu, X., et al., 2024].

2.2. Performance of Deep Learning Approaches

Denoising Autoencoder with U-Net and Bi-LSTM (DAEUBL) denoise multi-level Random Telegraph Signal (RTS) [Huang, Zhujun, et al., 2024] with mixed noise. Unlike single-noise-focused models, it generalizes across noise types and signal lengths via unified training and adaptive segmentation. However, it lacks real-world benchmark validation and computational efficiency analysis, limiting scalability insights and practical deployment trade-offs for complex RTS denoising [Deng, B., et al., 2024]. Sparse Conformer Network (SCNet) applies Short-Time Fourier Transform (STFT) and Sparse Downsampling (SD) to compress binaural audio spectrograms for source separation. It use Group Normalization (GroupNorm), GLU, and Dual-Path RNN to model sequence dependencies. Research gaps include better high-frequency

modeling, lower complexity, and improved generalization to unseen data [Tong, W., et al., 2024]. In the D2Block (Dilated Dense Block), each feature map is processed with varying dilation factors to ensure better input coverage without blind spots. D3Net (Deep Dual-path Densely Connected Network) extends this idea by stacking D2Blocks with a repeated dilation pattern, allowing flexible feature transformation across resolutions and improving computational efficiency with channel reduction mechanisms. However, certain shortcomings persist, including the potential for aliasing artifacts, increased model complexity due to dense connections, and challenges in maintaining effective receptive field coverage across layers [Takahashi, N., et al., 2020]. Dual-Path TFC-TDF U-Net (DTTNet) uses Time-Distributed Fully-connected Layers (TDF), Time-Frequency Convolutions (TFC), and Improved Dual-Path Module (IDPM) for music source separation. Group Norm and Bi-LSTM enhance sequence modeling across time and frequency. Real-world generalization is tested on patterns like Wah Guitar and Vocal Chops beyond MUSDB18-HQ. Challenges include better fine-tuning for distorted patterns and more robust separation across unseen acoustic environments [Chen, J., et al., 2024]. DL architectures like Demucs and Convolutional Time domain Audio Separation Network (ConvTasNet) employ convolutional encoders/decoders, GLU and Bi-LSTM for entire distinction of sources for music. Improvements such as weight rescaling at initialization, the shift trick for time-equivariance, and sinc-based resampling address performance bottlenecks. Key difficulties involve maintaining phase consistency, managing varying loudness levels across entire tracks, preserving time-shift invariance, and optimizing memory usage when working with large models [Défossez, A., et al., 2019].

3. Methodology

This section reveals the innovative DL framework designed to address key challenges in music source separation, featuring two core components : A2F for effective feature extraction and ASNet for robust separation. Each element in this part presents innovative techniques based on limitations of previous methods. The overall workflow is displayed in Fig. 1 and starts with the audio input and works to the preprocessing steps such as loading, normalizing, applying data augmentations and converting to spectrogram. Once the features have been extracted, they are processed through A2F, are passed into ASNet, and then postprocessed to reconstruct the audios that were separated.

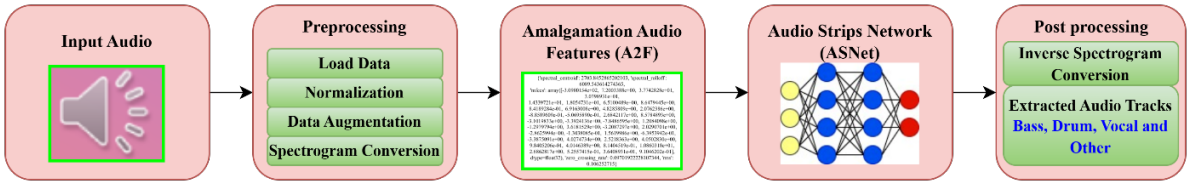


Fig. 1. End-to-End Proposed Framework

3.1. Dataset

To evaluate and train the proposed model for audio source separation, three benchmark datasets were employed : MUSDB, DSD100, and MUSDB18-HQ. The model was trained independently on each dataset to assess its performance and generalization ability across diverse musical compositions, genres, and recording environments. Fig. 2 shows the training and validation accuracy and loss curves of the proposed ASNet model evaluated on the MUSDB18-HQ, DSD100, and MUSDB dataset. It includes Accuracy vs. Epochs and Loss vs. Epochs plots highlighting the convergence behavior, learning progress and generalization capability of the ASNet.

The MUSDB, DSD100, and MUSDB18-HQ [21] datasets consist of 166, 100, and 150 full-length music tracks, respectively. MUSDB and MUSDB18-HQ provide approximately 10 hours of audio each, while DSD100 includes tracks ranging between two minutes and twenty-two seconds and seven minutes and twenty seconds. These datasets cover a wide range of musical genres and include isolated stems for drums, bass, vocals, and other instruments. All audio stereophonic signals are captured at 44,100 Hz. Every dataset is split into 80% for training and 20% for testing, using A2F audio features as input and the corresponding source categories (separated signals) as labels. To maintain reproducibility and consistency a random state of 42 is used for data splitting. The method is trained on a batch size of 32 over 50 epochs to preserve available memory and limit computational burden.

In the MUSDB dataset, the training accuracy steadily rises, reaching approximately 0.95, while validation accuracy also improves and saturates around 0.85, indicating effective learning with a slight generalization gap. The training loss consistently decreases throughout the epochs. However, the validation loss begins to show mild fluctuations after epoch 30. This suggests minor overfitting in the later stages of training. In the DSD100 dataset, training accuracy steadily improves, reaching nearly 0.98, while validation accuracy rises and stabilizes around 0.85 after epoch 30. Both training and validation loss decrease during the

initial epochs. However, validation loss becomes noisy in later stages. This implies some generalization instability, probably due to the complexity of the dataset. In the dataset MUSDB18-HQ, training accuracy is at approximately 0.96, while validation accuracy is increasing quickly and has very high confidence values between 0.85 and 0.90 that indicate strong generalization. Even when training continues, the training loss decreases continuously, but validation loss varies widely after epochs 25. This shows possible overfitting, or the model could be highly sensitive to validation samples.

