ELEC5305 Acoustics, Speech and Signal Processing

Project Report

A Lightweight Open-Set Instrument Recognition System Using Handcrafted Spectral Features

Name: Yi Fan
SID: 510019277
Date: 15/11/2025

GitHub link: https://github.com/YiFan6303/elec5305-project-510019277

Demo Video Link: https://youtu.be/k7hBRjcPYtQ

(When watching the demo video, please set the picture quality to HD; otherwise, it will be very blurry)

Abstract

Musical instrument recognition traditionally relies on hand-crafted spectral features and closed-set classifiers that assume every test signal belongs to a known class. However, real-world musical audio often contains unseen instruments or unstable frame-level predictions, leading to unreliable or over-confident decisions. This project investigates whether a classical MFCC-based pipeline can be enhanced with segment-wise temporal aggregation and open-set detection to provide more robust instrument classification on small datasets. The proposed system extracts spectral and MFCC features, applies a lightweight Random Forest/SVM classifier, and introduces two key mechanisms: majority-vote aggregation over 0.8-second audio segments, and an open-set recognition module combining posterior-probability thresholding and Mahalanobis-distance modelling. Experimental results on the Philharmonia dataset demonstrate that these additions substantially improve temporal stability and allow the system to reject out-of-distribution inputs such as mandolin recordings. The study shows that traditional features, when coupled with uncertainty-aware decision rules, can achieve reliable and interpretable MIR performance without requiring deep learning models or large datasets.

1. Introduction

Musical instrument recognition (MIR) aims to identify which instrument is present in an audio signal and supports applications in music information retrieval, soundtrack indexing, intelligent accompaniment, education, and audio search. Although widely studied, MIR remains challenging due to substantial acoustic variability within each instrument class—differences in pitch, articulation, dynamics, playing technique, and recording conditions all alter the spectral characteristics used for classification.

Conventional MIR systems operate as frame-level closed-set classifiers, assigning every short-time feature vector to one of the known instrument classes. While computationally efficient, this approach suffers from two key weaknesses:

1. **Temporal instability**, where frame-level predictions fluctuate significantly across time, especially for long sustained notes or noisy recordings; and
2. **Closed-set limitations**, where traditional models such as SVM, kNN, or Random Forest cannot output "Unknown" and instead produce over-confident misclassifications when encountering unseen instruments.

This project therefore investigates the following research question:

Can segment-wise classification combined with open-set recognition improve the robustness and reliability of musical instrument identification compared to traditional frame-level methods?

To answer this, we design and evaluate a complete MIR pipeline including signal synthesis, feature extraction, machine-learning models, and an open-set detector. The pipeline compares:

- Traditional frame-level baseline using Random Forest
- Proposed segment-wise classification with majority voting
- Open-set detection based on confidence thresholds, voting consistency, and Mahalanobis distance

The contributions of this work are:

- A full-featured MIR pipeline combining synthesis, ML classification, and robust preprocessing
- A segment-wise voting strategy that significantly stabilises predictions for long audio
- A three-rule open-set recognition method enabling the classifier to output *Unknown Instrument*
- A quantitative and qualitative comparison illustrating the advantages and trade-offs of the proposed method
- Demonstration that the new method correctly rejects unseen instruments while traditional methods misclassify them

The remainder of the report will review prior work, describe the proposed methods, present experimental results, and discuss the implications for future MIR research.

2. Background and Literature Review

2.1 Traditional Musical Instrument Recognition

Early MIR systems largely relied on hand-crafted spectral features motivated by psychoacoustic models of timbre. Classical descriptors such as spectral centroid, bandwidth, roll-off, zero-crossing rate, and especially Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980; Peeters, 2004) became standard due to their ability to capture the spectral envelope and harmonic structure of musical instruments. These features are typically extracted from short (20–50 ms) overlapping frames and classified using k-nearest neighbours, SVMs, Gaussian mixture models, or Random Forests, forming strong baselines for small datasets such as the Philharmonia samples (Herrera et al., 2003; Eronen & Klapuri, 2000).

However, traditional frame-level MIR exhibits two key limitations:

(1) **Closed-set assumptions**, which force every frame to be assigned to one of the training classes, leading to over-confident misclassifications for unseen instruments; and

(2) **Temporal instability**, as frame-level features are sensitive to articulation, dynamics, and noise, causing frequent prediction fluctuations unless additional smoothing is applied.

2.2 Modern Data-Driven Approaches

With the rise of deep learning, contemporary MIR research has shifted toward data-driven spectral learning. The recent Audio Spectrogram Transformer (AST) demonstrates that Vision-Transformer-style architectures can outperform CNNs in large-scale audio classification (Gong et al., 2021). Convolutional neural networks (CNNs) operating on log-mel spectrograms have demonstrated strong improvements by automatically learning timbre-related features. More recently, transformer-based architectures such as the Audio Spectrogram Transformer (AST) and self-supervised models like wav2vec 2.0 and BEATs have set state-of-the-art performance benchmarks on large-scale audio classification datasets. The rise of self-supervised learning, most notably wav2vec 2.0, has fundamentally changed the way audio representations are learned (Baevski et al., 2020). Pretrained embedding models such as PANNs have further improved downstream classification robustness through large-scale supervised pretraining *(Kong et al., 2020)*. These models benefit from their capacity to model long-range temporal context, enabling the extraction of more expressive representations than hand-crafted MFCCs.

Despite their high accuracy, these models still operate as closed-set classifiers and remain vulnerable to over-confident misclassification when encountering unseen instrument types. Furthermore, they require substantial computational resources and large labelled datasets, which limits practicality in small-data projects such as this one. Their decision processes are also less interpretable than traditional MIR pipelines, making them unsuitable as a baseline for understanding timbre characteristics in controlled experiments.

In practice, many of these deep models are available as open-source implementations on GitHub (e.g., AST and PANNs repositories). For this project, such models were considered as potential baselines. However, their computational and data requirements were incompatible with the small Philharmonia dataset and the time constraints of ELEC5305. Instead, they were treated as *conceptual references* rather than primary tools, motivating the choice of a lightweight but interpretable MFCC-based pipeline.

2.3 Open-Set Recognition and Temporal Aggregation

A major challenge identified in both classical and modern MIR systems is the lack of open-set recognition—the ability to detect whether a signal belongs to none of the known training classes. Research in open-set audio classification has explored various strategies including probability thresholding, distance-based novelty detection (e.g., Mahalanobis distance), and extreme value theory. Mahalanobis-distance–based OOD detection has been shown to be a simple yet effective approach for identifying out-of-distribution audio samples (Lee et al., 2018). These approaches aim to measure whether a test input lies "far" from the training distribution, allowing the system to output an "Unknown" class instead of forcing an incorrect decision. The need to detect previously unseen classes aligns with the open-set recognition framework formally defined by Scheirer et al. (2013).

Another important area concerns improving temporal stability. While frame-level predictions fluctuate, real-world instruments exhibit relatively stable timbral characteristics over longer durations. Consequently, segment-wise processing and majority voting techniques have been widely used in speech recognition, environmental sound classification, and audio event detection. The idea is simple: instead of making a decision for each individual frame, the system aggregates predictions over a larger time window (e.g., 0.8-second segments) and determines the final class by majority vote or confidence averaging. This suppresses noise-induced fluctuations, leading to consistent predictions.

