Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 5, October 2025, pp. 3903~3912
ISSN: 2302-9285, DOI: 10.11591/eei.v14i5.9962

Feature separation of music across diverse dataset: a comparative perspective

Sakthidevi Shunmugalingam Parvathi[1], Divya Chandrasekar[2]
[1]Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India
[2]Centre for Information Technology and Engineering, Faculty of Science, Manonmaniam Sundaranar University, Tirunelveli, India

Article Info

Article history:

Received Jan 28, 2025
Revised Aug 18, 2025
Accepted Sep 11, 2025

Keywords:

Application
Audio datasets
Feature extraction
Metrices
Neural networks

ABSTRACT

In music, feature separation is the process of separating distinguishable auditory characteristics, such as pitch, timbre, rhythm, and harmonic content, from a complicated, mixed signal. Virtual reality (VR), gaming, music transcription, karaoke systems, audio restoration, music information retrieval (MIR), music education, and audio forensics, are just a few of the areas where the topic has attracted a lot of attention. Feature extraction is crucial in music separation as it identifies and isolates sound elements, improving accuracy, and reducing noise. It simplifies raw audio into meaningful data for efficient processing and effective model learning. Without it, clean separation of audio components is very difficult. In this research, extracting features from mixed audio sources enables clean and accurate isolation of musical elements, enhancing quality, supporting precise evaluations, and boosting neural network performance across varied datasets including DSD100, MUSDB, and MUSDB18-HQ, which collectively afford rich musical content for making evaluations and benchmarks. Evaluation metrics, such as F1-score, precision, and recall, are utilized to demonstrate the performance data of the extracted features. The MUSDB18-HQ dataset yielded an overall increase of 17.86% in the F1-score metrics with significant increases in drums (+25.05%) and vocals (+20.04%), showing that the dataset was highly effective for feature separation.

Corresponding Author:

Sakthidevi Shunmugalingam Parvathi
Centre for Information Technology and Engineering, Manonmaniam Sundaranar University
Tirunelveli, Tamil Nadu, India
Email: sakthidevisp@gmail.com

1. INTRODUCTION

Feature separation in music involves separating a composite sound into its various elements such as pitch and timbre to analyze and manipulate each component separately [1]. Feature separation makes many modern audio applications possible, from improving music information retrieval (MIR), instrument recognition, genre classification, better transcription accuracy, and adaptive learning applications in music education. A feature-based approach remains relevant as it simplifies complex audio data into interpretable components, enhancing separation accuracy and computational efficiency. By reducing noise and improving model learning, it ensures cleaner isolation of musical elements across diverse datasets. This approach also supports robust evaluation, making it indispensable for relevant applications.

Recent breakthroughs in deep learning [2] and advances in signal-processing techniques [3] allow many modern techniques to access existing neural network architectures to learn complex feature representation from the data directly. By leveraging multiple richly annotated datasets in the training of these models, they have achieved previously unrealized accuracy in the extraction of audio features even from

heavily mixed audio. As a result, feature separation has served as a foundation for intelligent audio systems from everything digital, real-time performance feedback, interactive virtual reality (VR) soundscapes, and advanced forensic audio analysis [4], to extensive workflows in customizable music production. This research aims to evaluate the effectiveness of feature separation in music across diverse datasets. It seeks to benchmark performance using precision, recall, and F1-score to identify which dataset best supports accurate separation of audio features. The challenges of feature separation are discussed below.

Challenges feature separation with music poses a number of challenges, many of which affect the ability and speed of extracting meaningful features from audio signals [5], [6]. Even when you extract a meaningful feature, often they contain distortion or low fidelity, which consequently reduce the reliability of the feature for your downstream task [7]. Also, the variabilities in recordings [8] through various genres, recordings, instrumentations, recording setups and effects introduces instability to the meaning of features, even if they are based on the same pattern. The amount of processing needed for separating the desired features in real-time [9] while performing gigs, being an important part of performance or in someone playing an adaptive music system is normally quite high as well; often requiring very high amounts of processing and random-access memory (RAM), causing access issues for audio performers alike. The expense to set up deep learning models that have the ability to separate multiple meaningful features can be significant, and are limited even more by the unavailability of large diversified datasets of labelled and annotated features; downloadable datasets are often incomplete which also skews the viability of the machine learning, limiting the relevance of the extracted features, even in terms of what features to extract can sometimes be subjective, based on arbitrary criteria; moreover, feature sets could vary widely from one context to another [10]. Using separated features may also create a few legal issues surrounding unauthorized remixing, reproduction and/or analysis of copyrighted material. Overcoming these barriers necessitates improvements in algorithm design, more data availability, and trade-offs between existing technology and ethical realities.