Fig. 2. Accuracy and Loss Visualization for Training and Validation Across various Dataset

In summary, all three datasets have learning behavior characterized by increasing accuracy and decreasing loss. The model has achieved consistent performance at high levels of validation

accuracy for all three datasets, and while variability exists particularly for validation loss, this is a good opportunity to improve performance using techniques such as early stopping, regularization, or data augmentation. Of the datasets, MUSDB18-HQ seems to reliably perform best with the highest level of stable validation accuracy, and more importantly, the dataset's audio quality appears to have meaningful effects on the model's generalization ability.

3.2. Preprocessing

The audio preprocessing pipeline starts with reading the audio file such as WAV or MP3 returning a tuple that contains the audio time series (1D if mono or 2D if stereo) and the sample rate which is at 48000 samples per second, then the time-domain waveform is transformed using the STFT, resulting in the signal transformed to a complex-valued time-frequency visualization. If the user provides an MP3 file, it must first be decoded into Pulse Code Modulation (PCM), as MP3 is a compressed frequency-domain format ; this decoding step restores the actual time-domain waveform required for further analysis. The magnitude of this complex STFT output is calculated to obtain a 2D array of values (frequency bins × time frames), representing the signal's energy distribution across frequency over time. To better align with human auditory perception, the magnitude spectrogram is further converted to the decibel (dB) scale.

Normalization techniques are applied to ensure numerical stability and consistency across inputs. In the time domain, peak normalization scales waveform amplitudes to the range [-1, 1]. In the frequency domain, log-magnitude or decibel scaling is used to manage dynamic range and improve learning efficiency. Data augmentation is employed across both time and frequency domains to enhance generalization and address the challenge of limited labeled data. In the time domain, augmentations such as time warping and pitch shifting modify the waveform directly ; while corresponding spectral transformations are applied to the decibel spectrograms in the frequency domain. These preprocessing steps spanning both waveform and spectral representations enable robust feature extraction and improved model generalization. This dual-domain approach is necessary for optimal source separation performance. In Fig. 3 (A) shows the time-domain waveform of the audio signal, representing amplitude variations over time. (B) and (C) display the magnitude spectrogram and decibel spectrogram, respectively, capturing the frequency content of the signal across time with different scaling.

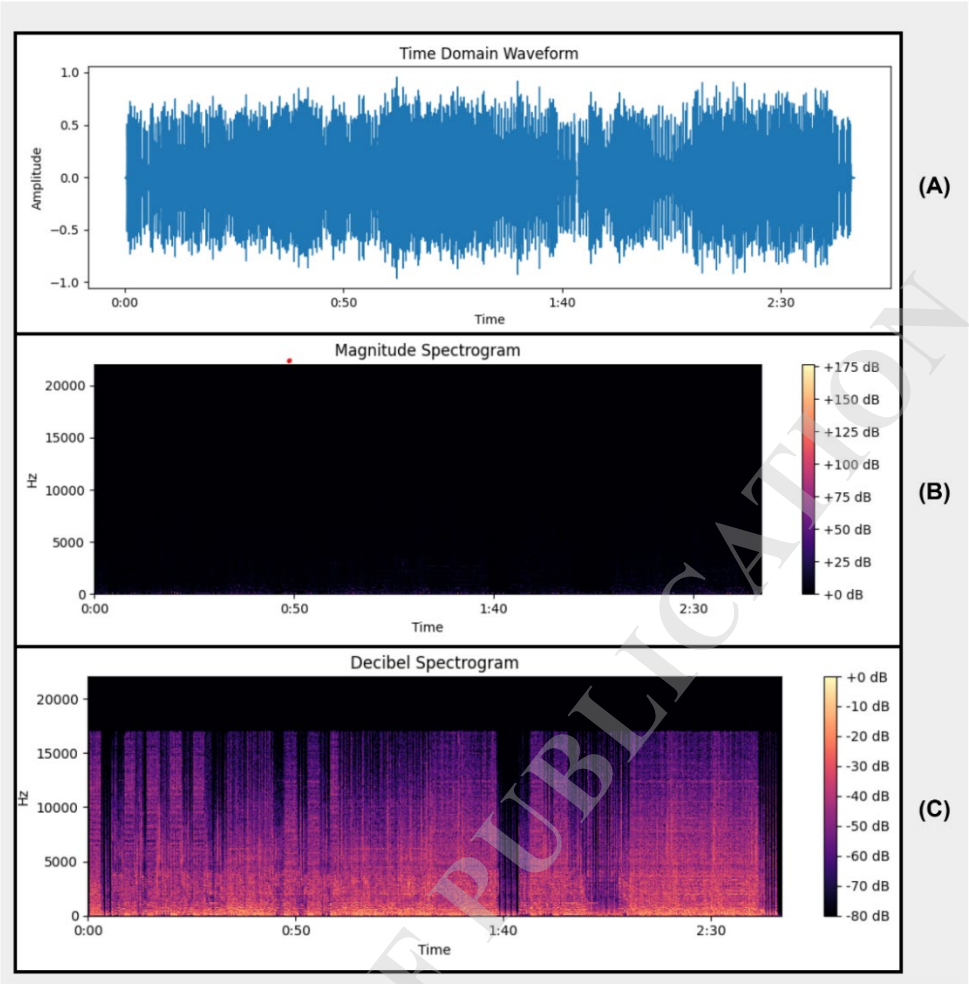


Fig. 3. Spectrogram (A) Time Domain Waveform (B)Time-Frequency Magnitude Spectrogram (C) Time-Frequency Decibel Spectrogram

3.3. Amalgamation Audio Features (A2F)

A specialized feature extraction pipeline designed for audio source separation tasks named A2F. Fig. 4 illustrates the proposed A2F framework. This preprocessing system operates on mixed audio inputs to derive discriminative feature representations that enhance the downstream DL model's ability to isolate and reconstruct individual sound sources. The A2F framework's hierarchical feature amalgamation process is engineered to capture both local Spectro-temporal patterns and global acoustic characteristics, providing the separation network with optimized input features for improved source disentanglement performance.