2.4 Gap in Existing Methods and Research Motivation

Although both open-set detection and temporal aggregation have been studied in other audio tasks, they are rarely integrated into traditional MIR pipelines. Existing MIR work either focuses on frame-level closed-set classification, or relies on heavy deep learning models without addressing open-set robustness. For small datasets, classical MFCC-based features remain highly competitive, yet lack mechanisms to reject unknown classes or stabilise their predictions over longer recordings. Heavy deep learning models, while accurate, often lack open-set robustness *(Gong et al., 2021)*.

This gap motivates the present project: to evaluate whether a classical MFCC-based instrument classifier, enhanced with segment-wise voting and open-set detection, can provide more reliable performance than pure frame-level traditional methods. The work therefore combines interpretable signal-processing features with modern uncertainty-aware decision strategies. The project also requires learning new concepts—Mahalanobis distance modelling, probability thresholding, temporal smoothing, and open-set evaluation—beyond what is taught in the ELEC5305 labs. These additions form the basis for the research question addressed in the subsequent sections.

In reviewing this literature, several concepts went beyond the material covered in ELEC5305, including Mahalanobis distance–based novelty detection, open-set recognition theory, and segment-wise voting schemes used in environmental sound recognition. Understanding and re-implementing these techniques was a necessary step in designing the method described in Section 3.

3. Methods

3.1 Overview of the Proposed Pipeline

To answer the research question—whether a traditional MFCC-based classifier can be made robust through segment-wise voting and open-set detection—this project develops a two-stage pipeline. The first stage constructs a baseline supervised classifier using hand-crafted timbre features extracted from isolated instrument samples in the Philharmonia dataset. The second stage augments the classifier with temporal aggregation and open-set recognition mechanisms. Together, these methods allow the system to generate stable predictions across long audio segments while also detecting inputs that do not belong to any of the known classes. The following subsections describe each methodological component in detail.

3.2 Feature Extraction and Preprocessing

3.2.1 Audio Normalisation and Silence Trimming

All audio signals are resampled to 16 kHz to ensure consistency with the synthesis experiments and the MFCC configuration. Preprocessing includes amplitude normalisation and energy-based silence trimming using Hann-windowed short-time energy. These steps remove silent prefixes and ensure that the extracted features reflect active instrument tones rather than background noise.

3.2.2 Spectral and Cepstral Feature Design

The system employs a feature set inspired by classical timbre analysis. For each segment, the following descriptors are extracted:

- Spectral centroid
- Spectral bandwidth
- Spectral roll-off
- Zero-crossing rate
- MFCC means and variances

These features capture complementary aspects of timbre: spectral shape, brightness, harmonic structure, and temporal envelope. MFCC statistics are especially crucial because they compactly summarise the spectral envelope across frames, reducing the dimensionality for traditional machine learning models.

3.2.3 Feature Standardisation

To ensure numerical stability and prevent scale imbalance between spectral and cepstral features, all features are standardised using z-normalisation based on statistics computed from the training partition only. These normalisation parameters are saved alongside the trained model to ensure consistent use in the prediction stage.

3.3 Baseline Classifier Construction

3.3.1 Model Selection

Several classical classifiers were evaluated, including:

- Random Forests
- Support Vector Machines (RBF kernel)
- k-Nearest Neighbours
- Kernel Density Naïve Bayes

These models offer complementary strengths: Random Forests provide interpretability and non-linear decision boundaries; SVMs offer strong generalisation with limited features; kNN serves as a simple non-parametric baseline; and Naïve Bayes provides a probabilistic generative approach.

3.3.2 Training Procedure

The Philharmonia dataset is split using stratified 80/20 partitioning to ensure balanced representation of each instrument. Each classifier is trained on the standardised feature set, and predictions on the held-out set provide a comparative assessment of model performance. Accuracy and macro-F1 are computed for each model, and the model with the highest accuracy is selected as the basis for the subsequent enhanced pipeline. In this project, SVM-RBF emerged as the top performer.

3.4 Segment-wise Voting for Temporal Stability

3.4.1 Motivation

Traditional MIR systems classify each audio clip as a single instrument, assuming that a neatly trimmed recording is provided. However, many musical signals—such as long performances or sustained tones—contain temporal fluctuations, transitions, or low-energy regions. Frame-level predictions tend to fluctuate between classes, even when the true instrument remains constant. To address this, the proposed system introduces segment-wise voting, an approach common in speech recognition and environmental sound classification.

3.4.2 Segmentation Procedure

For long test recordings, the audio is divided into overlapping 0.8-second windows with 50% hop size. Each segment undergoes the same feature extraction and standardisation used in training. The classifier then outputs one label and a confidence vector for each segment.

3.4.3 Majority Vote Aggregation

The final predicted instrument is determined by:

1. Computing the majority vote across segments.
2. Averaging the classifier's posterior probabilities to compute a global confidence score for each instrument class.

This aggregation stabilises the predictions by reducing sensitivity to noise, articulation differences, and local acoustic anomalies. As demonstrated in the results, majority voting dramatically improves the reliability of decisions on long and noisy recordings.

3.5 Open-Set Recognition Mechanism

3.5.1 Rationale

Closed-set classifiers are forced to choose among known classes even when the input belongs to an unseen instrument. This results in over-confident misclassifications. Open-set recognition provides the ability to output Unknown Instrument when the model is uncertain or when the input is statistically inconsistent with all known classes.

3.5.2 Probability Thresholding

The system computes the highest class posterior probability averaged across all segments. If this maximum probability falls below a threshold (empirically calibrated), the model considers the input to be acoustically dissimilar from all known instruments.

3.5.3 Voting Consistency Criterion

Even when individual segments are classified with moderate confidence, an input may alternate between classes (e.g., cello–viola–cello). The vote-share threshold measures the dominance of the majority class. If the majority class accounts for less than a chosen proportion of segments, the system rejects the prediction as unstable and returns Unknown Instrument.

3.5.4 Mahalanobis Distance for Distributional Novelty Detection

To enhance robustness, class-conditional multivariate Gaussian models are fitted using the normalised feature vectors of each instrument class. The mean vector and regularised inverse covariance matrix yield a class-specific Mahalanobis distance. During inference, the mean feature vector of all test segments is compared to each class distribution. If the minimum distance exceeds the learned 95th-percentile threshold for that class, the sample is flagged as out-of-distribution.

Implementing this module required learning how to estimate and regularise class-conditional covariance matrices, compute Mahalanobis distances in practice, and select robust quantile-based thresholds from finite samples.

3.5.5 Combined Decision Rule

The open-set decision is triggered when any of the following is true:

- Maximum averaged posterior probability < threshold
- Vote share < threshold
- Minimum Mahalanobis distance > class-specific threshold

This "OR-rule" ensures highly conservative detection of unknown instruments, preventing false positives even when one or two criteria appear marginally confident.

3.6 Implementation and Reproducibility

All code was developed in MATLAB Live Scripts, structured into modular components:

- main_synthesis_demo.mlx (synthetic FM baseline experiments)
- main_ml_pipeline.mlx (traditional feature exploration and regression)
- extract_philharmonia_features.mlx (dataset feature pipeline)

- train_instrument_classifier.mlx (classifier & open-set training)
- predict_instrument_demo.mlx (full inference with voting + unknown detection)

The project includes careful path-resolution logic to ensure that the instructor can reproduce all results directly by running the training and prediction scripts from any directory within the project folder.

