Team 9

Synthetic Speech Detection

GITHUB REPOSITORY

Skoltech

Research problem

The key **problems** are:

- growing generation of highly realistic fake speech
- poor generalization of synthetic speech detection systems.

spoof (ASVspoof21)

bonafide (ASVspoof21)

spoof (ASVspoof21)

Goal and Objectives

Developing a comprehensive fake speech detection system using state-of-the-art feature extraction and neural network architectures.

Approach:

- Implement multiple audio feature extraction methods
- Develop a transformer-based neural network
- Adopt base model ResNetX
- Train and evaluate on ASVspoof19 and ASVspoof21 datasets

Foundation: Building on the framework from "Fake Speech Detection Using Residual Network with Transformer Encoder" [1].

[1] Zhenyu Zhang, Xiaowei Yi, and Xianfeng Zhao. Fake speech detection using residual network with transformer encoder. In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec '21, pages 13–22, New York, NY, USA, 2021.

Dataset for training and Evaluation

- ASVspoof 2019 LA:
  - Training, development, and evaluation partitions
    - Spoofed utterances generated using 19 VC and TTS algorithms

- ASVspoof 2021 DF:
  - Evaluation set similar to ASVspoof 2019 LA
  - Exhibits audio coding and compression artifacts
  - No new training or development data released

Compare and describe datasets

Feature	ASVspoof 2019 LA [1]	ASVspoof 2021 DF [2]
Speech Types	TTS (Text-to-Speech) and VC (Voice Conversion)	Deep fake audio generated using advanced models
Data Composition	Synthetic and bonafide (real) speech	Deep fake and bonafide (real) speech
Total Utterances	Approx. 23,000+	Approx. 120,000+
Bonafide Utterances	Train: 2,580, Dev: 1,484, Eval: 7,355	Eval: 14,859
Spoofed Utterances	Train: 22,800, Dev: 22,296, Eval: 63,882	Eval: 519, 059
Duration of Utterances	3-10 seconds	Variable durations
Sample Rate	16,000 Hz	16,000 Hz
Audio Format	FLAC	FLAC
Reference	ASVspoof 2019	ASVspoof 2021
Citation	Todisco et al., Interspeech 2019	Kinnunen et al., Interspeech 2021

[1] X. Liu et al., "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2507-2522, 2023, doi: 10.1109/TASLP.2023.3285283.

[2] Yamagishi, Junichi; Todisco, Massimiliano; Sahidullah, Md; Delgado, Héctor; Wang, Xin; Evans, Nicolas; Kinnunen, Tomi; Lee, Kong Aik; Vestman, Ville; Nautsch, Andreas. (2019). ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2555.

Pipeline

Feature Extraction

- Log Power Spectrum (LPS)
- Mel-Frequency Cepstral Coefficients (MFCCs)
- Constant Q Cepstral Coefficients (CQCC)

Model

- ResNet18
- TE-ResNet

Training

- Use MFCC, CQCC, LPS features
- Train TE-ResNet and ResNet18
- Cross-entropy loss, Adam optimizer

Testing

- ASVspoof19
- ASVspoof21

Check for AI Music Contest

- CtrSVDD Track, Test Set
- Bonafide logit used rather than classification

Feature extraction modes

01

$$LPS(w)=log|F\left(x\left(t\right)\right)|^2$$

02

$$MFCC(q)=\sum_{m=1}^{M} log|MF\left(m\right)|cos\left[\frac{q\left(m-1/2\right)\pi}{M}\right]$$

03

$$CQCC(p)=\sum_{l=1}^{L} log|X^{CQ}\left(l\right)|^2 cos\left[\frac{p\left(l-1/2\right)\pi}{L}\right]$$

