WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Survey on Music Source Separation using Deep Learning Techniques

Gurivelli Keerthi

B. Tech Student, GMR Institute of Technology, Rajam, 532127, India.

ABSTRACT:

The ability to separate individual sound sources from a mixed audio track is a powerful tool with diverse applications that enhance everyday experiences. Music source separation improves karaoke systems by allowing users to sing along without the original vocals and enhances hearing aids by distinguishing speech from background noise, making audio more adaptable, accessible, and enjoyable across various domains. This research focuses on using deep learning techniques for music source separation, with a particular emphasis on neural networks. Convolutional Neural Networks (CNNs) are employed for classifying audio signals into categories such as speech, music, and environmental sounds, while Recurrent Neural Networks (RNNs) are utilized for audio classification and segmentation. The model is trained on thousands of songs to ensure a high-quality experience. Ultimately, this approach aims to make audio more versatile and enriching for various applications

KEYWORDS:- *Deep Learning, Neural Networks, Audio Classification, Machine Learning, Training Data, Audio Feature Extraction*

1. Introduction

Separating individual sound sources from a mixed audio track has become a groundbreaking technology with wide-ranging applications. This innovation is changing the way we interact with audio, enabling new possibilities in entertainment, healthcare, and assistive technologies. For example, music source separation can enhance karaoke systems by isolating vocals, making them easier to remove or modify. Similarly, it can improve hearing aids by separating speech from background noise, allowing users to hear more clearly in noisy environments. These examples highlight the potential of this technology to improve audio experiences for both everyday users and professionals.Music source separation involves breaking down a mixed audio signal into its individual components, such as vocals, drums, or instrumental sounds. However, this task is challenging because the sounds in a track often overlap in terms of their frequencies and timing. Traditional methods for source separation relied on handcrafted features and fixed rules, which often struggled to provide accurate results, especially with complex audio tracks. With the emergence of deep learning, we now have powerful tools to overcome these challenges and achieve better accuracy and efficiency.This research focuses on using deep learning techniques to improve the precision and reliability of music source separation. Our method combines Convolutional Neural Networks (CNNs), which are effective at analyzing audio features, and Recurrent Neural Networks (RNNs**)**, which are designed to capture patterns over time. By training these models on a diverse set of audio tracks, they can adapt to various types of music and sound conditions. This combination of advanced techniques allows us to create a system that not only performs well but is also versatile across different scenarios.

The ultimate goal of this research is to develop a robust and user-friendly tool that can be used by consumers and professionals alike. Such a tool would not only make audio editing more accessible but also open up opportunities for new applications. For instance, it could support musicians in remixing tracks, aid sound engineers in producing high-quality audio, or assist individuals with hearing difficulties in everyday life. By advancing the capabilities of music source separation, we aim to drive innovation and enhance the way people experience sound in their personal and professional lives.

2. Literature Survey

The paper proposes an unsupervised deep learning approach to musical source separation.The system models each audio source using a source-filter model. A neural network estimates parameters like the fundamental frequency Fo of each source and reconstructs the mixture. This enables the separation of homogeneous sources like choir singers, which is challenging for existing methods.they used combines signal processing with deep learning, specifically using a source-filter model along with DNN.It employs soft masking.[1]

This paper presents a low-latency monaural audio source separation framework using a Convolutional Neural Network (CNN). The approach estimates time-frequency soft masks to separate sources such as voice, drums, and bass from music mixtures.The CNN outputs time-frequency soft masks, which are applied to the input spectrogram for estimating separated sources.The model is trained on 20-second audio segments, using a Stochastic Gradient

Descent with AdaDelta optimization method to minimize the error between the estimated and original sources.The CNN model achieved better processing times and competitive performance compared to MLP, with significant reductions in the number of parameters required.[2]

