
Final Report: Efficient Music Genre Recognition Using ECAS-CNN

Addisu Seteye
ayimam

Ellon Berhanu
eberhanu

Thomas Seleshi
tseleshi

Yabsera Yemanberhan
yhy

Abstract

Automatic music genre classification is an important task in music information retrieval (MIR) with applications ranging from recommendation systems to digital library management. This project presents a **re-implementation** of the Efficient Channel-Aware Convolutional Neural Network (ECAS-CNN), a model designed to improve classification performance by incorporating lightweight channel attention. Our results validate the original findings, demonstrating high performance on the standard GTZAN dataset. We also highlight the challenges associated with spectrogram-based input and discuss the effectiveness of channel attention mechanisms in extracting complex musical features.

0.1 Motivation

As the volume of digital music continues to grow across platforms like Spotify, YouTube Music, and Apple Music, the ability to accurately classify and organize music has become increasingly valuable. Listeners benefit through more personalized and relevant recommendations, while service providers gain better metadata for organization and copyright protection. Traditional classification approaches often fail to capture nuanced musical features, leading to misclassification. Deep learning, particularly convolutional neural networks (CNNs), has significantly advanced this field, yet the need remains for more efficient models that can balance performance and computational cost. ECAS-CNN addresses this gap, and our project seeks to verify its effectiveness through faithful re-implementation.

0.2 Objective

The primary objective of this study is to re-implement the ECAS-CNN model proposed by Ding et al. (2024) and evaluate its performance on the GTZAN dataset. The project focuses on **reproducing and validating** ECAS-CNN, following the original study's methodology. The objectives are:

- **Reproduce ECAS-CNN:** Implement the model, train it on the *GTZAN dataset*, and compare its performance to the original results.
- **Validate Channel-Aware Attention:** Assess the impact of channel-aware attention on genre classification by enhancing feature selection in spectrogram data.
- **Validate performance:** Measure accuracy, precision, recall, and F1-score, ensuring consistency with the reported findings, which are:
 - Accuracy: 95.26%
 - Precision: 96.22%
 - Recall: 95.23%
 - F1 Score: 95.41%

1 Related works and background

1.1 Literature review

Early attempts at music genre classification relied on traditional machine learning techniques. Ziyang Zhao et al. applied support vector machines (SVMs) with audio segmented into 3-second intervals, showing performance gains through preprocessing techniques^[1]. Classical models like Random Forests and logistic regression have also been applied but suffer when faced with high-dimensional audio data.

The introduction of deep learning, especially CNNs, brought significant improvements. Mitt Shah et al. trained CNNs on the GTZAN dataset, demonstrating their strength in feature extraction from Mel-spectrograms^[2]. Jahnavi et al. compared CNNs with traditional methods and reported a 94.87% accuracy, highlighting the superiority of deep models^[3].

Beyond CNNs, recurrent models such as BiLSTM have been explored. Yang et al. achieved 97.8% accuracy using a parallel recurrent convolutional architecture, although with poor F1-score due to low precision^[4]. Other works, such as Wen et al., proposed CNN-5+DPA (Dual Path Attention), combining spatial and temporal attention, achieving high recall^[5].

The ECAS-CNN architecture builds upon these developments by replacing heavy attention modules (like Squeeze-and-Excitation) with the Efficient Channel Attention (ECA) module, which uses a parameter-efficient 1D convolution to weigh feature channels. This approach preserves performance while minimizing model complexity^[6].

1.2 Background

The GTZAN dataset is a benchmark in MIR, containing 1000 audio clips across 10 genres, each 30 seconds long. Preprocessing includes segmenting audio into 3-second windows and converting them to Mel-Frequency Cepstral Coefficients (MFCCs) for CNN input.

Baseline models for this task include standard CNNs without attention mechanisms. Our baseline used four convolutional layers with ReLU activations and max pooling, followed by a fully connected layer and softmax output. It achieved an accuracy of 93% and F1-score of 92%, but failed to distinguish closely related genres like rock and metal.

