Wavelet Based Feature Extraction for Musical Source Identification

Aneesh Ojha
A69032335
anojha@ucsd.edu

Ketki Patankar
A69032361
kpatankar@ucsd.edu

Abstract

Musical mixtures contain overlapping instruments and vocals that occupy similar time–frequency regions, making source-related classification a challenging problem. Many instruments share frequency bands, limiting the effectiveness of simple spectral or energy-based separation. Moreover, music is highly non-stationary, and fixed-resolution transforms such as the STFT struggle to capture rapid transients alongside slowly varying components. To address these challenges, we employ a multi-resolution representation based on a 6-level Discrete Wavelet Transform (DWT) with Haar wavelets, which better captures both transient and steady-state structures in audio. Using the MUSDB18 dataset, we generate additional training examples by constructing controlled two and three-component mixtures from available stems (e.g., vocals + drums, bass + drums + other), providing a diverse set of polyphonic scenarios. Wavelet-domain features are fed into a CNN with residual (ResNet-style) blocks to perform multi-label classification of the stems present in each mixture. Our results show that wavelet representations combined with deep convolutional architectures enable reliable detection of multiple overlapping musical components within a single audio signal.

1 Introduction

Music signals are fundamentally complex, consisting of multiple sound sources that interact in rich and often unpredictable ways. Even within a single genre, instruments and vocals vary widely in their temporal behavior, harmonic structure, and dynamic range. These characteristics make it difficult not only to separate sources but even to determine which sources are present in a mixture. Reliable detection of musical components is valuable for downstream tasks such as automated mixing, music information retrieval, content-based recommendation, and instrument-aware audio effects.

A central difficulty in this setting is that the observable mixture reflects the combined behavior of several stems whose contributions may vary rapidly over time. This is further complicated by the fact that mixtures in real music are rarely clean additions of isolated elements; they include overlapping harmonics, correlated rhythmic patterns, and production artifacts. As a result, conventional source separation assumptions—such as harmonic stacking, sparsity, or independence—are frequently violated.

Recent work has explored deep learning approaches for identifying or isolating musical sources, often relying on time–frequency representations as network inputs. However, the effectiveness of such approaches strongly depends on how well the representation captures both transient and sustained elements under realistic polyphonic conditions. This motivates exploring alternative representations that retain fine temporal detail without sacrificing low-frequency resolution.

In this paper, we investigate the use of wavelet-domain features for multi-label detection of musical components. Rather than performing full source separation, we address the simpler but still challenging task of determining which stems (vocals, bass, drums, others) contribute to a given mixture. This

framing allows us to focus on discriminative structure in the representation itself, independent of explicit reconstruction constraints.

To support this study, we construct a dataset of controlled mixtures derived from MUSDB18 stems. By systematically combining two and three components per mixture, we create diverse examples that reflect common musical configurations while allowing precise control over source presence. These mixtures form the basis for training a convolutional model designed to recognize multiple overlapping sources.

Our experiments demonstrate that wavelet-based representations provide discriminative structure that supports accurate multi-label detection, especially in scenarios where traditional fixed-resolution transforms may fail. The results highlight the potential of wavelet-domain features as a foundation for robust, source-aware music analysis.

2 Related Works

Musical instrument recognition has been studied across a wide range of settings, from isolated notes to fully polyphonic real-world music. Early research focused primarily on clean, studio-recorded sounds, while more recent work has attempted to address the challenges of polyphonic mixtures. However, the task of identifying predominant instruments in professionally produced music remains largely underexplored.

Han et al. (1) proposed a deep convolutional neural network (ConvNet) framework for predominant instrument recognition in polyphonic music recordings. Their approach leveraged mel-spectrograms as input features and trained on single-labeled, fixed-length audio excerpts to identify multiple predominant instruments in variable-length music clips. The study demonstrated that very deep ConvNets could automatically learn effective spectral representations, outperforming previous state-of-the-art methods on the IRMAS dataset without relying on source separation. They also investigated different activation functions and aggregation strategies, showing that leaky ReLU improved performance over standard ReLU and that class-wise sum followed by normalization was more robust than majority voting for combining short-time window predictions. Despite these advances, their approach has limitations: training on single-labeled excerpts does not fully exploit the polyphonic nature of music, temporal overlap between instruments is not explicitly modeled, and the aggregation method does not adapt to instrument-specific characteristics, which may limit recognition accuracy in real-world, highly polyphonic recordings.

