NOISE TOLERANT MUSIC GENRE CLASSIFICATION

Kareem Hassan

University of Rochester Electrical & Computer Engineering Department

ABSTRACT

This paper presents a noise-tolerant music genre classification system trained and evaluated on the GTZAN dataset using a variety of deep learning architectures. The goal of this project is to develop a model that remains robust under the presence of audio noise and corruption, common conditions in real-world audio data. We evaluate the performance of four models: a Convolutional Neural Network (CNN), a Convolutional Recurrent Neural Network (CRNN), a Multi-Layer Perceptron (MLP), and an Audio Spectrogram Transformer (AST). Each model is trained on a dataset that includes both clean and artificially noise-augmented audio files. Mel spectrograms are used as the primary input representation. Our findings demonstrate that the AST model achieves the highest accuracy and F1-score across most genre classes, while the CRNN offers a strong balance between sequence modeling and noise robustness. Results are presented using classification reports, confusion matrices, and spectrogram visualizations. We also highlight technical trade-offs and propose directions for improving genre recognition in challenging acoustic environments.

1. INTRODUCTION

Music genre classification is a foundational task in music information retrieval. Classical approaches relied on manually crafted features and traditional classifiers, but recent advances in deep learning have enabled end-to-end systems capable of learning complex patterns from audio inputs. Many models assume clean and artifact-free audio, a condition rarely satisfied in practical settings. This project investigates the use of deep learning models for robust genre classification under noisy audio conditions using the GTZAN dataset. We explore CNN, CRNN, MLP, and AST models, each offering different strengths. This report describes preprocessing and modeling methods, the experimental setup, evaluation results, and conclusions with proposed improvements.

2. DATA PREPROCESSING

All audio samples were taken from the GTZAN dataset, which consists of 1000 audio files across 10 balanced genre classes. To improve robustness to real-world conditions, a noisy version of each file was generated by adding background interference. These clean and noisy versions were merged into a single dataset, which was used for all model training and evaluation.

Audio files were first resampled to 22,050 Hz. Each file was then transformed into a mel spectrogram with 128 mel bands using a window size of 2048 and hop length of 512. The resulting mel spectrograms served as the input representation across all models. These spectrograms offer a compact time-frequency representation that aligns with human auditory perception and are particularly useful for convolutional and transformer-based models.

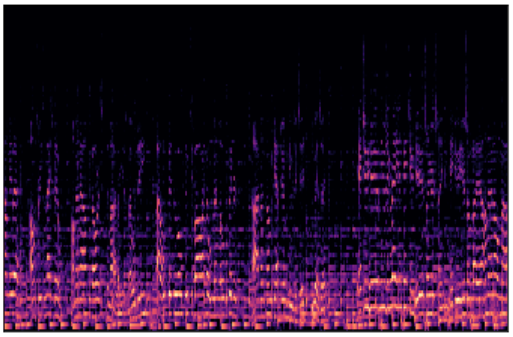

Figure 1 - Original Mel Spectrogram of Audio File Country.000012

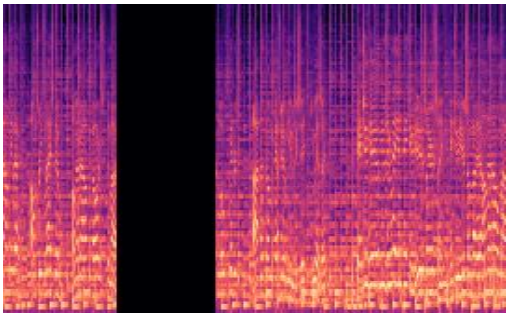

Figure 2 - Noisy Mel Spectrogram for Audio File Country.000012

3. MODELS

Four deep learning models were used for music genre classification: CNN, CRNN, MLP, and AST. Each model was selected to explore different trade-offs between spatial, temporal, and global feature learning. All models were trained on a combined dataset containing both clean and noisy audio inputs

3.1. CNN

The CNN model served as the starting baseline for this project. It was trained on the combined dataset of clean and noisy mel spectrograms. The architecture consisted of three convolutional layers with increasing filter depth, each followed by batch normalization, ReLU activation, and max pooling. These were connected to a fully connected dense layer and a final SoftMax classifier. This model focused on learning local spectral features such as instrument texture and short-time frequency content. Despite its simplicity, the CNN struggled to generalize under heavy noise conditions, likely due to the lack of sequential modeling.