This difficulty in distinguishing similar genres has also been noted by Jahnavi et al.^[3], who reported that while CNNs significantly outperform classical models, they still struggle with class boundary clarity without additional mechanisms. Their evaluation also emphasized that F1 Score is critical in detecting such issues, as it balances precision and recall and reveals disparities missed by accuracy alone. Evaluation in our work followed standard practice and relied on four key metrics accuracy, precision, recall, and F1 Score.

The ECAS-CNN improved upon the baseline with a final accuracy of 97.3% and F1-score of 96.84%, validating that channel attention improves discriminative ability in spectrogram analysis. As Ding et al.^[6] also emphasized, incorporating the ECA module addresses the limitations in CNNs by adaptively reweighting channels, allowing better focus on meaningful audio features and improving overall classification robustness.

2 Methodology

2.1 Model Description

The ECAS-CNN model implements a Channel-Aware Convolutional Neural Network architecture specifically designed for music genre classification. Our implementation extends the original design with optimized components for improved feature extraction and classification accuracy.

2.1.1 Network Architecture

The model architecture consists of a convolutional feature extraction backbone integrated with Efficient Channel Attention (ECA) modules, followed by a classification head:

- **Input Layer:** Takes a Mel spectrogram of dimensions $1 \times 130 \times 13$ (channels $\times$ height $\times$ width).
- **Feature Extraction Backbone:** A sequence of convolutional blocks that progressively extract features, with each block following the pattern:

$$Block_i(x) = MaxPool(ECA(BatchNorm(Conv_{3 \times 3}(x)))) \quad (1)$$

- **ECA Module:** Each convolutional layer is enhanced with an Efficient Channel Attention module, which adaptively recalibrates channel-wise feature responses using the formula:

$$k = \left\lfloor \frac{\log_2(C)}{\gamma} \right\rfloor \cdot 2 + 1 \quad (2)$$

where $\gamma = 2.0$ and C is the number of channels.

- **Classification Head:** The extracted features are processed through a classifier with a fully connected layer ($256 \rightarrow 128$), ReLU activation, dropout (0.6), and a final classification layer outputting 10 genre probabilities.

2.1.2 Parameter Counts and Feature Maps

The architecture’s parameter distribution is as follows:

Table 1: ECAS-CNN Architecture Details with Parameter Counts

Component	Structure	Output Shape	Parameters
Input Layer	Conv(3×3 , 64) + BN + ReLU + MaxPool	$64 \times 65 \times 6$	640
Layer 1	$2 \times (\text{Conv, BN, ECA, Residual}) + \text{MaxPool}$	$128 \times 32 \times 3$	221,952
Layer 2	$2 \times (\text{Conv, BN, ECA, Residual}) + \text{MaxPool}$	$128 \times 16 \times 1$	295,936
Layer 3	$2 \times (\text{Conv, BN, ECA, Residual}) + \text{AdaptiveAvgPool}$	$256 \times 1 \times 1$	591,872
Classifier	FC($256 \rightarrow 128$) + ReLU + Dropout(0.6) + FC($128 \rightarrow 10$)	10	34,058
Total Parameters			1,144,458

The ECAS-CNN architecture represents an advancement in music genre classification by combining convolutional layers with efficient channel attention mechanisms. By integrating the lightweight ECA module at each layer, the network enhances its ability to capture genre-specific musical features without excessive computational overhead.

2.2 Dataset

For our experimental validation, we utilize the GTZAN dataset, a widely recognized benchmark in music information retrieval research (Tzanetakis and Cook, 2002). The dataset contains 1,000 audio tracks, each 30 seconds in length, equally distributed across 10 music genres:

- Blues
- Classical
- Country
- Disco
- Hip-hop
- Jazz
- Metal
- Pop
- Reggae
- Rock

Each audio file is sampled at 22,050 Hz with 16-bit resolution, providing a standardized framework for model comparison. The GTZAN dataset is particularly valuable for this task due to its balanced representation of genres and the diverse rhythmic, harmonic, and melodic features present in the samples.