Hung and Lerch (2) proposed a multitask learning framework that integrates instrument activation detection (IAD) with music source separation to improve separation quality. Their model jointly learns to predict instrument activations and estimate source signals in an end-to-end manner, using the predicted activation labels during inference to weight time frames and suppress those not containing the target instrument. They evaluated the approach on six instruments using a combination of MedleyDB and Mixing Secrets datasets (referred to as MM), demonstrating that their multitask model outperforms the baseline Open-Unmix model on these datasets while maintaining comparable performance on MUSDB. However, the approach has some limitations: the multitask framework increases computational complexity, which may limit its applicability to real-time or large-scale music processing scenarios and the study lacks discussion on how the system behavior would change in the presence of noise or degraded audio quality.

Düzenli and Özkurt (3) investigated the use of discrete wavelet transform (DWT) and dual-tree complex wavelet transform (CWT) features for real-time speech/music discrimination (SMD). They extracted statistical features from wavelet coefficients and compared their performance with conventional time, frequency, and cepstral-domain features, using artificial neural networks (ANNs) for classification. Principal component analysis was applied to reduce feature dimensionality, and the Daubechies8 wavelet was found to yield the best performance among DWT options. Their results showed that DWT-based features outperformed CWT-based ones, and that ratio parameters contributed modestly (1–1.5%) to discrimination performance. The study demonstrated the feasibility of implementing a real-time SMD system using DWT features. However, the approach has limitations: the evaluation was limited to clean speech and music recordings without mixed or noisy scenarios, and the binary classification setup does not generalize to more complex real-world audio, suggesting the need for adaptive filter design and dataset expansion for multi-class or noisy conditions.

A multi-input deep convolutional neural network (deep-CNN) architecture that combines discrete wavelet transform (DWT) features with traditional spectral features was proposed by Dash et al. (4) for predominant instrument recognition in polyphonic music. Their model employs a dual-input approach, with statistical features taken from DWT-decomposed signals in the second input and Mel-frequency cepstral coefficients (MFCCs) and Mel-spectrograms in the first. For feature selection, particle swarm optimization (PSO) is used to eliminate unnecessary features and minimize dimensionality. The model outperforms the benchmark CNN model by 12.3% and 23.0% with micro and macro-level F1 scores of 0.695 and 0.631, respectively, following evaluation on variable-length, multi-labeled extracts from the IRMAS dataset and training on fixed-length, single-labeled data. The strategy has drawbacks despite these advancements: Its evaluation is restricted to the IRMAS dataset without testing in noisy or real-world polyphonic recordings; it depends on meticulous feature selection and hyperparameter adjustment, which may reduce generalizability.

Overall, these works demonstrate the effectiveness of deep learning and wavelet-based features for instrument recognition, source separation, and speech/music discrimination. Yet, challenges remain in handling noisy or overlapping audio, generalizing across diverse datasets, and designing adaptive, computationally efficient models suitable for real-time or large-scale applications.

3 Dataset

MUSDB18 (5) is the base dataset used for our study. It is one of the widely used datasets primarily used for music separation tasks.

3.1 Key Statistics and Features:

- Contains 150 diverse music tracks from various genres ,referred to as mixtures
- Each music track has a duration of 3-4 minute.
- Each mixture is a combination of 4 stems: drums, vocals, bass, other instruments
- Each mixture has its corresponding isolated stems
- High quality, clean stereo audio

The main reason behind selecting this particular dataset for our task of music source identification is the flexibility that it offers in recombining the isolated stems to create different level mixture components. In our work, we curated our own dataset by generating mono/ single stem audio, 2-stem mixtures and 3-stem mixtures for each music track in original dataset.

Our curated dataset contains the single/multi-stem audio track and the target label which is essentially a 4 dimensional binary array, indicating the presence of each of the 4 stems in the current audio sample.

4 Method

Our project is structured as a comparative study. Here, we have designed a neural network based model architecture that is evaluated using two different types of input. The model takes either the raw audio samples (RawWave-CNN) or their corresponding Discrete Wavelet Transform (DWT-CNN) as input. Following this, a thorough study is carried out to understand the performance of the two approaches.

