Deep Learning Project Report

Aldridge Albert Abaasa

Abstract

This deep learning project aims to evaluate the impact of various preprocessing techniques on audio data using three different models: Simple Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). The study evaluates three different preprocessing techniques: no preprocessing, fixed sample rate and duration using trimming, padding, and resampling techniques. The models were trained on the GTZAN music genre dataset, which contains 100 audio clips of ten different music genres.

To evaluate the impact of preprocessing techniques, several hyperparameters were adjusted, including batch sizes, epochs, number of layers, number of neurons, and learning rate. The accuracy of each model was evaluated using the mean accuracy and standard deviation for each preprocessing technique. The results were analyzed to determine the most effective preprocessing technique for each model.

The experiments were implemented in TensorFlow, and the dataset was preprocessed using the Librosa audio processing library in Python. The models were trained using the Adam optimizer with a cross-entropy loss function, and the accuracy metric was used to evaluate model performance.

The study's findings have important implications for audio data processing and deep learning applications, such as music genre classification and speech recognition. By comparing the accuracy of each model under different preprocessing techniques, the study provides insights into which approach is most effective for analyzing audio data. This information can be used to improve the accuracy of audio-based applications.

1 Introduction

Audio data processing using deep learning has become increasingly important in many different applications due to the availability of large-scale datasets and advances in technology. Deep learning algorithms can automatically learn complex features from raw audio data without the need for manual feature engineering. This study aims to comprehensively evaluate the impact of different preprocessing techniques on the accuracy of deep learning models for audio data classification using the widely used GTZAN music genre dataset. The study seeks to provide insights into the best practices for preprocessing audio data and building deep learning models for accurate audio classification, with potential applications in speech recognition, music genre classification, and audio-based surveillance systems. The dataset consists of 1,000 audio clips divided into 10 genres. The study seeks to answer research questions related to the impact of different preprocessing techniques and the performance of different types of deep learning models on the dataset with different preprocessing techniques.

The study aims to provide insights into the strengths and weaknesses of different types of deep learning models for audio classification. Specifically, the study examines the impact of different preprocessing techniques on the accuracy of four models: MLP, CNN, RNN, and CRNN, using the GTZAN music genre dataset. The contribution of this study lies in its evaluation of the effectiveness

Preprint. Under review.

Table 1: Contents within the GTZAN Audio Dataset:

Item	Range
Genre	10
Length (seconds)	29.9- 30.6
Sample rate (Hz)	22,050 - 44100
Corrupt file	1
Total audio clips	1000

of various preprocessing techniques in improving model performance, and its comparison of three different types of deep learning models with varying degrees of complexity. Ultimately, the study aims to provide guidance on the most effective approach to audio classification for music genre recognition.

2 Related Work

Audio data processing is crucial for speech recognition, music genre classification, and sound event detection. Deep learning is a powerful tool for analyzing audio data, but different preprocessing techniques can affect the accuracy of deep learning models. This study explores the impact of various preprocessing techniques on the accuracy of four deep learning models in music genre classification. Feature extraction, signal processing, and preprocessing are essential techniques in audio data processing. Deep learning models, including MLPs, CNNs, RNNs, and CRNNs, have unique strengths and limitations. Understanding and appropriately applying audio data processing techniques and deep learning models can improve the performance of audio analysis systems and enable a wide range of applications. A study by Dhall et al. (2021) compared the performance of deep learning models on speech emotion recognition tasks with and without preprocessing techniques such as normalization, segmentation, and feature extraction. The study found that preprocessing techniques significantly improved the accuracy of deep learning models for speech emotion recognition. Another study by Cheng and Kuo (2022) investigated the impact of various preprocessing techniques on the accuracy of deep learning models for music genre classification tasks. The study evaluated the performance of deep learning models on the GTZAN dataset with preprocessing techniques such as spectrogram conversion, mel-frequency cepstral coefficients (MFCC), and chroma feature extraction. The study found that MFCC and chroma features improved the accuracy of deep learning models for music genre classification. However, these studies did not investigate the impact of preprocessing techniques on different types of deep learning architectures, such as MLP, CNN, RNN, and CRNN. There is also a need to investigate the impact of preprocessing techniques on audio data with varying sample rates and durations.