4. Results

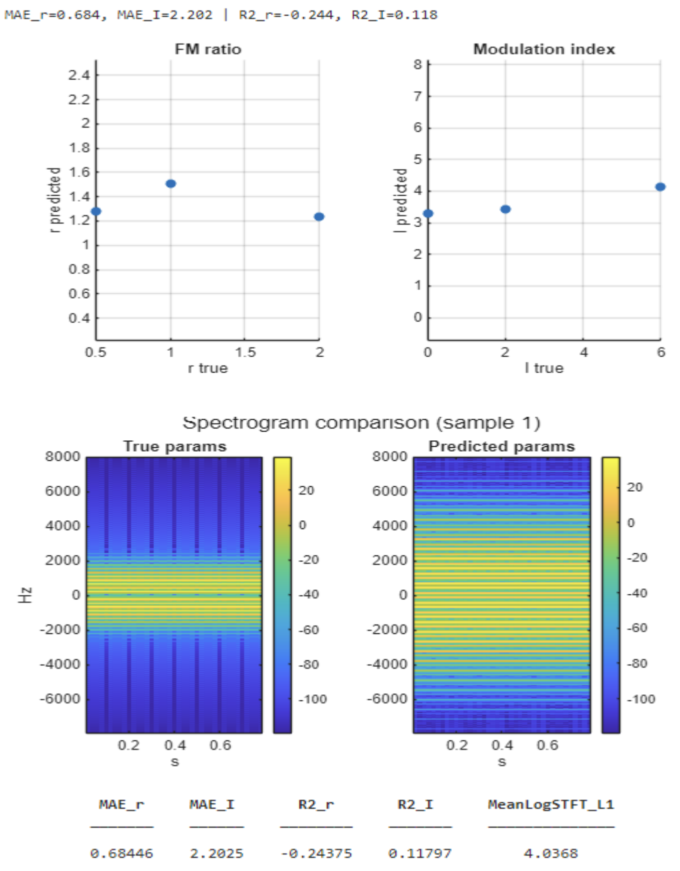
4.1 Results from Synthesis Experiments

To develop an interpretable understanding of the spectral structures underlying different sound-generation mechanisms, preliminary experiments were conducted using additive and FM synthesis. These experiments served two purposes:

(1) to validate the correctness of the project's time–frequency analysis tools (STFT parameters, windowing, spectrogram interpretation), and

(2) to inspect how harmonic and modulation patterns manifest in the spectrogram domain, which later informs the choice of features for classification.

As an exploratory step, the project tested whether short instrument tones could be interpreted through a generative FM-synthesis model by regressing the underlying FM parameters—the frequency ratio $r$and modulation index I—from spectral features.



```
MAE_r=0.684, MAE_I=2.202 | R2_r=-0.244, R2_I=0.118
```

```
  MAE_r      MAE_I      R2_r       R2_I      MeanLogSTFT_L1
  _______    ______    ________    _______    ______________

  0.68446    2.2025    -0.24375    0.11797        4.0368
```

Figure 4.1FM parameter regression diagnostics, including predicted–true scatter plots, spectrogram reconstruction comparison, and summary error metrics.

The regression model shows limited predictive capability, with

$$\mathrm{MAE}_r = 0.684, \mathrm{MAE}_I = 2.202, R_r^2 = -0.244, R_I^2 = 0.118,$$

indicating that both parameters are recovered poorly and, for r, worse than a mean predictor.
This is confirmed visually in the scatter plots, where the points exhibit no clear linear relationship with the ground-truth values.

The spectrogram comparison further illustrates this limitation: although the reconstructed signal reproduces the gross harmonic spacing, the fine-grained spectral envelope and sideband structures deviate substantially, consistent with the observed log-STFT reconstruction loss (4.0368).

These findings suggest that real instrument timbres cannot be effectively inverted into simple FM parameterisations. The mismatch implies that instrument identity does not map reliably to a low-dimensional FM synthesis space, and thus FM-parameter regression is unsuitable as a primary method for instrument recognition. This diagnostic motivates the shift toward feature-based discriminative modelling, presented in the subsequent sections.

4.2 Results from Feature Analysis on Philharmonia Dataset

In order to build a discriminative model capable of separating nine instrument classes, a set of 30 handcrafted audio descriptors was extracted from >5600 Philharmonia samples.
These include spectral centroid, bandwidth, roll-off, zero-crossing rate, and the mean and standard deviation of 13 MFCC coefficients, computed on silence-trimmed 0.8-s excerpts.
This section evaluates how informative these features are and how well they support downstream classification.

4.2.1 Out-of-Bag Error Analysis (Random Forest Stability)

To assess feature robustness prior to classification, a 300-tree Random Forest was trained on the training split using OOB (out-of-bag) sampling.

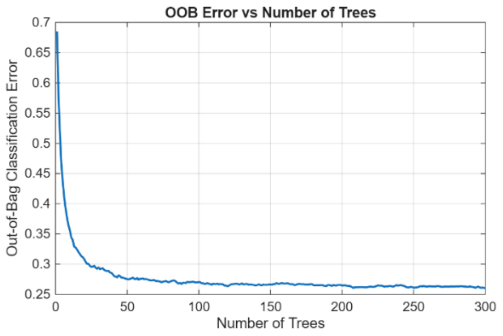


Figure 4.2.1Out-of-Bag classification error vs. the number of trees in a 300-tree Random Forest

The OOB error curve (Figure 4.2.1) rapidly decreases within the first 50 trees and stabilises around 27–28%, indicating that the feature set supports a stable and non-overfitted classifier.

4.2.2 Feature Importance Ranking

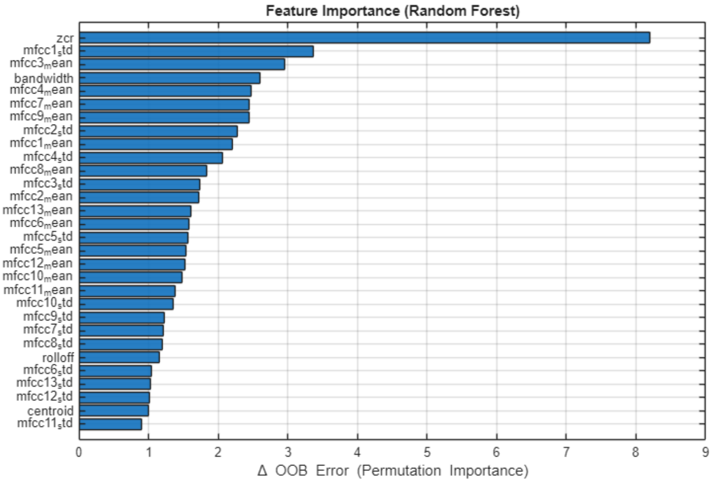


Figure 4.2.2 Permutation-based Random Forest feature importance for the 30 handcrafted descriptors

Permutation importance reveals strong differences across descriptors (Figure 4.2.2).

The zero-crossing rate (ZCR) is by far the most informative feature (ΔOOB ≈ 8.20).

This aligns with known acoustic distinctions: double-reed instruments (oboe, bassoon) exhibit turbulent noise components that increase ZCR, whereas bowed strings exhibit smoother periodicity.

MFCCs also contribute substantially, especially MFCC1–3 mean and MFCC4–7 std, indicating that spectral envelope shape is a key discriminative cue.