4.1 Feature Extraction

The raw audio track itself is a high dimensional and highly redundant signal. As a result of which, directly extracting meaningful temporal or spectral structure from it, would be a complex task. To address this challenge, our study explores a more compact and informative representation of the audio signal, using Discrete Wavelet Transform (DWT).

While earlier studies in this field mostly used spectrograms from the Short-Time Fourier Transform (STFT), these representations have limitations because of their fixed time and frequency resolution. This makes them less effective at capturing both sharp transients and slowly changing harmonic

structures at the same time, which are common in musical mixtures. In contrast, the Discrete Wavelet Transform (DWT) provides a multi-resolution analysis that adjusts to the signal's frequency content.

DWT is particularly well-suited for our study, as it is capable of capturing non-stationary and transient signals like the musical tracks. It is capable of capturing high-frequency components such as drum hits, as well as low-frequency harmonic content like the bass. It does this through Multi-Resolution Analysis (MRA), where the signal is decomposed into different frequency bands/scales using an iterated filter bank structure repeated on the low-frequency channel. In contrast to the standard Fourier Transform, which uses infinite-duration sinusoidal waves, DWT uses a localized and finite duration basis which can be easily translated in time and stretched or compressed in frequency to be able to capture signals at different resolutions and time locations. For instance, to capture the high frequency drum hits or vocal onsets it uses a short-duration/fine-scale wavelet, while capturing the structure of bass it relies on longer-duration/coarse-scale wavelet.

