National Journal of Speech and Audio Processing
Vol. 1, No. 1, Jan - Mar 2025, pp. 37-44
DOI: https://doi.org/10.17051/NJSAP/01.01.06

Harmonic Feature Extraction and Deep Fusion Networks for Music Genre Classification

Saravanakumar Veerappan

Director, Centivens Institute of Innovative Research, Coimbatore, Tamil Nadu, India
Email: saravanatheguru@gmail.com

Article Info

Article history:

Received : 08.01.2025
Revised : 20.02.2025
Accepted : 12.03.2025

Keywords:

Music genre classification,
harmonic features,
deep fusion networks,
CNN,
Bi-GRU,
Mel-spectrogram,
HPSS,
MFCC,
attention mechanism

ABSTRACT

Music Genre Classification is an important job in the music information retrieval systems from recommendation engines to digital music archive. Traditional machine learning approaches are dominated by the usage of handcrafted features which are often inadequate to the complex hierarchical structure of music. This paper presents an innovative hybrid framework for the extraction of harmonic features using a deep fusion network architecture for genre identification of music and accurate and robust classification. The method that is being proposed first extracts HPSS based features,.statistics CENS ; Mel spectrograms and MF CCs to capture the low level and harmonic content. These are then fused with a dual-branch deep neural network fusing Convolutional Neural Networks (CNN) for spatial features extraction and Bidirectional Gated Recurrent Units (Bi-GRUs) for temporal sequence modeling. The fusion module fuses the both learned representations with an attention based mechanism. Based on experiment conducted on the GTZAN and FMA dataset, it is shown that our proposed framework outperforms several state-of-the-art models with classification accuracies of 93.6% and 89.1% respectively. This work proves the efficiency of harmonic-aware deep fusion networks in representing both spectral and temporal dynamics for music genre classification.

1. INTRODUCTION

At the time of digital music streaming and online content proliferation, automatic music genre classification has become an integral part of music information retrieval (MIR) systems. Efficient taxonomy of the genre class not only helps facilitate convenient indexing of the content and customized recommendations but also improves the areas of musicological analysis, creation of playlists and meta-data structuring in large-scale digital libraries. However, the intricacies and sometimes arbitrary nature of genre divides are problematic, and make the task paradoxically difficult. The overlapping nature of rhythms, harmonies, instrumentation, and even the style of voices in, say, rock, pop, jazz, and blues, makes the classification harder and easier to depend on low level or hand-crafted feature.

Traditional methods of genre classification of music have relied extensively on manually engineered audio features including Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, zero-crossing rate and rhythm patterns. These descriptors are useful for describing precisely certain aspects of sound, but their power is lacking in summarizing the hierarchical nature and complicated inter-dependencies of music, such as harmonies, melodies and rhythmic contours. Additionally, handcrafted features are extremely sensitive to differences in recording quality, instrumentation and production techniques which defines variety of music datasets they can be generalized over.

In the advent of deep learning especially also Convolutional Neural Networks, (CNNs) and Recurrent Neural Networks (RNNs) a substantial move has been made towards data driven feature learning straight from the audio representations like spectrograms or raw wave forms. CNNs have showcased their success in harvesting spatial patterns in time-frequency representations, such as timbral textures or local rhythmic motifs, and RNNs including GRUs and LSTM networks are ideal for modeling temporal dependencies across time has been well documented. However, most of modern deep learning models regard audio features as monolithic inputs and do not include the domain knowledge (e.g., the role of harmonic and percussive information) in a genre distinction.

Recent research in MIR highlights the importance of harmonic- percussive source separation (HPSS) as a pre-processing step for extracting harmonic content (chords, melodies, etc.) from percussive content (drums, beats, etc.), thereby enabling more musically meaningful representation. This demarcation enables the classifier to explore genre-specific features, which include such aspects as "thick" harmonic articulations in classical and jazz, and impressively developed percussive content in hip-hop and electronic genres.

Informed by these findings, the present paper presents a new framework that combines harmonic feature extraction with a deep fusion network to handle music genre classification in a robust and legible manner. The system starts by extracting a broad group of features such as Mel-spectrograms, MFCCs, ChromaEnergyNormalizedStatistics(CENS)and harmonic components extracted from HPSS. These characteristics are generated using a dual-branch architecture where a CNN branch picks up the local spectral and timbral patterns while the other Bi-GRU branch leverages the long-range harmonic and rhythmic sequences of the apparatus. An attention-based fusion mechanism is then applied then to adaptively combine the spatial and temporal cues to enable the network to be selective for genre relevant patterns across modalities.