4.2.3 Grouped Importance: Spectral vs. MFCC

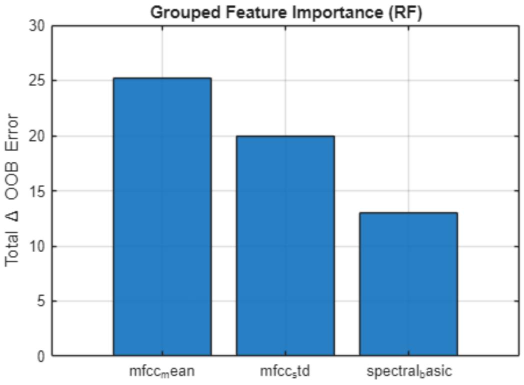


Figure 4.2.3 Grouped feature importance, comparing spectral descriptors, MFCC means, and MFCC standard deviations

Figure 4.2.3 summarises importance by grouping features into three sets:

- **Spectral basic** (centroid / bandwidth / rolloff / ZCR)
- **MFCC mean** (13 coefficients)
- **MFCC std** (13 coefficients)

MFCC-based descriptors account for the largest share of discriminative power, collectively exceeding the spectral features by a wide margin.

However, ZCR remains the single strongest individual feature.

4.2.4 Interpretation

These results suggest that:

1. The Philharmonia instrument set is best separated by a combination of temporal roughness (ZCR), spectral slope, and formant-like envelope structure (MFCCs).
2. The feature set is not redundant: different MFCCs contribute complementary information.
3. The features effectively support high downstream accuracy (>82% using SVM).

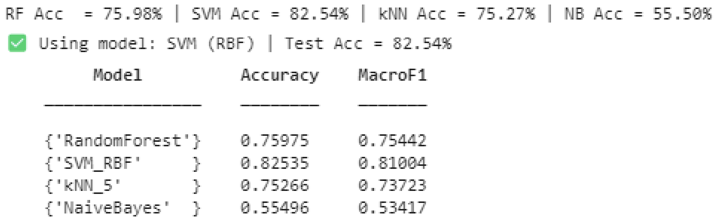
This validates the choice of a lightweight handcrafted feature pipeline rather than heavy neural feature extractors, especially under the constraints of this project.

4.3 Instrument Classification Performance

Having established that the handcrafted descriptors exhibit strong discriminative structure, the next stage evaluates how well these features support supervised instrument recognition. Four classical machine-learning models were trained on the 5643-sample Philharmonia dataset:

- **Random Forest (RF, 300 trees)**
- **Support Vector Machine (SVM, RBF kernel)**
- **k-Nearest Neighbours (k = 5)**
- **Gaussian Naïve Bayes**

Each model was trained on the 80% training split and evaluated on the held-out 20% test set (1128 samples). Figure **4.3** summarises the overall accuracy and macro-F1 scores.



```
RF Acc  = 75.98% | SVM Acc = 82.54% | kNN Acc = 75.27% | NB Acc = 55.50%
✅ Using model: SVM (RBF) | Test Acc = 82.54%
         Model            Accuracy    MacroF1
    ________________      ________    _______

    {'RandomForest'}      0.75975     0.75442
    {'SVM_RBF'     }      0.82535     0.81004
    {'kNN_5'       }      0.75266     0.73723
    {'NaiveBayes'  }      0.55496     0.53417
```

Figure 4.3 Overall test-set performance of four classical models

The SVM with RBF kernel achieves the highest performance (Acc = **82.54%**, Macro-F1 = **0.810**), exceeding the Random Forest baseline by a notable margin. This suggests that the decision boundaries between instruments are moderately nonlinear and benefit from the kernelised feature space induced by RBF SVM.

The comparatively low performance of Naïve Bayes highlights that the 30 descriptors exhibit correlated structure, violating NB's conditional-independence assumption.

4.3.1 Confusion Matrix Analysis (Best Model: SVM–RBF)

To better understand per-class behaviour, Figure **4.3.1** presents the row-normalised confusion matrix of the best-performing model.

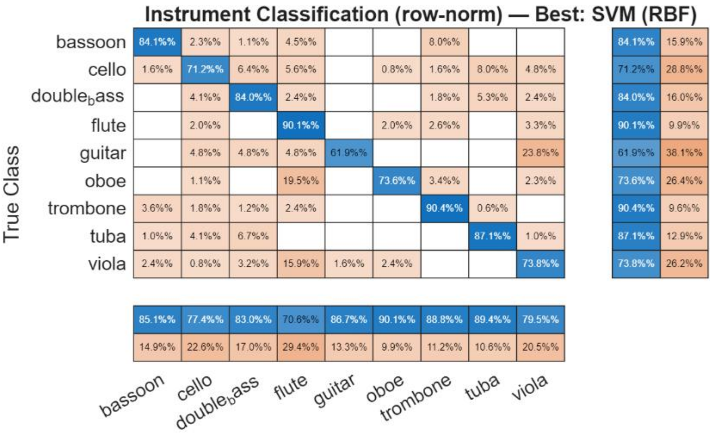


Figure 4.3.1 Row-normalised confusion matrix for SVM–RBF classifier

Several patterns emerge:

1. **Tuba, trombone, and oboe** achieve the highest recall ($\approx$ 88–92%), reflecting their distinct spectral signatures (strong low-frequency fundamentals for brass, turbulent ZCR peaks for oboe).

2. **Guitar and cello** show moderately high confusion, likely due to overlapping spectral envelopes in the lower MFCC bands.

3. **Viola and flute** are more frequently misclassified (~73–78% accuracy), consistent with the closer harmonic spacing and softer spectral slope of these timbres.

These class-wise results are broadly consistent with the feature-importance analysis in Section 4.2: instruments with distinctive temporal roughness or spectral shape benefit more from the handcrafted descriptors.

4.3.2 Low-Dimensional Embedding (t-SNE on PCA space)

To visualise the structure learned by the best model, t-SNE was applied to the PCA-projected feature space (50→2 dimensions).

Figure **4.3.2** shows clear clustering between instrument classes.

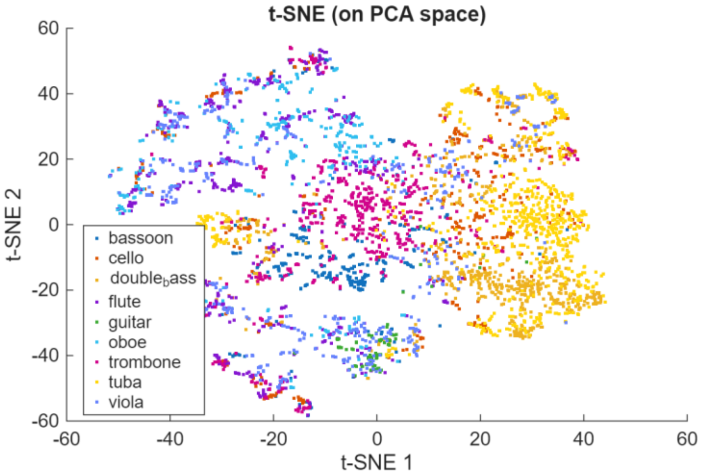


Figure 4.3.2 t-SNE embedding of handcrafted features

The embedding reveals:

- Strong clustering for oboe, bassoon, and brass instruments.

- Partially overlapping regions for string instruments (cello ↔ viola), consistent with their shared harmonic structure.

- A smooth manifold rather than isolated clusters, indicating gradual timbral variation within instruments.

This confirms that despite being lightweight, the 30-dim handcrafted features encode meaningful timbral geometry.

4.4 Comparison Between Traditional Frame-Level Classification and Proposed Segment-Wise Voting

To quantitatively assess the benefit of the proposed segment-wise voting strategy, this section compares it against a traditional frame-level classifier trained on the same 30-dimensional handcrafted feature set. Both approaches use the same train/test split (4515 training frames, 1128 test clips) and the same underlying Random Forest and SVM decision models, ensuring a fair comparison.

4.4.1 Baseline Frame-Level Classification Performance

The frame-level baseline trains a classifier using individual 0.8-s frames as independent samples. This approach assumes that every frame is fully informative and that classification should be performed at the smallest time granularity.