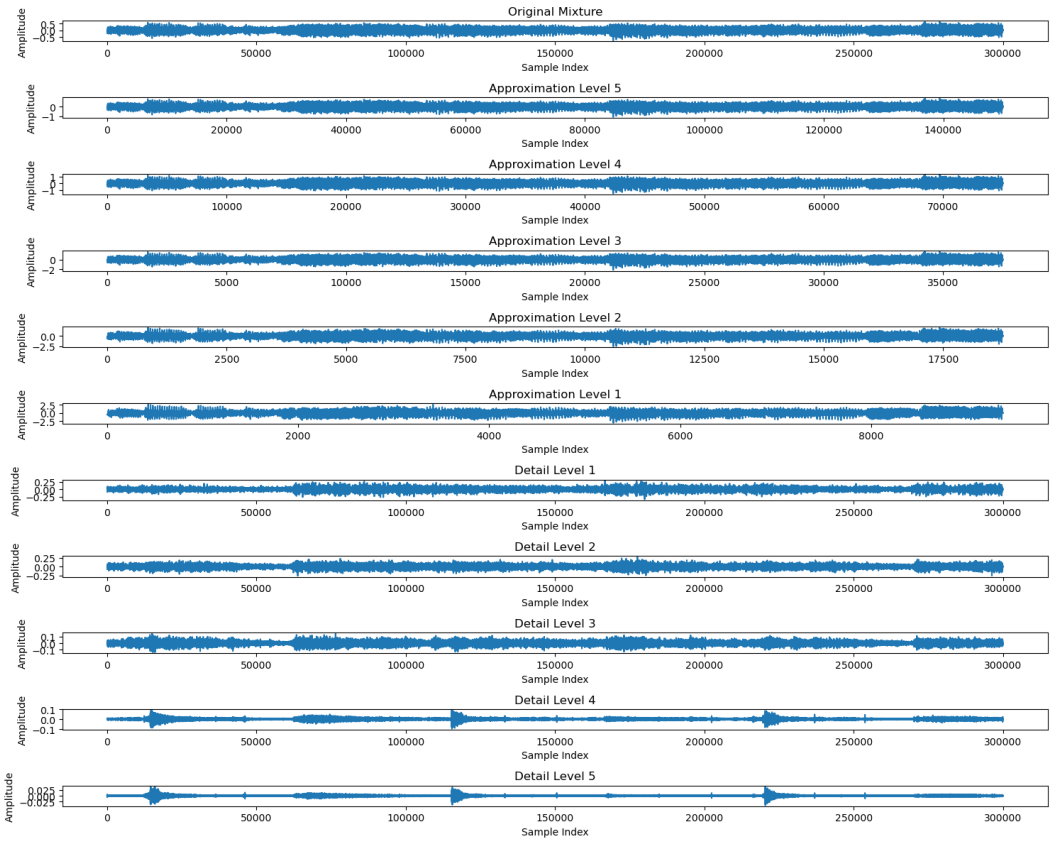

Figure 1: Waveforms of approximation and detail coefficients from MRA

On complete MRA analysis, we end up with lowest-resolution approximation coefficient and different scale wavelet/detail coefficients. Both the signal's high-frequency transients and low-frequency structure are captured by these coefficients. Figure 1 displays the waveforms for each of the approximation and detail coefficients for a sample audio stem, showing how different scales highlight different signal components.

Below section summarizes the related formulas used to calculate these coefficients.

Level-1 approximation coefficients (Haar DWT)

$$cA_1[k] = \sum_{n \in \mathbb{Z}} x_{\text{mix}}[n]\, \varphi_{\text{Haar}}(2k - n)$$

Level-1 detail coefficients (Haar DWT)

$$cD_1[k] = \sum_{n \in \mathbb{Z}} x_{\text{mix}}[n]\, \psi_{\text{Haar}}(2k - n)$$

6-level Haar DWT reconstruction

$$x_{\text{mix}}[n] = \sum_{k\in\mathbb{Z}} cA_6[k]\,\varphi_{\text{Haar},6}(n - 2^6 k) + \sum_{j=1}^{6}\sum_{k\in\mathbb{Z}} cD_j[k]\,\psi_{\text{Haar},j}(n - 2^j k)$$

Symbol Descriptions

- $x_{\text{mix}}[n]$: Discrete-time mono audio mixture at time index n.
- $cA_1[k]$: Level-1 approximation coefficient (low-frequency content) at index k.
- $cD_1[k]$: Level-1 detail coefficient (high-frequency / transient content) at index k.
- $cA_6[k]$: Level-6 approximation coefficient (coarsest, very low-frequency structure).
- $cD_j[k]$: Detail coefficient at scale $j \in \{1, \ldots, 6\}$ and index k.
- $\varphi_{\text{Haar}}(\cdot)$: Haar scaling basis function.
- $\psi_{\text{Haar}}(\cdot)$: Haar wavelet basis function.
- $\varphi_{\text{Haar},6}(\cdot)$: Haar scaling function at scale 6.
- $\psi_{\text{Haar},j}(\cdot)$: Haar wavelet at scale j.
- n: Time index in the discrete-time audio signal.
- k: Index of the wavelet coefficients at a given scale.
- j: Decomposition level (scale) in the 6-level DWT.

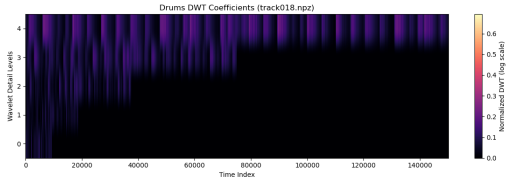

Figure 2: Wavelet coefficient visualization for Stem - Drums

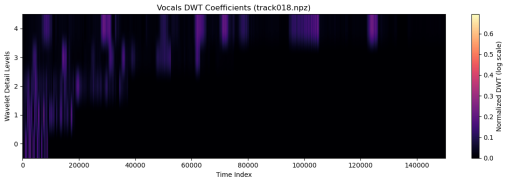

Figure 3: Wavelet coefficient visualization for Stem - Vocals

On further analyzing the wavelet coefficients for individual stems over time, we can observe a distinct frequency fingerprint for each stem. In the diagrams above, vertical lines correspond to transient or high-frequency events. In the plot for instrument drums, this may correspond to the drum hits while in the plot for vocals it may correspond to the vocal onsets.

As mentioned earlier, each of them have a distinct structure, like the drum hits appear bursty (see Figure 2) while vocals display more sustained energy (see Figure 3). This observation confirms DWT's capability of capturing instrument-specific structures from its time-frequency analysis.

4.2 Model Architecture

We have adopted a supervised learning approach on our dataset and designed a ResNet based architecture to map DWT features to stem-presence prediction. The image below shows an end-to-end model architecture that takes as input the one-dimensional DWT coefficients from the MRA and as output it generates a multi-label stem predictions. The first block in this architecture is the stem convolution block which directly operates on the DWT coefficients, converting input features into a feature map. During training, this block essentially learns a compact feature representation of input DWT array .