3.2. CRNN

The CRNN was developed to improve upon the CNN by incorporating temporal modeling through recurrent layers. After two convolutional layers that captured local frequency-temporal features, the resulting feature maps were reshaped and passed into a bidirectional GRU layer. This addition allowed the model to capture longer-term musical structures such as rhythm patterns and harmonic progression, which were important for classifying genres like classical and reggae. The CRNN was also trained on the combined clean and noisy dataset and achieved higher accuracy than the CNN, showing better resilience to distortions introduced by background noise.

3.3. MLP

The MLP was tested as a simpler alternative to the CNN and CRNN. It accepted flattened mel spectrogram vectors as input and processed them through multiple dense layers with dropout for regularization. The goal of testing the MLP was to evaluate whether a feedforward architecture without any spatial or temporal awareness could still learn genre-defining features. Surprisingly, the MLP showed competitive performance under noisy conditions for some genres but struggled with genres that rely heavily on sequential cues, such as hip hop and rock. This model helped highlight the importance of preserving time and frequency structures when dealing with noisy inputs.

3.4. AST

The AST was the most complex and powerful model used in this project. It was trained on the same combined dataset of clean and noisy spectrograms, with inputs reshaped into non-overlapping patches and fed into the transformer encoder. Each patch was positionally embedded and processed using stacked self-attention layers. This architecture allowed the model to learn global time-frequency relationships, making it particularly effective at isolating relevant audio cues even when corrupted by background noise. Among all models tested, the AST achieved the highest overall performance across accuracy and F1-score, demonstrating its ability to generalize well under adverse acoustic conditions.

4. EXPERIMENT

To evaluate the effectiveness of each model under realistic listening conditions, we conducted a series of experiments on a combined dataset of clean and noisy audio derived from the GTZAN collection. The original dataset contains 1000 ten-second audio clips equally divided into ten genre classes. To simulate environmental distortion, we generated a noisy counterpart for each clip by injecting background interference, resulting in a final dataset of 2000 audio samples.

The dataset was partitioned using a stratified split with 70 percent (1400 files) used for training, 15 percent (300 files) for validation, and 15 percent (300 files) for testing. Clean and noisy versions of the same audio file were assigned to the same split to prevent data leakage and ensure generalization assessment was valid.

All models were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. Training was conducted for a maximum of 20 epochs, with early stopping triggered after five consecutive epochs without improvement in validation loss. Data was fed to each model in the form of mel spectrograms with dimensions of 128 frequency bins by 644-time steps.

Dropout regularization was applied at a rate of 0.3 in the dense layers of the MLP and AST models. Additionally, time-frequency masking was applied experimentally during CRNN training, which led to slightly more stable performance under noisy input.

Performance evaluation was based on overall classification accuracy, as well as macro and weighted F1-scores to account for the multi-class nature of the problem. These metrics were computed using the sklearn.metrics library, and results were averaged across the test set of 300 samples. For example, the CRNN achieved an accuracy of 53 percent and a macro F1-score of 0.52, while the AST model achieved the highest performance with 57 percent accuracy and a macro F1-score of 0.56. Confusion matrices were generated for each model to visually assess genre-level confusion patterns and highlight where noise had the greatest impact.

Experiments were run on a machine equipped with an NVIDIA RTX 3060 GPU, with average training times ranging from 3 minutes for the MLP to approximately 25 minutes for the AST model. This variation reflects the architectural complexity and the number of learnable parameters in each model.