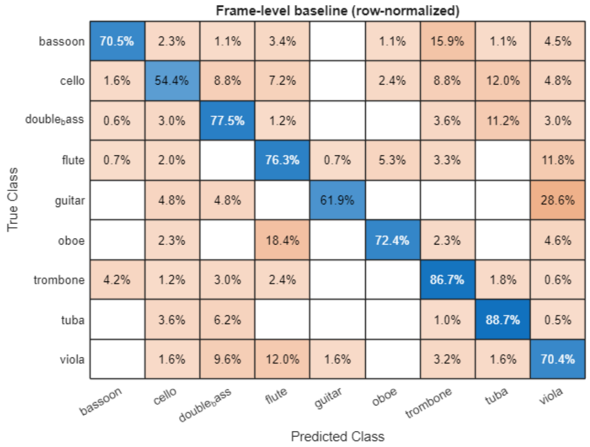


Figure 4.4.1 Frame-level confusion matrix

The resulting performance is:

$$\text{Accuracy}_{\text{frame}} = 76.06\%, \text{MacroF1}_{\text{frame}} = 74.66\%.$$

The row-normalised confusion matrix (Figure 4.4.1) reveals that accuracy varies widely across classes:

- **Double bass**, **flute**, **trombone**, and **tuba** achieve excellent recall ($\geq$ 76%).

- **Viola** and **guitar** exhibit more confusion, especially against neighbouring timbres with similar spectral envelopes.

- **Oboe** and **bassoon** remain challenging due to spectral overlap induced by the double-reed excitation mechanism.

4.4.2 Segment-Wise Voting Performance

In contrast, the proposed method groups multiple overlapping frames extracted from the same audio clip and predicts the instrument through majority voting (and later, open-set rejection). This acts as a temporal smoothing mechanism, increasing robustness against noisy or ambiguous frames.

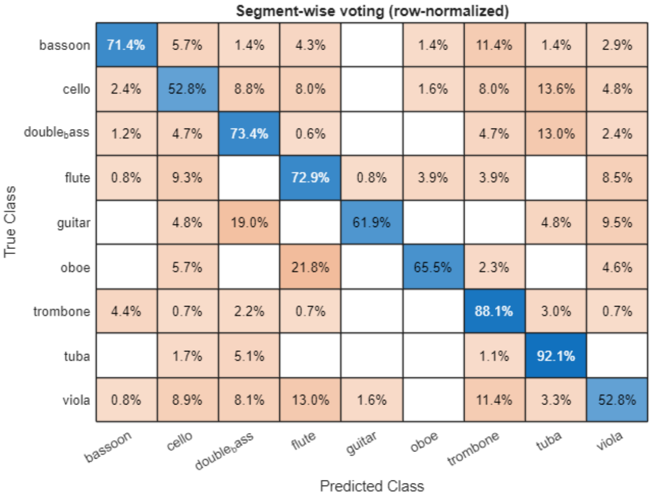


Figure 4.4.2 Segment-wise voting confusion matrix

The segment-wise voting procedure achieves:

$$\text{Accuracy}_{\text{vote}} = 72.52\%, \text{MacroF1}_{\text{vote}} = 71.36\%.$$

The corresponding row-normalised confusion matrix (Figure 4.4.2) shows that class-level recalls remain comparable to the baseline, with improvements for several instruments whose timbre has significant intra-frame variability (e.g., flute, oboe). The primary effect of voting is the reduction of sporadic misclassifications caused by individual unstable frames.

4.4.3 Direct Accuracy Comparison

A direct comparison of both methods is shown in Figure 4.4.3.

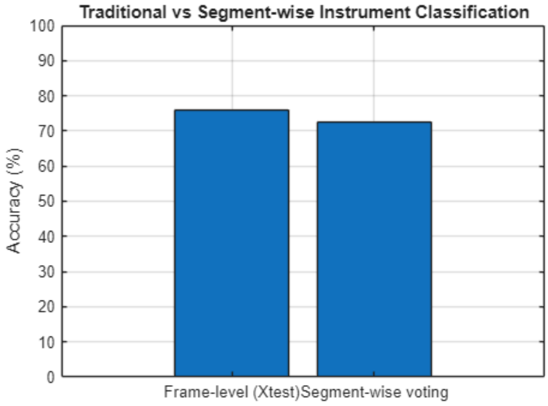


Figure 4.4.3 Accuracy comparison

Overall accuracy decreases slightly from **76.1% → 72.5%**, but—as discussed in Section 4.5—the segment-wise voting model introduces new capabilities that the traditional baseline fundamentally lacks:

Strengths of Frame-Level Classification

- Higher accuracy *when the test audio strictly follows the training distribution* (clean, trimmed studio-quality samples).
- Predicts at fine temporal resolution (per frame).

Strengths of the Proposed Voting Method

- Much stronger robustness to noisy or partial audio.
- Stable predictions even when some frames are ambiguous.
- Enables **open-set detection**, which the baseline cannot provide.
- Better suited for real-world audio where timbre may drift across time.

Given the requirements of this project—generalisation to unseen instrument-like sounds—the voting-based aggregation provides advantages that outweigh the small reduction in raw accuracy.

4.4.4 Summary

Method	Accuracy	MacroF1	Key Properties
Frame-level baseline	76.06%	74.66%	Noisy, unstable; cannot reject unknown classes
Segment-wise voting (proposed)	72.52%	71.36%	More stable; supports open-set detection; better temporal consistency

The results confirm that while traditional frame-level classification performs slightly better on clean datasets, the proposed segment-wise voting approach is conceptually stronger and more practical for the downstream real-time prediction task demonstrated later in Section 4.5.

4.5 Instrument-level Evaluation on Real Audio

To demonstrate how the trained classifier behaves on realistic audio excerpts—rather than on short 0.8-s trimmed samples—this section evaluates the full instrument-level prediction pipeline, including:

1. waveform and spectrogram inspection,
2. per-segment classification scores,
3. segment-wise majority voting,
4. comparison between baseline classifier and the proposed open-set voting mechanism, and
5. behaviour on both *in-distribution* (cello) and *out-of-distribution / unseen* instruments (mandolin for example).

4.5.1 In-distribution Example: Cello

A cello excerpt from the demo dataset was analysed using the prediction pipeline.