3 Experimental Setup

All audio data used in this study was passed through a segmentation and feature extraction pipeline, and raw data was not preprocessed. The audio duration was fixed at 30 seconds through trimming and padding, and the sample rate was fixed at 22050 Hz. All audio clips were converted to the same format, and jazz track number 54 was converted to WAV format. Segmentation was performed on all audio data to ensure that each segment was entirely in the training, validation, or testing set to prevent data leakage and correlation between samples in the same segment. Preprocessing Techniques:

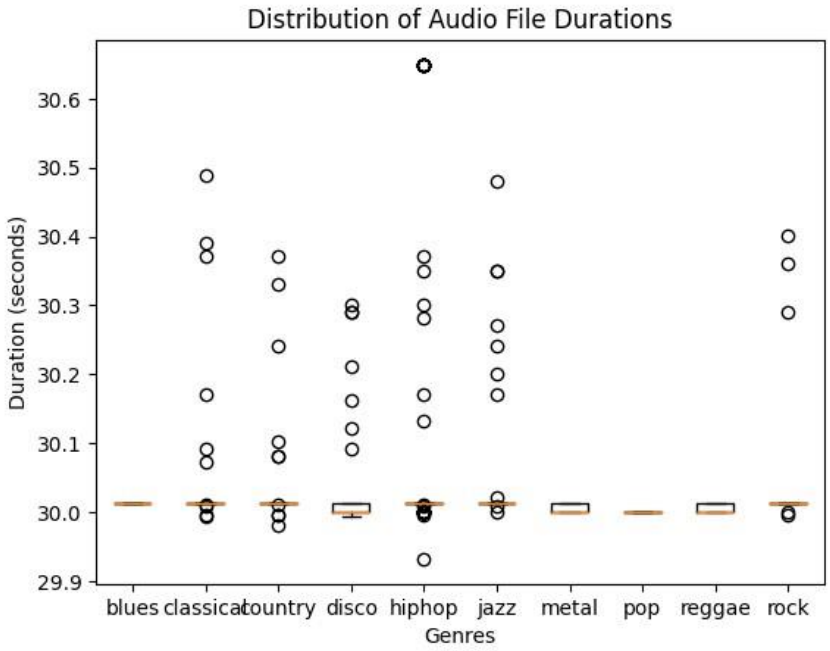

Figure 1: Distribution of audio clip duration.

Three preprocessing techniques were used in this study. The first technique involved using the raw audio data as input without any modification and only apply a segmentation and feature extraction pipleline. The second technique is to align the timing of audio samples with different durations and resampling the audio clips to a uniform sample rate of 22050 Hz and truncating them to a fixed duration of 30 seconds. The audio was then converted to Mel-spectrogram images using the Librosa library, and the Mel-spectrogram images were normalized using mean and standard deviation normalization. Models: Three models were used in this study. The first model was a Simple Multilayer Perceptron (MLP) with three hidden layers with 256 and 128 and 64 units, respectively, and used Rectified Linear Unit (ReLU) activation functions. The output layer had ten units corresponding to the ten music genres with a softmax activation function. The second model was a Convolutional Neural Network (CNN) with two convolutional layers with 32 and 64 filters, respectively, followed by a max-pooling layer. The output of the convolutional layers was flattened and fed into a fully connected layer with 128 units. The output layer had ten units. The third model was a Recurrent Neural Network (RNN): This model used a single Long Short-Term Memory (LSTM) layer with 128 units followed by a fully connected layer with 128 units. The output layer had ten units. The model was trained using the Adam optimizer with a learning rate of 0.001 set in the parameters. Training and Testing: The GTZAN music genre dataset was split into training, validation, and test sets with a ratio of 60:15:25, respectively. The models were trained using a batch size of 16 and 64 and 25 and 100 epochs. The performance of the models was evaluated using the accuracy metric, which is the percentage of correctly classified samples in the test set.

4 Results

Below are the results from the experiments carried out with the above setup: Based on the table, it

Table 2: Test Accuracy Scores from Models:

Runs	MLP-R	MLP-P	CNN-R	CNN-P	LSTM-R	LSTM-P
25 epochs / 16 batch size	21.3	31.4	62.2	58.4	54.8	51.3
25 epochs / 64 batch size	38.9	29.2	61.7	61.3	44.3	44.9
100 epochs / 64 batch size	35.5	36.4	62.8	62.6	56.9	54.8

appears that processing the data (P) generally leads to higher accuracy scores for MLP and LSTM models, while for the CNN model, there is not a clear trend. In terms of batch size and number of epochs, the impact on accuracy varies by model and data processing technique. Overall, the results suggest that the choice of deep learning model, batch size, and preprocessing technique can all have a significant impact on the accuracy of audio data classification tasks.

