Audio Deepfake Detection

Feature Engineering, Model Selection & Zero-Shot Evaluation

George Nazos

MSc Project

February 10, 2026

Outline

Problem Statement

- Audio deepfakes generated by modern TTS and voice-cloning systems are increasingly realistic and difficult to detect by ear.
- **Goal:** Build a classifier that distinguishes *real* human speech from *fake* (machine-generated) speech using handcrafted audio features.
- We train on the **Fake-or-Real (FoR)** dataset and evaluate zero-shot generalization on two unseen datasets.

Datasets

Dataset	Samples	Role	Source
FoR (train)	53 864	Training + Validation	Controlled TTS
FoR (test)	31 526	In-distribution test	Controlled TTS
In-the-Wild	—	Zero-shot evaluation	Various real-world
ElevenLabs	—	Zero-shot evaluation	Single TTS provider

- **In-the-Wild**: diverse sources, varied recording conditions.
- **ElevenLabs**: single synthetic voice provider — narrow domain.

Extracted Audio Features

For every audio clip we extract frame-level features and aggregate them:

Feature Group	**Description**	**Coefficients**
MFCC	Mel-frequency cepstral coefficients	20
MFCC Δ	First-order temporal derivatives	20
MFCC Δ^2	Second-order temporal derivatives	20
Mel Spectrogram	Energy per mel-frequency band	128
Spectral	Centroid, bandwidth, flatness, rolloff	4
Other	RMSE, ZCR, pitch (YIN)	3

Each feature was computed as both **mean** and **standard deviation** across all frames ($\sim$394 total features).

Feature Selection via Logistic Regression

We trained a Logistic Regression model on three feature subsets and compared prediction scores on the FoR validation set:

Feature Set	Features	F1	ROC AUC
All (mean + std)	∼394	0.72	0.80
Without mel spectrogram	∼138	0.55	0.56
Mean only (selected)	**195**	**0.78**	**0.85**

Key findings:

- Removing mel spectrogram features collapsed performance ⇒ they are essential.
- Dropping std aggregations *improved* every metric ⇒ std added noise.
- Final set: **195 mean-only features** used for all models.

Training Data: Correlation Heatmaps

- Pearson and Spearman correlation matrices reveal **strong internal correlation** within the 128 mel spectrogram bands.
- MFCC, Δ, and Δ^2 features show moderate cross-group correlation.
- This redundancy explains why some models overfit when mel features are included at full dimensionality.

(See notebook output for full heatmap visualizations.)

Models & Grid Search

Nine models were trained using `grid_search_joblib` with F_1 (macro) scoring:

Model	Key Hyperparameters Searched
Logistic Regression	C (logspace 10^{-3}–10^{2}), penalty (ℓ_1, ℓ_2)
Linear SVM	$C \in \{0.001 \ldots 1000\}$
RBF SVM	C, γ
Polynomial SVM	C, γ, degree, coef0
Sigmoid SVM	C, γ, coef0
Decision Tree	max_depth (5–19), criterion, ccp_alpha
Random Forest	n_estimators, max_depth, max_samples
XGBoost	depth, LR, regularization, subsampling
XGBoost (no mel)	Same as XGBoost, 67 features

- Best model retrained on **train + validation**, evaluated on **test**.
- Linear models use `StandardScaler`; tree-based use `SimpleImputer(median)`.

Best Hyperparameters (Selected Models)

Model	**Best Parameters**
Logistic Reg.	$C = 2.15$, penalty $= \ell_2$, solver = saga
Linear SVM	$C = 0.01$
RBF SVM	$C = 0.1$, $\gamma = 0.01$
Poly SVM	$C = 1$, degree $= 2$, $\gamma = 0.01$, coef0 $= 0.0$
Sigmoid SVM	$C = 0.1$, $\gamma =$ scale, coef0 $= -1.0$
Random Forest	500 trees, max_depth $= 4$, max_samples $= 0.6$
XGBoost	depth $= 2$, LR $= 0.03$, 800 estimators
XGBoost (no mel)	depth $= 3$, LR $= 0.05$, 700 estimators

FoR Test Set – All Models

Model	Acc.	Prec.	Recall	F1	AUC
Logistic Reg.	0.716	0.706	0.718	0.707	0.777
Linear SVM	0.720	0.707	0.718	0.710	0.781
RBF SVM	0.706	0.697	0.709	0.697	0.763
Poly SVM	0.674	0.650	0.614	0.614	0.691
Sigmoid SVM	0.660	0.677	0.686	0.658	0.704
Decision Tree	0.590	0.550	0.545	0.545	0.522
Random Forest	0.640	0.601	0.566	0.555	0.651
XGBoost	0.665	0.696	0.565	0.526	0.728
XGBoost (no mel)	**0.741**	**0.731**	**0.698**	**0.706**	**0.786**

- **XGBoost (no mel)** achieves the best overall FoR test performance.
- Linear models (LR, Linear SVM) are the most consistent.
- Decision Tree and Random Forest show signs of overfitting.