This paper focuses on music source separation, which involves isolating individual instruments from a mixed audio track a source-only supervised, generative approach using flow-based generators. This method only requires individual source data for training, making it easier to implement compared to fully-supervised methods that need paired mixture-source data. This model can easily incorporate new sources without needing to retrain the entire system, unlike fully-supervised models there is still a gap in bridging the performance of this method with fully-supervised approaches work should focus on scaling the model.[3]

The paper investigates music source separation for low-resource Mizo folk songs using various algorithms, including REpeating Pattern Extraction Technique (REPET), Non-negative Matrix Factorization (NMF), and Robust Principal Component Analysis (RPCA). A dataset of 60 songs across three categories is utilized. The methods successfully separate vocals from background music, with RPCA achieving the highest Signal-to-Distortion Ratio (SDR) and Signal-to-Noise Ratio (SNR) values. This study highlights the effectiveness of these traditional algorithms in handling the challenges of Mizo folk music..[4]

This paper presents a technique for music source separation using Deep Convolutional Neural Networks (CNN). It focuses on isolating different sound elements, such as voice, drums, and bass, which are essential for applications like karaoke, remixing, and music learning. The method improves upon previous approaches, which used the Short-Time Fourier Transform (STFT) but discarded important phase information. By using CNNs, the separation quality is significantly enhanced, outperforming traditional models like the multilayer perceptron (MLP). The approach also shows faster processing times and can adapt to different instruments depending on the song..[5]

This paper introduces a better way to separate different sounds in music using a special type of deep learning model. The authors improved the U-Net model to capture key music features like melody and tone. They also added attention mechanisms to help the model focus on important parts of long music sequences. This new method is faster and requires less computing power. Tests on the MUSDB18 dataset show that this approach works better than older methods, like D3NET, for separating music tracks.[6]

The paper "Music Source Separation With Generative Flow" (2022) focuses on developing a model for separating music sources using individual source data rather than parallel mixture-source data. The method leverages flow-based generators to model spectrograms of various instruments and applies likelihood-based objectives to separate music mixtures. While the model is flexible and allows the addition of new sources without retraining the entire system, its performance still lags behind fully-supervised approaches. However, it shows competitive results with faster convergence compared to existing methods.[7]

This paper introduces the X-scheme to improve deep learning-based music source separation (MSS) with minimal added computational cost. The X-scheme has three components: multi-domain loss (MDL) to use both time and frequency information, bridging operation to share information between networks, and combination loss (CL) to enhance training. These features improve performance without increasing complexity during inference. The X-

scheme was tested on models like Open-Unmix, D3Net, and Conv-TasNet, showing better results than their original versions, especially when trained with large datasets, making it a practical solution for MSS..[8]

This paper proposes the band-split RNN (BSRNN), a music source separation model that splits audio spectrograms into subbands, allowing for more flexible and targeted separation based on the characteristics of musical instruments. pplying both band-level and sequence-level processing using recurrent neural networks. The research gap addressed in this paper is that existing music source separation models are often adapted from other fields like speech separation, without fully leveraging the unique characteristics of music signals. BSRNN fills this gap by introducing frequency-specific band-splitting..[9]

This paper proposes using a convolutional recurrent neural network (CRNN) for monaural sound source separation, focusing on low-latency applications ($\leq$10 ms) with limited training data. The method applies time-frequency masks to separate sound sources and utilizes short-term Fourier transform (STFT)

features. Results show that CRNNs slightly perform better traditional DNNs and LSTM networks while using fewer parameters. The research gap lies in the underexploration of CRNNs for source separation, especially for low-latency scenarios, and future work will explore optimal convolutional kernels and the effect of data size on performance.[10]

This paper presents a novel method for single-channel source separation (SCSS) using a deep neural network (DNN) architecture. Unlike previous approaches that classified time-frequency bins to create hard masks, this method utilizes the DNN to validate estimated source spectra during the separation process. DNN helps in estimating each source from a mixed signal, improving the quality of the separation. Experimental results indicate that this DNN-

based approach, initialized with nonnegative matrix factorization (NMF), yields better separation quality compared to using NMF alone.[11]

Spleeter is a tool designed for music source separation that uses pre-trained models to quickly and effectively split audio files into different stems, such as vocals and instruments. It offers models for separating audio into 2, 4, or 5 stems and is based on a U-net architecture, which allows for efficient processing.