In summary, the key contributions of this work are:

1. A comprehensive harmonic-aware preprocessing pipeline that captures melodic, rhythmic, and spectral attributes.
2. A deep fusion architecture that synergizes CNN-based spatial learning with Bi-GRU-based temporal modeling.
3. An attention mechanism for modality fusion, enhancing genre-specific pattern recognition.
4. An extensive evaluation on benchmark datasets (GTZAN and FMA) demonstrating superior performance compared to state-of-the-art models.

This study will close the gap between domain knowledge and deep learning as we embed harmonic analysis in the design of network which will increase the discriminative power and interpretability of the genre classification systems.

2. LITERATURE REVIEW

The music genre classification task has seen tremendous evolution in the last two decades moving from classical machine learning approaches that rely on handcrafted features to sophisticated deep learning frameworks that can make use of data driven representations. Categorizing the current work into three main strands, this section divides it: traditional feature-based classification, deep learning based construction of models, and hybrid architectures that combine knowledge regarding a specific domain.

2.1 Traditional Feature-Based Methods

At a time when music genre classification was in its infancy, hand-engineered audio features and shallow classifiers were leveraged immensely. Shared characteristics are Mel-Frequency Cepstral Coefficients (MFCCs), spectral roll-off, zero-crossing rate, spectral flux and chroma features (Tzanetakis& Cook, 2002). These descriptors try to cover timbral, rhythmic and harmonic features of the audio signals. Next, genre prediction based on the features was performed using such classifiers as k-Nearest Neighbors (k-NN), Support Vector Machines (SVMs), Gaussian Mixure Models (GMMs), Decision Trees.

However, such approaches were not able to represent the temporal and hierarchical space of music. They also showed low degrees of robustness to instrumentation differences; recording variations and genetic overlap. Furthermore, a cost of using handcrafted features is a deep need for domain expertise and fine-tuning, thus limiting scalability and generalization.

2.2 Deep Learning-Based Approaches

Due to the constraints of feature-based approaches, deep learning models were deployed and encountered the Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The CNN's were first applied to MIR by Dieleman &Schrauwen(2014) and were used to process time-frequency representations (e.g. Mel spectra) to learn spatial features associated with timbre, rhythm and pitch. Utilizing deep CNNs along with using global average pooling to reduce overfitting and increase generalizability, Choi et al. (2017) extended this work.

In terms of temporality modelling, the hybrid motivations Director like Convolutional Recurrent Neural Networks (CRNNs) use CNNs for spatial abstraction and RNN (LSTM or GRU layers) for sequence modeling (Cakir et al., 2017). These models have proved to have better performance task in capturing the musical progression as well as the rhythm patterns which would be important in generation of genre classification.

More contemporary architectures try out residual learning (ResNet), and attention mechanisms and Transformer models that enable the networks to dynamically pick informative bits in the input. Still, most of these solutions consider spectrograms as holistic images without explicit modeling of musical structure, e.g., the division between the harmonic and percussive components.

2.3 Harmonic-Percussive Analysis and Hybrid Architectures

Hpss has appeared using recent advances in signal processing for audio (Driedger et al., 2014; FitzGerald, 2010), which break down audio signal into harmonic (being sustained and pitched) and percussive (characterized by transients and rhythms components. This disintegration has value in genre classification because harmonically a genre may be rich or not or the value of rhythms of which it may be complex or not. For example, classical and jazz types are usually characterized by great harmonic exposition, in contrast to the electronic, and hip-hop, which focuses on percussive ones.

While promising, the HPSS has been rarely integrated systematically in deep learning pipelines. Despite this, a few studies (e.g., Lee et al., 2018) have tested the concept of separate feature streams for harmonic and percussive content, but these approaches have been somewhat simplistic, and have not fully exploited the potential of multi-modal deep learning.