Following this, we have three Residual blocks stacked one after the other. For complex feature inputs like DWT, it is not surprising that the model might need multiple trainable layers to capture appropriate features for music source predictions. However, deeper networks face the problem of vanishing gradients. Hence, the use of ResNet architecture here is of vital significance to maintain gradient flow across deep neural networks. ResNets are characterized by skip/residual connections that carry the activation of a layer further ahead into the network, bypassing several layers via shortcut paths, that helps improve gradient flow.

Figure 4: DWT-CNN Model Architecture for Stem-Presence Prediction

Internally a Residual Block contains standard convolution and max-pooling layers. They progressively increase channels simultaneously, simultaneously reducing the temporal resolution. This allows the network to capture more features or patterns over longer time spans. The output of the residual blocks are then given to a global average pooling layer, which then collapses the time dimension into a single 64-dimensional feature vector. Each audio clip, regardless of its length, is reduced to this 64-dimension representation, making the model invariant to clip length. At the end, we have a fully-connected layer followed by a sigmoid activation producing four independent probabilities, one for each stem. They highlight the presence or absence of that stem in the input audio track. In this way, this model is able to identify presence of a particular musical source/ instrument in a mixture.

5 Model Evaluation

5.1 Training and Validation Loss Curves

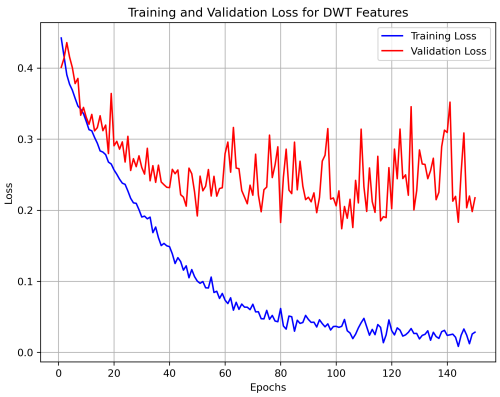

Figure 5: Training and Validation Loss for DWT-CNN

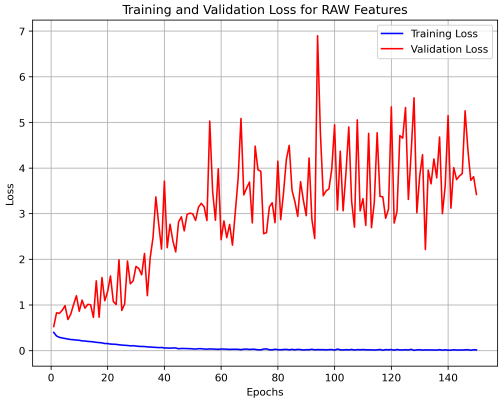

Figure 6: Training and Validation Loss for RawWave-CNN

To analyze the stability of the training, we closely observed the loss curves for training and validation datasets. From Figure 5 and Figure 6 we can observe that the DWT-CNN model leads to faster and more stable convergence as compared to RawWave-CNN. The model trained on DWT features achieves significantly lower loss values much earlier in training, indicating that this compact representation of the audio signal aids the model in learning meaningful patterns efficiently. In contrast, the raw audio input, being high dimensional and redundant, causes slower and more unstable training process. Raw waveform is extremely high-dimensional, noisy, and highly redundant, forcing the network to learn all relevant filters (time–frequency decomposition) from scratch. With a small

dataset, this causes overfitting: the training loss decreases, but the model fails to generalize, so the validation loss keeps increasing instead of converging.

5.2 Evaluation and Metrics

To evaluate the performance of our proposed DWT-CNN model for multi-label instrument recognition, we conducted experiments on the test set extracted from the MUSDB18 dataset. The model outputs a set of probabilities for each instrument in a music excerpt, which are converted into binary predictions using a threshold τ (default $\tau = 0.5$):

$$\hat{y}_{i,j} = \begin{cases} 1, & \text{if } p_{i,j} > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $p_{i,j}$ is the predicted probability for instrument j in sample i, and $\hat{y}_{i,j}$ is the corresponding binary prediction.

