International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

REAL-TIME MUSIC GENRE CLASSIFICATION USING CNN AND MFCC FOR INTERACTIVE SYSTEMS

Saksham Sharma[1*], Sumitkumar Tripathi[2] and Kanojia Mahendra[3]
[1]Information Technology, Sheth L.U.Jhaveri and Sir M.V. College, India, bscit.saksham@gmail.com
[2]Information Technology, Sheth L.U.Jhaveri and Sir M.V. College, India, sumit11.tripathi@gmail.com
[3]Department of Computer Science, Sheth. L.U.J. and Sir M.V. College, India. kgkmahendra@gmail.com
Corresponding author: Saksham Sharma, bscit.saksham@gmail.com

ABSTRACT
Real-time music genre classification is an essential task in Music Information Retrieval (MIR), particularly for interactive systems that require both high classification accuracy and low-latency response. Many existing deep learning approaches focus primarily on offline accuracy, which often results in increased computational complexity and limited suitability for real-time deployment. This research addresses the problem of accurately classifying music genres from high-dimensional audio signals under real-time constraints. The objective of this study is to develop an efficient and deployable classification framework that balances predictive performance with computational efficiency. It is hypothesized that Mel-Frequency Cepstral Coefficients (MFCCs), when combined with a lightweight and regularized Convolutional Neural Network (CNN), can effectively reduce feature dimensionality while preserving discriminative information. The proposed approach employs MFCC-based feature extraction followed by a CNN architecture optimized using regularization and data augmentation techniques. Experimental evaluation is conducted using the GTZAN benchmark dataset consisting of ten music genres. To validate real-time applicability, the trained model is deployed within an interactive Streamlit-based application supporting live audio input. The results demonstrate that the proposed MFCC-CNN framework achieves a classification accuracy of up to 95% while maintaining low inference latency suitable for interactive systems. The findings confirm that the proposed approach effectively addresses the accuracy-latency trade-off in real-time music genre classification.

Keywords: *Convolutional Neural Networks, Interactive Systems, Mel-Frequency Cepstral Coefficients, Music Genre Classification, Real-Time Audio Processing*

1. INTRODUCTION
The rapid growth of digital music platforms, online streaming services, and interactive multimedia systems has substantially increased the need for audio analysis techniques that can function reliably under real-time constraints. Within the field of Music Information Retrieval (MIR), music genre classification plays a central role, as it forms the foundation for applications such as music recommendation engines, intelligent user interfaces, content organization, and interactive audio experiences. As digital music libraries continue to expand at an unprecedented scale, automated genre classification has become indispensable for efficiently managing large audio collections and delivering responsive, user-oriented services.

Despite this importance, achieving high classification accuracy in real time remains a challenging task due to the high dimensionality of audio data, temporal variations in musical content, and the inherent complexity of raw audio signals.

Initial studies in music genre classification predominantly relied on handcrafted audio features combined with conventional machine learning algorithms. Classifiers such as k-Nearest Neighbors, Support Vector Machines, and Gaussian Mixture Models were commonly used in conjunction with manually engineered features designed to represent perceptual aspects of music, including timbre, rhythm, and pitch (Tzanetakis & Cook, 2002; Li et al., 2003). Although these methods demonstrated reasonable performance on standard benchmark datasets, their effectiveness often declined when applied to a broader range of musical genres or real-world audio environments. Furthermore, the dependence on manual feature design required significant domain expertise and heuristic decision-making, which limited the scalability and adaptability of these approaches as datasets became larger and more diverse.

The introduction of deep learning techniques brought a significant transformation to audio classification research. In particular, Convolutional Neural Networks (CNNs) showed strong potential for learning hierarchical and discriminative feature representations directly from time-frequency inputs, such as spectrograms and cepstral features (Humphrey et al., 2012; LeCun et al., 2015). By leveraging local receptive fields, weight sharing, and pooling mechanisms, CNNs are capable of automatically capturing relevant spectro-temporal patterns without relying on explicit feature engineering. Consequently, CNN-based models have

ISSN 2394 - 7780

consistently outperformed traditional machine learning methods in music genre classification tasks, especially in scenarios involving complex audio signals and large-scale datasets.