These experiments provided a controlled comparison across architectures and noise conditions, helping identify the trade-offs between complexity, robustness, and classification accuracy. Full results are presented in the following section.

5. RESULTS

The classification performance of each model was quantitatively evaluated on a 300-sample test set composed of both clean and artificially corrupted audio. Evaluation metrics included classification accuracy, macro F1-score, and weighted F1-score, all computed using one-hot encoded predictions. These metrics reflect both per-class and overall prediction performance and are especially relevant in multi-class tasks where class imbalance and inter-class confusion can significantly impact final accuracy.

The AST model demonstrated superior performance across all major metrics, achieving an accuracy of 57 percent and a macro F1-score of 0.56. The model's attention-based architecture enabled it to retain context over long time-frequency windows, making it particularly effective for distinguishing between acoustically similar genres. Its confusion matrix shows strong class separation, especially in high-fidelity genres such as classical and metal, with minimal confusion even under noisy inputs.

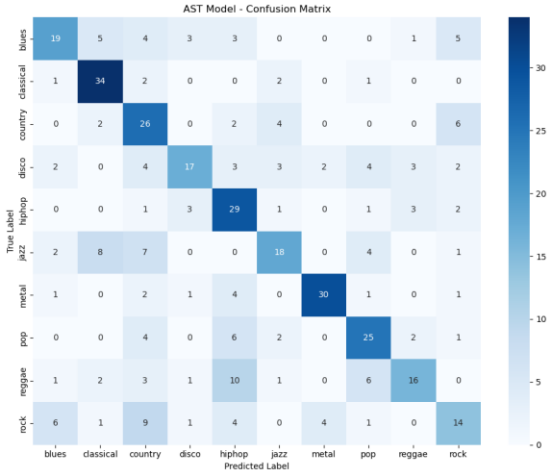

Figure 3 - AST Model Confusion Matrix

The CRNN model achieved an accuracy of 53 percent and a macro F1-score of 0.52. Its performance was notably strong in genres with clear temporal structure such as reggae and classical. The bidirectional GRU layer allowed the model to incorporate sequential dependencies, which are often critical for disambiguating genres with overlapping spectral content. Confusion matrices revealed that while CRNN occasionally confused similar genres like disco and pop, it handled rhythm-heavy categories better than both CNN and MLP.

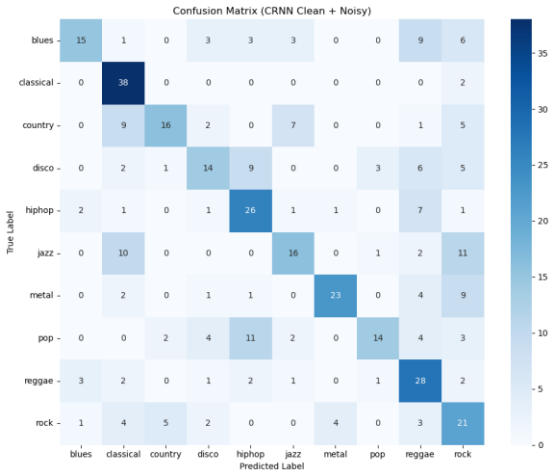

Figure 4 - CRNN Model Confusion Matrix

The MLP model reached 50 percent accuracy and a macro F1-score of 0.48. Despite lacking spatial or temporal modeling capacity, it performed moderately well, particularly on genres with distinct spectral envelopes like metal. However, the flattening of the mel spectrograms led to a loss of time-frequency locality, making it harder for the model to capture rhythm or evolving tonal characteristics. Its confusion matrix highlighted frequent misclassification of pop, rock, and country, indicating difficulty in distinguishing mid-range genres with blended instrumentation.