2. ANALYSIS OF MUSIC ISOLATION APPLICATIONS

Applications supported by music separation include audio repair, karaoke systems, music remixing, and cleaning. Additionally, it facilitates the following activities: music education and practice, forensic audio analysis, music transcription, VR and gaming, musician performance analysis, music sampling and licensing, and MIR. The use of artificial intelligence (AI) and deep learning ensures more accurate and efficient handling of complex audio.

Table 1 (see in Appendix) [11]-[19] investigates different applications of music isolation technology, offering a detailed comparison about the specific purpose of each application and the potential benefits of applications: ranging from the enhancement of creativity to audio quality improvements. The table also mentions the limitations of the applications.

3. DATASET ACQUISITION

Choosing a dataset for music separation is arguably the most important step in determining if the results are accurate and meaningful. It is imperative that the dataset is diverse, high quality, and representative of the audio contexts it is trying to solve. Information regarding the datasets, DSD100, MUSDB, MUSDB18, and MUSDB18-HQ will be presented below.

The Table 2 provides an overview of four popular music datasets used in music source separation research: DSD100, MUSDB, MUSDB18, and MUSDB18-HQ. It highlights the availability of isolated stems for specific musical components (bass, drum, vocal, and other) and whether the dataset includes mixtures. '✓' indicates the availability of a specific component or mixture in the dataset, while a '-' signifies that the component is not provided. This comparison helps researchers select the most suitable dataset based on the requirements of their music separation tasks.

Table 2. Overview of music dataset

Dataset	Bass	Drum	Vocal	Other	Mixtures
DSD100	✔	✔	✔	✔	✔
MUSDB	✔	✔	✔	✔	✔
MUSDB18	-	-	-	-	✔
MUSDB18-HQ	✔	✔	✔	✔	✔

3.1. Dataset description

The demixing secrets dataset 100 (DSD100) is a dataset with 100 entire length music recordings that was created for music split-source research. The four components of each track—vocals, drums, bass, and other instruments—are given independent stems and are supplied as a blend. The dataset has been separated into 50 testing tracks and 50 training tracks, each in 44.1 kHz high-quality WAV format. DSD100 is widely used to benchmark source separation algorithms, as it offers a diverse collection of genres and ensures a standardized framework for evaluating separation performance [20], [21]. MUSDB is a widely used dataset regarding music isolation. It includes 150 entire length compositions from four different genres: hip-hop, pop, rock, and electronic. Separated into 50 testing and 100 training tracks, the dataset offers separated stems for bass, drums, vocals, and other components in addition to high-quality, 16-bit WAV files at 44.1 kHz. It is specifically designed to support research in separating different musical sources from mixed audio tracks, offering both mixed and individual source files for comprehensive evaluation in various music separation tasks [22]. The MUSDB18 dataset is one of the most recognized benchmarks in the field of music separation research. It consists of 150 full tracks in four different genres: hip-hop, pop, rock, and electronic. It has 50 evaluations and 100 training tracks; all provided in high-resolution a 16-bit WAV format at 44.1 kHz. Each file includes individual stems for bass, drums, vocals, and other instruments, allowing researchers to work with specific elements in a mixed audio context. MUSDB18 is a valuable resource for both academic research and practical music processing, given its popularity in training and evaluation of source separation algorithms [23]. The high-quality MUSDB18 dataset, known as the MUSDB18-HQ dataset, is frequently utilized in the field of music source separation research, containing the same 150 full length tracks in genres such as pop, rock, electronic, and hip-hop. The difference with MUSDB18-HQ is that the tracks are high-resolution uncompressed WAVs unlike the original MUSDB18 dataset. Since many separation applications require high fidelity audio for effective processing, having access to high-resolution files is helpful where the sound quality is a paramount reason to ensure the source separation is conducted effectively. The dataset includes separated stems for voices, other instruments, bass, and drums, as well as mixed audio tracks, supporting a range of music separation and evaluation tasks. Its higher quality makes MUSDB18-HQ particularly suitable for research where audio quality is critical [24].

4. METHOD

The method entails extracting comprehensive audio features from preprocessed datasets and subsequently training a deep learning model for multi-class classification using standardized features, as detailed in the following section.

The Figure 1 presents the overall processing workflow which shows the process starting with a mixed audio file containing overlapping sources, which is converted into a spectrogram using Mel-frequency cepstral coefficients (MFCC) [4] or short time fourier transform (STFT) [25] or time-domain waveform [26], [27] to obtain the depiction of time frequency. Deep learning models, such as convolutional neural network (CNN) [28] or transformer-based architectures [29] extract distinguishing features from the spectrogram. These features are utilized specifically in music separation methods, to obtain separated audio components such as bass, drum, vocals, and other instruments. Additionally, the separated sources can then be reconstructed as the audio signal using the inverse short-time fourier transform (ISTFT) to obtain the output extracted audio sources. Figure 2 illustrates feature classification pipeline where raw audio is converted into spectrograms, followed by extraction of features like MFCC, Chroma, and Tonnetz. A SoftMax classifier is the result of feeding these features into a neural network that has several ReLU and dropout layers. The 80-20 train-test split is used to train the model for 100 epochs. Finally, the system classifies the four categories features like bass, drum, vocal, and other.