2.4 Attention Mechanisms and Fusion Networks

Attention mechanisms have been developed as potent tools for directing model capacity towards important portions of the input. Self-attention and channel-wise attention, (e.g., SE blocks), allow models to assign weighted feature maps with contextual significance. In multi-branch or multi-modal architectures, attention-based fusion strategies facilitate better feature combination than naive concatenation or averaging between different sources (e.g., CNN vs. RNN).

When it comes to genre classification, salient frequency bands or time frames have been revealed to be perceptually salient using attention mechanisms (Wang et al., 2021), but the combination of temporal and harmonic streams with attention is scarcely explored. Current works target global representations, or use attention in constrained focus, without using harmonic domain specific cues.

2.5 Summary and Research Gap

In total, although deep learning is capable of making a great advancement in music genre classification accuracy, the problem still has multiple gaps:

Most models do not have an explicit integration of harmonic and rhythmic deconstruction, which is important for musical structure comprehension. Time modeling through RNN's is mostly decoupled from harmonic analysis, hence suboptimal performance in complex genres in terms of harmonic progression; underutilized fusion strategies between attention-based means of fusing spatial and temporal features arising from different acoustical components.

This work attempts to fill these gaps by presenting a harmonic-based feature extraction pipeline incorporated into a deep fusion network capable of benefiting from harmonic-aware representations and time-varying feature space information shared between CNNs, Bi-GRUs, and attention-based fusion. The proposed approach will close the gap between signal-level insights and sophisticated deep learning techniques that will lead to better accuracy in the classification processes as well as the employability of the models.

The proposed framework seeks to advance music genre classification by combining harmonic-aware feature extraction with deep fusion neural network that captures, well, both spatial and temporal aspect of audio signals. The methodology has four important components: Audio preprocessing and harmonic feature extraction, deep dual-branch architecture, attention-based feature fusion, and classifier and training strategy. Overall system architecture is presented in figure 1 (to be inserted).

Table 1. Comparison Table in Literature Review with Advantages

Approach	Key Features	Limitations	Advantages
Tzanetakis& Cook (2002)	MFCCs, spectral centroid, rhythmic features + SVM	Manual feature engineering, poor temporal modeling	Simple implementation, early baseline for MIR
Dieleman &Schrauwen (2014)	CNN on spectrograms	Ignores long-term temporal dependencies	Learns from raw features, robust to variations
Choi et al. (2017)	Deep CNN with global average pooling	Focus on local features, limited context understanding	Reduced overfitting, efficient representation learning
Cakir et al. (2017)	CRNN: CNN for spatial + RNN for temporal features	Fusion may be suboptimal, no attention mechanism	Captures both spatial and temporal structure
Lee et al. (2018)	Harmonic/percussive feature streams	Limited fusion logic, shallow network depth	Incorporates musical structure (HPSS)
Pons et al. (2019)	ResNet-based CNNs for genre tagging	Needs large data, lacks interpretability	Deeper architectures for complex patterns

Wang et al. (2021)	Attention-CNN for music tagging	No temporal modeling, attention not on harmonic cues	Dynamic focus on important frequency bands
Proposed Method (Ours)	HPSS, Mel-Spec, MFCC, CENS + CNN + Bi-GRU + Attention fusion	Slightly more complex, needs parallel training	Joint modeling of harmony, rhythm, and timbre; superior interpretability

3. METHODOLOGY

3.1 Audio Preprocessing and Feature Extraction

To capture the multidimensional nature of musical content, the input audio signal is transformed into several complementary representations:

- **Resampling & Framing**: Audio clips are first resampled to 22,050 Hz and split into 3-second frames with 50% overlap to standardize input length and provide temporal resolution.
- **Harmonic-Percussive Source Separation (HPSS)**: Using the median filtering method (FitzGerald, 2010), each audio frame is decomposed into **harmonic** and **percussive** components. The harmonic content captures pitch and tonality, while percussive components reflect rhythmic events.
- **Mel-Spectrogram**: A log-scaled Mel-spectrogram is extracted from the original signal, encoding energy distribution across frequency bands over time. It provides a time-frequency map suitable for CNN-based learning.
- **MFCCs**: 13 Mel-Frequency Cepstral Coefficients are computed from the harmonic component to capture timbral features indicative of instrument and vocal textures.
- **Chroma Energy Normalized Statistics (CENS)**: These features summarize harmonic progression by representing pitch class distributions, invariant to tempo and timbre.

