REI505M Machine Learning - Interims Report

Due: Wednesday 12.11.2025

Project: Music genre classification

Group: Spirou (Benedikt Baumgarten, Erik Schwaar, Inken Hofbauer)

Possible Audio Data Augmentation

Based on: Audio Data Augmentation Techniques: The Theory by Valerio Velardo - The Sound of AI

Data augmentation for audio can be applied in two principal ways: directly on the waveform, or on the spectrogram (time–frequency) representation.

Velardo (2020) presents a comprehensive overview of augmentation techniques, though not all are equally suitable for musical data. The two methods identified as most effective for music are time stretching, which alters the playback speed without changing pitch, and pitch scaling, which modifies the pitch while maintaining the original tempo. Additional approaches that can be beneficial include noise addition (e.g., white, pink, or background noise) and impulse response addition (introducing various types of reverberation).

Care must be taken to avoid excessive distortion, as overly aggressive augmentation can render audio samples unrepresentative of their original genre.

Further experimentation will be required to determine which augmentation techniques are most effective in improving genre classification performance within this project.

Possible libraries are librosa, torchaudio.

We have already implemented and trained a model following the network architecture described in the assignment. For a two-class (two-genre) setup, the model achieved strong results - around 80 % accuracy on the training set and 90 % accuracy on the test set - indicating that the architecture and data preprocessing are working well.

However, we encountered difficulties when scaling the model to the full dataset, as training and inference times became significantly longer. To address this issue, we experimented with reducing the length of the audio samples used for training from 30 seconds to 15 seconds per wav-file. This change greatly reduced computational cost and training time, but unfortunately led to a noticable drop in accuracy, suggesting that shorter clips lose important temporal and spectral information needed for reliable genre classification.

Our next steps will focus on optimizing the trade-off between performance and runtime. We plan to explore data augmentation strategies, overlapping segments and more efficent model configurations to maintain high accuracy while keeping training time manageable.