We use standard multi-label evaluation metrics to assess the performance of the model:

Accuracy (Subset Accuracy)

Measures the fraction of samples where all predicted labels exactly match the ground-truth labels:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{\mathbf{y}}_i = \mathbf{y}_i\} \tag{2}$$

where N is the number of test samples, $\mathbf{y}_i$ is the ground-truth label vector for sample i, and $\mathbf{1}\{\cdot\}$ is the indicator function.

Precision (Sample-wise)

Computes the fraction of correctly predicted labels among all predicted labels for each sample:

$$\text{Precision}_{\text{samples}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{\mathbf{y}}_i \cap \mathbf{y}_i|}{|\hat{\mathbf{y}}_i| + \epsilon} \tag{3}$$

Table 1: Performance comparison between DWT-CNN and RawWave-CNN models.

Model	Accuracy (%)	Precision	Recall	F1 Score
DWT-CNN	69.31	0.81	0.82	0.80
RawWave-CNN	41.86	0.66	0.83	0.69

As observed in the Table 1, the model trained on raw audio input suffers with low accuracy as well as F1 score as it has to learn the complex time-frequency structure of the audio signal from scratch. This is considerably difficult as this audio input is a high dimensional complex data. When we switch to DWT-CNN we see better performance as we use the more-informative, compact and easy to learn DWT representation that has already encoded the time-frequency information. Hence, we can conclude by this analysis that DWT features not only make learning easier, but also give consistently better performance for instrument identification task.

5.3 Rationale for Sample-wise Metrics

In multi-label instrument recognition, each music excerpt may contain a variable number of active instruments. Sample-wise metrics evaluate performance at the level of individual music excerpts, capturing the ability of the model to correctly predict the full set of active instruments per sample. This is more informative than label-wise metrics, which compute averages across instruments and may overestimate performance when some instruments dominate the dataset. Sample-wise evaluation

thus provides a realistic measure of the model's performance on polyphonic music, where the co-occurrence and overlap of instruments vary widely.

In addition to quantitative evaluation, we inspect a few predictions to qualitatively verify the model's behavior on real-world polyphonic music excerpts.

5.4 Noise Sensitivity Analysis

In our study, we have not restricted ourselves to just model architecture and performance evaluation, but have also performed a thorough noise sensitivity analysis for our classifier. With the aim to assess our stem presence prediction classifier's robustness to noise, we added additive white gaussian noise at a range of SNR levels spanning 30 dB, 40 dB and 50 dB to the audio mixtures. The model was then re-evaluated to check whether it could still produce reliable instrument presence predictions under noisy recording environments, in contrast to the perfectly clean studio-quality audio in the dataset.

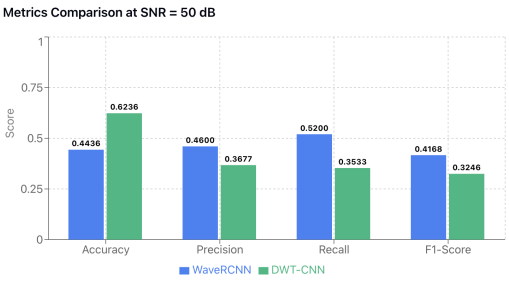

Figure 7: Performance Comparison at SNR of 50 dB

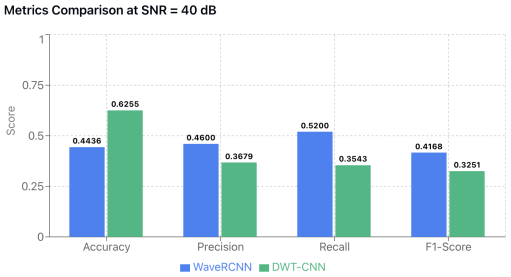

Figure 8: Performance Comparison at SNR of 40 dB

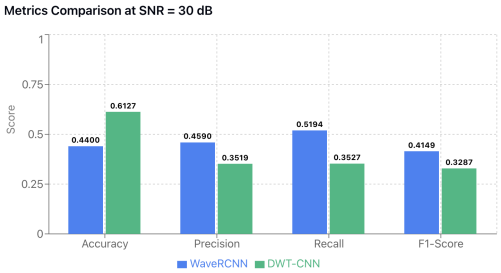

Figure 9: Performance Comparison at SNR of 30 dB

Figure 10: Performance metrics across different SNR levels for RawWave-CNN and DWT-CNN models.