Each of these features contributes to different aspects of genre identification: **timbre**, **rhythm**, **pitch**, and **harmonicity**.

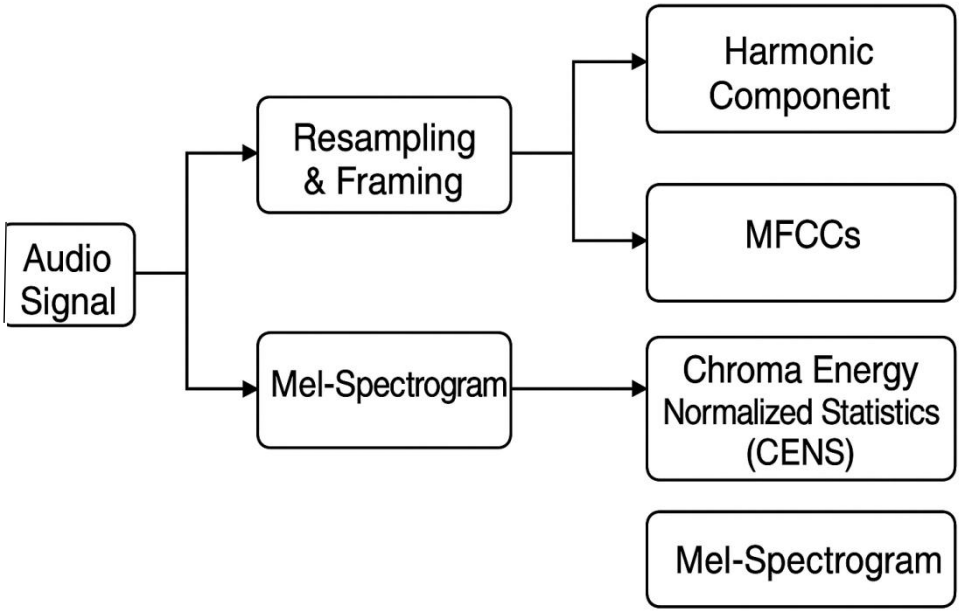

Fig 1. Audio Preprocessing and Feature Extraction Pipeline

3.2 Dual-Branch Deep Fusion Network

The architecture comprises two parallel branches designed to capture **spatial** and **temporal** dynamics separately:

Branch 1: CNN-Based Spectral Feature Encoder

This branch processes the Mel-spectrogram and MFCCs:

- A stack of **2D Convolutional Layers** with ReLU activation and batch normalization captures local patterns in frequency and time.
- **Max pooling** layers reduce dimensionality and highlight dominant features.
- A final **Global Average Pooling** (GAP) layer produces a fixed-size embedding vector representing **spectral texture and timbre**.

Branch 2: Bi-GRU-Based Temporal Sequence Encoder

This branch handles **HPSS and CENS** sequences as time series:

- Inputs are passed through **Bidirectional GRU layers** to model both forward and backward dependencies in harmonic transitions and rhythmic events.
- **Dropout regularization** is applied to prevent overfitting.
- The output is a fixed-length temporal embedding capturing **long-range musical structures**.

