THE UNIVERSITY OF
SYDNEY

ELEC5305

Project Proposal

Instrument Family Classification in Music Recordings Using Audio Signal Processing and Machine Learning

---

Student:
Dawod Ghifari 520140154

Github Username:
dawodghifari

Teacher:
Craig Jin

September 7, 2025

Contents

1 Project Overview

This project addresses the challenge of automatically classifying the instrument families present in music recordings. The goal is to identify broader categories such as strings, percussion, winds, keyboards, and voice instead of recognizing individual instruments like violin or guitar. Focusing on families simplifies the task while still providing musically meaningful information that highlights the distinct acoustic qualities of each group.

The problem is important because instrument family classification supports a range of applications in music information retrieval, recommendation systems, and educational tools. By reducing the complexity of fine-grained instrument recognition, the project remains feasible within the semester while still engaging with real-world audio data. It also allows the exploration of both traditional signal processing and modern machine learning approaches in a practical and creative setting.

The proposed solution involves building an audio classification pipeline. The system will preprocess recordings by converting them to a uniform format and segmenting them into short clips. From each segment, discriminative features such as MFCCs, chroma vectors, and spectral descriptors will be extracted. Alternatively, mel-spectrograms will be used as direct input to a convolutional neural network for automated feature learning. These features will then be passed to a classification model, ranging from traditional machine learning methods such as Support Vector Machines to deep learning approaches using pretrained audio embeddings like VGGish. The output will be family-level predictions that can be aggregated to describe the dominant instrument families across longer recordings.

2 Background and Motivation

Automatic recognition of instruments in music is an important area of research within music information retrieval (MIR). Many existing systems attempt to identify specific instruments such as a violin, a trumpet, or a guitar. While these approaches demonstrate strong results on isolated notes or small datasets, they often struggle with real-world recordings that contain multiple instruments and varying recording conditions (Eronen, 2007). This fine-grained classification also requires large amounts of labeled data, which can make the task impractical for smaller projects.

An alternative approach is to classify instruments into families such as strings, percussion, winds, keyboards, and voice. Families share distinct acoustic characteristics that are easier to capture using signal processing features. For example, spectral descriptors and MFCCs can represent the timbre of strings or winds, while chroma features highlight harmonic content typical of keyboards. By focusing on families instead of individual instruments, the system reduces complexity while still producing musically meaningful insights [1].

This problem is relevant for applications such as music recommendation systems, which can benefit from knowing the dominant timbres in a track, and educational tools, where students may want to analyze which families appear in a recording. It also provides a manageable way to explore modern audio classification techniques, from traditional machine learning to deep learning methods that use spectrogram representations [2].

The motivation for this project is therefore twofold. First, it allows the development of a practical system that demonstrates the core principles of audio signal processing and machine learning. Second, it highlights the creative and interpretive aspects of music analysis by providing outputs that are easy for listeners to understand and relate to.

3 Proposed Methodology

The project will be implemented in **Python**, making use of widely adopted libraries for audio analysis and machine learning, including `Librosa` for signal processing, `scikit-learn` for traditional classifiers, and `PyTorch` or `TensorFlow` for deep learning models.

3.1 Preprocessing

Audio recordings will first be standardized to a **mono channel** format and resampled to a fixed sampling rate (e.g., 16 kHz) to ensure consistency across the dataset. Each recording will be segmented into short fragments (e.g., 2–4 seconds), which provide manageable inputs for feature extraction and classification.

3.2 Feature Extraction

Two feature extraction approaches will be explored:

1. **Handcrafted features:** Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral descriptors (centroid, bandwidth, roll-off, zero-crossing rate) will be computed to capture timbral and harmonic content [1].
2. **Learned features:** Mel-spectrograms will be generated and used directly as inputs to convolutional neural networks, enabling the model to automatically learn discriminative time–frequency patterns [2].

3.3 Classification Models

Two types of classifiers will be tested:

- **Traditional methods:** Support Vector Machines (SVMs) and Random Forests will be applied to the handcrafted feature set.
- **Deep learning methods:** A lightweight 2D CNN will be trained on mel-spectrograms. In addition, pretrained audio embeddings such as **VGGish** may be incorporated to leverage transfer learning [3].

3.4 Datasets

The project will draw on publicly available datasets such as:

- **IRMAS** (Instrument Recognition in Musical Audio Signals) for polyphonic music with labeled instrument annotations.
- **NSynth** for isolated notes grouped by instrument family.

These datasets provide a suitable mix of controlled and real-world recordings for training and evaluation.

3.5 Evaluation

Performance will be assessed using metrics such as accuracy, precision, recall, and F1-score. A **confusion matrix** will be used to identify common misclassifications (for example, strings mistaken for keyboards). Visualization techniques, such as spectrograms with predicted labels and family activity timelines across songs, will provide additional interpretability.

4 Expected Outcomes

The expected outcome of this project is a functional audio classification system capable of identifying instrument families from music recordings. The system will demonstrate the complete pipeline from preprocessing and feature extraction through to classification and visualization.

From a technical perspective, the project will deliver:

- An **implementation of feature-based classification** using traditional machine learning methods such as Support Vector Machines and Random Forests.
- An **implementation of deep learning classification** using convolutional neural networks trained on mel-spectrograms, with the option of transfer learning through pretrained embeddings such as VGGish.
- A **comparison of performance** between these approaches, highlighting the trade-offs between interpretability, computational complexity, and accuracy.

The project will be evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. A confusion matrix will be presented to show common misclassifications between instrument families.

In addition to numerical evaluation, the project will deliver **visualization tools**. These include spectrograms annotated with predicted instrument families and timeline plots that illustrate family activity across longer recordings. Such outputs will make the system more interpretable and engaging for non-technical audiences.

Finally, the project will provide a **GitHub repository** containing the source code, documentation, and example demonstrations. A simple demo interface will allow users to upload a music clip and view predicted instrument families in real time. This will make the project both technically rigorous and accessible to a wider audience.

5 Timeline

The project will be completed according to the following schedule:

Weeks	Task
6–7	Literature review and dataset collection. This includes identifying suitable datasets (IRMAS, NSynth), reviewing prior research on instrument classification, and setting up the project repository.
8–9	Initial implementation and testing. Preprocessing pipeline and feature extraction (MFCCs, chroma, spectral descriptors, mel-spectrograms) will be implemented, followed by baseline classifiers such as SVM and Random Forest.
10–11	Optimization and evaluation. Develop and train convolutional neural networks, apply transfer learning with pretrained embeddings (e.g., VGGish), and conduct experiments to compare models using classification metrics.
12–13	Final report and documentation. Results will be compiled, visualizations will be prepared, and a demo interface will be implemented. The GitHub repository will be finalized with code, instructions, and documentation.

Table 1: Project timeline from Weeks 6 to 13.

References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 293–302, 07 2002.

[2] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," *arXiv:1609.04243 [cs]*, 12 2016.

[3] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," *arXiv:1609.09430 [cs, stat]*, 01 2017.