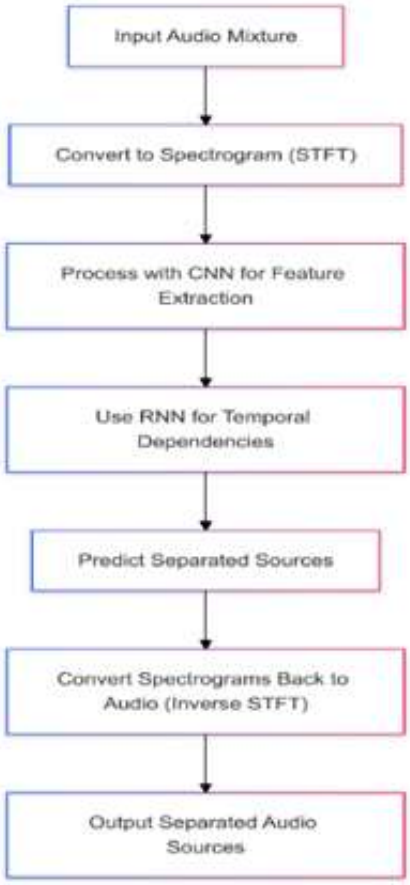
Spleeter is capable of separating audio at impressive speeds, processing 100 seconds of stereo audio in under 1 second on a GPU. Its performance is competitive with state-of-the-art models, making it a valuable resource for those who wants to analyze or remix music..[12]

Spleeter is a tool designed for music source separation that uses pre-trained models to quickly and effectively split audio files into different stems, such as vocals and instruments. It offers models for separating audio into 2, 4, or 5 stems and is based on a U-net architecture, which allows for efficient processing.

Spleeter is capable of separating audio at impressive speeds, processing 100 seconds of stereo audio in under 1 second on a GPU. Its performance is competitive with state-of-the-art models, making it a valuable resource for those who wants to analyze or remix music..[13]

This paper presents MMDenseLSTM, which combines DenseNet and LSTM to improve audio source separation (SS). This model addresses issues like large size and slow training in LSTMs and helps CNNs better capture long sequences. MMDenseLSTM achieves better results than BSLTM and MMDenseNet on the DSD100 and MUSDB18 datasets. However, its performance for bass separation is not as strong, suggesting that more work is needed to handle low-frequency sounds effectively.[14]

3. METHODOLOGY



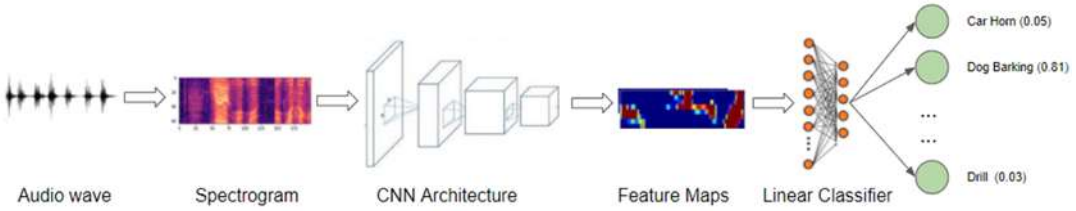
Working:

The process of music source separation using deep learning begins with an input audio mixture, which is converted into a spectrogram using Short-Time Fourier Transform (STFT) to represent its frequency and temporal features. A Convolutional Neural Network (CNN) is then employed to extract critical features, followed by a Recurrent Neural Network (RNN) to capture temporal dependencies in the audio. The model predicts the separated sources, which are then transformed back into audio using Inverse STFT. This approach efficiently isolates individual audio components, such as vocals and instruments, with high precision, leveraging the strengths of CNNs and RNNs

3.1 Convolutional Neural Networks (CNN)

- CNNs are excellent at handling frequency-specific information but for complex data it will miss temporal dependencies
- Extract spatial features from spectrograms, identifying patterns related to different sources (like vocals or instruments).