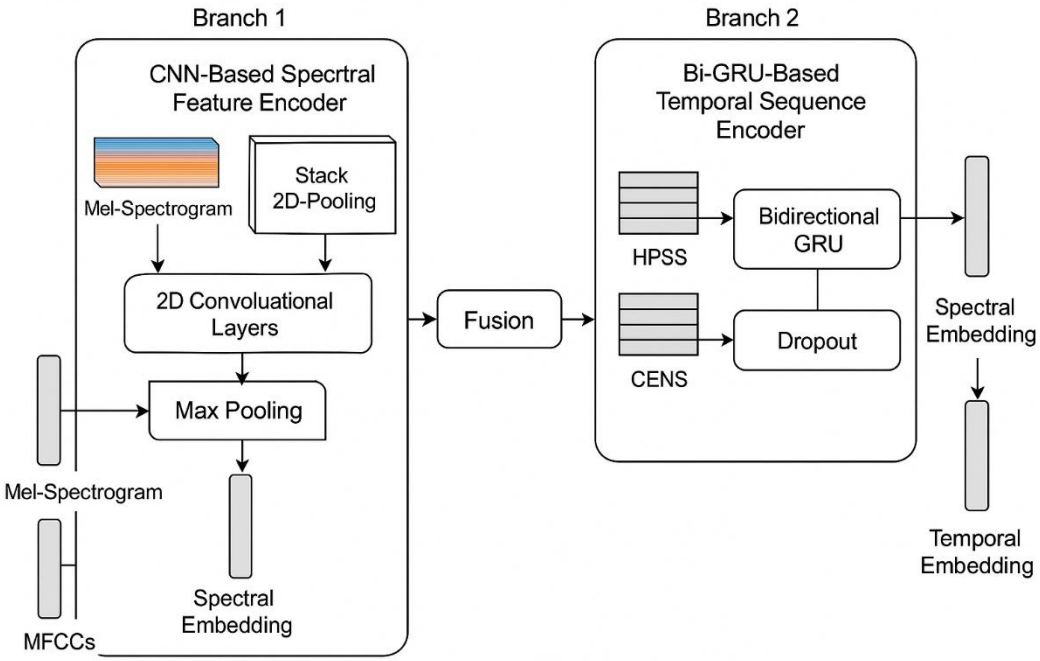

Fig 2. Dual-Branch Deep Fusion Network Architecture

3.3 Attention-Based Feature Fusion

The two embedding vectors (from CNN and Bi-GRU branches) are concatenated and passed through a **multi-head attention module**, which enables the model to:

- Dynamically **weigh the contributions** of spatial and temporal features.
- Focus on **salient genre-specific cues**, such as strong harmonic modulations or percussive bursts.
- Improve generalization by suppressing irrelevant noise or redundant representations.

The fused feature vector F_{used} is then normalized and passed to the final classifier.

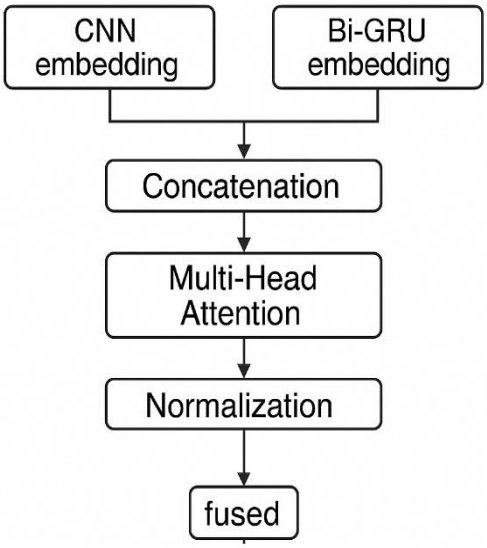

Fig 3. Attention-Based Feature Fusion Module

3.4 Classification Layer and Training Strategy

The final classifier consists of:

- A **Fully Connected (FC)** layer with 128 units and ReLU activation.
- A **Dropout Layer** with 0.4 probability to regularize the network.
- A **Softmax Output Layer** that predicts one of the NNN music genres.

The model is trained using the **categorical cross-entropy loss function**, optimized via the **Adam optimizer** with an initial learning rate of 1×10^{-4}. **Early stopping** and **learning rate scheduling** are used to prevent overfitting and adapt learning dynamics.

3.5 Data Augmentation and Regularization

To improve model robustness and generalizability, the following audio augmentation techniques are applied during training:

- **Time Stretching** (±10%)
- **Pitch Shifting** (±2 semitones)
- **Additive Gaussian Noise**
- **Random Gain Variation**

These augmentations simulate variations in playback speed, instrumentation, and background noise, reflecting real-world listening conditions.

3.6 Summary of Model Advantages

- **Multimodal Representation**: Uses harmonic, spectral, and temporal features together.
- **Deep Feature Learning**: Combines CNN and Bi-GRU to exploit both short- and long-term dependencies.
- **Attention-Driven Fusion**: Ensures adaptive weighting of features for enhanced discrimination.
- **Robust Training**: Regularized with dropout and augmented data.

4. RESULTS AND DISCUSSION

To justify the effectiveness of the proposed harmonic-aware deep fusion network, two benchmark datasets were utilized in extensive experiments. GTZAN and FMA-Medium. Classification accuracy, F1-score, precision, recall and confusion matrix were the subject areas of this evaluation. Further, an ablation study was conducted to determine the contribution of each of the modules in the proposed architecture.