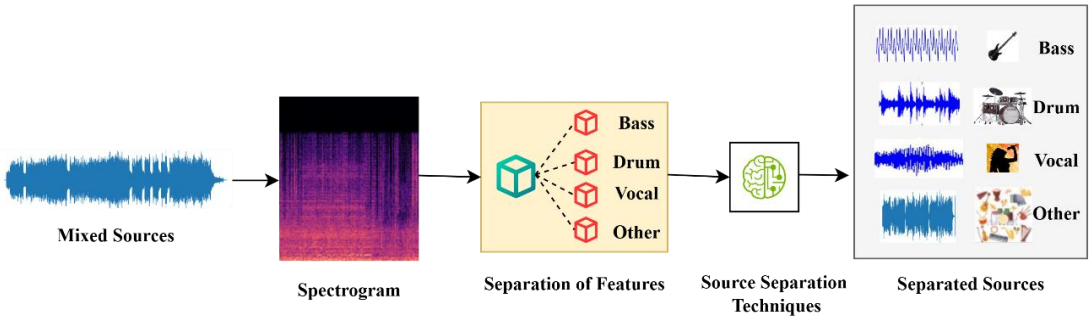

Figure 1. workflow of music separation

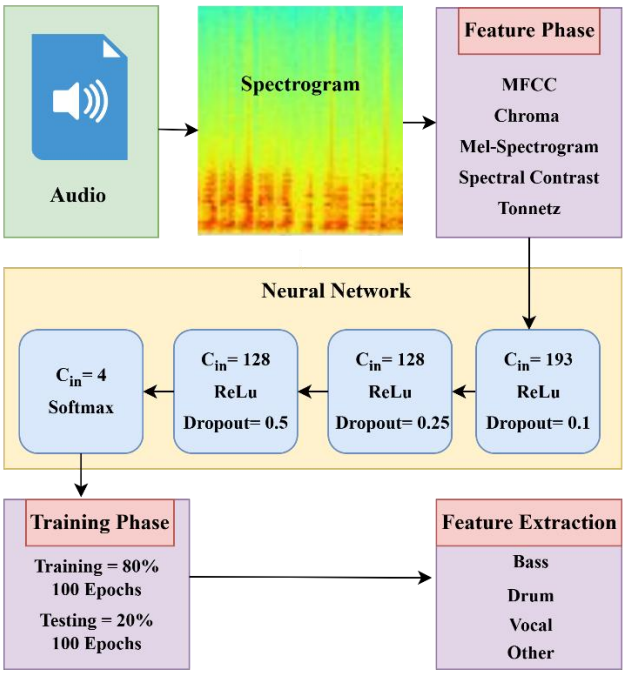

Figure 2. Neural network-based feature separation

4.1. Feature separation techniques

MFCC frequently used in speech and music recognition tasks because they can capture the timbre of audio. It emphasises perceptually relevant frequencies by applying a Mel filter bank to the audio signal's power spectrum. The discrete cosine transform (DCT) is then used to convert this filtered spectrum into the cepstral domain, producing a compressed representation of the spectral envelope. In order to accurately depict the audio characteristics, 40 MFCC features are usually extracted. Chroma characteristics show the harmonic and tonal content of audio by mapping frequencies onto 12 semitone bins that correspond to musical pitches. This makes them particularly useful for recording harmonic frameworks and chord progressions in music. The features are typically derived from the STFT of the audio signal and consist of twelve chroma features, one for each chroma bin. Mel-spectrogram captures the spectral energy distribution of audio on a perceptual Mel scale, which approximates how humans perceive frequency. It works by computing the power spectrogram of the audio signal and then mapping it onto the Mel scale using a filterbank. This representation is widely used in deep learning models for various audio processing tasks. The default number of Mel bands, usually 128; this determines how many features are extracted. Spectral contrast, where one measures the contrast between peaks and valleys in the frequency spectrum, is a helpful technique for identifying different musical instruments, or textures, by separating the frequency spectrum into sub-bands and calculating the contrast between the loudest component and the softest component for each sub-band. Tonnetz features express the tonal connections in audio, such as key and harmony, by projecting the harmonic content into a 6-dimensional tonal space. These features are particularly useful for tasks like chord recognition and key detection. The input is derived from the harmonic part of the audio signal, which can be isolated. A total of 6 Tonnetz features are extracted to capture these tonal characteristics. These audio features were selected as they represent distinct but complementary aspects of sound: MFCC capture the timbre of audio by modeling the spectral envelope on a perceptual Mel scale, making them essential for distinguishing voices, instruments, and phonemes. Chroma features map frequencies to 12 semitone bins, enabling robust detection of harmonic structures and chord progressions regardless of octave shifts. Mel-spectrograms preserve detailed time–frequency energy patterns on the Mel scale, providing rich input for deep learning models. Spectral contrast measures the difference between spectral peaks and valleys, aiding in distinguishing instrument types and textures. Tonnetz features encode tonal relationships in a 6D space, supporting key and harmony recognition. Together, these features complement each other by capturing timbre, pitch, harmony, texture, and detailed spectral dynamics, providing a comprehensive representation for music and speech analysis. Their combination ensures that both perceptual and structural elements of audio are captured, improving model accuracy. This makes them well-suited for deep learning tasks like source separation and transcription.