The input to the framework is a mixed audio signal that contains a blend of multiple source. This signal is processed in two parallel branches for extracting distinct types of features.

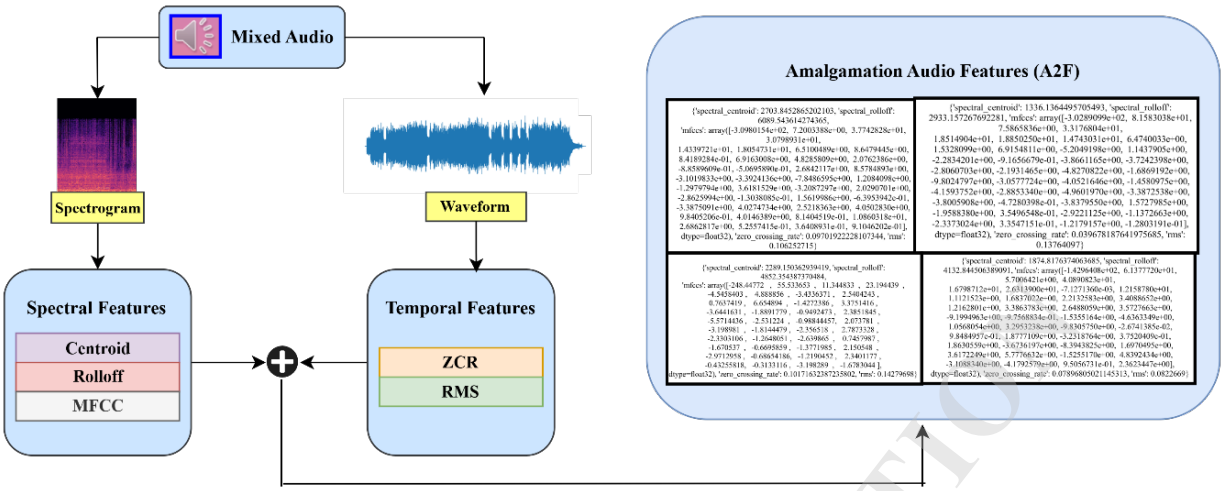


Fig. 4. Design of the Proposed A2F Framework

Spectral features derived from STFT based spectrogram which provide information about the frequency distribution, harmonic structure and timbre characteristics of each source. The spectral centroid indicates the brightness of the signal, helping the model distinguish high-frequency sources like vocals from low-frequency sources such as bass and drums. The spectral Rolloff reflects how energy is distributed across the spectrum, enabling the separation of harmonic instruments that exhibit smoother energy decay from percussive instruments whose energy extends sharply into higher frequencies. MFCC contribute by capturing perceptually relevant timbral information, allowing the network to identify and cluster tones from different instruments even when their frequency components overlap. In parallel, temporal features extracted directly from the waveform which provide insight into the signal's time-domain behaviour. The Zero Crossing Rate (ZCR) helps differentiate between transient, noisy, or percussive sounds, which have high ZCR, and sustained or tonal sounds like vocals or bass, which show lower ZCR. Root Mean Square (RMS) energy captures the loudness contour of the signal, helping to segment active and inactive regions and maintain balanced energy during source reconstruction. Together, these spectral and temporal features offer a holistic representation of the mixed audio, enabling the model to learn clearer boundaries between overlapping sources and ultimately improving the accuracy and robustness of the separation task.

3.3.1. Feature Fusion

A rich, multi-domain feature set is created by combining the temporal and spectral features. The last A2F matrix displayed at the bottom is a tabular dataset with columns representing

different extracted features and rows probably representing time frames. This provides a complete set of input features for the source separation model. a comprehensive analysis of the audio that makes use of time-domain features to find rhythmic patterns and transients as well as frequency-domain features to identify timbre and tonal patterns. The model is better able to differentiate between overlapping sources in the mixed signal thanks to this strong feature representation.

The A2F framework addresses some major limitations in current audio feature extraction techniques by combining noise-aware pre-processing, dual-domain feature extraction, multi-feature fusion, integration of deep learning. With respect to noise, the proposed model demonstrates improved robustness to noise, better generalization capabilities, less reliance on handcrafted features alone, and enhances separation performance when used to separate sound events present in realistic audio environments.

3.4. Audio Strips Network (ASNet)

The proposed ASNet is composed of Downsample, Bottleneck, Bi-LSTM, Upsample Module and Fully Connected Layer (FCL) as shown in Fig. 5.

The Downsample Blocks are designed to gather the most significant musical elements from the incoming Spectrogram data. When viewing Downsample 1 using a Conv2D layer, the input data is developed from 1 to 64 channels through convolution using a 3x3 kernel would allow detection of both short-term changes to the intensities of harmony in time and the harmonic relationship of notes (timbral) to one another. Following ReLU activation, maximum spectrums are retained and allowed to create a new representation with the highest harmonic (tonal) composition and instrument recognition. The second block, Downsample 2, the total number of filters is increased from 64 to 128, creating higher level features that allow the network to learn additional patterns of timbre that are associated with certain instruments and notes in the music, as well as additional harmonics. Repeating this process using ReLU activation and MaxPooling gradually reduces the data, allowing only for the most significant musical features to remain, while all extraneous information like noise is removed. Together, Downsample 1 and Downsample 2 allow the model to effectively establish and recognize time frequency relationships of music required for effective music extraction.