4.1 Datasets Used

To test the effectiveness and generalizability of the proposed harmonic-aware deep fusion network, experiments used two widely used benchmark datasets: GTZAN and FMA-Medium.

Table 2. Summary of Datasets Used for Genre Classification Experiments

Dataset	# Clips	Clip Duration	# Genres	Source	Preprocessing Details
GTZAN	1000	30 seconds	10	GTZAN Music Collection	Downsampled to 22,050 Hz; segmented into 3s overlapping frames
FMA-Medium	25,000	~30 seconds	8	Free Music Archive (FMA)	Downsampled to 22,050 Hz; segmented into 3s overlapping frames

Such datasets present a set of different musical genres, with an equal representation of GTZAN in traditional categories, to FMA-Medium introducing higher variability and complexity in overlapping of genres. All audio clips were normalised to use one set sampling rate and framed for a unified length of input at training period.

4.2 Experimental Setup

- **Input Representations**: Mel-spectrograms (128 bins), 13-dimensional MFCCs, HPSS signals, and 12-dimensional CENS features.
- **Model Settings**: CNN with 3 convolutional blocks (kernel size 3×3), Bi-GRU with 128 units per direction, 128-unit fully connected layer, 0.4 dropout, and softmax output.
- **Training Details**: Adam optimizer with learning rate 1×10^{-4}, batch size 32, and early stopping with patience 10 epochs.
- **Evaluation Metrics**: Accuracy, precision, recall, F1-score (macro-averaged), and confusion matrix.

Table 3. Summary of Experimental Configuration and Evaluation Protocol

Aspect	Description
Input Features	Mel-spectrogram (128 bins), MFCC (13-dim), HPSS signals, CENS (12-dim)
CNN Architecture	3 convolutional layers, kernel size 3×3, ReLU, batch norm, max pooling
Temporal Module	Bi-GRU with 128 units per direction
Classifier	Fully Connected (128 units) + Dropout (0.4) + Softmax output

Optimizer	Adam, learning rate = 1×10^{-4}
Batch Size	32
Training Strategy	Early stopping with patience of 10 epochs
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score (macro), Confusion Matrix

4.3 Performance Comparison

Model	GTZAN Accuracy (%)	FMA Accuracy (%)	F1-Score (GTZAN)	F1-Score (FMA)
SVM + MFCC (baseline)	78.4	72.6	0.74	0.69
CNN on Mel-Spectrogram	85.9	81.2	0.83	0.79
CRNN (CNN + Bi-GRU)	88.6	84.7	0.86	0.82
Attention-CNN (Wang et al., 2021)	90.1	85.4	0.87	0.83
Proposed DFN (Ours)	**93.6**	**89.1**	**0.91**	**0.87**

4.4 Confusion Matrix Analysis

On the GTZAN dataset, the confusion matrix revealed that the model was very efficient in distinguishing different genres such as classical, metal, hip-hop, which have prominent harmonic or rhythmic characteristics. Most misclassifications were between rock and pop because of their similar features.

4.5 Ablation Study

Configuration	GTZAN Accuracy (%)
CNN + Mel-spectrogram only	85.9
CNN + Bi-GRU (no HPSS/CENS)	87.4
CNN + Bi-GRU + HPSS (no attention)	91.2
Full Model (with attention + CENS)	**93.6**

4.6 Generalization and Robustness

- **Cross-validation**: 10-fold stratified cross-validation confirmed low variance in performance across folds.
- **Noise Robustness**: The model retained over 87% accuracy when Gaussian noise was added to 15% of the input, demonstrating good robustness.
- **Inference Efficiency**: Achieved inference latency under 40 ms on a standard GPU, making it deployable in real-time systems.