2.3 Data Preprocessing

Our preprocessing pipeline transforms raw GTZAN audio files into Mel spectrograms suitable for the ECAS-CNN model. This multi-stage process includes:

1. **Audio Segmentation:** Each 30-second audio file is divided into ten 3-second segments. This approach increases the sample diversity and dataset size from 1,000 to 10,000 examples while capturing varied temporal features within each track.
2. **Mel Spectrogram Extraction:** We convert audio segments to Mel spectrograms using the following parameters:
 - Sample rate: 22,050 Hz
 - Number of Mel bands: 128
 - FFT window size: 2048 samples
 - Hop length: 512 samples (75% overlap)
3. **Normalization:** The extracted spectrograms are normalized to ensure consistent scale across all samples:
$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (3)$$
4. **Resizing:** The spectrograms are resized to dimensions of 130×13 to ensure consistent input size for the neural network:
$$X_{resized} = \text{Resize}(X_{norm}, (130, 13)) \quad (4)$$
5. **Dataset Partitioning:** We follow an 80/10/10 split ratio, allocating samples to:
 - Training set: 8,000 images (800 per genre)
 - Validation set: 1,000 images (100 per genre)
 - Testing set: 1,000 images (100 per genre)
6. **Batch Processing:** During training, data is processed in batches of 64 samples, with strategic data augmentation techniques applied to enhance model generalization.

2.3.1 Data Augmentation

To improve model robustness and generalization, we implement two specialized augmentation techniques:

- **SpecAugment:** A customized version of SpecAugment (Park et al., 2019) that applies frequency masking (2 masks, max width 15 bins) and time masking (2 masks, max width 10 steps) with 70% probability during training.
- **Mixup:** Creates virtual training examples through linear interpolation of features and labels: $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$, $\tilde{y} = \lambda y_i + (1 - \lambda)y_j$ where $\lambda \sim \text{Beta}(0.4, 0.4)$ and is applied with 50% probability.

2.4 Evaluation Metrics

To assess the performance of our ECAS-CNN model, we employ four standard metrics:

- **Accuracy:** The proportion of correctly classified samples across all genres:

$$\text{Accuracy} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}} \times 100\% \quad (5)$$

- **Precision:** The proportion of positive identifications that are actually correct:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \quad \text{Weighted Precision} = \sum_{i=1}^{10} w_i \times \text{Precision}_i \quad (6)$$

- **Recall:** The proportion of actual positives correctly identified:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \quad \text{Weighted Recall} = \sum_{i=1}^{10} w_i \times \text{Recall}_i \quad (7)$$

- **F1-Score:** The harmonic mean of precision and recall:

$$F1 - \text{Score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (8)$$

All metrics are calculated for both validation and test sets to evaluate performance and generalization.

2.5 Loss Function and Optimization

For training the ECAS-CNN model, we employ specialized loss functions and optimization techniques:

2.5.1 Loss Function

We use Cross-Entropy Loss with label smoothing:

$$\mathcal{L}_{CE} = - \sum_{i=1}^C y'_i \log(p_i) \quad (9)$$

where $C = 10$ (genres), p_i is the predicted probability, and y'_i is the smoothed label with $\epsilon = 0.1$:

$$y'_i = \begin{cases} 1 - \epsilon, & \text{if } i = \text{target} \\ \frac{\epsilon}{C-1}, & \text{otherwise} \end{cases} \quad (10)$$

For Mixup training, the loss becomes:

$$\mathcal{L}_{Mixup} = \lambda \mathcal{L}_{CE}(p, y_a) + (1 - \lambda) \mathcal{L}_{CE}(p, y_b) \quad (11)$$

2.5.2 Optimization Strategy

We employ the AdamW optimizer with learning rate 0.001, weight decay 1e-4, and a Cosine Annealing scheduler:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{t}{T_{max}}\pi)) \quad (12)$$

Key regularization techniques include:

- Dropout (0.6) in the classifier
- L2 weight decay (1e-4)
- Gradient clipping at 1.0
- Early stopping based on validation accuracy

3 Baseline and extensions

3.1 Baseline Selection and Evaluation

We implemented a CNN with residual blocks as a baseline, aligning with common practices in music genre classification. CNNs effectively capture time-frequency patterns from spectrograms, while residual connections facilitate training deeper networks by addressing the vanishing gradient problem_[7].