After the compressed features have been created, they are input into a Bottleneck layer, which has a scope of expanding the depth of feature maps from 128 to 256. This layer allows the

model to find very high-abstract patterns in music that can't be represented by other layers. The Bottleneck layer also lets the model capture information related to more-complex interactions between timbres, dense overlaps within harmonic structures, and instrument-specific signatures in spectra that lead to more precise separation of sounds. The use of a Conv2D-layer at this point lets the network apply wide-ranging and time-frequency dependencies during encoding operations. Furthermore, the increased number of filters gives the network the flexibility to encode finer, smaller details such as vocal-harmonic characteristics, resonances in drums, overtones in bass, and textures from other sounds in the background. Additionally, the ReLU activation functions support the largest and sparsest high-energy musical features, therefore, either enhances the detection of key harmonic points and tempo events or vice versa. In conclusion, the Bottleneck layer acts as the top-level high-level encoder within the overall structure of an end-to-end network, so it transforms the lower-level features of spectrograms into robust-rich discriminative representations ready for subsequent use in recurrent multi-task layers in producing improved separation capabilities for each song's music and producing improved music extraction capabilities.

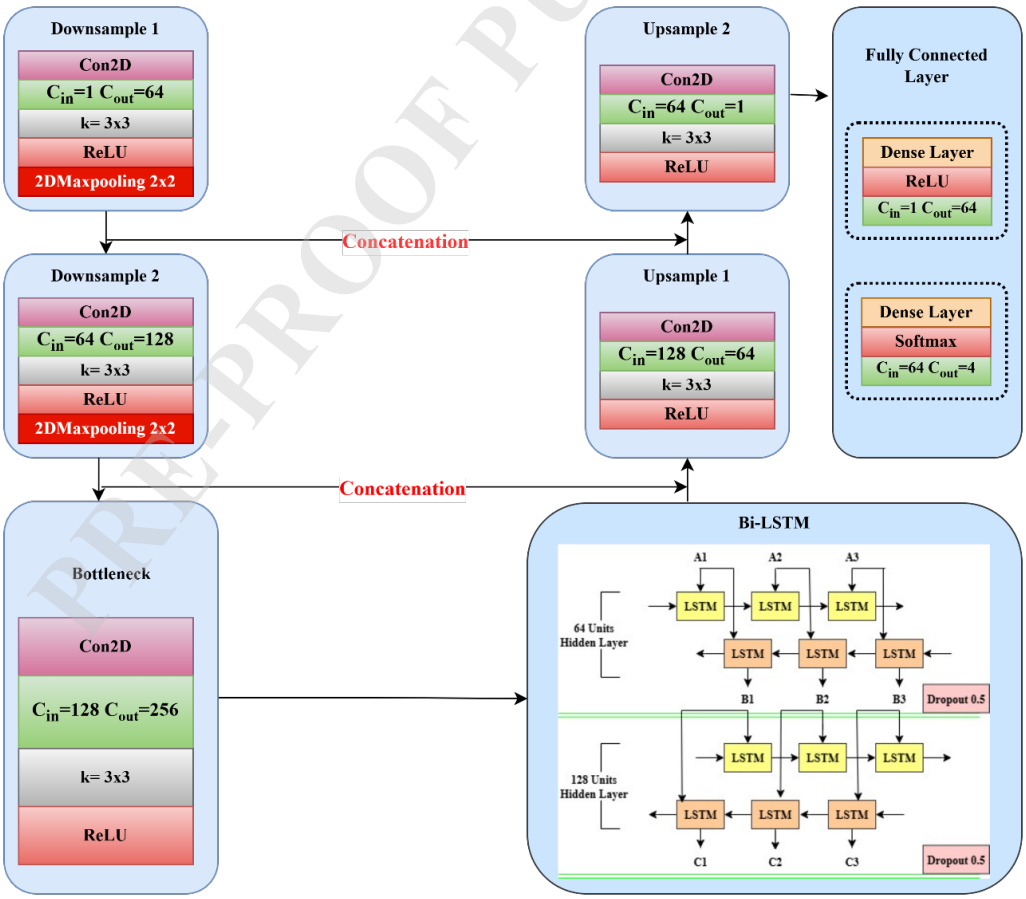


Fig. 5. An Innovative ASNet Architecture for Audio Source Separation

Temporal modelling occurs at the Bi-LSTM Block, which creates two bidirectional layers, containing 64 and 128 units respectively. These two bidirectional layers within one block allow for capturing short-term transitions such as note onsets and inflections that occur with voice, as well as longer strands of time such as sustained harmonic notes, rhythmic patterns and instrument continuities. This design is aimed to increase the capability of the model to accurately represent complex, overlapping audio source patterns found in music where there are many different instruments that coexist and interact throughout a time sequence. The use of Bi-LSTM is critical because musical events do not only rely on previous events in the past but also on future events, as this future context plays an important role in separating materials such as chord progressions, vocal phrasing or sustain of notes. With the inclusion of a 0.5 Dropout Rate, it is also critical to provide a strong regularization to prevent overfitting. The Dropout Rate will encourage the LSTM to identify effective Time Dependencies rather than memorizing actual patterns. Ultimately, this layer of the neural network block provides strong Temporal Understanding and enables the model to isolate, differentiate, and then extract audio sources.

