Anomaly Detection from Stochastic Differential Equation realizations

Written by Nathan Abell

Generative AI technologies can be maliciously used for misinformation, therefore detection tools need to keep up with the rapid evolution of generative AI. We frame the problem of detecting fake data in time series as an anomaly detection problem. This project intends to develop a pipeline for detecting synthetic audio using stochastic differential equation (SDE)-based modeling and interpretable machine learning.

0 INTRODUCTION

The proliferation of generative audio models has introduced new challenges in verifying the authenticity of spoken content. Detecting synthetic audio is critical in applications such as prevention of misinformation, biometric security, and the integrity of digital media. Traditional audio classification relies heavily on spectral descriptors, however, these may overlook subtle temporal anomalies present in fake generated audio.

This project introduces a hybrid approach that combines stochastic differential equation (SDE)-based modeling with spectral feature extraction to identify fake audio by recognizing it as an anomaly. By analyzing the dynamics of Mel spectrogram time series and augmenting them with statistical and machine learning classifiers, we construct an interpretable and effective detection framework. The approach emphasizes mathematical rigor through residual analysis of SDEs while leveraging data-driven insights through SHAP-explained ML models.

1 Stochastic Differential Equations

Stochastic Differential Equations (SDEs) are differential equations that incorporate random noise, commonly modeled by Brownian motion or Wiener processes. They are used to describe systems where evolution over time is influenced by both deterministic trends and inherent randomness. One of the most common uses is the Black-Scholes model, which is a mathematical model for the dynamics of a financial market containing derivative investment instruments.

Stochastic Differential Equations (SDEs)

A **Stochastic Differential Equation (SDE)** models the evolution of a variable over time under both deterministic and random influences. The general form is:

$$dX_t = \mu(X_t,t)\,dt + \sigma(X_t,t)\,dW_t \tag{1}$$

where:

- X_t is the stochastic process
- $\mu(X_t,t)$ is the *drift term*, representing deterministic trends,
- $\sigma(X_t,t)$ is the *diffusion term*, representing random fluctuations,
- W_t is a Wiener process (standard Brownian motion),
- dt is an infinitesimal time increment,
- dW_t is a random increment with mean 0 and variance dt.

An **SDE realization** is a single sample path (trajectory) generated by solving the SDE numerically, typically using methods like Eulertextbf–Maruyama. Each realization reflects a possible evolution of the process over time.

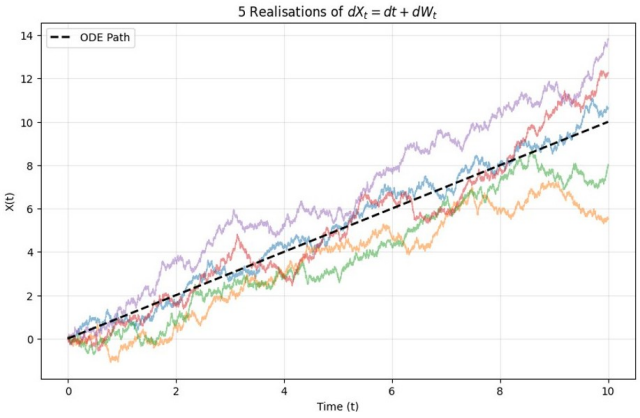

Fig. 1. Example of anomaly detection feature distribution

1.1 Processing

To prepare the audio signals for analysis, each waveform is first normalized and converted into a Mel spectrogram, which represents the signal's energy distribution across perceptually spaced frequency bands over time. This transformation captures both spectral and temporal dynamics of the audio in a compact, interpretable format. From the Mel spectrogram, statistical features are extracted for each Mel band, such as range, smoothness ratio, and increment statistics. In parallel, a SDE is fit to each band's time series to estimate dynamic properties like drift and noise residuals. Additionally, global spectral features such as MFCCs, spectral centroid, bandwidth, flatness, roll-off, and zero-crossing rate are computed directly from the waveform to characterize the signal's timbre and structure. Together, these features form a high-dimensional representation that encapsulates both deterministic trends and subtle stochastic variations useful for anomaly detection.