Results, ResNet18

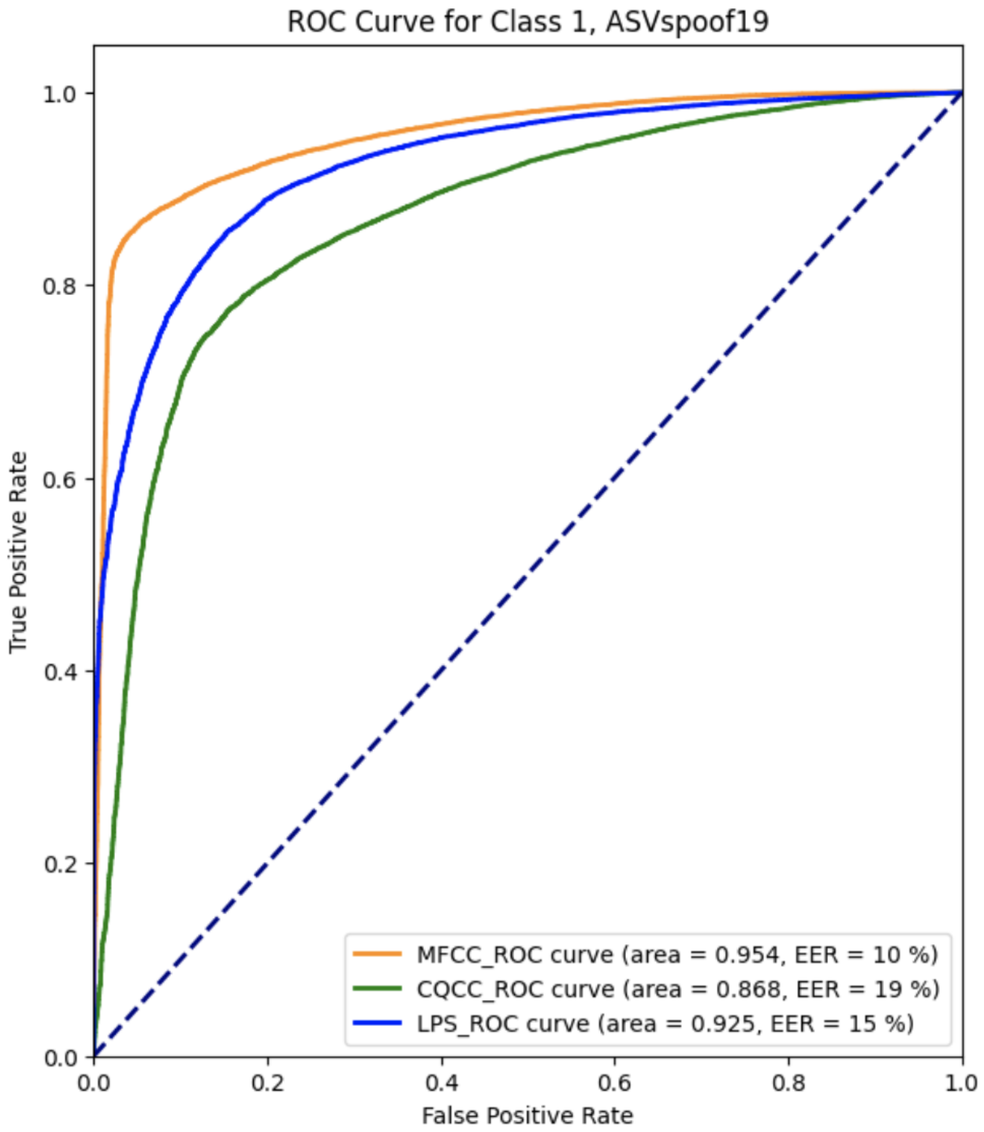
Training				
Feature Extractor	Loss	Balanced Accuracy	Precision	Recall
MFCC	0.025	0.981	0.995	0.996
CQCC	0.079	0.923	0.981	0.986
LPS	0.019	0.981	0.996	0.996