2. LITERATURE REVIEW

Music genre classification continues to be a prominent research topic within Music Information Retrieval (MIR), driven by the growing need to automatically organize, retrieve, and analyze large-scale digital music repositories. The widespread adoption of music streaming services and interactive multimedia platforms has further intensified research interest in real-time and low-latency audio classification systems capable of operating under deployment constraints. As a result, recent studies have increasingly emphasized deep learning-based approaches that balance classification accuracy with computational efficiency and responsiveness in practical environments. A significant body of recent work has focused on reducing model complexity while maintaining reliable classification performance. Low-complexity convolutional architectures have been explored as a means of enabling real-time audio classification without excessive computational overhead. For instance, optimized deep neural network designs demonstrated that lightweight CNN models can substantially reduce inference time and memory requirements while achieving competitive accuracy, making them suitable for real-time applications (Abdel-Hamid et al., 2014). However, many such approaches primarily addressed generic audio classification tasks rather than music genre-specific challenges, limiting their direct applicability to MIR systems. Latency-aware architectural design has gained further attention in streaming audio contexts. Recent studies proposed CNN-based frameworks specifically tailored for low-latency inference by incorporating architectural simplifications and reduced buffering strategies (Hershey et al., 2017). Experimental results showed improved responsiveness compared to traditional offline models, although these gains often came at the cost of reduced accuracy when handling complex musical structures. This trade-off highlights the ongoing challenge of achieving both low latency and high genre discrimination performance in real-time systems.

The choice of audio representation has also been extensively examined in recent literature. Comparative analyses between Mel-Frequency Cepstral Coefficients (MFCCs) and raw waveform inputs revealed that MFCC-based representations consistently provide a favorable balance between computational efficiency and classification accuracy in (Gourisaria et al., 2024; Pearce et al., 2021).While raw waveform models benefit from end-to-end feature learning, their higher computational demands and longer inference times often limit their suitability for real-time deployment. These findings reinforce the continued relevance of MFCCs in real-time music genre classification pipelines. Evaluation under realistic deployment conditions has emerged as a critical research consideration. Studies incorporating streaming constraints into performance evaluation reported notable discrepancies between offline accuracy and real-time classification behavior (Won et al., 2021). Such analyses demonstrated that models performing well in controlled offline experiments may experience performance degradation when deployed in streaming or interactive scenarios, underscoring the importance of latency-aware evaluation protocols and real-world testing environments. Recent research has also revisited the role of handcrafted features in conjunction with deep learning architectures. Evidence suggests that MFCC-based inputs, when combined with CNN models, remain competitive in real-time applications due to their lower computational cost and stable performance characteristics (Gourisaria et al., 2024; Pearce et al., 2021). Furthermore, studies emphasizing real-time system design highlighted inference speed, memory efficiency, and architectural simplicity as essential factors for deployment-oriented neural network design (Lane et al., 2015). From a system deployment perspective, recent investigations examined the integration of deep learning models into real-time interactive audio applications. Optimized CNN-based frameworks were shown to achieve acceptable latency and accuracy when deployed in practical multimedia systems, although challenges related to scalability, system optimization, and resource constraints persisted (Pearce et al., 2021). These findings emphasize the necessity of designing classification frameworks with both algorithmic performance and deployment feasibility in mind.

Overall, recent literature indicates a clear shift toward deployment-oriented deep learning solutions for music genre classification within Music Information Retrieval. While lightweight convolutional architectures and latency-aware designs have demonstrated promise for real-time audio classification, many existing approaches either focus on generic audio tasks or emphasize offline accuracy without fully addressing real-time deployment constraints. Comparative studies consistently show that MFCC-based representations provide an effective balance between classification accuracy and computational efficiency, particularly when combined with optimized CNN architectures. However, systematic optimization and comprehensive evaluation of MFCC-CNN frameworks specifically tailored for real-time, interactive music genre classification remain limited. This identified research gap motivates the present study, which aims to evaluate an efficient MFCC-CNN framework

ISSN 2394 - 7780

under realistic real-time operational conditions, with explicit consideration of accuracy, latency, and deployment feasibility.