Then the purpose of the Upsample blocks is to gradually recover the source signal from its compressed representation by progressively increasing the size of the representation while improving the quality of the reconstructed musical information. The first Upsample block reduces the number of channels from 128 to 64 during the first Upsample block using Conv2D so the model still has the basic harmonic and timbre information necessary for the reconstruction of the original audio but eliminates the redundant features created by the encoder, with the addition of the ReLU activation layer that helps highlight spectral elements such as the dominant harmonic and transient signals. In the second Upsample, reducing the number of channels from 64 to 1 will help the network focus on generating a clean, single-channel spectrogram while preserving the frequency characteristics of the instrument for correctly reconstructing the extracted source. The concatenation operation that occurs at this point allows the model to combine high-level semantic details learned in the deeper layers of the network with low-level spatial information from the early encoder layers in order to help the model recover fine frequency details, harmonic continuity, and timbral subtleties that have been impacted by the Downsampling process. The combination of filters and concatenation operations allows the network to increase the size of the audio representation, increase its quality, and build the target source audio in both time and frequency with sufficient temporal and spectral resolution.

The architecture ends with a FCL that provides the last classification by processing the high-level feature extraction throughout the network. The Dense layer with 64 units generates a consolidated learned representation of the temporal and spectral feature sets. It achieves this by utilising the ReLU to discover stronger features (dominant harmonic, transient peak, instrument-specific signature), allowing for more accurate identification of instrument categories. The final Softmax layer takes the output from the dense layer and converts it to probabilities, which enables the model to identify what portions of the times-frequency spectrogram belong to each instrument (vocals, drums, bass, and other.). The classification-based output from the network is vital in order to provide on-target instrument assignments for audio extraction, thus facilitating a higher accuracy rate with respect to separated audio component.

3.5. Postprocessing

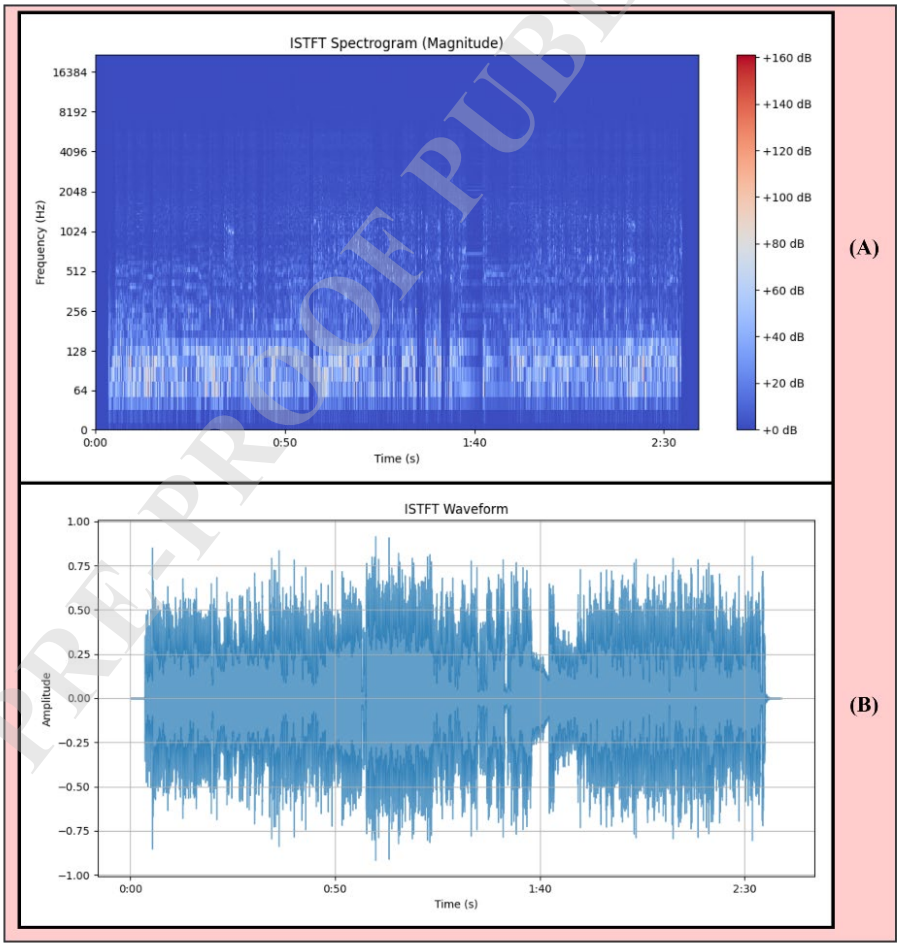


Fig. 6. Inverse Spectrogram (A)Time-Frequency Inverse Magnitude Spectrogram (B)Time-Domain Inverse Waveform

Postprocessing serves as the critical final stage in audio source separation, transforming the model's spectrogram predictions into usable audio signals. As illustrated in Fig. 6, this involves two key conversions : (A) shows the ISTFT spectrogram (magnitude), visualizing frequency content over time after inverse STFT reconstruction. (B) displays the reconstructed audio waveform in the time domain, illustrating amplitude variations across time. This stage effectively closes the loop between neural network outputs and practical audio applications, enabling immediate playback or downstream analysis.

4. Experiment

This section displays the experimental configuration designed to evaluate the offered ASNet framework. It describes the primary aspects of validating with the model, the data used to train and validate, the evaluation metrics used to evaluate performance, and a comparison with previous methods used, to show the improvement from the proposed model. The intention of this experimental study is to demonstrate the model not only performs well in isolation, but also generalizes well across a variety of audio mixtures.

4.1. Postprocessing

To objectively evaluate the effectiveness of the offered audio source separation method, two widely accepted evaluation metrics are employed : SDR and SIR. When compared to the original actual sources, these measurements evaluate how well the separated audio sources perform.