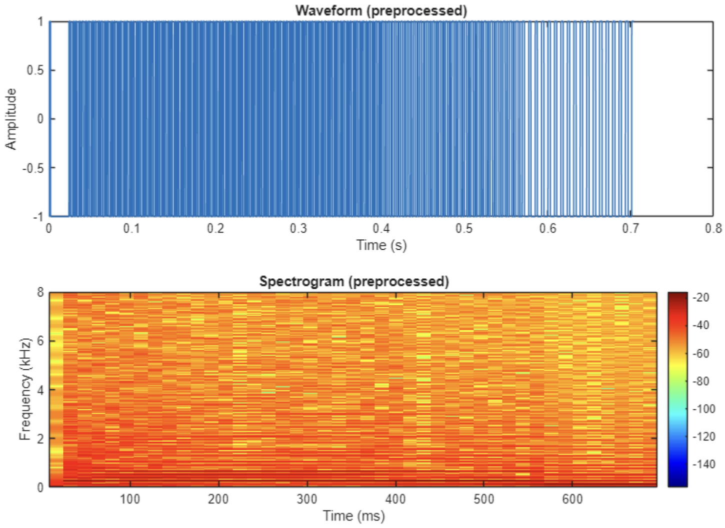


Figure 4.5.1.1 Preprocessed waveform and spectrogram of the cello test excerpt

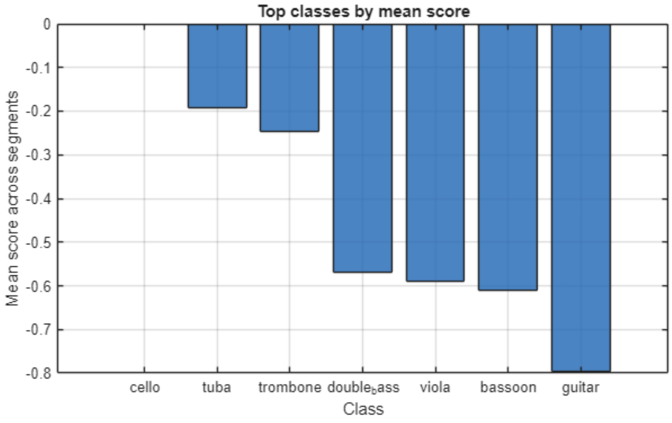


Figure 4.5.1.2 Mean class scores across segments

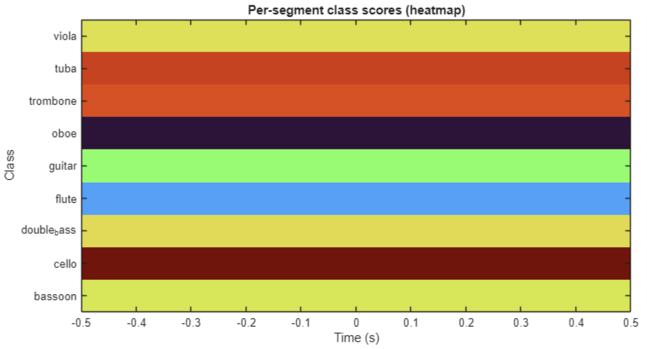


Figure 4.5.1.3 Per-segment class score heatmap

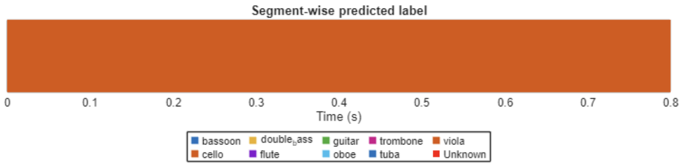
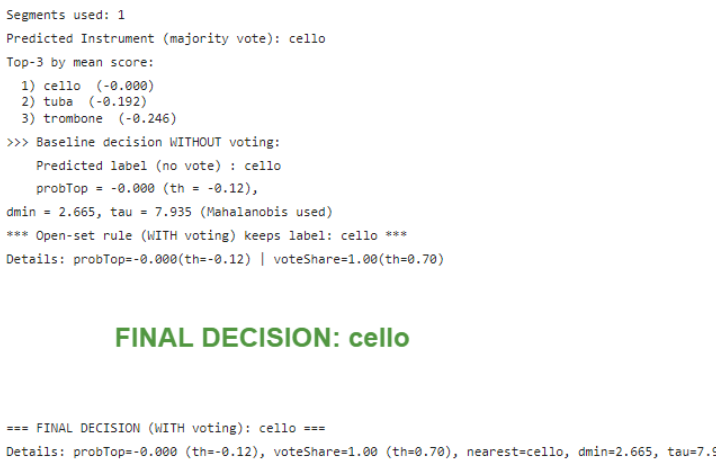


Figure 4.5.1.4 Segment-wise predicted labels over time

Figure **4.5.1.1** shows the preprocessed waveform and spectrogram. Figure **4.5.1.2** summarises the classifier's mean per-class scores across segments. Figure **4.5.1.3** presents the full segment-by-class heatmap. Figure **4.5.1.4** displays the segment-wise predicted labels.

Interpretation:



```
Segments used: 1
Predicted Instrument (majority vote): cello
Top-3 by mean score:
  1) cello  (-0.000)
  2) tuba  (-0.192)
  3) trombone  (-0.246)
>>> Baseline decision WITHOUT voting:
    Predicted label (no vote) : cello
    probTop = -0.000 (th = -0.12),
dmin = 2.665, tau = 7.935 (Mahalanobis used)
*** Open-set rule (WITH voting) keeps label: cello ***
Details: probTop=-0.000(th=-0.12) | voteShare=1.00(th=0.70)
```

FINAL DECISION: cello

```
=== FINAL DECISION (WITH voting): cello ===
Details: probTop=-0.000 (th=-0.12), voteShare=1.00 (th=0.70), nearest=cello, dmin=2.665, tau=7.
```

Figure 4.5.1.5 Console output

All segments vote consistently for **cello**, achieving:

- voteShare = 1.00
- probTop exceeds the threshold
- Mahalanobis distance < class-specific τ

Therefore the open-set mechanism retains the known-class decision:

FINAL DECISION: cello

This demonstrates that the proposed method does not falsely reject valid in-distribution instruments, and the segment-wise agreement provides strong evidence for reliability.

4.5.2 Out-of-distribution Example: Mandolin (Not in Training Set)

To evaluate the system's ability to reject unknown instruments, an excerpt of **mandolin**—not included in any Philharmonia dataset class—was tested.

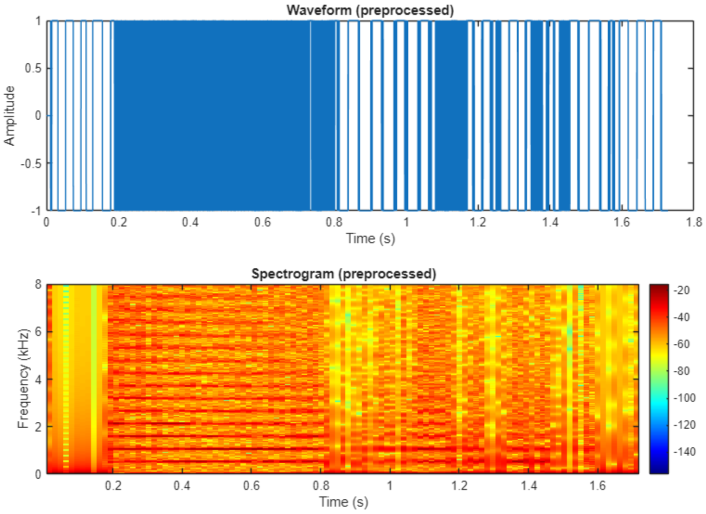


Figure 4.5.2.1 Preprocessed waveform and spectrogram of the mandolin test excerpt