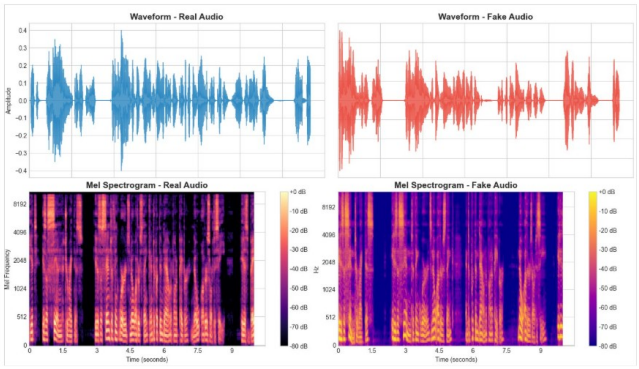

Fig. 2. Example of time series audio data going through mel-spectrography

1.1.1 Models

Following a construction of a dataset from the extracted features a few models can be developed to solve this anomaly detection problem. To classify audio clips without using machine learning, we implemented a rule-based algorithm based on statistical deviation. First, we computed the mean (μ) and standard deviation (σ) of each audio feature across a corpus of real clips. For every new clip, we calculated a Z-score for each feature using $Z = \frac{|x-\mu|}{\sigma}$ to measure how far it deviated from the real-audio norm. Features with Z-scores exceeding a defined threshold were flagged as abnormal. If the number of abnormal features in a clip surpassed a second threshold, the clip was classified as fake. The two tunable parameters—the Z score threshold and the feature count threshold—were optimized in a validation set to balance sensitivity and specificity. This approach offers interpretable anomaly detection based solely on statistical principles. We secondly also used a Random Forest algorithm to also simulate what an approach would be if using ML-based models. A Random Forest is an ensemble-based machine learning method that constructs multiple decision trees during training and aggregates their predictions for classification. Each tree is trained on a random subset of the data with replacement, and at each node, a random subset of features is considered for splitting. For classification tasks, such as distinguishing real from fake audio, the Random Forest outputs the class chosen by the majority of trees. This method is robust to overfitting, handles high-dimensional data effectively, and provides interpretable outputs through feature importance scores, making it well-suited for tabular data with complex feature interactions like those extracted from audio signals.

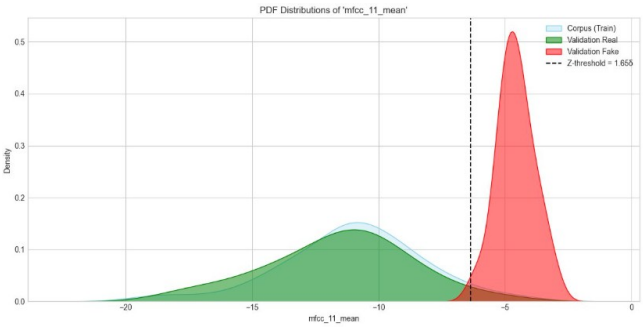

Fig. 3. Distribution of Z-scores of a singular feature

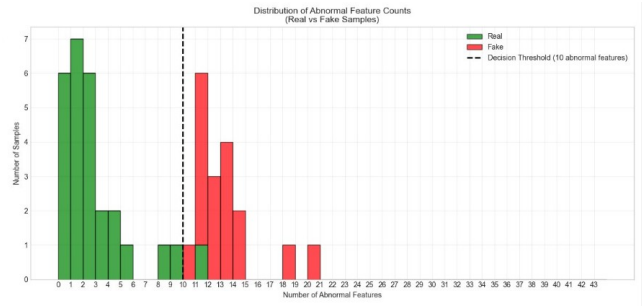

Fig. 4. Distribution of predictions

We can then test our models on a test data and construct a confusion matrix and calculate confidence or shap values to see our results and which features contribute the most to the final model. These results will be displayed in a confusion matrix and the feature importance will then be graphed.

2 CONCLUSION

This project presents a hybrid approach to synthetic audio detection by integrating stochastic differential equation (SDE)-based time series analysis with spectral feature extraction. The combination of mathematically interpretable SDE residual modeling and statistical or machine learning classifiers enables effective identification of fake audio content. Results show that both rule-based and Random Forest models can successfully distinguish real from generated speech, with SHAP analysis enhancing model transparency. This methodology not only provides strong detection performance but also maintains interpretability—making it a promising framework for secure and trustworthy audio authentication systems.