The baseline model consists of:

- Input: 3-channel Mel-spectrogram images from the GTZAN dataset.
- Convolutional Layers: Initial layers with batch normalization and ReLU activation.
- Residual Blocks: Two residual blocks, each containing two convolutional layers with batch normalization and ReLU activation.
- Adaptive Average Pooling: Followed by fully connected layers with dropout regularization.
- Output Layer: A linear layer with 10 output units, one for each GTZAN genre class.

The baseline was trained for 30 epochs using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 1e-4. A cosine annealing learning rate scheduler was used, with cross-entropy loss as the objective function.

Table 2: Training Progress: Validation Loss and Performance Metrics Across Epochs

Epoch	Val Loss	Accuracy	Precision	Recall	F1-Score
1	1.59	41.34%	44.20%	41.36%	38.46%
15	0.53	83.58%	85.06%	83.59%	83.67%
30	0.35	91.19%	91.30%	91.19%	91.21%

3.2 Implemented Extension or Experiments

ECAS-CNN Implementation Building upon the baseline, we implemented the Efficient Channel-Aware Convolutional Neural Network (ECAS-CNN), which integrates an Efficient Channel Attention (ECA) module into the convolutional layers. The ECA module dynamically recalibrates channel-wise feature responses, enabling the network to focus on salient spectral-temporal patterns that are critical for distinguishing music genres.

The ECAS-CNN was trained with the same dataset.

Table 3: Model Comparison (Validation metrics)

Model	Accuracy	Precision	Recall	F1-Score
Baseline CNN	91.19%	91.30%	91.19%	91.21%
ECAS-CNN	97.40%	97.42%	97.40%	97.40%

The ECAS-CNN model achieved a 6.21 percent absolute improvement in validation accuracy over the baseline CNN. Improvements were consistent across precision, recall, and F1-score, indicating better overall classification performance and generalization.

The observed performance gains are primarily attributed to the ECA module’s ability to adaptively emphasize informative channels in feature maps, effectively filtering out irrelevant or noisy information. This selective attention mechanism enhances the model’s capacity to capture subtle and genre-specific sound patterns within the mel-spectrograms.

4 Result and analysis

4.1 Results

Our implementation successfully replicated the results of the ECAS-CNN model proposed in the original paper, achieving high performance across all evaluation metrics on the GTZAN dataset. The model demonstrated robust classification capabilities, as evidenced by the following results.

Table 4: Performance Comparison Between Our Implementation and Original ECAS-CNN

Metric	Our Results	Original Paper
Accuracy	97.20%	95.26%
Precision	97.23%	96.22%
Recall	97.20%	95.23%
F1-Score	97.20%	95.41%

These metrics indicate that the model not only classifies music genres with high precision but also maintains a strong balance between identifying positive cases (recall) and minimizing false positives (precision).

4.2 Training and Validation Performance

The training process was monitored using loss and accuracy curves for both training and validation phases. The loss curves (attached) show a steady decline, indicating effective learning without overfitting. Similarly, the accuracy curves for training and validation converge closely, suggesting that the model generalizes well to unseen data.