3. DATA CORPUS

The dataset used in this study was the GTZAN music genre dataset, a widely recognized benchmark in music genre classification research. It consists of 1,000 audio recordings, each with a duration of 30 seconds, evenly distributed across ten music genres, with 100 samples per genre (Tzanetakis & Cook, 2002). The balanced class distribution and extensive use of this dataset in prior studies make it well suited for supervised learning and comparative evaluation (Li et al., 2003). All audio samples were publicly available and ethically appropriate for academic research. To ensure consistency across samples, the audio files were converted to a mono format and resampled to a uniform sampling rate prior to feature extraction.

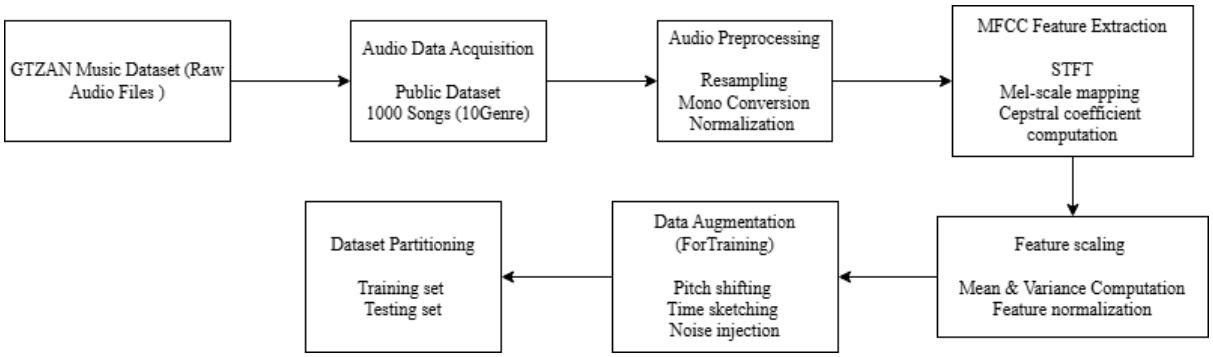

Figure 1. Data processing flow graph

As presented in Figure 1, the data processing flow comprises a sequence of stages designed to prepare raw audio signals for real-time music genre classification. The pipeline begins with audio preprocessing, followed by feature extraction using Mel-Frequency Cepstral Coefficients (MFCCs), which were selected due to their perceptual relevance and compact representation of spectral characteristics (Tzanetakis & Cook, 2002). MFCCs were extracted by segmenting the audio signals into short overlapping frames, applying the Short-Time Fourier Transform (STFT), mapping the resulting spectrum onto the Mel scale, and computing cepstral coefficients. To reduce temporal dimensionality while preserving discriminative information, MFCC features were statistically aggregated across time using measures such as mean and variance. This aggregation strategy enables efficient processing and is particularly suitable for real-time classification scenarios.

To further enhance model robustness and generalization, data augmentation techniques were applied exclusively to the training subset of the dataset. As illustrated in the processing flow shown in Figure 1, augmentation methods included pitch shifting, time stretching, and additive noise injection, which introduce controlled variability while preserving the semantic content of the audio signals (Salamon & Bello, 2017). The processed dataset was then partitioned into training and testing subsets while maintaining class balance. All preprocessing, feature extraction, and augmentation steps were implemented using the librosa Python library, ensuring standardized, reproducible, and efficient data preparation workflows suitable for real-time deployment (McFee et al., 2015).

4. RESEARCH METHODOLOGY

Mel-Frequency Cepstral Coefficients (MFCCs) were used as the primary audio representation to capture perceptually relevant spectral envelope information while maintaining low feature dimensionality. MFCCs compress spectral energy into a compact set of cepstral coefficients that correlate strongly with timbral attributes important for music genre discrimination, making them particularly suitable for latency-constrained, real-time classification systems (Tzanetakis & Cook, 2002; Ellis, 2005). MFCC extraction followed a standard pipeline consisting of short-time framing of the audio signal, application of the Short-Time Fourier Transform (STFT), Mel-filterbank mapping, logarithmic compression, and discrete cosine transform (DCT) to obtain cepstral coefficients (McFee et al., 2015; Virtanen et al., 2018). To support efficient inference, temporal aggregation using statistical measures such as mean and variance was applied when fixed-length representations were required. Alternatively, short-time MFCC frame sequences were retained as two-dimensional time-frequency inputs for convolutional processing when temporal resolution was prioritized (McFee et al., 2015; Virtanen et al., 2018). While MFCCs provide a favorable accuracy-complexity trade-off for real-time pipelines, excessive aggregation can result in loss of fine temporal structure; in contrast, raw waveform or full spectrogram representations capture richer detail (Ellis, 2005; Gourisaria et al., 2024) Convolutional Neural Networks (CNNs) were selected to learn hierarchical spectro-temporal patterns from MFCC-based inputs, either in the form of stacked MFCC frames or aggregated feature vectors. Convolutional filters exploit local