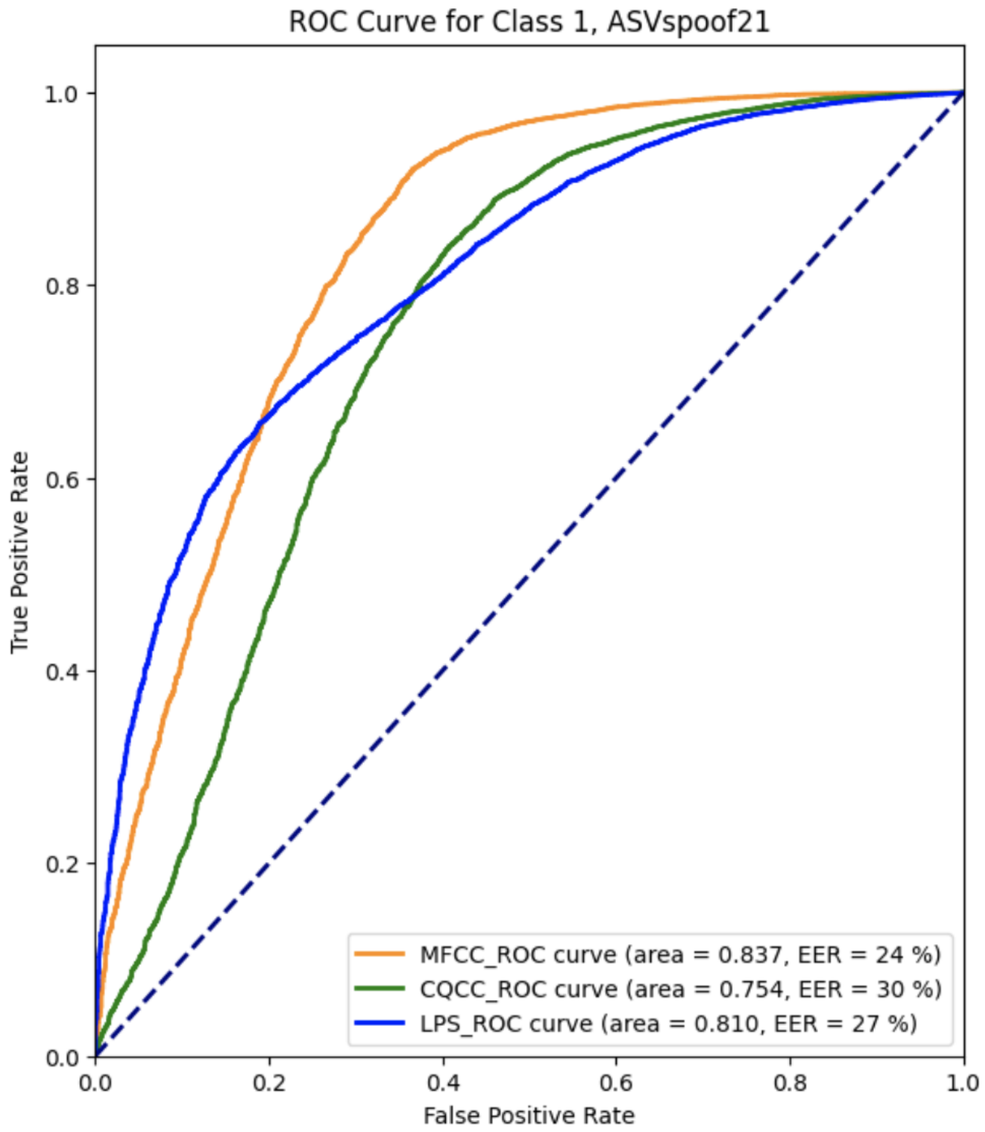
Results, ResNet18

Evaluation on ASVspoof19				
Feature Extractor	Loss	Balanced Accuracy	Precision	Recall
MFCC	0.411	0.893	0.987	0.893
CQCC	0.399	0.695	0.924	0.949
LPS	0.580	0.839	0.988	0.758

Evaluation on ASVspoof21				
Feature Extractor	Loss	Balanced Accuracy	Precision	Recall
MFCC	0.228	0.801	0.981	0.966
CQCC	0.242	0.656	0.968	0.99
LPS	0.534	0.739	0.982	0.813