4.2. Design and training of a neural network

The model is designed using a sequential architecture, forming a simple feed-forward neural network. The input layer consists of 193 neurons, corresponding to the 193 extracted audio features. It includes two hidden layers, each with 128 neurons and ReLU activation functions [27] to introduce non-linearity. To prevent overfitting, dropout layers are applied after the hidden layers, randomly dropping 10%, 25%, and 50% of the neurons during training. Four neurons in the output layer have a softmax activation function [30], which allows for multi-class classification into groups like drums, vocals, bass, and others. The Adam optimizer enabled effective training, dropout layers were included for regularization to reduce overfitting, and categorical cross-entropy was employed as the loss function. The model was trained with a batch size of 256, and early stopping tracked the validation loss and stopped training if no progress was seen in 100 epochs.

Vocal features include pitch (the degree to which a sound is perceived as low or high), recognition (the ability to identify gender), timbre (quality of a tone), dynamics (loudness or softness), onsets and offsets (the timing of notes), and harmonics (integer multiples of the fundamental frequency). The many features of vocal health include Jitter and Shimmer in pitch and amplitude. Bass features are characterized as frequency range (typically below 250 Hz), pitch (clear pitch contour which defines harmonic progression), timbre (defined tone quality, usually captured by MFCC), rhythm and onset (timing, and a significant contributor to rhythmic structure), and dynamics (loudness). Other features such as harmonic content (strong fundamentals and active harmonics) and correlation with other instruments (often collabs with drums or rhythm guitars) help to identify features of the bass as well. Drum features are defined by transient nature (sharp attacks to clearly distinguish them from sustained instruments), frequency range (kick- lows, snare- mid range, hi-hat-higher range, toms- mid range), rhythmic patterns (often specific grooves and rhythms). Other noteworthy features of the drum include onset detection (the placement in time of hits, using spectral flux and zero crossing rate to detect hits) as well as timbre (the quality of a sound that distinguishes it), often captured with MFCC, and also dynamics (the loudness of hits can vary to express uniqueness). Other instruments have frequency ranges that can be divided into low-heavy, medium-vocals, and high-medium-cymbals, violins, and flutes are examples of high; guitars, pianos, and saxophones as an example of mid; and bass guitars, kick drums, and tambourines as lower.

4.3. Training the dataset

The DSD100, MUSDB, and MUSDB18-HQ datasets were used for training, with 50 audio files selected for each class bass, drums, vocals, and others and all audio files resampled to a uniform rate of 44.1 kHz for consistency. From each file, a comprehensive set of features was extracted, such as Tonnetz, MFCCs, Chroma, Mel-spectrogram, and Spectral Contrast, forming 193-dimensional feature vectors. Labels were assigned accordingly and one-hot encoded to fit a softmax classification framework, producing target arrays with shape (number of samples 4). 80% of the data was used for training, and 20% was used for testing, and all features were standardized using StandardScaler for zero mean and unit variance to ensure stable model convergence. While the MUSDB18 dataset was excluded from training due to the absence of isolated sources, the DSD100, MUSDB, and MUSDB18-HQ datasets, which contain isolated instrument stems, were effectively employed. To switch between datasets during experimentation, only the folder path needs to be updated.

5. EVALUATION OF FEATURE SEPARATION OF MUSIC

Assess the capability of feature separation using DSD100, MUSDB, and MUSDB18-HQ datasets. The assessment is carried out using, precision, recall, and F1-score with consideration to accuracy, macro averages and weighted averages. The results for each dataset are evaluated thoroughly and highlight the key performance relating to the feature extraction.

Table 3 displays the DSD100 dataset showed moderate precision for bass and drums (0.775), while vocals had the highest precision (0.85). All classes indicated equivalent recall, meaning that relevant instances were consistently retrieved. The F1-score was also similar to recall and precision and showed satisfactory performance for all categories; particularly, vocal classification had the greatest performance.