THE AUTHORS

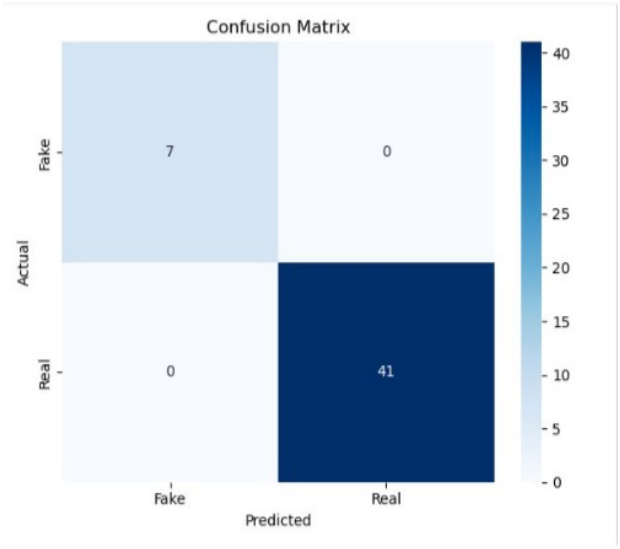

Fig. 5. Results of the ML Model

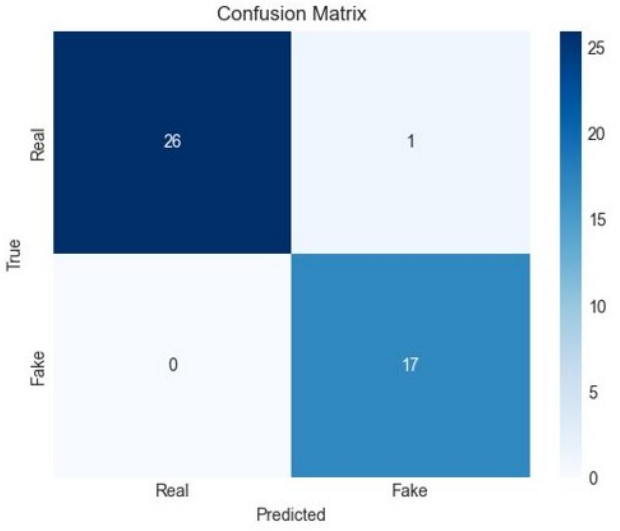

Fig. 6. Results of Non-ML Model

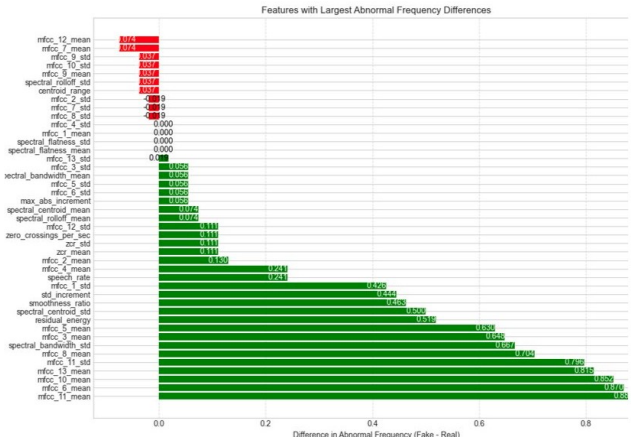

Fig. 7. Confidence (Fake - Real) of Non-ML Model

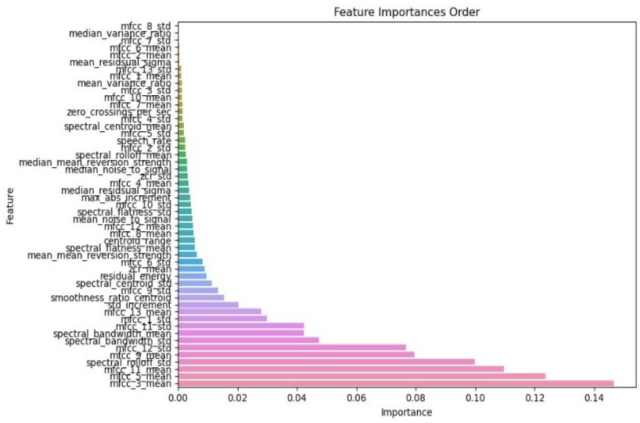

Fig. 8. Confidence (SHAP) of ML Model