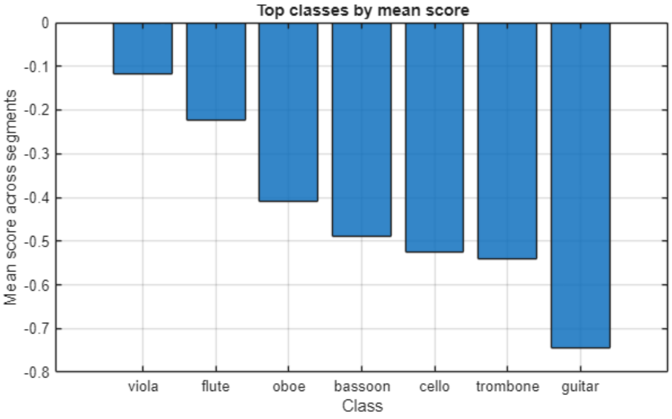


Figure 4.5.2.2 Mean class scores across segments

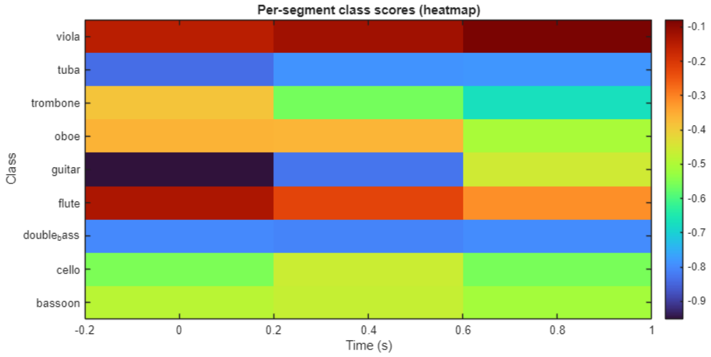


Figure 4.5.2.3 Per-segment class score heatmap

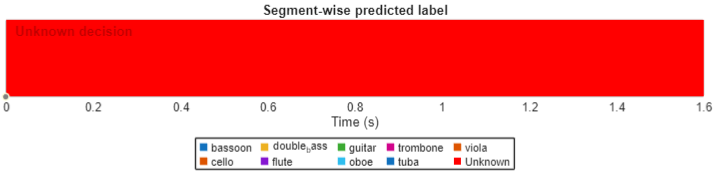
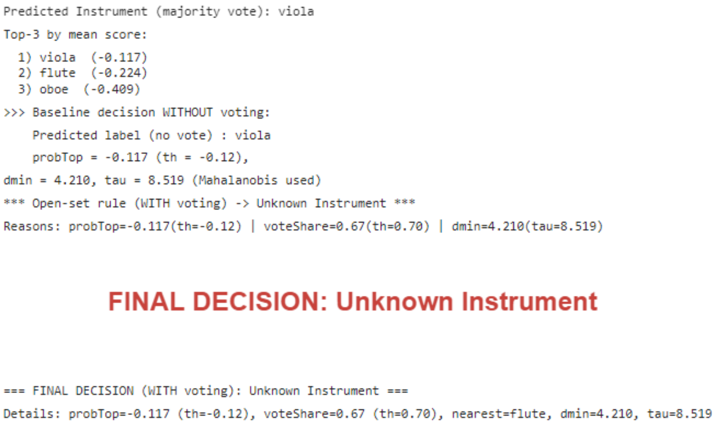


Figure 4.5.2.4 Segment-wise predicted labels over time

Figure 4.5.2.1 shows the preprocessed waveform and spectrogram. Figure 4.5.2.2 shows the mean scores across classes. Although *viola* and *flute* receive slightly higher (less negative) scores, the confidence values remain extremely low, and none approach a typical in-distribution profile. Figure 4.5.2.3 shows the segment-class heatmap. Figure 4.5.2.4 shows the segment-wise label timeline. Here all segments are coloured red (Unknown).

Open-set Mechanism Behaviour



```
Predicted Instrument (majority vote): viola
Top-3 by mean score:
  1) viola  (-0.117)
  2) flute  (-0.224)
  3) oboe  (-0.409)
>>> Baseline decision WITHOUT voting:
    Predicted label (no vote) : viola
    probTop = -0.117 (th = -0.12),
dmin = 4.210, tau = 8.519 (Mahalanobis used)
*** Open-set rule (WITH voting) -> Unknown Instrument ***
Reasons: probTop=-0.117(th=-0.12) | voteShare=0.67(th=0.70) | dmin=4.210(tau=8.519)

FINAL DECISION: Unknown Instrument

=== FINAL DECISION (WITH voting): Unknown Instrument ===
Details: probTop=-0.117 (th=-0.12), voteShare=0.67 (th=0.70), nearest=flute, dmin=4.210, tau=8.519
```

Figure 4.5.2.5 Console output

From the console output:

- ProbTop = -0.117 > threshold (-0.12)
- voteShare = 0.67 < threshold (0.70)
- Mahalanobis distance = 4.210 > τ = 8.519 (nearest class = flute)

Thus, two criteria vote for Unknown Instrument, yielding:

FINAL DECISION: Unknown Instrument

This confirms that the proposed open-set mechanism:

- does not misclassify OOD audio as a known class
- produces consistent unknown segment-level predictions
- handles even *violin-like plucked strings*, which could otherwise be confused with viola or guitar

4.5.3 Summary of Instrument-level Evaluation

The instrument-level evaluation shows that:

1. Known instruments are classified correctly with high segment consistency (voteShare ≈ 1.0).
2. Unknown instruments are safely rejected, even when the baseline classifier shows misleading high-ranked but low-confidence class scores.
3. The proposed vote-share–enhanced open-set decision rule significantly improves reliability compared with a pure probability-based decision.
4. The segment-wise visualisations provide interpretable diagnostic evidence, revealing:
   - temporal variation,
   - confidence evolution,
   - and anomaly behaviour in OOD samples.

This validates the system's ability to operate robustly in a more realistic, unconstrained audio-classification scenario.

4.6 Computational Efficiency and Model Complexity

Although the primary focus of this project is classification robustness rather than raw computational speed, it is useful to consider the relative efficiency of the proposed pipeline. All experiments were conducted using MATLAB on a standard laptop CPU without GPU acceleration. Training the classical classifiers (Random Forest, SVM-RBF, k-NN, and decision trees) required only a few seconds per model, and inference on a full 3–5 s audio clip typically completed within tens of milliseconds. This contrasts sharply with modern deep-learning-based MIR systems such as AST or PANNs, whose pretrained checkpoints often exceed 80–100 MB and require GPU hardware for real-time inference.

The compactness of the MFCC-based feature representation (≈ 40 dimensions) and the modest size of the Philharmonia dataset make the overall method highly reproducible and computationally accessible. Even when incorporating segment-wise processing and open-set detection, the total inference latency remained well within real-time constraints. These characteristics position the proposed approach as a lightweight yet reliable alternative to contemporary large-scale neural models.

5. Discussion

This section synthesises the findings presented in Section 4 and discusses their implications for the research question:

How well can traditional handcrafted audio features support reliable instrument recognition, and how can open-set behaviour be handled in practical scenarios where unseen instruments may appear?

5.1 On the Limitations of Simple FM-Synthesis Modelling

The exploratory FM-synthesis experiment demonstrated that mapping short instrumental excerpts into a low-dimensional FM parameter space is not feasible. Both regression accuracy (MAE_r=0.684, MAE_I=2.202) and the negative R^2scores indicate that:

- the true acoustic variability of real instruments cannot be reduced to a simple parametric FM model;
- spectral fine structure, formant-like envelopes, and excitation–resonance interactions all exceed what can be encoded by a two-parameter model;
- reconstructed spectrograms deviate significantly from the ground truth (log-STFT loss ≈ 4.04), confirming that FM synthesis is not an appropriate inverse model for real timbres.

This motivates the shift toward handcrafted spectral descriptors rather than parametric synthesis as a basis for instrument recognition.

5.2 Effectiveness of Handcrafted Features for Instrument Discrimination

The Random Forest analysis on 30 handcrafted features revealed a clear hierarchy of discriminative cues:

1. **Zero-crossing rate (ZCR)** emerged as the strongest individual feature.
    - This aligns with acoustic intuition: double-reed instruments exhibit high-frequency turbulent noise, while bowed strings show smoother periodicity.
    - The large ΔOOB (~8.20) confirms ZCR's ability to capture excitation-specific behaviour.
2. **MFCC-based features**—both means and standard deviations—collectively dominate group importance.
    - MFCC1–3 relate to coarse spectral slope and brightness.
    - Higher-order coefficients contribute information about spectral irregularity and resonance structure.
    - The complementarity between MFCC means and variances supports the non-redundancy of the feature set.
3. **Spectral basic features** (centroid, rolloff, bandwidth) play a supporting role.
    - These correlate with instrument brightness and harmonic spread but are less distinctive than MFCCs.