On taking a closer look at the results in Figure 7 , Figure 8 and Figure 9 we can confirm that the DWT-based model keeps relatively high accuracy across multiple noise levels. This indicates that the wavelet features are effective in preserving the instrument-specific structure even in the presence of noise.

As the same time, as observed, the DWT model's precision, recall and F1 score remain consistently lower than for the raw audio model across all SNR level. This is mainly because DWT is characterized by implicit denoising as it compresses and smooths out the audio signal, so broadband noise is suppressed more effectively in noisy conditions. This inherent denoising itself causes suppression of weaker signals (short instrument events) along with noise. This leads to more missed detections, thereby reducing both the recall and F1 score. This results into loss of some details in the DWT signal, which makes the model less discriminative and hence results into a lower precision.

6 Novelty

- Most prior work in Music Source Analysis has focused on full source separation. Our study shifts the focus to a relatively unexplored task: instrument presence identification. Interestingly, tackling this downstream task can also support the more complex challenge of source separation.
- Unlike earlier methods that depend on raw audio signals or spectrogram-based transforms such as the STFT, our approach uses one-dimensional DWT features to obtain a much more compact representation of the signal. This preserves the essential time–frequency structure while significantly lowering computational load. In contrast, training directly on raw waveforms is highly complex, resource-intensive, and slow, making it impractical for real-time or deployment-oriented systems.
- We use a compact ResNet for our analysis, enabling faster inference compared to large spectrogram-based CNNs.
- We have examined the model's sensitivity to Gaussian noise as an extension of our initial efforts.

7 Conclusion

- Specific to the task of music source identification, faster and more consistent training is made possible by DWT's compact, informative time–frequency representation, which minimizes redundancy and helps the model rapidly pick up important patterns.
- Model performance is greatly improved with DWT-based input: Accuracy (0.69 against 0.41) and F1-score (0.80 vs. 0.69) are much greater than using raw audio, suggesting that wavelet characteristics provide more discriminative knowledge.
- Across all SNRs, DWT-CNN surpasses WaveRCNN in accuracy. At the same time, it has lower sensitivity to weaker features, therefore lowering recall and F1-score.

8 Appendix

The implementation code for our work is available at `https://github.com/anojha12/WaveletBased_MusicalSource_Identification.git`, and the MUSDB18 dataset used in our experiments can be accessed at `https://zenodo.org/records/1117372`.

9 Contributions

Table 2: Team Member Contributions

Task	**Aneesh**	**Ketki**
Conceptualization	✓	✓
Data Curation	✓	
Feature Extraction		✓
Methodology	✓	✓
Model Training	✓	✓
Report Writing	✓	✓
Preparation	✓	✓

References

[1] Blaszke, M. Kostek, B. (2022). Musical Instrument Identification Using Deep Learning Approach. *Sensors*, 22(8), 3033.

[2] Hung, Y.-N. Lerch, A. (2020). Multitask Learning for Instrument Activation Aware Music Source Separation. arXiv:2008.00616.

[3] Düzenli, T. Özkurt, N. (2011). Discrete and Dual Tree Wavelet Features for Real-Time Speech/Music Discrimination.

[4] Dash, S. K., Solanki, S. S., Chakraborty, S. (2024). Deep CNNs for Predominant Instrument Recognition Using DWT.

[5] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., Bittner, R. (2017). The MUSDB18 corpus for music separation.

[6] Stöter, F.-R., Liutkus, A., Ito, N. (2018). The 2018 Signal Separation Evaluation Campaign.

[7] Tzanetakis, G., Essl, G., Cook, P. (2001). Audio Analysis using the Discrete Wavelet Transform.

[8] Aggarwal, R., Singh, J. K., Gupta, V. K., Rathore, S., Khare, A. (2011). Noise Reduction of Speech Signal using Wavelet Transform.

[9] Singh, V., Singh, S. (2015). Audio Noise Reduction using DWT.

[10] Šarić, M., Biličić, L., Dujmić, H. (2005). White Noise Reduction using Wavelet Transform.