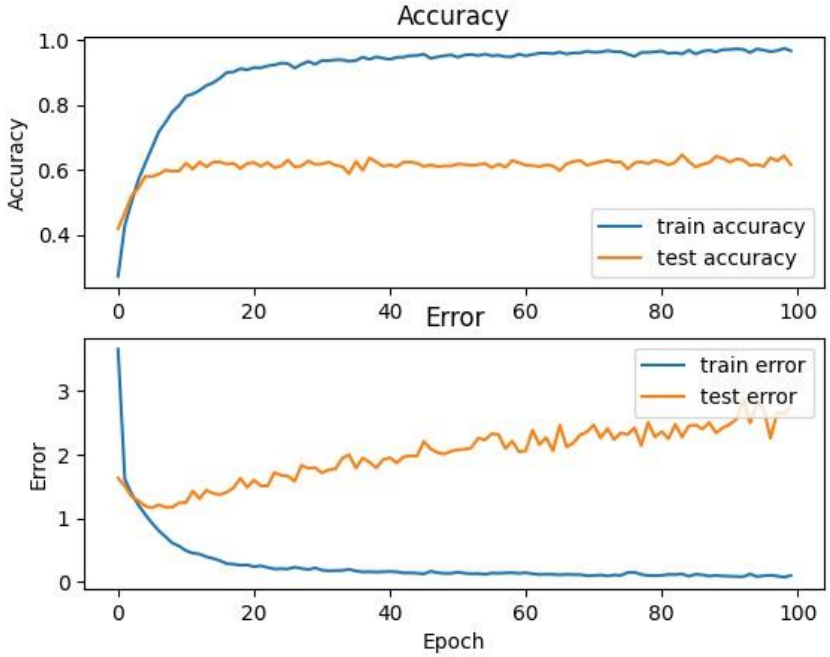

Figure 2: Accuracy and Loss chart for CNN on processed data.

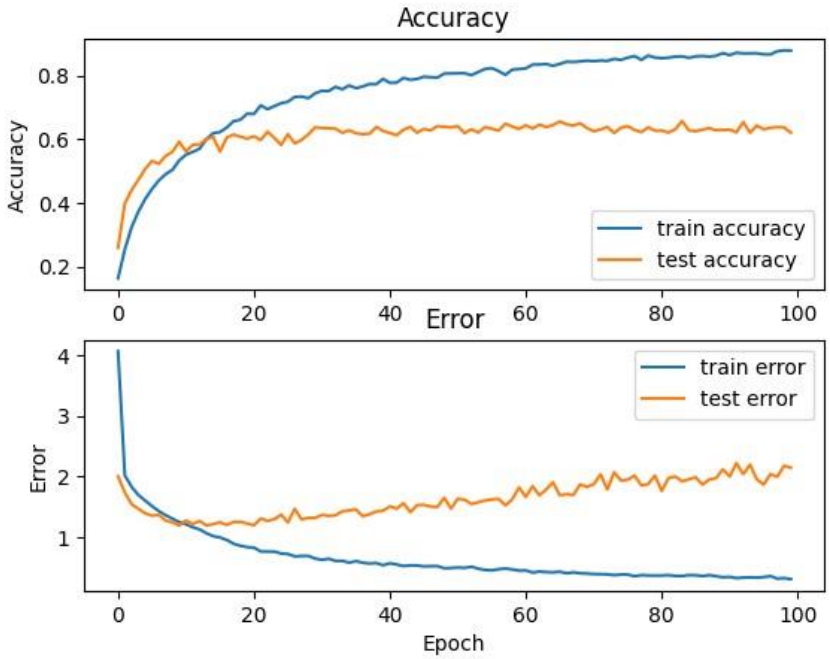

Figure 3: Accuracy and Loss chart for CNN on raw data.

We also see that the CNN trained on segmented raw data out performed all other models irrespective of the data being used to train them. Although we see more improvement in the score when Processed data is used.

