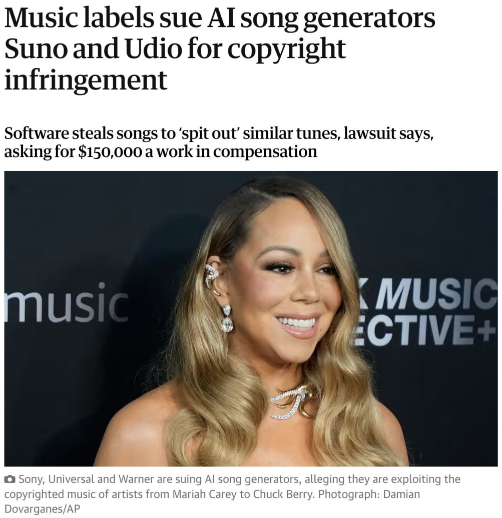
Detecting AI-Generated Music Using LambdaResNet and Swin Transformers on Mel-Spectrograms

ST311 Group 12
Candidate Numbers: 44535, 40538

Motivation

Why Detect AI-Generated Music?

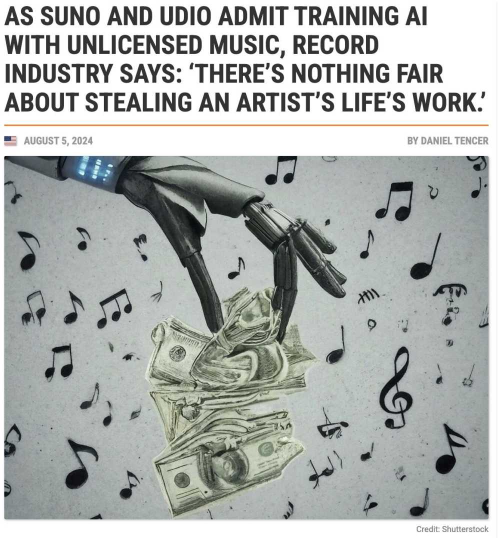
- AI music generators (e.g. Suno, MusicGen) can now mimic human composition with striking realism
- Threats to authenticity and copyright: Artists face risk of being mimicked or replaced
- Platforms may be flooded with synthetic content, risking monetization abuse and user trust



Music labels sue AI song generators Suno and Udio for copyright infringement

Software steals songs to 'spit out' similar tunes, lawsuit says, asking for $150,000 a work in compensation

Sony, Universal and Warner are suing AI song generators, alleging they are exploiting the copyrighted music of artists from Mariah Carey to Chuck Berry. Photograph: Damian Dovarganes/AP



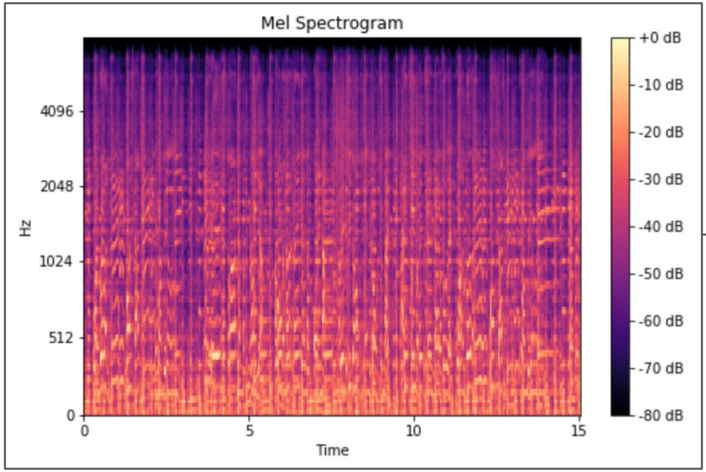
AS SUNO AND UDIO ADMIT TRAINING AI WITH UNLICENSED MUSIC, RECORD INDUSTRY SAYS: 'THERE'S NOTHING FAIR ABOUT STEALING AN ARTIST'S LIFE'S WORK.'

AUGUST 5, 2024 BY DANIEL TENCER

Credit: Shutterstock

Research Question

Can 5-second mel-spectrograms be classified using small vision models?