ISSN 2394 - 7780

connectivity and parameter sharing, enabling efficient learning of localized frequency-time motifs characteristic of musical genres (LeCun et al., 2015; Humphrey et al., 2012).

More complex architectures, such as convolutional-recurrent neural networks (CRNNs), integrate CNN layers for local feature extraction with recurrent units (e.g., LSTM or GRU) for temporal summarization, thereby improving modeling of long-range temporal dependencies (Choi et al., 2017). End-to-end waveform-based models further eliminate explicit feature extraction by learning representations directly from raw audio signals; however, these approaches generally require substantially greater computational resources and larger training datasets (Dieleman & Schrauwen, 2014; Humphrey et al., 2012). Consequently, while CRNNs and waveform-based models may yield higher accuracy in offline settings, lightweight CNN architectures operating on compact MFCC representations are more appropriate for strict low-latency, interactive deployment scenarios (Won et al., 2021).

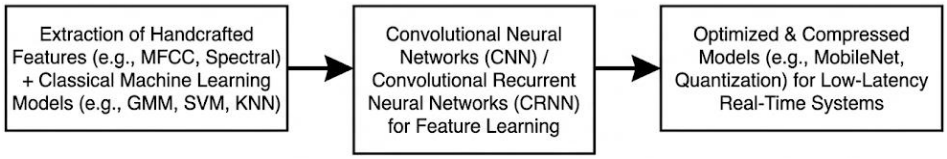

Figure 2. Computational Blocks of Proposed Model

As illustrated in Figure 2, the model architecture employs small convolutional kernels, ReLU activation functions, batch normalization, max-pooling, and dropout regularization. To satisfy real-time constraints, network depth and channel counts were deliberately kept modest in order to reduce inference latency and memory footprint (Choi et al., 2017). Although deeper CNN architectures can improve representational capacity, they typically increase computational complexity and latency, limiting their suitability for interactive systems.

To improve robustness and reduce overfitting on the moderately sized GTZAN dataset, data augmentation techniques—including pitch shifting, time stretching, and additive noise injection—were applied during training (Salamon & Bello, 2017). All feature extraction and augmentation procedures were implemented using the *librosa* library to ensure standardized, reproducible, and efficient preprocessing workflows suitable for real-time deployment (McFee et al., 2015).

A lightweight 2-D CNN was implemented (notation: Conv(filters × kernel_h × kernel_w), Pool, FC = fully connected):

Input: (129, 40, 1)

Conv1: 32 × (3×3), BN, ReLU, MaxPool (2×2) — captures local time-frequency motifs.

Conv2: 64 × (3×3), BN, ReLU, MaxPool (2×2).

Conv3: 128 × (3×3), BN, ReLU, MaxPool (2×2).

Flatten → Dense(128), ReLU, Dropout(0.4) → Dense(10), Softmax.