ROC/AUC ResNet18

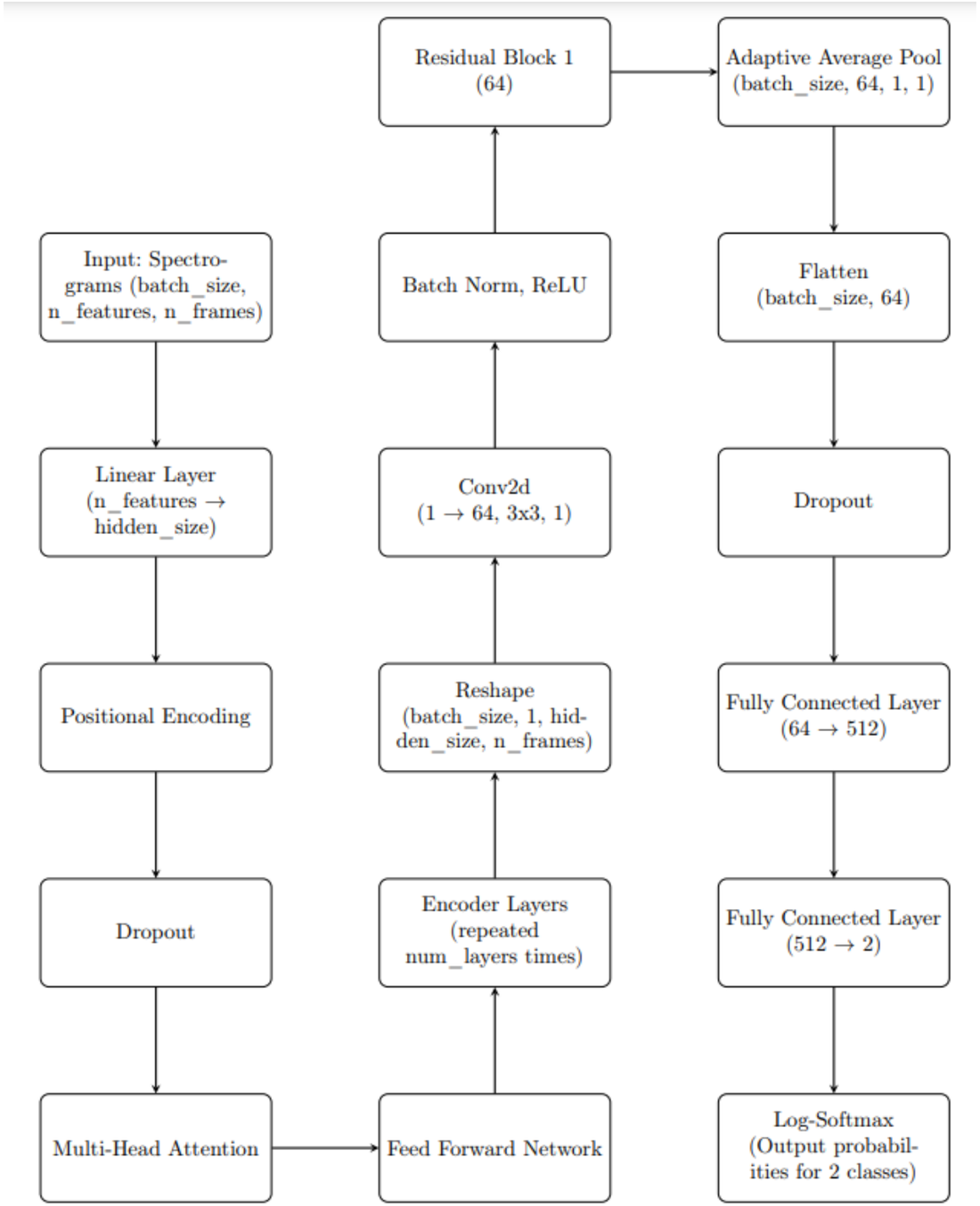




Skoltech

TE_ResNet

The TE-ResNet consists of two parts: transformer encoder and residual network. The transformer encoder is used to pre-process the acoustic features matrix to get the deep feature maps.