Table 3. Evaluation of the feature separation using DSD100 dataset

	DSD 100 dataset											
	Accuracy				Macro average				Weighted average			
	Bass	Drum	Vocal	Other	Bass	Drums	Vocal	Other	Bass	Drum	Vocal	Other
Precision	0.775	0.775	0.85	0.7	0.771	0.813	0.841	0.7	0.805	0.857	0.884	0.811
Recall	0.775	0.775	0.85	0.7	0.788	0.803	0.861	0.7	0.775	0.775	0.850	0.700
F1-score	0.775	0.775	0.85	0.7	0.759	0.775	0.836	0.7	0.774	0.791	0.857	0.711

According to Table 4, with respect to model evaluation, the MUSDB dataset showed improvements over DSD100 with improvements of note for precision scores in the category of other (0.875) bass (0.8) and a more consistent precision of 0.775 for vocals. Recall scores aligned with precision scores indicating balanced retrieval and the F1-score produced evidence to suggest that the model can distinguish most relevant features, harmonic and tonal in particular.

Table 4. Evaluation of feature separation using MUSDB dataset

	MUSDB											
	Accuracy				Macro average				Weighted average			
	Bass	Drum	Vocal	Other	Bass	Drums	Vocal	Other	Bass	Drum	Vocal	Other
Precision	0.8	0.7	0.775	0.875	0.810	0.731	0.771	0.861	0.858	0.792	0.809	0.897
Recall	0.8	0.7	0.775	0.875	0.836	0.747	0.778	0.889	0.800	0.700	0.775	0.875
F1-score	0.8	0.7	0.775	0.875	0.797	0.693	0.749	0.865	0.805	0.701	0.771	0.879

According to Table 5, the MUSDB18-HQ dataset performed the best of the three datasets with vocals performed the best in precision (0.9), followed by drums (0.875) and bass (0.95), while the "other" category performed well similarly at 0.825. High recall across all classes indicated reliable retrieval, and the F1-scores further confirmed the accuracy in classification of the dataset with bass, vocals, and drums having the most success in terms of feature separation.

Table 5. Evaluation of feature separation with MUSDB18-HQ dataset

	MUSDB18-HQ											
	Accuracy				Macro average				Weighted average			
	Bass	Drum	Vocal	Other	Bass	Drums	Vocal	Other	Bass	Drum	Vocal	Other
Precision	0.95	0.875	0.9	0.825	0.942	0.861	0.890	0.836	0.956	0.889	0.899	0.872
Recall	0.95	0.875	0.9	0.825	0.956	0.886	0.894	0.847	0.950	0.875	0.900	0.825
F1-score	0.95	0.875	0.9	0.825	0.946	0.867	0.889	0.820	0.950	0.876	0.897	0.834

6. DISCUSSION

A comparison study is presented between three significant datasets, DSD100, MUSDB, and MUSDB18-HQ, for audio feature classification. There are four categories of songs for audio feature classification: bass, drums, vocals, and others. To compare the performance of dataset DSD100, MUSDB, and MUSDB18-HQ analyzed their performance with metrics defined with Precision, Recall, and F1-score. Evaluating the datasets with regards to audio feature classification across four categories (bass, drums, vocals, and others): DSD100, MUSDB, and MUSDB18-HQ. While all datasets performed well for classifying audio features across varied levels, there was a meaningful difference through the datasets across categories.

MUSDB18-HQ yielded the best performance in the bass category with a precision, recall, and F1-score equal to 0.95, respectively. The MUSDB18-HQ significantly outperformed MUSDB and DSD100 which yielded a F1-score of 0.775. The robust F1-score indicates the MUSDB18-HQ has enhanced fidelity that allows accurate representation of the lower frequency components in the feature isolations. Compared to DSD100, the MUSDB18-HQ had improved F1-score by 22.58%, and over MUSDB improved 18.75% overall F1-score reflecting the expected superiority of MUSDB18-HQ separating bass signal. The drums categories MUSDB18-HQ achieved 0.875 for both precision and recall values, with a F1-score improvement of 25.05% over MUSDB (0.7), and 12.9% over DSD100 (0.775). The high expansion and coverage of MUSDB18-HQ likely yielded higher performance values as the method automatically reduced the number of false elements for this source group with periodic arrangements and variable rhythmic elements. Outperforming DSD100 and MUSDB in vocals category also were good metrics each respective created F1-score analysis 0.9, 0.85 for DSD100 and 0.775 for MUSDB, which created an F1-score improvement over DSD100 of 11.76% and 20.04% improvement over MUSDB. The high fidelity of the data set increased the possibility of isolating and confirming vocal components, portraying a valuable source for future study interested specifically in vocal separation activities. The classification grouping of other shows that MUSDB had a little better outcome of a F1-score of 0.8 shown as compared to DSD100 output of 0.775, and MUSDB18-HQ score was yet again improved with 0.875 F1-score.