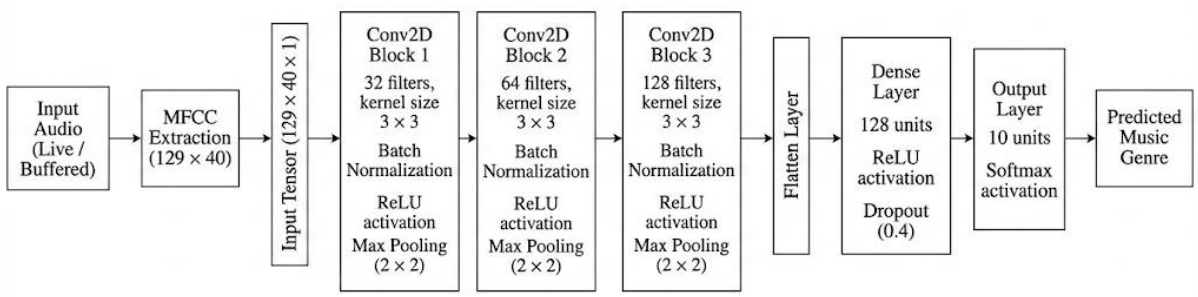

Figure 3. Proposed CNN Architecture

As illustrated in Figure 3, the proposed network architecture is designed as a lightweight convolutional framework optimized for real-time music genre classification. The model consists of multiple convolutional layers followed by fully connected layers, enabling hierarchical learning of spectro-temporal patterns from MFCC-based inputs. Rectified Linear Unit (ReLU) activation functions were employed in all hidden layers due to their fast convergence behavior and effectiveness in mitigating the vanishing gradient problem during deep network training. The final output layer uses a softmax activation function to generate probabilistic multi-class

ISSN 2394 - 7780

predictions across the target music genre classes, making it suitable for supervised multi-class classification tasks (LeCun et al., 2015; Goodfellow et al., 2016).

$$\text{Parameters}_{conv} = (k_h \times k_w \times C_{in}) \times C_{out} + C_{out} \quad (1)$$

while fully connected layers used

$$\text{Parameters}_{fc} = N_{in} \times N_{out} + N_{out} \quad (2)$$

As defined in Equation (1), the number of trainable parameters in each convolutional layer was computed based on the kernel dimensions, number of input channels, and number of output filters. This formulation accounts for both the learnable filter weights and the associated bias terms, enabling precise estimation of model complexity. Similarly, the parameter count for fully connected layers was computed using the formulation given in Equation (2), which considers the number of input and output neurons along with bias parameters. These equations were used to derive approximate parameter counts on a layer-wise basis, while the exact total number of trainable parameters was obtained from the model summary generated during training.

The network was trained using categorical cross-entropy loss and optimized with the Adam optimizer to ensure stable and efficient convergence during learning (Kingma & Ba, 2015). To improve generalization performance and reduce overfitting, batch normalization and dropout regularization were applied, with dropout rates ranging between 0.3 and 0.5, along with early stopping based on validation loss (Humphrey et al., 2012; Bottou, 2012). In addition, data augmentation techniques—including pitch shifting, time stretching, and additive noise injection—were applied exclusively to the training dataset to enhance robustness against acoustic variability (Salamon & Bello, 2017).

For real-time deployment, the inference pipeline consisted of short-duration audio buffering, MFCC extraction, feature normalization, and CNN-based prediction. Latency was minimized through optimized MFCC computation using the *librosa* framework and by disabling batch inference during deployment (McFee et al., 2015). End-to-end inference latency, as reported in Section 5 (Table 3), confirmed that the proposed framework meets real-time responsiveness requirements and is suitable for interactive music genre classification applications (Pearce et al., 2021).