Skoltech

Results for TE-ResNet

Training				
Feature Extractor	Loss	Balanced Accuracy	Precision	Recall
MFCC	0.331	0.514	0.898	0.7594

Evaluation on ASVspoof19				
Feature Extractor	Loss	Balanced Accuracy	Precision	Recall
MFCC	0.333	0.511	0.897	0.7587

Evaluation on ASVspoof21				
Feature Extractor	Loss	Balanced Accuracy	Precision	Recall
MFCC	0.183	0.626	0.963	1.000

Model Generability Check on Singing Voice Deepfake Detection Challenge

Attack Type	EER(%)
Pooled Attack	49.36
Attack A14	43.02
Attack A13	49.06
Attack A12	51.53
Attack A11	51.34

Conclusion

In the current work we have achieved the following results:

- **Successfully** load and work with ASVspoof dataset of audio files with high memory capacity (~40 Gb).
- **Successfully** implement audio preprocessing and feature extraction pipeline.
- **Successfully** calculate metrics, such as balanced accuracy, precision, recall, ROC/AUC, EER.
- **Successfully** train and evaluate ResNet18 and get good results (~15 % EER on ASVspoof19 and ~25% EER on ASVspoof21).
- Design TE-ResNet, train and evaluate and get results that **needed improvement.**