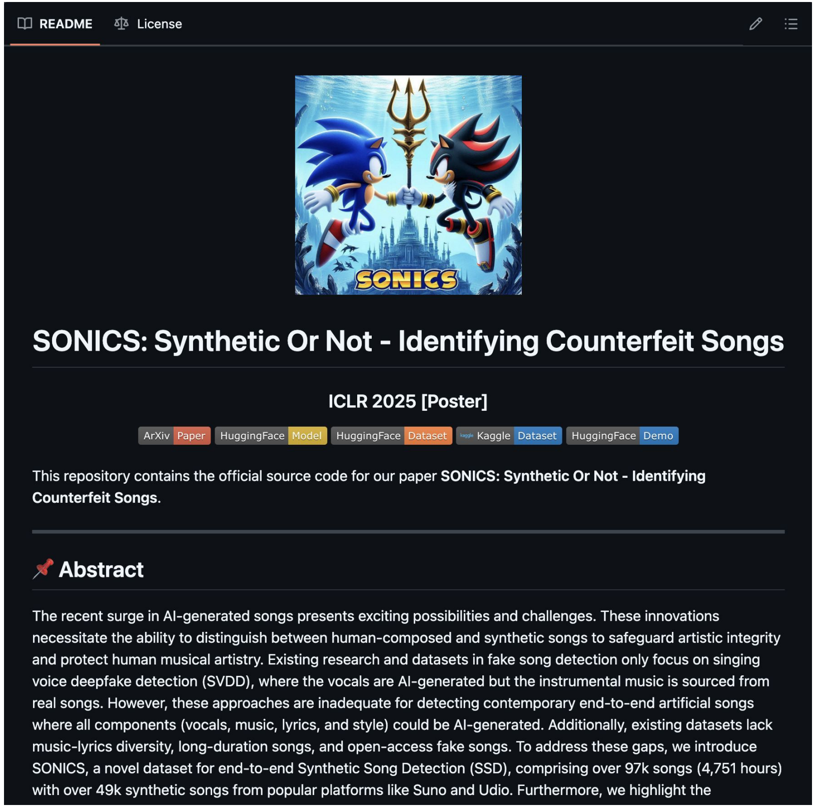
Input

Is the song AI-generated? (Yes/No)

Output

Dataset: SONICS

- Contains 50,000 real + 50,000 AI generated songs
- In our use case, we selected 10,000 real + 10,000 AI generated songs randomly, creating our own dataset
- Train / Val / Test ratio = 60:20:20



SONICS: Synthetic Or Not - Identifying Counterfeit Songs

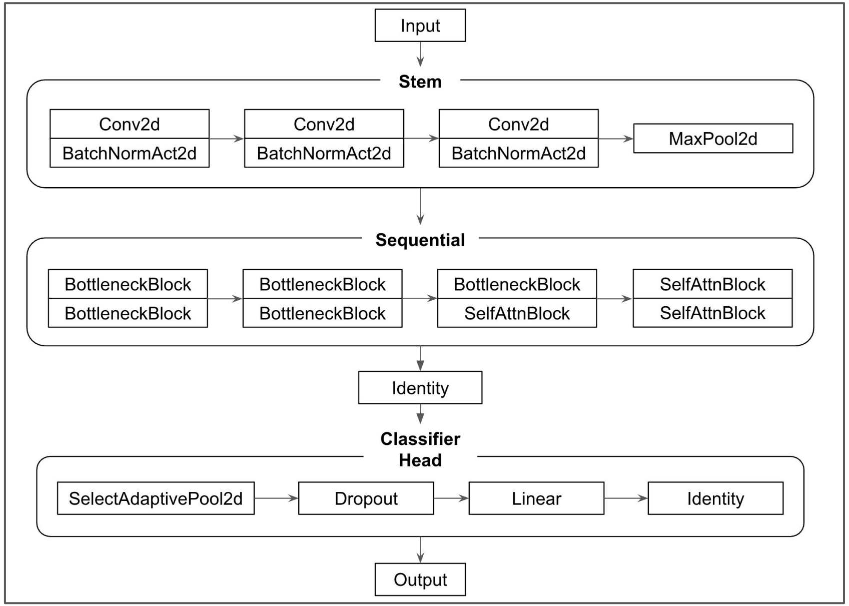
ICLR 2025 [Poster]

This repository contains the official source code for our paper **SONICS: Synthetic Or Not - Identifying Counterfeit Songs**.