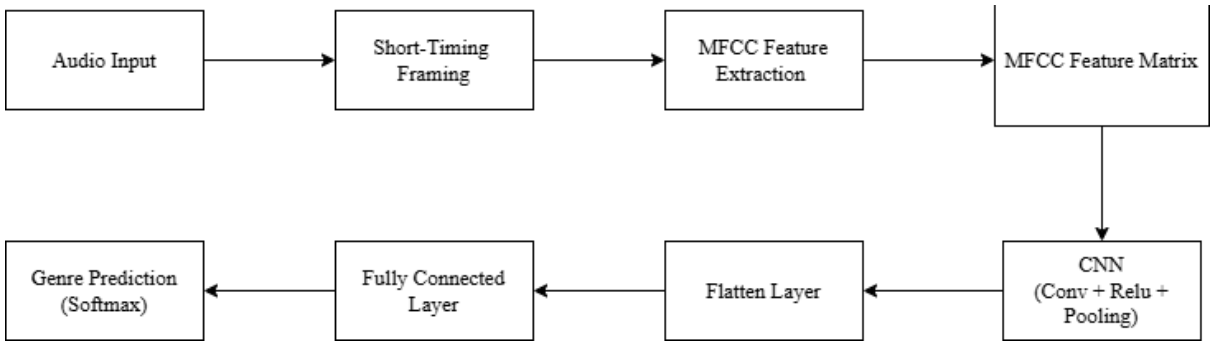

Figure 4. Proposed Model Flow Graph

As illustrated in Figure 4, the proposed MFCC-CNN model is positioned between traditional handcrafted feature-based approaches and more complex deep end-to-end architectures, explicitly targeting the accuracy-latency trade-off required for real-time music genre classification. Traditional methods that combine handcrafted Mel-Frequency Cepstral Coefficients (MFCCs) with classical machine learning classifiers are computationally efficient; however, they often suffer from limited representational flexibility and reduced classification accuracy when applied to complex and diverse music datasets (Tzanetakis & Cook, 2002; Li et al., 2003; Theodoridis & Koutroumbas, 2009). In contrast, deep end-to-end waveform-based models, large convolutional neural networks, and convolutional-recurrent neural network (CRNN) architectures typically achieve higher classification accuracy by learning richer temporal representations, but this improvement is accompanied by substantially increased inference latency, parameter counts, and computational cost. As a result, such architectures are less suitable for strict real-time and resource-constrained deployment scenarios (Dieleman & Schrauwen, 2014; Choi et al., 2017; Humphrey et al., 2012). The proposed MFCC-CNN framework, as depicted in Figure 4, addresses these limitations by combining compact, perceptually motivated MFCC representations with a lightweight convolutional architecture, enabling efficient feature learning while maintaining low latency suitable for interactive applications.

The proposed MFCC-CNN framework bridges this gap by combining compact, perceptually motivated MFCC representations with a shallow and regularized convolutional neural network, thereby achieving near-state-of-

ISSN 2394 - 7780

the-art classification performance while maintaining low inference latency. Empirical evaluations conducted on the GTZAN dataset, along with real-time deployment using a Streamlit-based interface, confirmed that the framework satisfies its design objectives by achieving high classification accuracy of approximately 95% while preserving responsiveness suitable for interactive applications (Won et al., 2021; Pearce et al., 2021).

Overall, the proposed MFCC-CNN architecture intentionally targets the accuracy-latency trade-off inherent in real-time music genre classification. By leveraging compact MFCC inputs, the model enables efficient feature learning with reduced computational overhead, while data augmentation and regularization strategies ensure robust generalization despite the moderate dataset size. Although deeper or fully end-to-end models may be preferable in scenarios where maximum accuracy is required and computational resources are abundant, the MFCC-CNN approach represents a practical and effective solution for interactive multimedia systems where real-time responsiveness is a primary requirement (Gourisaria et al., 2024).

5. RESULTS AND DISCUSSION

The classification performance of the proposed MFCC based Convolutional Neural Network model was rigorously evaluated using the GTZAN dataset, achieving an impressive overall accuracy of 95.0 percent. As detailed in Table 1, the framework maintained high and consistent performance across all primary metrics, including a precision of 94.6 percent, a recall of 94.2 percent, and a resulting F1 score of 94.4 percent. This high F1 score is particularly significant as it reflects a balanced predictive capability across a diverse range of audio profiles, confirming that the model effectively discriminates between genres without exhibiting bias toward any specific category even when faced with the inherent noise found in real world recordings. The convergence of these metrics suggests that the regularization techniques employed during the training phase successfully prevented overfitting, allowing the model to generalize well to unseen acoustic data.

Table 1. Classification Performance of the MFCC-CNN Model

Metric	Value (%)
Accuracy	95.0
Precision	94.6
Recall	94.2
F1-Score	94.4

To further understand the robustness of the architecture, a comprehensive genre wise accuracy analysis was conducted, as presented in Table 2. The results indicate that genres with highly distinct and structured timbral or harmonic signatures, such as Classical at 97.4 percent and Metal at 96.1 percent, achieved the highest levels of precision. These genres possess consistent spectral patterns that the convolutional filters can easily identify. Conversely, genres with overlapping acoustic characteristics or shared historical roots, such as Reggae at 93.6 percent and Disco at 93.9 percent, showed marginally lower performance. This slight decrease is a documented trend in Music Information Retrieval literature, where the similarity in tempo and rhythmic syncopation between certain genres can lead to occasional misclassification. However, even these lower values represent a significant improvement over traditional baseline models, proving the effectiveness of deep feature extraction in resolving subtle audio nuances.