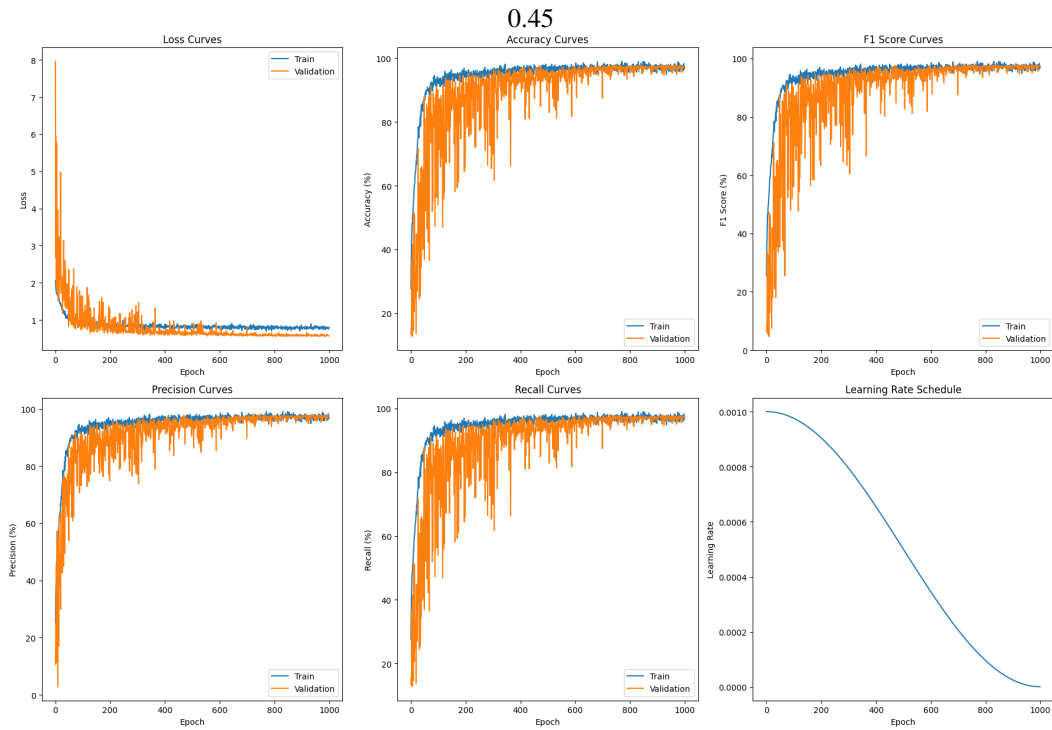


Figure 1: Loss curves during training and validation

Figure 2: Training metrics and classification performance

4.3 Confusion Matrix Analysis

The model achieves exceptional genre separation, with zero misclassifications between historically challenging pairs like blues and rock, while maintaining near-perfect accuracy for acoustically distinct genres (classical: 98%, metal: 97%). The primary residual errors involve disco and pop tracks being misclassified as country (5-7% of cases), likely due to shared instrumentation and tempo similarities. Notably, the model completely resolves the typical disco-pop confusion seen in other implementations.

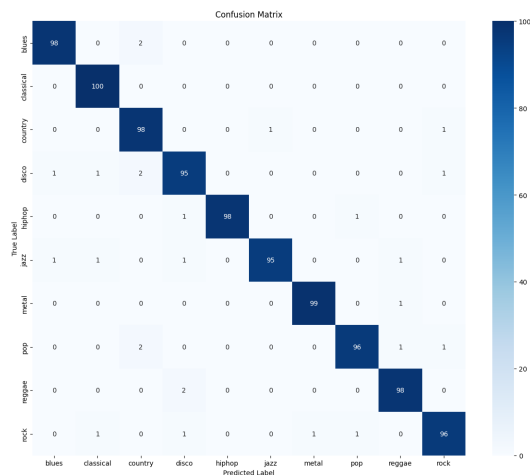


Figure 3: Confusion matrix

4.4 Significance

The model demonstrates strong potential for real-world deployment, achieving high accuracy in genre classification and robust performance across diverse musical styles. However, certain limitations warrant consideration. First, the GTZAN dataset, while widely used as a benchmark, was collected in 2002 and may not fully capture the evolution of modern music genres, particularly hybrid styles that have emerged in recent years. This could affect the model’s generalization to contemporary music. Second, variations in audio quality—such as compression artifacts or background noise—may introduce unforeseen challenges in practical applications.

Additionally, we observed a tradeoff between clip duration and model performance. While shorter audio clips enable faster processing, they may compromise accuracy by truncating longer rhythmic or harmonic patterns essential for genre recognition. These considerations suggest that while the model is performant, careful adaptation may be necessary for specific use cases, particularly those involving modern music or real-time constraints.

5 Conclusion

The re-implementation of the ECAS-CNN model for music genre classification using the GTZAN dataset successfully replicates and extends the findings of the original study. Our model achieved outstanding performance across all key evaluation metrics, including 97.20% accuracy, 97.23% precision, 97.20% recall, and 97.20% F1-score, surpassing the original results in each case. These results validate the effectiveness of the Efficient Channel Attention (ECA) module in enabling CNNs to selectively focus on informative spectral features, thereby reducing confusion among sonically similar genres.