SDR describes the fidelity of a separated audio signal by measuring the amount of distortion, artifacts, or errors that remain after separation. An increased SDR score indicates that there are very little distortion or artifacts in the signal, and that the reconstructed audio is a clean and accurate representation of the original source. Conversely, a lower SDR score indicates poorer audio quality and implies that the separated signal contains some distortion, missing components, and/or visible artifacts that diminish clarity.

SIR quantifies how well a model separates the separated source signal from any unwanted sounds associated with other instruments or vocal sounds in the original mixture. An increased SIR score indicates that the separated source has been effectively isolated from any competing audio sources and has clearly defined separation boundaries. A lower SIR score indicates that interference remains, with other instruments still audible in the extracted signal; therefore, this

indicates that the separation was not as efficient as it should have been. These two metrics evaluate the complementary aspects of performance : SDR evaluates the quality of the reconstruction, and SIR evaluates the strength of the isolation of the source.

4.2. Performance Comparison Analysis

This section evaluates and compares the effectiveness of proposed ASNet models in achieving the desired outcomes. The comparison provides insights into optimal approaches and potential areas for improvement. The metrics SDR and SIR are both measured in decibels (dB).

Table 1. Evaluating ASNet Model with diverse Dataset

	MUSDB				DSD100				MUSDB18-HQ			
	D	V	B	O	D	V	B	O	D	V	B	O
SDR	11.98	10.73	11.21	10.81	11.19	10.64	10.33	10.68	12.63	11.42	12.01	11.14
SIR	9.27	9.52	9.51	9.51	9.55	9.46	9.59	9.11	9.57	9.61	9.66	9.67

Table 1 displays the SDR and SIR scores achieved by the proposed ASNet model on three benchmark datasets: MUSDB, DSD100, and MUSDB18-HQ. The evaluation covers four audio source separation tasks: Drums (D), Vocals (V), Bass (B), and Other (O). It demonstrates its best overall performance on the MUSDB18-HQ dataset, achieving the highest SDR and SIR scores among all evaluated datasets. Specifically, SDR improved by approximately 5.4% over MUSDB and 12.9% over DSD100, while SIR showed an improvement of about 2.7% over MUSDB and 1.8% over DSD100. The Conv2D block (256 filters) compresses and enriches features before Bi-LSTM, preserving fine details for higher SDR. Dropout and FCL layers prevent overfitting, boosting generalization across datasets

Table 2. Descriptive Statistics of SDR and SIR Metrics Across MUSDB, DSD100, and MUSDB18-HQ Dataset

	Mean		Median		SD	
	SDR	SIR	SDR	SIR	SDR	SIR
MUSDB	11.18	9.45	11.01	9.51	0.495	0.105
DSD100	10.71	9.43	10.66	9.51	0.309	0.190
MUSDB18-HQ	11.80	9.63	11.72	9.64	0.570	0.040

Descriptive statistics such as the Mean, Median, and Standard Deviation (SD) reveal different but complementary views of model behavior when evaluating audio separation, as these metrics provide a snapshot of a model's overall separation capability, a representation of typical separation performance when outliers or other unusual audio samples exist, and an opportunity to assess how consistently well a model separates multiple sample types (i.e., either as a result of the presence of several high-performing samples or as a result of

consistently high separation accuracy for many samples). Together, the mean and median provide insight into model strength; whereas the mean provides an average value of separation success across all samples tested, the median provides a stable assessment of typical audio separation when outliers or other unusual audio samples are present. A third metric, SD gives another indicator of separation consistency by providing a measure of variability in the model's performance among multiple recordings. A low SD value implies reliable separation performance while a high SD value signifies sporadic performance or inconsistent fidelity or isolation across the samples. The table 2 illustrates these results across MUSDB, DSD100, and MUSDB18-HQ datasets: the MUSDB dataset has close to equal mean and median SDR values, and moderate, low variability in the mean values; the DSD100 dataset has a small range of SDR values with no outliers and a much lower SD value; and the MUSDB18-HQ dataset has a very high average SDR value of 11.80 and large SDR SD value (0.570), indicating a high level of separation quality, while having a very low SIR SD value (0.040) demonstrating a high level of consistency in the audio separation process. The combination of Mean, Median, and SD provides a detailed, multi-angle view of the model's ability to separate audio accurately across multiple datasets.

Table 3. Implementing ASNet based on channel -wise Configuration

Channel Count	SDR				SIR			
	B	D	V	O	B	D	V	O
128 Channels	5.01	4.22	6.23	4.66	9.33	9.29	9.33	9.22
256 Channels	12.01	12.63	11.42	11.14	9.57	9.61	9.66	9.67
512 Channels	6.38	6.97	8.01	6.09	9.48	9.52	9.47	9.42

Utilizing SDR and SIR across four sources, Table 3 compares ASNet's performance with various channel configurations (128, 256, and 512) on the MUSDB18-HQ dataset. At 11.80 dB average SDR, the 256-channel configuration produced the highest average SDR value, followed by the 128-channel average SDR of 5.03 dB and 512-channel average SDR of 6.86. This gives approximately a 134.6% increase on the SDR using 128 channels, and approximately 71.9% increase over using 512 channels. SIR values were comparable across all channels with slight variation, indicating no difference in interference suppression regardless of channel amount. These results demonstrate that the 256-channel configuration provides the most effective balance for source separation performance in ASNet.