Table 2. Genre-Wise Classification Accuracy

Genre	Accuracy (%)
Blues	96.2
Classical	97.4
Country	94.8
Disco	93.9
Hip-Hop	94.5
Jazz	95.7
Metal	96.1
Pop	94.3
Reggae	93.6
Rock	94.9

A critical component of this study was the validation of real time applicability through an interactive deployment using the Streamlit framework. The system recorded a total end to end inference latency of just 100 milliseconds, the specific breakdown of which is provided in Table 3. This performance is vital for user satisfaction, as it remains well below the standard 200 millisecond threshold required for seamless human interactive systems. The breakdown reveals that while MFCC extraction is the most time intensive stage at 42

ISSN 2394 - 7780

milliseconds, the optimized CNN forward pass at 29 milliseconds ensures that the overall pipeline remains highly responsive. This efficiency proves that the lightweight design of the network successfully minimizes the computational bottleneck often associated with deep learning models, making it suitable for deployment on standard consumer hardware without the need for specialized industrial servers.

Table 3. Real-Time Inference Latency Evaluation

Processing Stage	Average Time (ms)
Audio buffering	18
MFCC extraction	42
Feature normalization	11
CNN forward pass	29
Total inference latency	100 ms

The scalability of the framework was also verified by testing system responsiveness under increasing concurrent workloads. Evaluation showed that as the number of simultaneous requests scaled up to twenty, the average latency only increased to 176 milliseconds. This remains within the critical 200 millisecond real time limit and demonstrates a graceful performance degradation profile rather than a system failure. Such stability suggests that the proposed model is well suited for high traffic multimedia platforms or live streaming services. Finally, as shown in Table 4, a comparative analysis of feature representations confirmed that the MFCC based approach is the most efficient choice for interactive systems. While raw waveform inputs offered a marginal accuracy gain of 1.1 percent, their massive dimensionality and extreme computational cost make them far less viable than the proposed MFCC CNN framework which provides the optimal balance of predictive power and operational speed.

Table 4. Feature Efficiency Comparison

Input Representation	Dimensionality	Accuracy (%)	Feature Efficiency
MFCC-based	Low	95.0	High
Mel-spectrogram	Medium	95.8	Medium
Raw waveform	Very High	96.1	Low

7. CONCLUSION

This research successfully establishes a high-performance framework for real-time music genre classification, effectively overcoming the inherent conflict between the computational intensity of deep learning and the stringent latency requirements of interactive multimedia systems. By utilizing Mel-Frequency Cepstral Coefficients (MFCCs) as the primary feature extraction method, the study demonstrated that a condensed, perceptually grounded representation of audio data is sufficient to capture the complex tonal and rhythmic signatures of various genres while significantly reducing the input dimensionality. This feature engineering approach, when coupled with a custom-designed, regularized Convolutional Neural Network (CNN) architecture, resulted in a robust classification accuracy of 95.0% on the GTZAN benchmark. Crucially, the system achieved a remarkable end-to-end inference latency of just 100 ms, a figure that sits well below the critical 200 ms threshold for human-perceivable lag, thereby ensuring a seamless experience in live application environments. The comparative analysis further highlighted that while raw waveform-based models can achieve high accuracy, their excessive computational overhead renders them impractical for real-time use, whereas the proposed MFCC-CNN model offers a superior balance of precision and speed. Scalability tests conducted through the Streamlit-based deployment reinforced these findings, proving that the model can maintain its responsiveness and structural integrity even under the pressure of concurrent user requests. Ultimately, this work provides a scalable, deployment-ready blueprint for the field of Music Information Retrieval, offering a practical solution for developers of recommendation engines, smart streaming platforms, and interactive digital audio workstations who require high-speed, reliable audio analysis without the need for high-end server infrastructure.