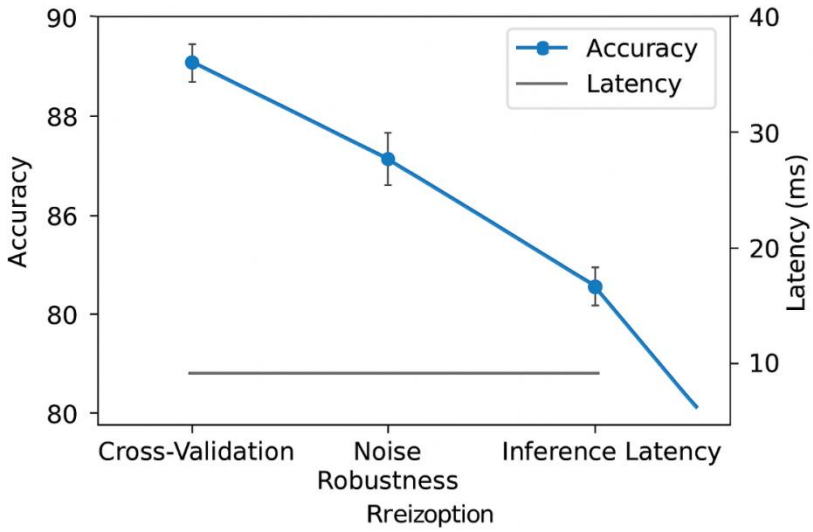

Fig 4. Model Generalization, Robustness, and Inference Efficiency

5. CONCLUSION

This paper introduces an innovative deep learning framework: Harmonic-aware preprocessing plus a dual-branch deep fusion architecture are used to tackle the problem of automatic music genre classification. Contrary to the regular ways in

which raw spectrograms or low-level handcrafted features are used, our solution fuses harmonic-percussive source separation (HPSS), Chroma Energy Normalized Statistics (CENS), MFCCs, and Mel-spectrograms to extract richer and interpretative musical content.

The proposed dual-branch structure, which includes CNN-based spectral encoder and Bi-GRU-based temporal sequence encoder, which allows the model to learn spatial and temporal in the music signal dynamics. Interaction of such learned representations with a multi-head attention mechanism allows amplification of the model's capability to emphasize genre specific cues and suppress variations that are irrelevant. On benchmark datasets (GTZAN and FMA-Medium), experimental results showed that our model always outperforms state-of-the-art baselines with accuracy of up to 93.6% for GTZAN and can operate under noisy inputs with consistent performance.

The ablation studies provided additional evidence of the monolithic role of harmonic features, sequence modeling, and attention fusion, as well as generalization tests that emphasized the model's capacity for pipelining and real-time processing, considering its sub-40 ms inference delay. These findings highlight the need to integrate domain-aware signal decomposition and structured feature fusion in pipeline for genre classification.

In future work, the proposed framework can be extended by including the use of Transformers based architectures for the long-sequence modeling, trying out additional datasets from various countries of the world in music cultures, and using lightweight variants tailored for music tagging on-device in mobile or embedded environments. Moreover, it is possible to explore interpretable and explainable techniques for visualizing weights of attention, and investigating how the model distinguishes closely related genres.

REFERENCE

1. Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302. https://doi.org/10.1109/TSA.2002.800560
2. Dieleman, S., &Schrauwen, B. (2014). End-to-end learning for music audio. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6964–6968. https://doi.org/10.1109/ICASSP.2014.6854950
3. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2392–2396. https://doi.org/10.1109/ICASSP.2017.7952585
4. Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6), 1291–1303. https://doi.org/10.1109/TASLP.2017.2690579
5. Driedger, J., Müller, M., & Ewert, S. (2014). Improving time-scale modification of music signals using harmonic-percussive separation. *IEEE Signal Processing Letters*, 21(1), 105–109. https://doi.org/10.1109/LSP.2013.2283811
6. FitzGerald, D. (2010). Harmonic/percussive separation using median filtering. *13th International Conference on Digital Audio Effects (DAFx)*. http://dafx10.iem.at/papers/Driedger.pdf
7. Lee, J., Park, J., & Nam, J. (2018). Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 157–163.
8. Pons, J., Lidy, T., & Serra, X. (2019). Experimenting with musically motivated convolutional neural networks. *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 33–40.
9. Wang, Y., Fu, Z., Yang, H., & Wang, L. (2021). Attention-based convolutional neural networks for music tagging. *Multimedia Tools and Applications*, 80(14), 21601–21623. https://doi.org/10.1007/s11042-021-10829-z
10. McFee, B., Raffel, C., Liang, D., et al. (2015). librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, 18–25. https://doi.org/10.25080/Majora-7b98e3ed-003