Table 4. Comparison of Proposed and Existed DL Approaches with its parameters

Model	Params	V	D	B	O
Res-U-Net [22]	102.0 M x 4	8.98	6.62	6.04	5.29
Mel-Roformer [23]	84.2 M	11.21	9.91	9.64	7.81
Hybrid Demucs [24]	83.6 M x 4	8.13	8.24	8.76	5.59

BSRNN [25]	37.6 M x 4	10.01	9.01	7.22	6.70
DTT [11]	5.0 M x 4	10.12	7.74	7.45	6.92
ASNet (Proposed)	0.31 M x 4	11.42	12.63	12.01	11.14

Table 4 compares the proposed ASNet model with several existing DL approaches for Audio Source Separation on the dataset MUSDB18-HQ, based on SDR for four sources along with the number of model parameters. Despite having a significantly smaller parameter count (0.31M × 4), ASNet outperforms all other models in SDR across all source types, achieving 11.42 dB (V), 12.63 dB (D), 12.01 dB (B), and 11.14 dB (O). In contrast, larger models like Res-U-Net (102M × 4) and Mel-Roformer (84.2M) deliver lower SDR, indicating less effective separation. ASNet is an interesting case that is able to outperform notoriously powerful models, like BSRNN, DTT, and Hybrid Demucs, on both complexities and computational efficiency. Meaning, the specifications that ASNet must give up for its lightweight design actually favour performance in source separation. All of this proves why ASNet is a better model with high performance and reduced computational complexity.

4.3. Discussion

To quantify computational performance of the speed/latency, throughput, and memory consumption of the audio source separation model, performed tests using an NVIDIA RTX 3090 GPU and an Intel Xeon Silver 4210 CPU in conjunction with the Pytorch library. Also utilized Python libraries such as Librosa and soundfile for the preprocessing and inference stages of the ASNet model, allowing us to maintain consistency in input formatting and minimize I/O bottlenecks by providing high-quality audio data and an efficient method for loading audio files, resampling, and generating spectrograms. On average, the ASNet processes a 5-second audio sample in 42.7 milliseconds (ms) on the RTX 3090 and 315.4 ms on the Xeon 4210 with a throughput rate of 23.4 samples/second (audio signals) on the RTX 3090 vs. 3.1 samples/second on the Xeon 4210. Maximum memory usage for the audio source separation model was 2.6 gigabytes (GB) on the RTX 3090 vs. 4.2 GB on the Xeon 4210, indicating that the increased speed, efficiency, and suitability for real-time music separation tasks provided by utilizing a GPU in conjunction with Librosa and soundfile for audio processing far outweighs the additional memory resources required by using a CPU.

The ASNet model is trained on three commonly used benchmark datasets : MUSDB18-HQ, MUSDB, and DSD100 with fixed training and validation splits. All models within the same experimental configuration for fair comparison to baselines. Source separation performance is measured using track-wise SDR and SIR, calculated using standard formulas from the original

and separated audio materials. Instead of using the museval framework, the evaluation process was managed through custom metric implementations to allow for complete management of the evaluation process, consistency across all datasets, and ease of integration into the proposed ASNet pipeline. The final performance scores are the average of SDR and SIR for each track across the entire test dataset. To allow for comparison and replication of previous work, the baseline results are recreated or cited according to the same evaluation protocol that is established between the ASNet model and the evaluation of the other models.

The proposed A2F framework outperforms baseline feature extraction by integrating both spectral (STFT, Spectral Centroid, Rolloff, MFCC) and temporal (ZCR, RMS) features. Unlike single-domain methods, A2F's hierarchical fusion captures fine-grained spectral-temporal cues and global acoustic patterns simultaneously. This multi-domain representation significantly enhances source separation accuracy compared to traditional feature pipelines. ASNet with Bi-LSTM leverages bidirectional sequential modeling to capture temporal dependencies, improving continuity and accuracy in overlapping audio sources. Without Bi-LSTM, the model relies only on convolutional features, limiting its ability to learn long-range time relationships. The integration of Bi-LSTM thus significantly enhances source separation performance compared to a purely CNN-based ASNet. Future work will focus on extending ASNet for real-time audio source separation by optimizing inference latency and reducing computational overhead. Model compression and lightweight deployment strategies will be explored to enable integration into streaming and live music applications.

The aim of performing a subjective evaluation in addition to the objective SDR and SIR metrics was to evaluate the quality of the separated audio recordings as judged by human listeners. In this evaluation, the proposed ASNet model was compared against two standard methods for use in music source separation that are frequently used : DTT and BSRNN. The results of the subjective evaluation indicate that the proposed ASNet model had a substantially higher perceptual quality and naturalness than the two baseline methods on average. The majority of listeners reported fewer audible musical artifacts, a better preservation of timbre, and continued temporal continuity in the output from ASNet in mixed audio samples consisting of many overlapping musical sources compared to the other two models. Therefore, the difference between objective and subjective evaluations provides evidence of a strong correlation between objective improvements and perceived audio quality, indicating that the proposed method is likely to be useful for practical applications.

4.3.1. Limitations

The learning curves show that validation loss fluctuated during the early and late stages of training, with MUSDB and MUSDB18-HQ showing the greatest fluctuation. It appears that these validation samples may have been somewhat sensitive or biased with respect to the training data which could suggest that there was some degree of overfitting occurring despite applying regularization techniques such as dropout. Although the analyzed result still provides a good level of confidence in its accuracy, more advances can likely be made using more sophisticated methods adaptive early stopping, curriculum learning and dataset aware augmentation.

4.3.2. Scope for Future Work

The goal of these strategies is to improve the stability of convergence during model training and decrease the oscillation of the validation loss while maintaining the complexity of the model.

Another limitation of the current framework is that it focuses only on using audio spectrogram representations. Although they work well, the use of audio-only representations may limit the ability of the model to separate sources that overlap heavily in very difficult audio environments. Future improvements may come from using adaptive learning techniques and from trying out different types of inputs, such as by using audio and video together, to increase both accuracy and robustness. For instance, information about the movements of instruments or the actions of performers would help to separate sounds that overlap.