Zero-Shot: In-the-Wild

Models evaluated on the In-the-Wild dataset **without any retraining**:

Model	Acc.	Prec.	Recall	F1	AUC
Logistic Reg.	—	—	—	—	—
Linear SVM	—	—	—	—	—
RBF SVM	—	—	—	—	—
Poly SVM	—	—	—	—	—
Sigmoid SVM	—	—	—	—	—
Decision Tree	—	—	—	—	—
Random Forest	—	—	—	—	—
XGBoost	—	—	—	—	—
XGBoost (no mel)	—	—	—	—	—

Fill in after running the notebook — values are computed dynamically.

Zero-Shot: ElevenLabs

Models evaluated on the ElevenLabs dataset **without any retraining**:

Model	Acc.	Prec.	Recall	F1	AUC
Logistic Reg.	—	—	—	—	—
Linear SVM	—	—	—	—	—
RBF SVM	—	—	—	—	—
Poly SVM	—	—	—	—	—
Sigmoid SVM	—	—	—	—	—
Decision Tree	—	—	—	—	—
Random Forest	—	—	—	—	—
XGBoost	—	—	—	—	—
XGBoost (no mel)	—	—	—	—	—

Fill in after running the notebook — values are computed dynamically.

Zero-Shot: F1 Comparison

(Insert grouped bar chart from notebook output: F1 scores across all models on FoR Test, ITW, and ElevenLabs.)

Key observations:

- All models experience a **performance gap** moving from FoR to unseen datasets.
- Linear models generalize more **consistently** across domains.
- Non-linear / tree-based models show higher variance in zero-shot performance.

Confusion Matrix Highlights

(Insert confusion matrix grids from notebook: all 9 models on ITW and ElevenLabs datasets.)

- Decision Tree performs better on **ElevenLabs** (narrow, single-source domain) but worse on **ITW** (diverse domain).
- Regularized ensembles (RF, XGBoost) show the opposite pattern: better ITW transfer, weaker on ElevenLabs.
- This reflects the **specificity vs. generality trade-off** inherent in model complexity.

Impact of Mel Spectrogram Features

XGBoost (195 features, with mel):

- FoR test F1 = 0.526
- ROC AUC = 0.728
- High mel correlation $\Rightarrow$ overfitting

XGBoost (67 features, no mel):

- FoR test F1 = **0.706**
- ROC AUC = **0.786**
- Better generalization

- Removing mel features **improved** XGBoost performance by +0.18 F1.
- For Logistic Regression, mel features were **essential** (F1 dropped from 0.78 to 0.55 without them).
- **Conclusion:** The effect of mel features is model-dependent — linear models benefit, tree-based models overfit on them.

Specificity vs. Generality Trade-off

- **Decision Tree** (deep, specific rules):
  - Captures narrow artifacts $\Rightarrow$ good on single-source ElevenLabs.
  - Fails on diverse ITW data — overly specific.
- **Ensemble models** (RF, XGBoost — regularized, smooth):
  - Tuned aggressively *against* overfitting (shallow depth, subsampling).
  - Handle ITW diversity better, but miss narrow provider-specific signatures.
- **Linear models** (LR, Linear SVM):
  - Simple decision boundaries — least susceptible to dataset-specific artifacts.
  - Most **consistent** cross-domain generalization.

Key Findings

1. **Feature selection matters:** Mean-only aggregation (195 features) outperformed the full feature set (∼394 features).
2. **Best in-distribution model:** XGBoost (no mel) — 0.74 accuracy, 0.71 F1, 0.79 AUC on FoR test.
3. **Best generalizers:** Linear models (Logistic Regression, Linear SVM) show the most consistent zero-shot transfer.
4. **No single model dominates** across all zero-shot domains — the choice depends on deployment scenario (narrow vs. diverse).
5. **Mel features are double-edged:** Essential for linear models, but cause overfitting in tree-based models.

Recommendations & Future Work

- Prefer **linear models** when cross-domain robustness is the priority.
- Consider **PCA or feature selection** on mel spectrogram bands to reduce redundancy for tree-based models.
- Explore **ensemble stacking** (linear + tree-based) for improved robustness.
- Investigate **domain adaptation** techniques to bridge the in-distribution vs. zero-shot performance gap.
- Evaluate on additional unseen datasets to validate generalization claims.

Thank you

Questions?

Source code: `https://github.com/g-nazos/audio-deepfake-detection`