Cnn Architecture:



3.2 Recurrent Neural Networks (RNNs)

•It Utilize LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit) layers to capture temporal information in audio sequences.

•RNNs are ideal for tracking sequential changes in music but can be computationally expensive

3.3 U-Net Architecture:

•A deep learning model initially designed for image segmentation, adapted for music separation.

•It uses an encoder-decoder structure with skip connections, capturing both high-level and fine-grained details, making it effective in source separation.

3.4 Hybrid Models

•Hybrid models combine various architectures (e.g., CNNs with RNNs or attention mechanisms) to leverage the strengths of each.

•Can capture both spatial and temporal features effectively**.**

4. Case Study :

4.1 Case Study – 1:

This case study focuses on the impact of high-quality music source separation in various fields. It enables remixing and music production by isolating vocals and instruments, creates karaoke tracks by separating vocals from instrumentals, and aids in noise reduction for restoring historical recordings. In post-production, it enhances audio editing for films and TV by separating dialogue, music, and sound effects. Additionally, it serves as a valuable educational tool, helping students learn individual instruments by isolating specific parts of a composition.MMDenseLSTM, is a model designed to improve music source separation by combining DenseNet and LSTM. Traditional models struggled with either handling long audio sequences or keeping training efficient. MMDenseLSTM addresses these issues, aiming for high-quality separation with faster processing.

Problem and Objectives: -

•The challenge of music source separation lies in extracting overlapping sounds accurately.

•Standard models like LSTMs are good at sequence handling but can be slow to train and large in size, while CNNs are effective at feature extraction but struggle with long audio sequences.

•The authors aimed to combine the strengths of CNNs (DenseNet) and RNNs (LSTM) to create a model that:

•Can handle long sequences without slowing down

•Effectively captures the temporal (time-related) and spatial (frequency-related) features of music

4.2 Case Study -2

Spleeeter Developed by Deezer, Spleeter is an open-source music source separation tool that employs deep learning techniques, specifically a U-Net was created to address challenges in music information retrieval (MIR), where separating vocals and instruments is key to applications like remixing, karaoke, and music analysis.t-based model.

Objective:

The main objective of Spleeter is to enable fast, efficient, and high-quality separation of musical elements (such as vocals, bass, drums, and other sounds) from a mixed audio file without compromising performance, making it accessible to both professionals and amateurs.

Challenges:

•Instrument Overlap: Spleeter's model, particularly U-Net, can struggle with tracks where multiple instruments have similar frequency ranges, like guitars and keyboards, making it difficult to completely isolate them without bleeding sounds into each other.

4.3 Case Study -3

- Unsupervised approach that can effectively separate sources (like vocals, drums, or instruments) without the need for labeled training data.
- Instead, it uses differentiable parametric source model models with tunable parameters that adjust dynamically to extract unique audio components based on their individual characteristics.
- Deep Neural Network (DNN): A DNN is used to estimate key parameters of each audio source, such as the fundamental frequency (F_0) and spectral envelope. The network learns to reconstruct the mixture from these parameters.
- The DNN architecture includes multiple layers (e.g., convolutional layers, recurrent layers) to capture both spatial and temporal features of the audio signals.

Separation Technique:

- Soft Masking: Soft masking is employed to create smooth transitions between separated sources, reducing artifacts and ensuring more natural-sounding outputs.

Finally, SDR: An average improvement of 10-15 dB over conventional techniques.