Figure 3 displays the accuracy performance of the model across the four sources (bass, drum, vocal, and other) using three datasets (DSD100, MUSDB, and MUSDB18-HQ) for validation. The Y-axis represents accuracy (ranging from 0 to 1), while the X-axis represents the separated sources: bass, drum,

vocal, and other, with the performance of the three datasets represented by colored bars (Green: DSD100 Accuracy, Red: MUSDB Accuracy, Blue: MUSDB18-HQ Accuracy). We observe that MUSDB18-HQ is consistently the top performer with the highest accuracy scores for “bass” (~0.93), “drum” (~0.88), and “vocal” (~0.90), highlighting the advantage high-quality data affords for source separation evaluation. For the “other” category, MUSDB returns slightly higher performances than MUSDB and DSD100 (~0.88). However, in all cases, DSD100 performs the worst across all four sources. These findings highlight that dataset quality and characteristics strongly influence the accuracy of different sources, with MUSDB18-HQ being particularly effective for instrument-specific separation.

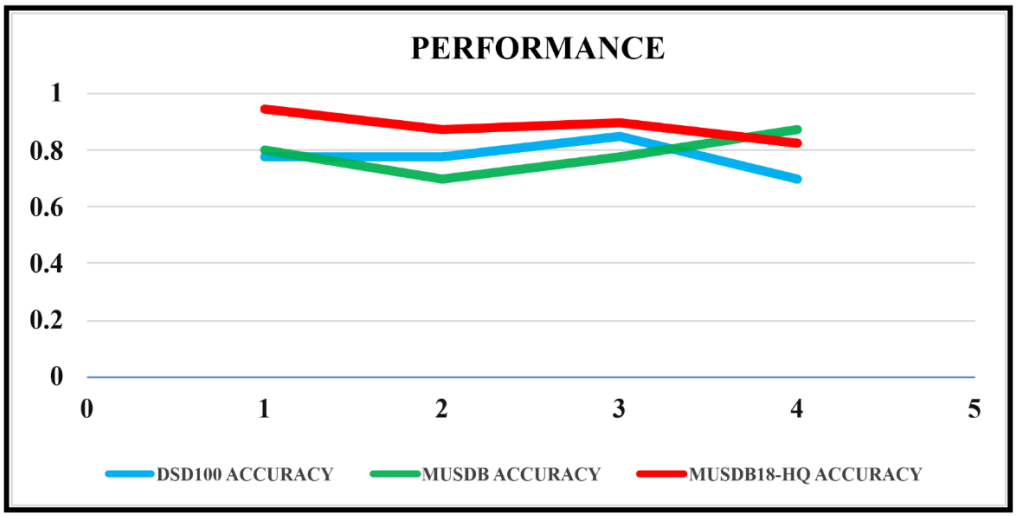

Figure 3. Feature separation performance for bass, drum, vocal, and other categories

7. CONCLUSION

This research establishes that MUSDB18-HQ can be considered the most rigorous and reliable dataset for the task, especially when it comes to bass, vocals, drums, and other. It significantly outperformed all other datasets based on all key metrics of F1-score, recall, and precision but especially when it comes to the bass, vocals, and drums datasets. These three categories exhibited high levels of capture and separation in audio that validated the potential of this dataset to effectively save audio components. Although overall, MUSDB had a slightly better performance on the "other" category, in overall F1-score improvements, and consistent across datasets, MUSDB18-HQ is more advantageous for being utilized across music remixing, audio restoration, and other music technology applications. This continues to show that MUSDB18-HQ has so much potential in improving the implementation of more effective music separation systems, as well as being able to apply to other various music practices. Future work will include experimenting and applying new feature extraction approaches, as well as refining the model architectures of the models as well as finding more datasets with diverse applications to further advance music source separation. The ongoing advancement in research will broaden the doors for creative and analytical approaches to music and audio processing across multiple industries.

ACKNOWLEDGMENTS

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

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sakthidevi Shunmugalingam Parvathi	✓	✓		✓		✓		✓	✓	✓				
Divya Chandrasekar					✓	✓	✓			✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] F. Alías, J. C. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Applied Sciences*, vol. 6, no. 5, pp. 1-44, 2016, doi: 10.3390/app6050143.

[2] J. Zhang, "Music Feature Extraction and Classification Algorithm Based on Deep Learning," *Scientific Programming*, vol. 1, pp. 1–9, 2021, doi: 10.1155/2021/1651560.

[3] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011, doi: 10.1109/JSTSP.2011.2112333.

[4] H. Dzulfikar, S. Adinandra, and E. Ramadhani, "The Comparison of Audio Analysis Using Audio Forensic Technique and Mel Frequency Cepstral Coefficient Method (MFCC) as the Requirement of Digital Evidence," *Jurnal Online Informatika*, vol. 6, no. 2, pp. 145–154, 2021, doi: 10.15575/join.v6i2.702.

[5] E. Yücesoy, "Gender Recognition Based on the Stacking of Different Acoustic Features," *Applied Sciences (Switzerland)*, vol. 14, no. 15, pp. 1-13, 2024, doi: 10.3390/app14156564.