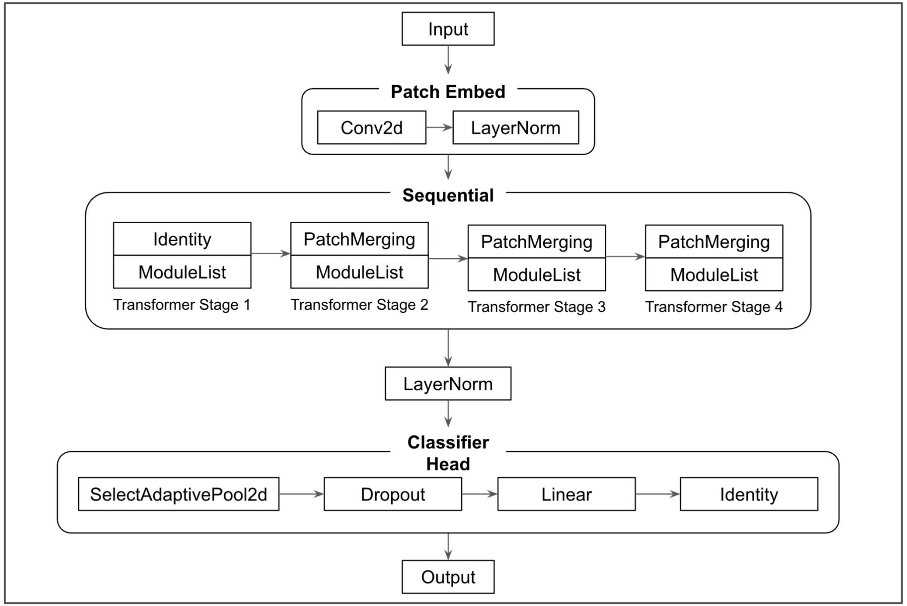
📌 Abstract

The recent surge in AI-generated songs presents exciting possibilities and challenges. These innovations necessitate the ability to distinguish between human-composed and synthetic songs to safeguard artistic integrity and protect human musical artistry. Existing research and datasets in fake song detection only focus on singing voice deepfake detection (SVDD), where the vocals are AI-generated but the instrumental music is sourced from real songs. However, these approaches are inadequate for detecting contemporary end-to-end artificial songs where all components (vocals, music, lyrics, and style) could be AI-generated. Additionally, existing datasets lack music-lyrics diversity, long-duration songs, and open-access fake songs. To address these gaps, we introduce SONICS, a novel dataset for end-to-end Synthetic Song Detection (SSD), comprising over 97k songs (4,751 hours) with over 49k synthetic songs from popular platforms like Suno and Udio. Furthermore, we highlight the

Model Architectures

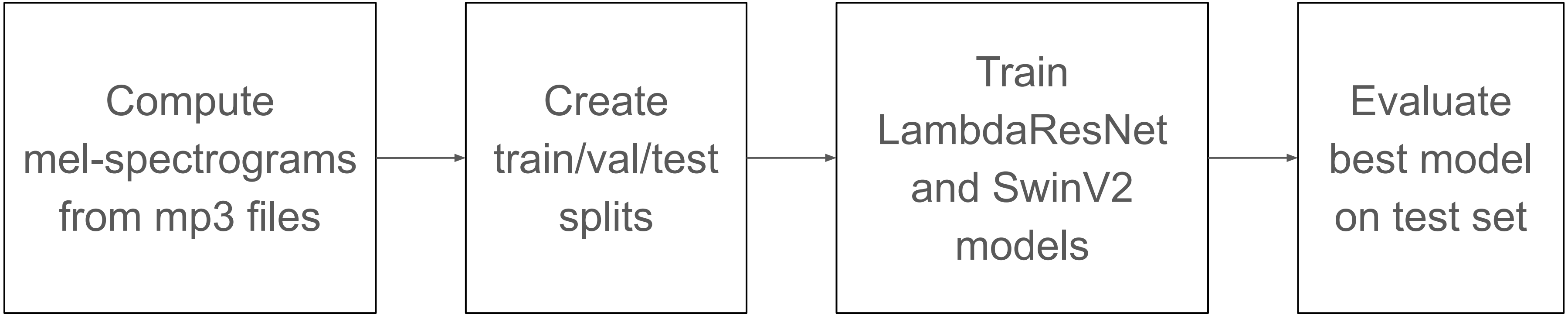


LambdaResNet26rp_256 (~11M params)



Swin Transformer V2 (~51M params)

Algorithm Pipeline



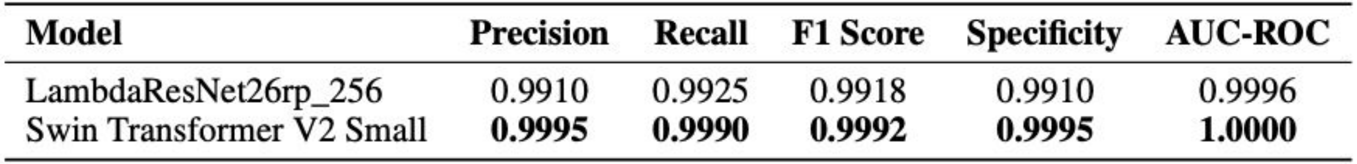
Compute mel-spectrograms from mp3 files

Create train/val/test splits

Train LambdaResNet and SwinV2 models

Evaluate best model on test set

Results



Model	Precision	Recall	F1 Score	Specificity	AUC-ROC
LambdaResNet26rp_256	0.9910	0.9925	0.9918	0.9910	0.9996
Swin Transformer V2 Small	**0.9995**	**0.9990**	**0.9992**	**0.9995**	**1.0000**

- **Swin Transformer** achieves **near-perfect classification** across all metrics:
  - 99.9%+ precision → Very few false positives
  - AUC-ROC = 1.000 → Perfect separation between classes
- **LambdaResNet** is slightly **less accurate** but still exceptional:
  - Over 99% across all metrics
  - Best for speed and efficiency (fewer params, faster inference)
- Both models are well-calibrated with balanced specificity and recall
- Swin Transformer is ideal for high-stakes, high-accuracy settings
- LambdaResNet is better for real-time or edge deployment

Limitations

- Used only SONICS dataset (might be biases in music genre)
- Binary classification only
- Trained on shorter music clips

Future Work

- Multi-class detection
- Longer audio modelling (using full songs)
- More explainability (why the model classified a clip as real / fake)