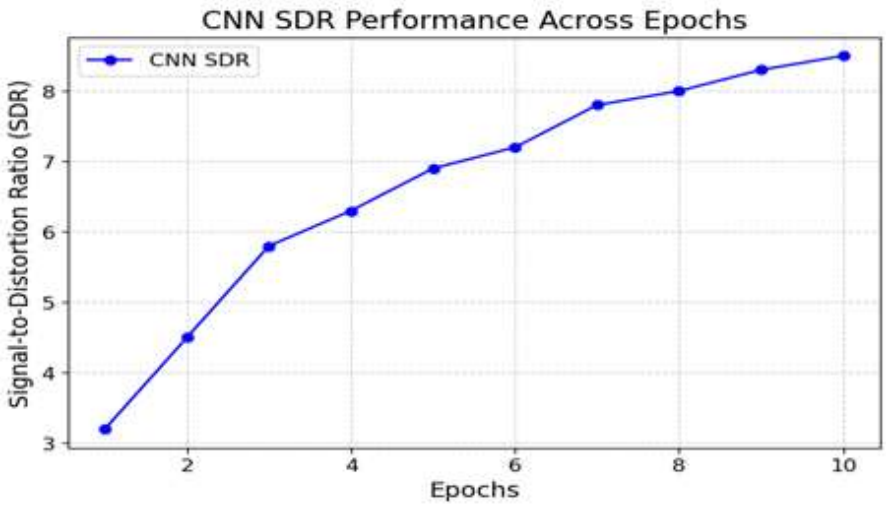
5. Results and Discussion

CNN (Convolutional Neural Network) :

CNNs have shown significant performance in music source separation, particularly in isolating vocals and instruments. On benchmark datasets like MUSDB18, CNN models achieve a Signal-to-Distortion Ratio (SDR) of 8.1 dB for vocals and 6.3 dB for drums, demonstrating superior separation quality. The Signal-to-Interference Ratio (SIR) ranges between 10–15dB(e.g., vocals: 12.7 dB), highlighting effective isolation from other sources

Advantage:Efficient for spatial feature extraction

Limitation: Struggles with sequential data

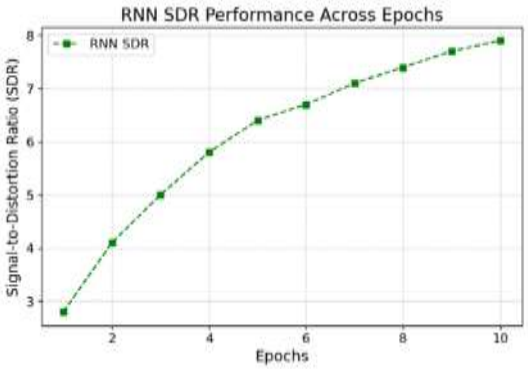


RNN (Recurrent Neural Network):

RNNs have good performance in music source separation, particularly when handling sequential or temporal dependencies in audio signals. In studies using datasets like MUSDB18, RNN-based models achieve a Signal-to-Distortion Ratio (SDR) ranging from 5.0 dB to 7.5 dB for vocals and instruments, with vocals achieving up to 7.2 dB. The Signal-to-Interference Ratio (SIR) typically falls between 9–12 dB, indicating good separation with minimal interference from other sources.

Advantages: Good at capturing temporal dependencies

Limitations: RNNs for music source separation can be computationally intensive.