[6] M. Mirbeygi, A. Mahabadi, and A. Ranjbar, "Speech and music separation approaches - a survey," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 21155–21197, 2022, doi: 10.1007/s11042-022-11994-1.

[7] S. George, S. Zielinski, and F. Rumsey, "Feature extraction for the prediction of multichannel spatial audio fidelity," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1994–2005, 2006, doi: 10.1109/TASL.2006.883248.

[8] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 424–434, 2008, doi: 10.1109/TASL.2007.909434.

[9] M. Mirbeygi, A. Mahabadi, and A. Ranjbar, "RPCA-based real-time speech and music separation method," *Speech Communication*, vol. 126, pp. 22–34, 2021, doi: 10.1016/j.specom.2020.12.003.

[10] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR*, 2017, pp. 316–323, doi: 10.48550/arXiv.1612.01840.

[11] M. A. M. Ramírez, W. H. Liao, G. Fabbro, S. Uhlich, C. Nagashima, and Y. Mitsufuji, "Automatic music mixing with deep learning and out-of-domain data," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022*, 2022, pp. 411–418.

[12] P. Patel, A. Ray, K. Thakkar, K. Sheth, and S. H. Mankad, "Karaoke Generation from songs: recent trends and opportunities," in *Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2022*, 2022, pp. 1238–1246, doi: 10.23919/APSIPAASC55919.2022.9980133.

[13] J.-M. Lemercier, J. Richter, S. Welker, E. Moliner, V. Välimäki, and T. Gerkmann, "Diffusion Models for Audio Restoration: A review*," IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 72–84, 2024, doi: 10.1109/MSP.2024.3445871.

[14] M. Furner, M. Z. Islam, and C. T. Li, "Knowledge discovery and visualisation framework using machine learning for music information retrieval from broadcast radio data," *Expert Systems with Applications*, vol. 182, pp. 1-11, 2021, doi: 10.1016/j.eswa.2021.115236.

[15] J. Nissen, "Aspirations and limitations: the state of world music education in secondary schools in multicultural Manchester," *British Journal of Music Education*, vol. 40, no. 3, pp. 385–396, 2023, doi: 10.1017/S0265051723000098.

[16] X. Gu, L. Ou, W. Zeng, J. Zhang, N. Wong, and Y. Wang, "Automatic Lyric Transcription and Automatic Music Transcription from Multimodal Singing," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 7, pp. 1–29, 2024, doi: 10.1145/3651310.

[17] J. G. R. Borquez, C. D. V. Soto, J. A. D. P. Flores, R. A. Briseño, and J. V. Aldás, "Neurogaming in Virtual Reality: A Review of Video Game Genres and Cognitive Impact," *Electronics*, vol. 13, no. 9, pp. 1-39, 2024, doi: 10.3390/electronics13091683.

[18] M. S. Zelenak, "Self-efficacy and music performance: A meta-analysis," *Psychology of Music*, vol. 52, no. 6, pp. 649–667, 2024, doi: 10.1177/03057356231222432.

[19] J. Watson, "Copyright and the Production of Hip Hop Music," *SSRN Electronic Journal*, 2024, doi: 10.2139/ssrn.4739736.

[20] W. H. Heo, H. Kim, and O. W. Kwon, "Source separation using dilated time-frequency DenseNet for music identification in broadcast contents," *Applied Sciences (Switzerland)*, vol. 10, no. 5, pp. 1-18, 2020, doi: 10.3390/app10051727.

[21] N. Takahashi and Y. Mitsufuji, "Multi-Scale multi-band densenets for audio source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE*, Oct. 2017, pp. 21–25, doi: 10.1109/WASPAA.2017.8169987.

[22] R. Zafar, A. Liutkus, F. Robert, A. Stöter, S. I. Mimilakis, and R. Bittner, "*The MUSDB18 corpus for music separation*," 2017.
[23] W. H. Heo, H. Kim, and O. W. Kwon, "Integrating dilated convolution into denseLSTM for audio source separation," *Applied Sciences (Switzerland)*, vol. 11, no. 2, pp. 1–19, 2021, doi: 10.3390/app11020789.
[24] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, *"MUSDB18-HQ - an uncompressed version of MUSDB18*," Zenodo, 2019.
[25] P. Magron, R. Badeau, and B. David, "Model-Based STFT Phase Recovery for Audio Source Separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 6, pp. 1091–1101, 2018, doi: 10.1109/TASLP.2018.2811540.
[26] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music Source Separation in the Waveform Domain," *arXiv preprint*, 2019, doi: 10.48550/arXiv.1911.13254.
[27] S. Sarkar, "Time-domain music source separation for choirs and ensembles," 2024.
[28] P. Mangal and R. Deolalikar, "Music Source Separation with Deep Convolution Neural Network," in *Lecture Notes in Networks and Systems*, pp. 199–206, 2023, doi: 10.1007/978-981-19-5331-6_21.
[29] S. Rouard, F. Massa, and A. Defossez, "Hybrid Transformers for Music Source Separation," *in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, IEEE*, Jun. 2023, pp. 1–5, doi 10.1109/ICASSP49357.2023.10096956.
[30] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-Margin Softmax Loss for Convolutional Neural Networks," in *Proceedings of the 33 rd International Conference on Machine Learning*, New York, NY, USA, 2016, pp. 1–10.