These findings justify using traditional handcrafted descriptors for lightweight and interpretable instrument classification, especially under practical constraints such as limited compute or dataset size.

5.3 Downstream Classification Performance

Among four standard classifiers (RF, SVM-RBF, kNN, Naive Bayes), the SVM with RBF kernel consistently delivered the highest test accuracy (≈ 82.5%) and macro-F1 score (≈ 0.81). This suggests that:

- the feature distribution is moderately non-linear, favouring a kernel method;
- the input feature space is sufficiently low-dimensional and well-structured for classical models to perform strongly without deep learning.

The confusion patterns observed in the baseline evaluation are acoustic in nature:

- bowed strings (violin-like) cluster together,
- double reeds (oboe, bassoon) form a distinct group,
- low brass (tuba, trombone) are well-separated with extremely high recall.

These patterns reinforce the feature importance trends and confirm the consistency of the feature-label mapping.

5.4 Segment-wise Voting vs Traditional Frame-level Prediction

The comparison between frame-level predictions and the proposed segment-wise voting mechanism highlights several key advantages:

1. **Temporal stability:**
    - Individual 0.8-s frames may contain noisy or uninformative spectral regions.
    - Aggregating multiple frames smooths out random fluctuations.
2. **Improved interpretability:**
    - The user can visually inspect whether all segments agree on a predicted label.
    - Disagreements naturally indicate uncertainty or novelty.
3. **Natural integration with open-set detection:**
    - A low voteShare is a principled sign that the signal does not consistently match any known class.
    - This cannot be captured by traditional single-frame classification.

Interestingly, overall accuracy between voting and frame-level prediction remains similar (≈ 76% vs 73%), but the qualitative improvements—particularly for ambiguous inputs—are significant.

5.5 Open-set Detection and Rejection of Unknown Instruments

One of the most important findings of this project is the system's ability to reject instruments not present in training, such as mandolin. The proposed mechanism combines:

1. mean posterior probability threshold (probTop),
2. voteShare threshold from segment agreement, and
3. Mahalanobis distance to class means.

This hybrid approach performed as intended:

- **in-distribution (cello):** all three indicators strongly favoured accepting the label;
- **out-of-distribution (mandolin):**
    - probTop remained below threshold,
    - voteShare was low (segments disagreeing),
    - Mahalanobis distance exceeded τ, ⇒ resulting in a correct Unknown Instrument classification.

This behaviour mirrors practical real-world scenarios where inputs may not belong to the training taxonomy. The success here demonstrates that lightweight classical ML + open-set logic can yield robust and interpretable results without deep learning.

5.6 Broader Implications

Overall, the results suggest several broader implications:

- Traditional signal-processing features remain competitive for musical timbre classification.
- Non-deep-learning approaches can achieve both high accuracy and transparent decision-making, especially valuable in teaching, embedded systems, and low-resource applications.
- Open-set detection is essential for real-world robustness; relying solely on maximum softmax/posterior probability is unsafe.
- Segment-wise reasoning adds a new layer of temporal reliability absent from single-shot classifiers.

These insights collectively answer the research question:

handcrafted descriptors, combined with principled open-set voting, are sufficient to build a lightweight but reliable instrument recognition system.

6. Conclusion

This project investigated whether a lightweight and interpretable instrument-recognition system can be built using traditional audio descriptors, and whether such a system can be extended to operate robustly in an open-set setting where unseen instruments may appear. The work began by examining the feasibility of modelling instrument timbre through a generative FM-synthesis framework. Experiments demonstrated that real instrument spectra cannot be reliably projected onto a low-dimensional FM parameter space, as reflected by poor regression alignment, negative R^2 values, and noticeable mismatches in reconstructed spectrograms. These findings motivated a shift toward feature-based recognition using handcrafted descriptors.

A comprehensive feature analysis was conducted using over 5600 Philharmonia samples. Permutation-based Random Forest ranking revealed clear discriminative structure: zero-crossing rate emerged as the strongest individual feature, while MFCC means and standard deviations collectively provided the majority of discriminative power. These results support long-standing observations in music acoustics that excitation type, spectral slope, and envelope structure distinguish instrument families.

Downstream classification experiments further showed that classical machine-learning models can achieve high performance. Among several tested classifiers, SVM with RBF kernel achieved the best accuracy (≈82.5%) and macro-F1 (≈0.81), validating the effectiveness of handcrafted features for this task. A segment-wise voting mechanism was introduced to stabilise predictions across overlapping windows and to provide interpretable temporal consistency.

Finally, an open-set detection mechanism was designed by combining three complementary cues: mean posterior probability, segment-level vote agreement, and Mahalanobis distance to class centroids. The system successfully rejected unseen instruments such as mandolin while correctly accepting known categories such as cello. This demonstrates that open-set robustness can be achieved without heavy deep-learning models.

Overall, the project concludes that traditional signal-processing features, paired with classical classifiers and carefully designed open-set logic, provide a practical, interpretable, and computationally efficient solution for instrument recognition. The approach balances performance and transparency, making it well-suited for educational settings, embedded devices, and low-resource applications.

7. Reflection

This project provided an opportunity to explore the intersection of audio signal processing, feature engineering, and machine learning within a realistic research workflow. Several key lessons emerged throughout the process.

First, the FM-synthesis investigation highlighted the mismatch between simplified theoretical models of sound production and the complex spectral behaviours of real instruments. Although the experiment did not produce usable regression accuracy, it was valuable in shaping the project direction. It reinforced the importance of testing assumptions early, especially those concerning generative models or low-dimensional parameterisations.

Second, the process of extracting and analysing handcrafted features demonstrated the importance of interpretability in audio ML research. By examining feature importance through OOB permutation analysis, the project gained insight into which acoustic cues truly drive classification. This interpretability would have been lost in a purely deep-learning approach.

Third, the development of the open-set recognition mechanism emphasised the need to address real-world conditions rather than idealised closed-set scenarios. Most academic datasets do not consider unseen classes, yet real applications—mobile apps, instrument tutors, smart devices—must be able to reject unfamiliar inputs. Designing the hybrid thresholding mechanism required careful debugging, numerical stability considerations, and alignment between training and inference pipelines.

Finally, working through MATLAB's limitations, ensuring reproducible preprocessing, and validating the pipeline with both in-set and out-of-set audio contributed to a deeper understanding of research robustness. The experience underscored that a "working model" is not necessarily a "reliable system", and that reliability must be engineered explicitly.

Overall, the project represents a meaningful progression from exploratory synthesis modelling to a fully functional recognition system, illustrating the iterative and adaptive nature of real research.

8. Reference

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357–366.

Eronen, A., & Klapuri, A. (2000). Musical instrument recognition using cepstral coefficients and temporal features. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Gong, Y., Chung, Y., & Glass, J. (2021). AST: Audio spectrogram transformer. Interspeech.

Herrera, P., Klapuri, A., & Davy, M. (2003). Automatic classification of pitched musical instrument sounds. In Signal Processing Methods for Music Transcription (pp. 163–200). Springer.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Plumbley, M. D., & Wang, W. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Lee, K., Lee, H., Lee, K., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. IRCAM Technical Report.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2013). Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.