Project Team

Inna Larina

Folu Obidare

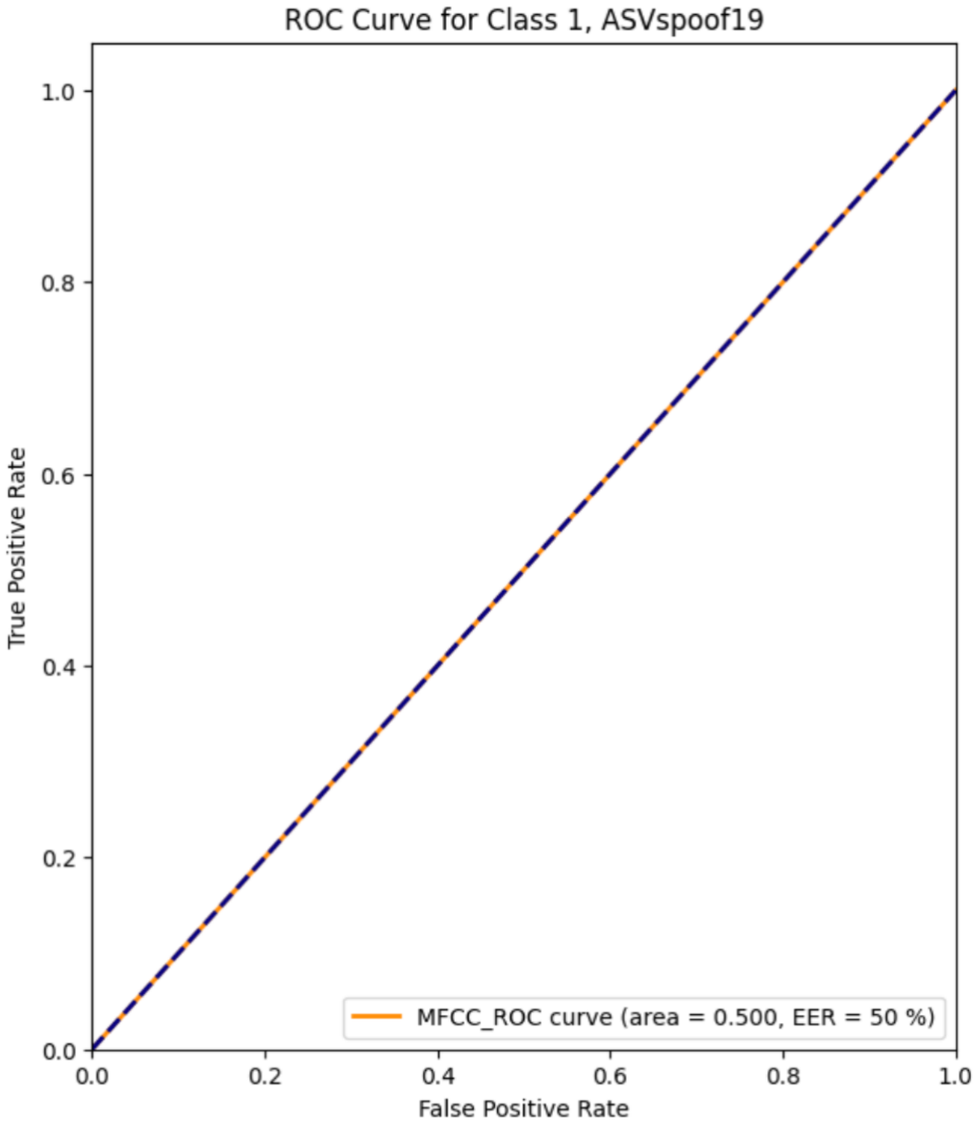
Maksim Komiakov

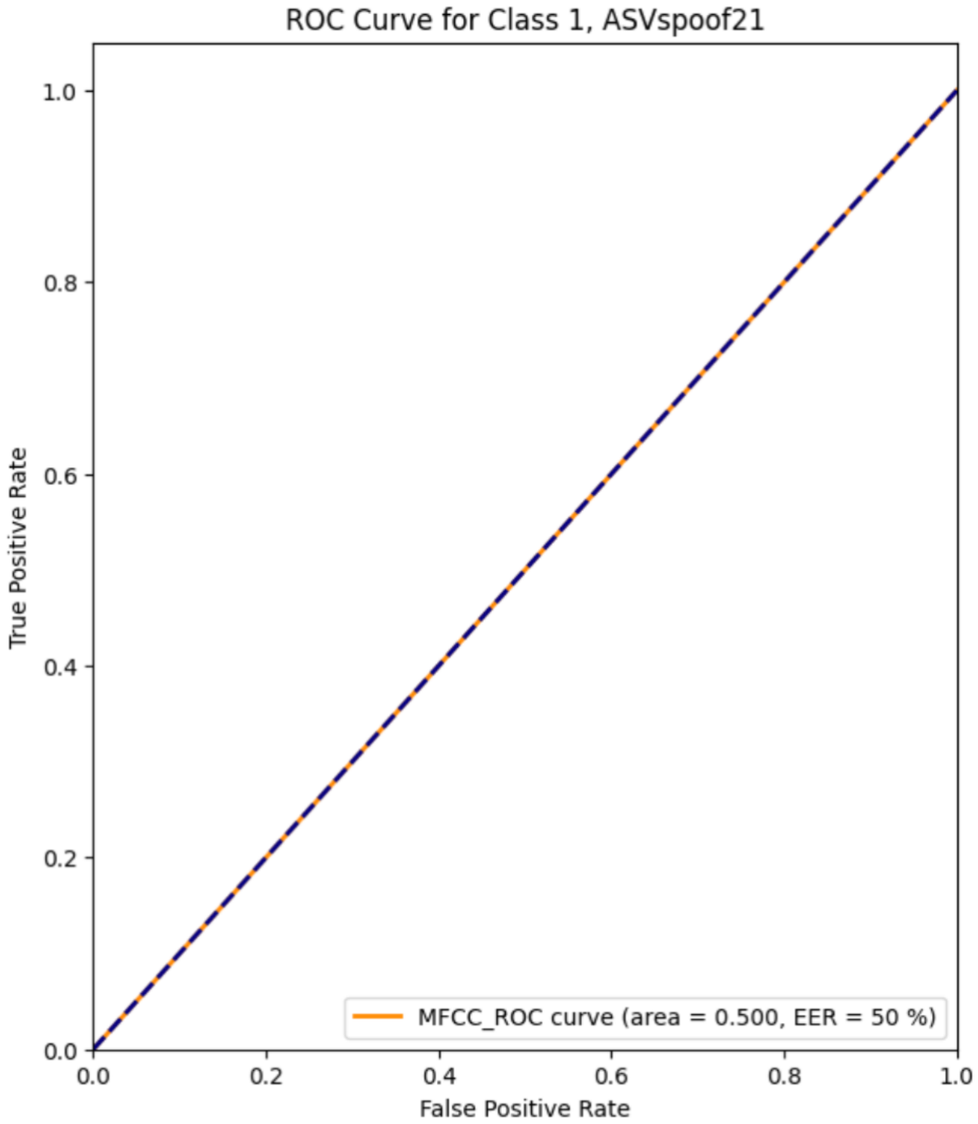
Ilona Basset

Hernán Nenjer

Skoltech

ROC/AUC TE-ResNet





Skoltech