APPENDIX

Table 1. Impacts of music separation applications

Application	Description	Purposes	Benefits	Challenges
Music remixing [11]	Transforms existing musical compositions into fresh versions by modifying tempo, instrumentation, or structure and adding new elements.	Artistic expression club and DJ culture	Enhanced creative freedom. Flexible reimagining of music with isolated components.	Risk of losing the essence of the original track. Potential copyright and licensing challenges.
Karaoke systems [12]	It allows individuals to sing along to instrumental versions of songs with lyrics displayed on a screen, leveraging music extraction to create high-quality instrumental tracks.	Entertainment social bonding practice, and performance	Creation of clean, customizable instrumentals. Real-time volume control and dynamic adjustments.	Computationally demanding processes. Challenges in achieving artifact-free separation.
Audio restoration and cleaning [13]	Enhances the quality of audio recordings by removing noise and distortions while maintaining sound integrity through music extraction.	Removing noise, distortions restoring lost frequencies, and preserve originality	Targeted noise reduction with minimal impact on quality. Flexibility in recombining and using restored elements.	Risk of introducing artifacts during extraction. High computational requirements.
MIR [14]	Uses computational techniques to extract, analyze, and organize musical information, integrating music extraction for deeper analysis.	Music recommendation and analysis Music synthesis and composition	Enhanced accuracy in musical analysis. Broad application scope in research and industry.	Computationally intensive tasks. Difficulties in processing highly complex compositions.
Music education and practice [15]	Structured teaching and learning of music enhanced by tools that simplify complex compositions through music extraction.	Emotional expression and cultural understanding Creativity and imagination	Precision in analyzing musical elements. Accessibility to high-quality educational resources.	Dependency on technology for educational enhancements. Potential loss of traditional teaching nuances.
Music transcription [16]	Converting musical sounds into written notation enhanced by separating components for precision.	Arrangement and composition Education and learning	Precise identification of overlapping frequencies. Efficient handling of complex arrangements.	Dependence on accurate extraction for quality transcription. Limitations in processing certain intricate audio files.
Gaming and VR [17]	Music extraction enhances gaming and VR experiences by creating dynamic, adaptive soundscapes.	Real-time feedback Spatial audio for sound localization	Improved user experience and interactivity. Creative, adaptive soundscapes.	Technical challenges in real-time implementation. High resource and computational requirements.
Performance analysis for musicians [18]	Systematic evaluation of a musician's performance using extracted audio components.	Interpretation refinement Impact assessment	Data-driven, objective insights. Real-time feedback for precise corrections.	Relies on advanced music extraction tools. Costs associated with implementing technology.

Table 1. Impacts of music separation applications (*continued*)

Application	Description	Purposes	Benefits	Challenges
Music sampling and licensing [19]	Reusing segments of existing music and ensuring legal compliance through extraction techniques.	Creative expression Genre blending	Efficient identification of licensable materials. Efficient identification of licensable materials.	Legal disputes over unauthorized sampling. High costs and time for securing licenses.

BIOGRAPHIES OF AUTHORS

Sakthidevi Shunmugalingam Parvathi is a full-time Research Scholar at the Centre for Information Technology and Engineering, Manonmaniam Sundaranar University. She completed her postgraduate studies (M.Sc. in Information Technology) at the same university in 2023, following her undergraduate degree (B.Sc. in Information Technology) from Sri Ram Nallamani Yadava College of Arts and Science, affiliated with Manonmaniam Sundaranar University, in 2021. She authored one book chapter indexed in Scopus. Also, she authored one journal paper indexed in Scopus. As a Reviewer for Elsevier and Springer Publications, she has reviewed 23 research papers. Her research interests lie in the fields of deep learning and artificial intelligence. She can be contacted at email: sakthidevisp@gmail.com.

Dr. Divya Chandrasekar an Assistant Professor at the Centre for Information Technology and Engineering within Manonmaniam Sundaranar University, has made significant contributions to the field. She holds a Ph.D. from the same university and has authored more research papers in International/National Journals/Proceedings/Books. She actively participates in scholarly activities, serving as a reviewer for international journals and being part of editorial boards. Her current research interests include data analytics, cyber security, nanodevices and low power VLSI circuits wireless sensor networks, and communication networks. With a strong background in engineering, she continues to impact the academic community through her research and teaching. She was awarded the Young Scientists Fellowship by TNSCST. She can be contacted at email: cdivyame@gmail.com.