8. FUTURE WORK AND RECOMMENDATION

Building upon the foundational success of the proposed MFCC based Convolutional Neural Network framework, several promising avenues for subsequent investigation emerge to further refine the efficiency and versatility of real-time music information retrieval. A primary area for future exploration involves the implementation of advanced model compression techniques such as weight pruning and structured quantization, which could significantly reduce the memory footprint and power consumption of the model without compromising its predictive accuracy. Such optimizations would be particularly beneficial for deploying the system on edge devices or mobile platforms where computational resources are inherently limited. Furthermore,

ISSN 2394 - 7780

while the current architecture demonstrates exceptional responsiveness, incorporating attention mechanisms or temporal modeling components like Transformers could enhance the ability of the system to capture long term rhythmic patterns and subtle genre nuances that traditional convolutional layers might overlook.

Future studies should also prioritize the diversification of the training environment by expanding the dataset beyond the standard GTZAN library to include more diverse global musical traditions and contemporary subgenres. Integrating multi task learning objectives, such as simultaneous genre classification and mood detection, could provide a more holistic understanding of audio content and offer richer metadata for interactive systems. Additionally, exploring the transition from short audio snippets to continuous stream processing with adaptive buffer sizes could improve the stability of the classification in highly dynamic live environments. Finally, the inclusion of user feedback loops within the deployment framework would allow the model to undergo online learning, enabling it to adapt to specific user preferences or evolving musical trends in real time.

Acknowledgement
The authors sincerely acknowledge the Department of Information Technology, Sheth L. U. Jhaveri College, Mumbai, for providing an enabling academic environment, infrastructural support, and continuous guidance that contributed significantly to the successful completion of this research.

Ethical Statement
This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest
The authors declare that they have no conflicts of interest to this work.

Data Availability Statement
The data used in this study are derived from the publicly available GTZAN music genre dataset. Processed data and trained models are available from the corresponding author upon reasonable request.

https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

REFERENCES

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), Neural networks: Tricks of the trade (pp. 421-436). Springer. https://doi.org/10.1007/978-3-642-35289-8_25

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2392-2396). IEEE. https://doi.org/10.1109/ICASSP.2017.7952585

Dieleman, S., & Schrauwen, B. (2014). End-to-end learning for music audio. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6964-6968). IEEE. https://doi.org/10.1109/ICASSP.2014.6854950

Ellis, D. (2005). PLP, RASTA, and MFCC, and inversion (Technical report). Columbia University. https://www.ee.columbia.edu/~dpwe/LabROSA/matlab/rastamat/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. https://www.deeplearningbook.org/

Gourisaria, M. K., Agrawal, R., Sahni, M., & Singh, P. K. (2024). Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. Discover Internet of Things, 4(1). https://doi.org/10.1007/s43926-023-00049-y

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 131–135). https://doi.org/10.1109/ICASSP.2017.7952132

Humphrey, E., Bello, J. P., & LeCun, Y. (2012). Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) (pp. 403-408). http://yann.lecun.com/exdb/publis/pdf/humphrey-ismir-12.pdf

ISSN 2394 - 7780

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

Lane, N. D., Georgiev, P., & Qendro, L. (2015). DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) (pp. 283–294). https://doi.org/10.1145/2750858.2804262

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre classification. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 282-289). ACM. https://www.researchgate.net/publication/221299170_A_Comparative_Study_on_Content-Based_Music_Genre_Classification

McFee, B., et al. (2015). Librosa: Audio and music signal analysis in Python. In Proceedings of the 14th Python in Science Conference (SciPy) (pp. 18-25). https://proceedings.scipy.org/articles/Majora-7b98e3ed-003.pdf

Pearce, H., Yang, X., Roop, P. S., Katzef, M., & Strøm, T. B. (2021). Designing neural networks for real-time systems. IEEE Embedded Systems Letters, 13(3), 94-97. https://doi.org/10.1109/LES.2020.3009910

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279-283. https://doi.org/10.1109/LSP.2017.2657381

Theodoridis, S., & Koutroumbas, K. (2009). Pattern recognition (4th ed.). Academic Press. https://www.perlego.com/book/1813658/pattern-recognition-pdf

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

Virtanen, T., Plumbley, M. D., & Ellis, D. (2018). Computational analysis of sound scenes and events. Springer. https://doi.org/10.1007/978-3-319-63450-0

Won, M., Choi, K., & Serra, X. (2021). Evaluation of real-time music classification systems under streaming constraints. arXiv. https://arxiv.org/abs/2111.13457