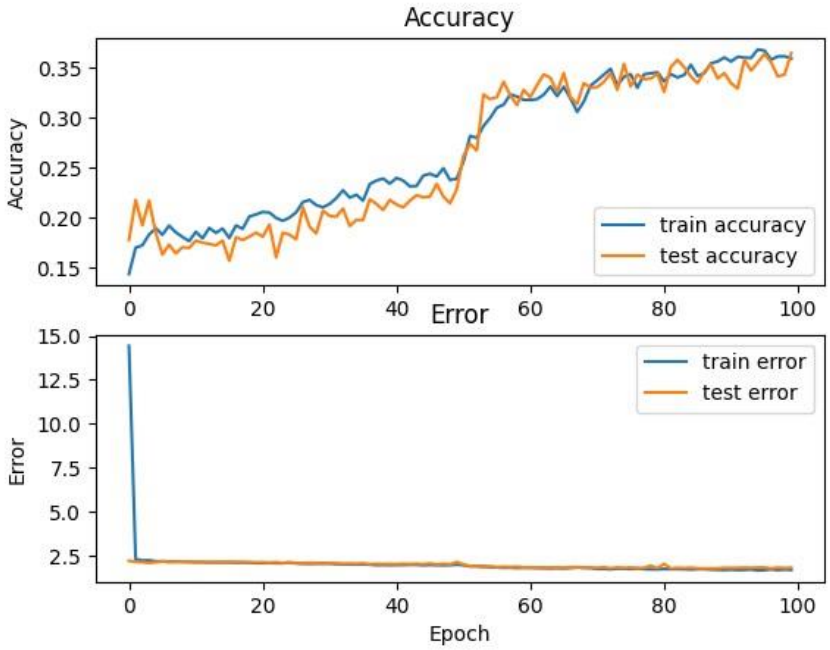

Figure 4: Accuracy and Loss chart for MLP on processed data.

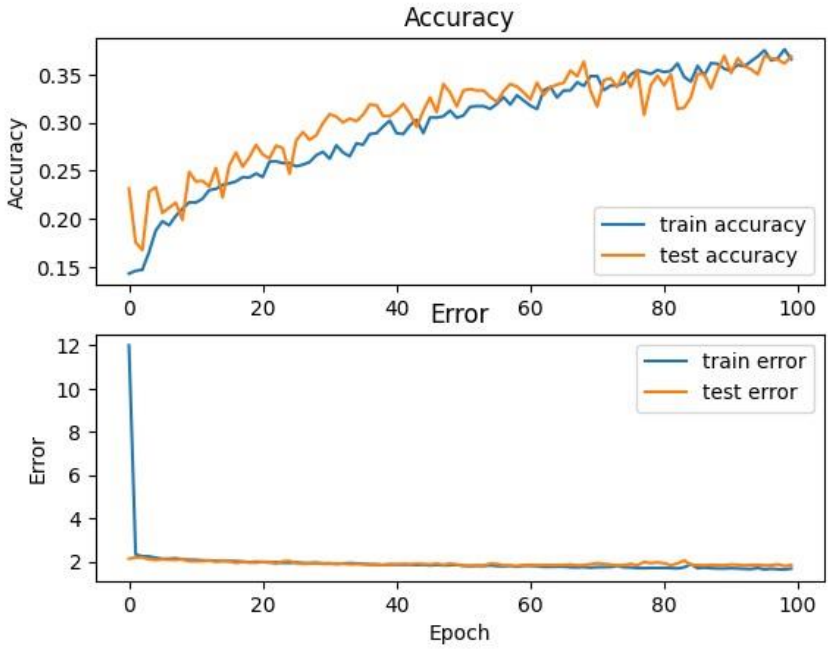

Figure 5: Accuracy and Loss chart for MLP on raw data.

For the MLP model, significantly poor test accuracy scores for both the raw and processed data, this means that further adjustments need to be made to the model architecture to ensure a higher test accuracy. This will require adding more layers and neurons at each layer. The model has failed in this experiment.