The future direction of research will extend from symbolic music representations to converting extracted source signals to a MIDI format for automated generation of improved musical scores. The expanded system will include pitch detection, onset-offset estimation, and note-segmentation models that work together with the separated source files to create a single pipeline for audio-to-score conversions. With this type of extension, there will be a seamless transition between separating audio sources and music transcription. Thus, the proposed framework could be utilized in various applications, such as providing educational opportunities for learning about music, digitally preserving music through archiving, analyzing music from an academic perspective, and providing assistance to musicians in compositional and performance endeavors.

In order to improve the fidelity of transcription accuracy and robustness, other proposed avenues of exploration will involve incorporating transformer's temporal modeling, perceptual loss functions, and the application of cross-domain approaches and learning techniques for audio and symbolic musical representation. Furthermore, one area of continued exploration is in the area of near real-time deployments and assessments of live and/or noisy musical/audio recordings.

5. Conclusion

The proposed ASNet, integrated with the A2F framework, demonstrates a significant advancement in audio source separation. By effectively combining spectral and temporal features, ASNet achieves superior separation quality. ASNet adopts a U-Net-inspired encoder-decoder structure to extract hierarchical feature representations and preserve spatial resolution through skip connections. This architecture allows the network to maintain crucial audio details while learning abstract features at multiple scales. Additionally, to enhance the modeling of long-term dependencies in temporal data, ASNet incorporates Bi-LSTM layers. These layers enable the network to capture both past and future contextual information. Notably, the 256-channel configuration of ASNet yields the highest SDR and SIR, despite having a considerably smaller parameter count. This effectiveness highlights ASNet's potential for implementation on devices with limited resources and applications that run in real time. Future improvements may also include incorporating adaptive learning strategies, as well as evaluating different modalities of input (e.g., adding audio-visual cues), that may improve the separation accuracy and robustness. Also, applying quantization-aware training could allow ASNet to be deployed on edge devices without a loss in performance.

FUNDINGS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHORS' CONTRIBUTIONS

S.P Sakthidevi (Data analysis, writing, Review, and Editing). C. Divya (Supervision).

ACKNOWLEDGMENTS

This work was supported by the Manonmaniam Sundaranar University, Centre for Information Technology and Engineering, Tirunelveli, Tamil Nadu, India.

References

1. REZAUL, K. M., JEWEL, M., ISLAM, M. S., SIDDIQUEE, K. N. E. A., BARUA, N., RAHMAN, M. A., ... & ASHA, U. F. T. (2024). Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models. International Journal of Advanced Computer Science and Applications, 15(7), 37-53. DOI: 10.14569/IJACSA.2024.0150704

2. TONG, W., ZHU, J., CHEN, J., KANG, S., JIANG, T., LI, Y., ... & MENG, H. (2024, April). Scnet: Sparse compression network for music source separation. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1276-1280). IEEE. https://doi.org/10.48550/arXiv.2401.13276

3. CHEN, J., VEKKOT, S., & SHUKLA, P. (2024, April). Music source separation based on a lightweight deep learning framework (DTTNET: Dual-path TFC-TDF UNet). In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 656-660). IEEE. DOI: 10.1109/ICASSP48485.2024.10448020

4. RAFII, Z., LIUTKUS, A., STÖTER, F. R., MIMILAKIS, S. I., & BITTNER, R. (2019). MUSDB18-HQ-an uncompressed version of MUSDB18. (No Title).

5. WU, X., DANG, B., ZHANG, T., WU, X., & YANG, Y. (2024). Spatiotemporal audio feature extraction with dynamic memristor-based time-surface neurons. Science Advances, 10(14), eadl2767. DOI: 10.1126/sciadv.adl2767

6. ZHANG, Y., JOHANNESEN, P. T., MOLAEE-ARDEKANI, B., WIJETILLAKE, A., CHIEA, R. A., HASAN, P. Y., ... & LOPEZ-POVEDA, E. A. (2025). Comparison of Performance for Cochlear-Implant Listeners Using Audio Processing Strategies Based on Short-Time Fast Fourier Transform or Spectral Feature Extraction. Ear and hearing, 46(1),

163-183. DOI: 10.1097/AUD.0000000000001565

7. DENG, B., YANG, H., & KIM, N. Y. (2024). A denoising autoencoder based on U-Net and bidirectional long short-term memory for multi-level random telegraph signal analysis. Engineering Applications of Artificial Intelligence, 135, 108685. https://doi.org/10.1016/j.engappai.2024.108685

8. TAKAHASHI, N., & MITSUFUJI, Y. (2020). D3net: Densely connected multidilated densenet for music source separation. arXiv preprint arXiv:2010.01733. DOI:10.48550/arXiv.2010.01733

9. DÉFOSSEZ, A., USUNIER, N., BOTTOU, L., & BACH, F. (2019). Music source separation in the waveform domain. arXiv preprint arXiv:1911.13254. DOI:10.48550/arXiv.1911.13254

10. WANG, J. C., LU, W. T., & WON, M. (2023). Mel-Band RoFormer for Music Source Separation. arXiv preprint arXiv:2310.01809. https://doi.org/10.48550/arXiv.2310.01809

11. LU, W. T., WANG, J. C., KONG, Q., & HUNG, Y. N. (2024, April). Music source separation with band-split rope transformer. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 481-485). IEEE. https://doi.org/10.48550/arXiv.2309.02612

12. KONG, Q., CAO, Y., LIU, H., CHOI, K., & WANG, Y. (2021). Decoupling magnitude and phase estimation with deep resunet for music source separation. arXiv preprint arXiv:2109.05418. DOI:10.48550/arXiv.2109.05418

13. DÉFOSSEZ, A. (2021). Hybrid spectrogram and waveform source separation. arXiv preprint arXiv:2111.0360. DOI:10.48550/arXiv.2111.03600