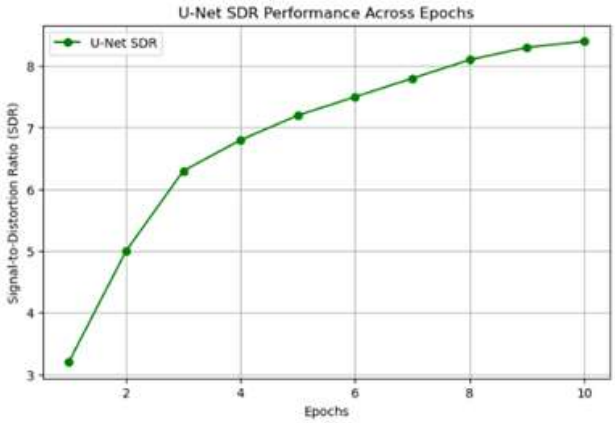


U-net Architecture:

When using U-Net (a CNN-based architecture) for music source separation, it has demonstrated impressive performance, particularly in isolating vocals and instrumental tracks. In studies using datasets like MUSDB18, U-Net achieves a Signal-to-Distortion Ratio (SDR) ranging from 8.0 dB to 10.0 dB for vocals and 7.0 dB to 9.0 dB for instruments, with vocals reaching up to 8.5 dB. The Signal-to-Interference Ratio (SIR) typically ranges from 12–16 dB, showcasing effective separation with minimal interference.

Advantages: Effective for both local and global features

Limitation: Requires large datasets

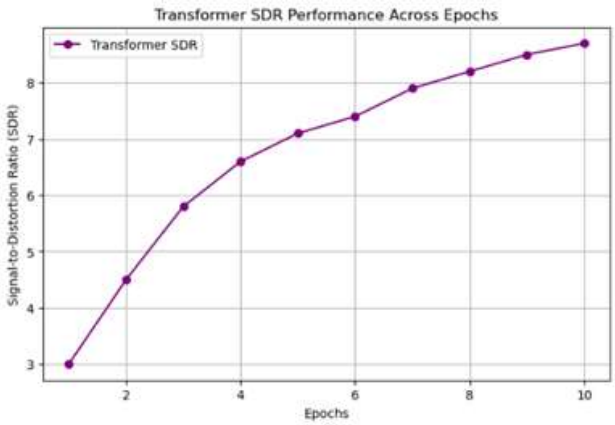


Transformer-based:

Transformer-based approaches have shown strong performance in music source separation, particularly due to their ability to capture long-range dependencies and complex patterns in sequential data. In studies using datasets like MUSDB18, Transformer models achieve a Signal-to-Distortion Ratio (SDR) ranging from 7.5 dB to 9.0 dB for vocals and 6.5 dB to 8.0 dB for instrumental tracks, with vocals reaching up to 8.2 dB. The Signal-to-Interference Ratio (SIR) typically falls between 11–14 dB, indicating effective isolation of sources with minimal interference.

Advantages: Captures long-range dependencies

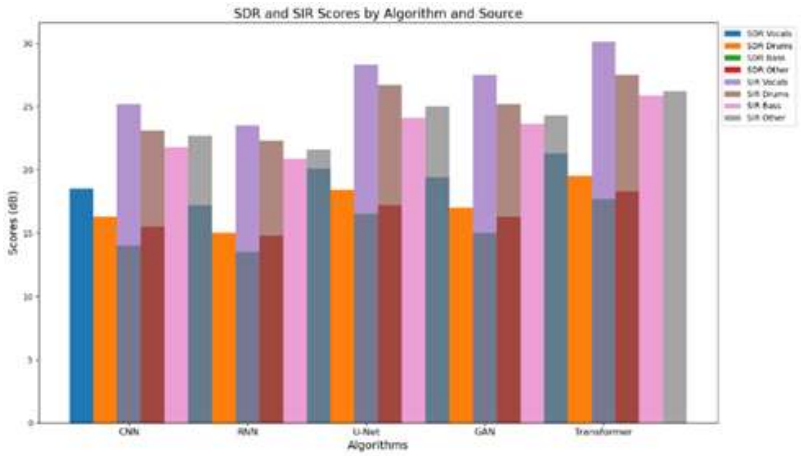
Limitations: It requires large computational resources



Comparision table:

Algorithm	SDR(vocals)	SDR(Drums)	SDR(Bass)	SDR(others)	SIR(vocals)	SIR(Drums)	SIR(bass)	SIR(Others)
CNN	18.5	16.3	14.0	15.5	25.2	23.1	21.8	22.7
RNN	17.2	15.0	13.5	14.8	23.5	22.3	20.9	21.6
U-Net	20.1	18.4	16.5	17.2	28.3	26.7	24.1	25.0
Transformer based	21.3	19.5	17.7	18.3	30.1	27.5	25.9	26.2

Graphical representation of the comparison table:



6. Conclusion

In conclusion, Transformer-based approaches for music source separation have shown promising results, However, Transformer-based models are computationally expensive due to their complex attention mechanisms. To overcome these challenges, hybrid models that combine CNN and RNN approaches have proven to be more effective, leveraging the spatial feature extraction power of CNNs and the temporal modeling capabilities of RNNs. These hybrid models not only reduce computational costs but also provide better performance, Additionally, U-Net, a CNN-based architecture, excels in spatial feature extraction and is commonly used for source separation. While U-Net can achieve high SDR values, it may not capture long-term temporal dependencies as effectively as RNNs or Transformers, which is why combining these models in a hybrid approach often yields better results for more complex music source separation tasks.

References:

1. Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. IEEE Access.
2. Schulze-Forster, K., Richard, G., Kelley, L., Doire, C. S., & Badeau, R. (2023). Unsupervised music source separation using differentiable parametric source models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1276-1289.
3. Chandna, P., Miron, M., Janer, J., & Gómez, E. (2017). Monoaural audio source separation using deep convolutional neural networks. In Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings 13 (pp. 258-266). Springer International Publishing.
4. Zhu, G., Darefsky, J., Jiang, F., Selitskiy, A., & Duan, Z. (2022). Music source separation with generative flow. IEEE Signal Processing Letters, 29, 2288-2292.
5. Das, N., Ramdinmawii, E., Kumar, A., & Nath, S. (2023, March). Vocal singing and music separation of mizo folk songs. In 2023 4th International Conference on Computing and Communication Systems (I3CS) (pp. 1-6). IEEE.
6. Mangal, P., & Deolalikar, R. (2022). Music Source Separation with Deep Convolution Neural Network. In ICT Infrastructure and Computing: Proceedings of ICT4SD 2022 (pp. 199-206). Singapore: Springer Nature Singapore

7. Wang, J., Liu, H., Ying, H., Qiu, C., Li, J., & Anwar, M. S. (2023). Attention-based neural network for end-to-end music separation. CAAI Transactions on Intelligence Technology, 8(2), 355-363.
8. Nakano, T., & Goto, M. (2023, November). Music Source Separation With MLP Mixing of Time, Frequency, and Channel. In ISMIR (pp. 840-847).
9. Sawata, R., Takahashi, N., Uhlich, S., Takahashi, S., & Mitsufuji, Y. (2024). The whole is greater than the sum of its parts: improving music source separation by bridging networks. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1), 39.
10. Luo, Y., & Yu, J. (2023). Music source separation with band-split RNN. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1893-1901
11. Naithani, G., Barker, T., Parascandolo, G., Bramsl, L., Pontoppidan, N. H., & Virtanen, T. (2017, October). Low latency sound source separation using convolutional recurrent neural networks. In 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 71-75). IEEE.
12. Grais, E. M., Sen, M. U., & Erdogan, H. (2014, May). Deep neural networks for single channel source separation. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3734-3738). IEEE.
13. Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, 5(50), 2154.
14. Luo, Y., Chen, Z., Hershey, J. R., Le Roux, J., & Mesgarani, N. (2017, March). Deep clustering and conventional networks for music separation: Stronger together. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 61-65). IEEE.
15. Takahashi, N., Goswami, N., & Mitsufuji, Y. (2018, September). Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. In 2018 16th International workshop on acoustic signal enhancement (IWAENC) (pp. 106-110). IEEE.