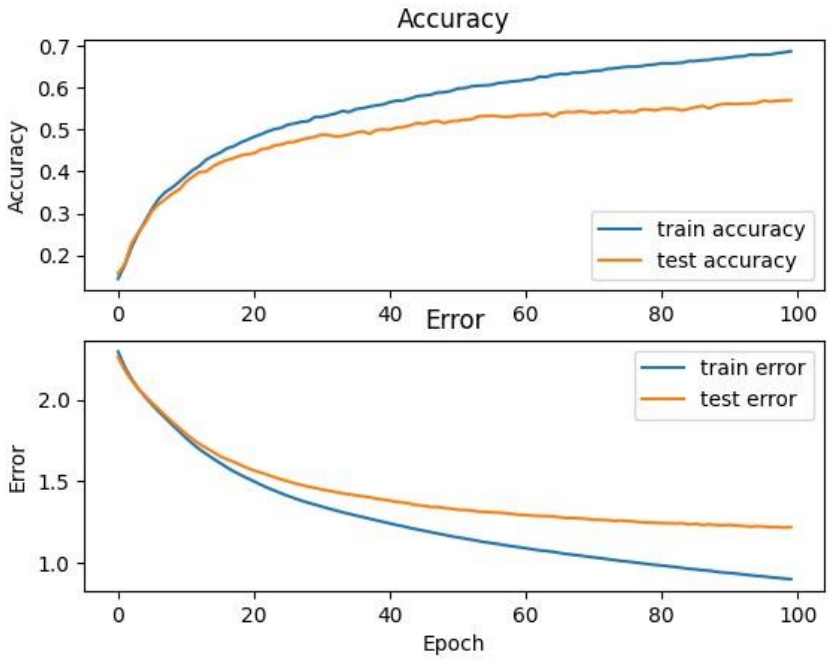

Figure 6: Accuracy and Loss chart for RNN on processed data.

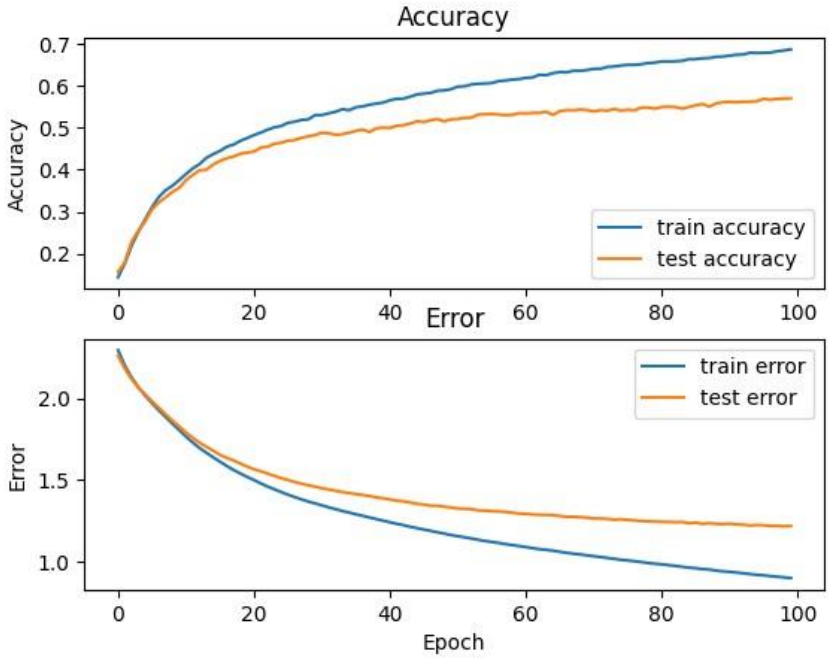

Figure 7: Accuracy and Loss chart for RNN on raw data.

For the RNN model it has not performed well in this experiment scoring within the range of random guessing for both datasets.

5 Conclusion & Future Work

The table shows the results of experiments conducted on four deep learning models (MLP, CNN, LSTM) in music genre classification, with two types of data (raw and processed) and different configurations of epochs and batch size. The results suggest that preprocessing can have a

significant impact on the performance of the models. In particular, MLPs tend to perform better on raw data, while CNNs and LSTMs tend to perform better on processed data.

Moreover, the batch size appears to have a noticeable effect on the performance of the models, with smaller batch sizes often leading to better results. However, the optimal batch size may depend on the specific dataset and model used. Overall, these results suggest that careful consideration of preprocessing and batch size can improve the accuracy of deep learning models for music genre classification.

Further experiments could investigate the impact of other preprocessing techniques, such as data augmentation, on the performance of deep learning models. Additionally, it would be interesting to explore the generalizability of the models to other datasets and tasks, such as speech recognition or sound event detection.

Future work could also involve developing new deep learning architectures specifically designed for audio data processing. This could include models that incorporate prior knowledge about audio signals, such as their temporal or spectral properties. Furthermore, techniques from other domains, such as attention mechanisms or reinforcement learning, could be adapted for audio data processing to improve the accuracy and efficiency of models.

References

Cheng, Y.-H. and Kuo, C.-N. (2022). Machine learning for music genre classification using visual mel spectrum. *Mathematics*, 10(23).

Dhall, A., Srinivasa Murthy, Y. V., and Koolagudi, S. G. (2021). Music genre classification with convolutional neural networks and comparison with f, q, and mel spectrogram-based images. In Biswas, A., Wennekes, E., Hong, T.-P., and Wieczorkowska, A., editors, *Advances in Speech and Music Technology*, pages 235–248, Singapore. Springer Singapore.