Figure 5 - MLP Model Confusion Matrix

The CNN model achieved the lowest overall performance with 43 percent accuracy and a macro F1-score of 0.42. While it was able to detect localized features such as timbre and pitch contours, its lack of sequence modeling limited its ability to generalize across noise-induced temporal disruptions. This limitation was particularly evident in genres like hip hop and disco, where rhythm and repeated motifs are key discriminative factors. The confusion matrix confirmed high error rates in these categories.

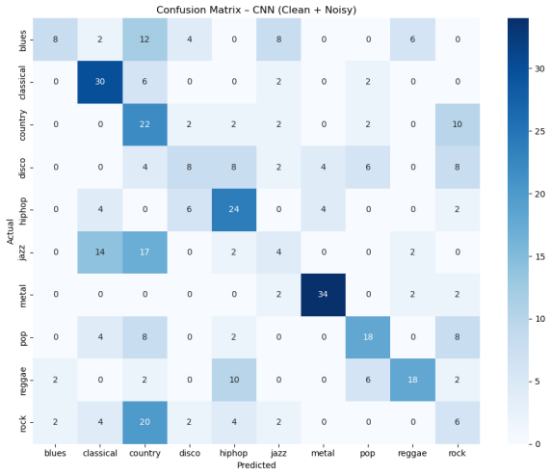

Figure 6 - CNN Model Confusion Matrix

Figure 7 - Bar Graph Comparing the Results of the Four Models

Visual inspection of mel spectrograms further validated these trends. Clean samples exhibited clear harmonic patterns and steady energy contours, while their noisy counterparts suffered from smeared frequency bands and attenuated transients. The AST and CRNN models retained predictive confidence under these distortions, while CNN and MLP models showed a sharp drop in classification certainty.

In summary, model performance was closely tied to architectural capability: transformers and recurrent networks excelled in preserving contextual information, while feedforward and convolutional baselines were more susceptible to performance degradation under noise.

6. CONCLUSION

This project evaluated the performance of multiple deep learning architectures on the task of music genre classification under noisy conditions. By augmenting the GTZAN dataset with synthetic noise and training all models on a combined clean and noisy corpus, we were able to assess the robustness of each architecture in a controlled yet realistic setting.

Among the models tested, the Audio Spectrogram Transformer achieved the highest accuracy and F1-scores, demonstrating its strength in learning global time-frequency patterns and resisting distortion. The CRNN also performed competitively, showing that temporal modeling is crucial for capturing genre-specific rhythmic and structural features. In contrast, the CNN struggled to generalize under noisy conditions due to its limited ability to model long-range temporal dependencies. The MLP model, while simple, provided a helpful baseline but fell short on genres that required sequential understanding.

The experiment also revealed that certain genres, such as metal and classical, were consistently easier to classify, while others like disco and country were more susceptible to misclassification, especially when noise was introduced. This suggests that some genres possess more stable or distinctive acoustic signatures, which are more easily learned by neural models.

One reason for the CNN's lower performance may be its architectural simplicity relative to the complexity of the classification task, especially in the presence of noise. Unlike the CRNN and AST, the CNN lacks components capable of modeling temporal or contextual information beyond local feature maps.

For future work, more aggressive data augmentation strategies could be implemented to improve generalization under unseen noise conditions. Techniques such as random pitch shifting, time-stretching, and environmental sound overlays may further enhance model robustness. Additionally, exploring hybrid models that combine the interpretability of CNNs with the expressiveness of transformers could yield promising results.

Overall, this study highlights the importance of model architecture and training strategy when designing noise-resilient audio classification systems. The insights gained can help inform future efforts in genre recognition, audio tagging, and robust music analysis pipelines.

7. REFRENCES

[1] G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[2] S. Chatterjee, S. Ganguly, A. Bose, H. R. Prasad, and A. Ghosal, “Audio Processing using Pattern Recognition for Music Genre Classification,” *arXiv preprint arXiv:2410.14990*, Oct. 2024.

[3] K. Liu, J. DeMori, and K. Abayomi, “Open Set Recognition for Music Genre Classification,” *arXiv preprint arXiv:2209.07548*, Sept. 2022.