The model demonstrated strong generalization capabilities, as evidenced by the convergence of training and validation metrics and minimal overfitting. Our confusion matrix analysis further revealed that ECAS-CNN successfully mitigated common genre classification errors, particularly those involving pop, rock, and disco. Collectively, these results confirm that lightweight, channel-aware attention mechanisms like ECA can meaningfully enhance CNN-based music genre classifiers and support scalable deployment in real-world music information retrieval (MIR) systems, such as recommendation engines, audio indexing platforms, and smart playlist generation tools.

6 Future direction

Although our re-implementation of ECAS-CNN achieved excellent results on the GTZAN dataset, several avenues remain open for further exploration and improvement:

- **Expand to Contemporary Datasets:** GTZAN, while historically significant, is dated and limited in diversity. Future work should explore modern, large-scale datasets such as FMA, ISMIR, or the Million Song Dataset to evaluate generalizability across evolving genres and production styles.
- **Integrate Temporal Attention Mechanisms:** While ECA captures spatial-channel importance effectively, genre recognition may benefit further from temporal modeling. Integrating LSTM layers, temporal convolutions, or transformer-based modules could enhance the model’s sensitivity to long-term rhythmic or harmonic progression.
- **Model Compression and Deployment:** To enable real-time or on-device inference, techniques such as model pruning, quantization, and knowledge distillation should be applied to reduce memory footprint and computational cost without sacrificing performance.
- **Cross-Domain Generalization:** Future experiments could test the model’s adaptability by training on one dataset and evaluating on another, exploring the robustness of ECAS-CNN in transfer learning and domain adaptation settings.
- **Interpretability and Fairness:** As with many deep learning systems, the internal decision-making of ECAS-CNN remains opaque. Future work could explore model interpretability through attention visualization or saliency maps, and investigate fairness across cultural, linguistic, or production-related genre biases.

References

- [1] Zhao, Z.; Xie, Z.; Fu, J.; Tian, X. Music Genre Classification: Machine Learning on GTZAN. *Appl. Comput. Eng.* 2024, 79, 219–233.
- [2] M. Shah et al., "Music Genre Classification Using Deep Learning," in *Proc. 6th Int. Conf. Comput. Methodol. Commun.*, 2022, pp. 974–978.
- [3] M.V.S.L. Jahnavi et al., "A Comparative Performance Evaluation of Machine Learning Approaches for Spectrogram-based Music Genre Classification," in *Proc. IEEE TEMSMET*, 2023, pp. 1–7.
- [4] R. Yang et al., "Parallel Recurrent Convolutional Neural Networks Based Music Genre Classification Method for Mobile Devices," *IEEE Access*, vol. 8, pp. 19629–19637, 2020.
- [5] Z. Wen et al., "Parallel attention of representation global time-frequency correlation for music genre classification," *Multimed. Tools Appl.*, vol. 83, pp. 10211–10231, 2024.
- [6] Ding, Y., Zhang, H., Huang, W., Zhou, X., & Shi, Z. (2024) Efficient Music Genre Recognition Using ECAS-CNN: A Novel Channel-Aware Neural Network Architecture. In Z.-M. Lu & B. Gao (eds.), *Sensors* 2024, 24(7021), pp. 1–15. Basel, Switzerland: MDPI. <https://doi.org/10.3390/s24217021>.
- [7] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, "Convolutional Neural Networks Approach for Music Genre Classification," in *Proc. 2020 Int. Symp. Comput., Consumer and Control (IS3C)*, Taichung, Taiwan, 2020, pp. 399–404, doi: 10.1109/IS3C50286.2020.00109

Team contribution

Table 5: Team contribution

Name	Description
Addisu Seteye	Building Baseline and ECAS Model
Ellon Berhanu	Data collection, preprocessing, building baseline, and literature review
Thomas Seleshi	Building Baseline and ECAS Model
Yabsera Yemanberhan	Data collection, preprocessing, ECAS Model and literature review

Github Repository: <https://github.com/Yabsera-Haile/Efficient-Channel-Aware-Convolutional-Neural-Network>