UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Audio Deepfake Detection

Digital Forensics and Biometrics

Ferrari Luca (id. 2166294)
Scalco Riccardo (id. 2155352)

UNIVERSITÀ DEGLI STUDI DI PADOVA

Project overview

objective

Implementation of a model that recognizes and separates bonafide audio from spoof audio

applications

voice authentication, fake news diffusions, digital forensics, voice cloning, plagiarism

Università degli Studi di Padova

Environment

software

Code Editor:

- Kaggle Notebook

Programming Language

- Python

key libraries

- **matplotlib** (used to create plots)
- **torch** (used for neural networks, tensors, and GPU computations)
- **torchaudio** (PyTorch extension for working with audio: loading files, transformations, feature extraction)
- **sklearn** (provides tools for classic machine learning)

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dataset

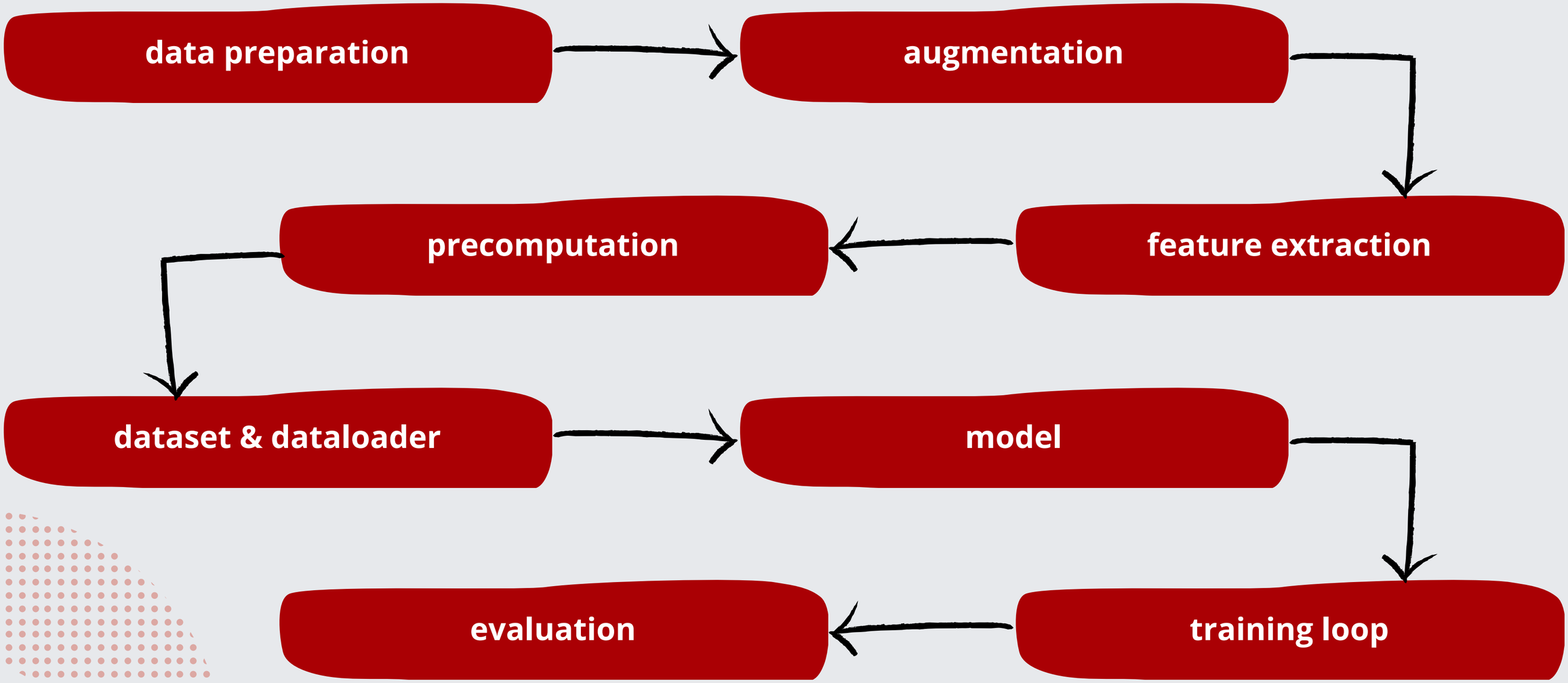
ASVspoof2019

- LA (Logical Access)
- PA (Pysichal Access)

./ASVspoof2019_root
- --> LA
  - --> ASVspoof2019_LA_asv_protocols
  - --> ASVspoof2019_LA_asv_scores
  - --> ASVspoof2019_LA_cm_protocols
    - --> ASVspoof2019_LA_dev
    - --> ASVspoof2019_LA_eval
  - --> ASVspoof2019_LA_train
- --> PA
  - --> ASVspoof2019_PA_asv_protocols
  - --> ASVspoof2019_PA_asv_scores
  - --> ASVspoof2019_PA_cm_protocols
    - --> ASVspoof2019_PA_dev
    - --> ASVspoof2019_PA_eval
  - --> ASVspoof2019_PA_train

Università degli Studi di Padova

Project Pipeline



UNIVERSITÀ DEGLI STUDI DI PADOVA

Project pipeline: PRO & CONS

- 👍 Just one time mel computation
- 👍 Targeted augmentation → improves robustness and reduces overfitting.
- 👍 On-the-fly integration → efficient variability without huge datasets.
- 👍 Cross-condition generalization → resistant to different spoofing attacks and environments.
- 👎 Higher training time → more augmentation mean longer epochs
- 👎 CPU/GPU overhead → some augmentations increase preprocessing load

UNIVERSITÀ DEGLI STUDI DI PADOVA

Metadata - LA

	speaker_id	audio_file_name	environment_id	attack_id	label	path	tag
0	LA_0079	LA_T_1138215	-	-	0	/kaggle/input/asvpoof-2019-dataset/LA/LA/ASVspoof2019_LA_train/flac/LA_T_1138215.flac	LA
1	LA_0079	LA_T_1271820	-	-	0	/kaggle/input/asvpoof-2019-dataset/LA/LA/ASVspoof2019_LA_train/flac/LA_T_1271820.flac	LA
2	LA_0079	LA_T_1272637	-	-	0	/kaggle/input/asvpoof-2019-dataset/LA/LA/ASVspoof2019_LA_train/flac/LA_T_1272637.flac	LA
3	LA_0079	LA_T_1276960	-	-	0	/kaggle/input/asvpoof-2019-dataset/LA/LA/ASVspoof2019_LA_train/flac/LA_T_1276960.flac	LA
4	LA_0079	LA_T_1341447	-	-	0	/kaggle/input/asvpoof-2019-dataset/LA/LA/ASVspoof2019_LA_train/flac/LA_T_1341447.flac	LA

- SPEAKER_ID: LA_****, a 4-digit speaker id
- AUDIO_FILE_NAME: LA_****, name of the file audio
- ENVIRONMENT_ID: not used for LA
- ATTACK_ID: ID of the speech spoofing system (A01 - A19), or, for bonafide speech SYSTEM-ID is left blank ('-')
- LABEL: 'bonafide' (0) for genuine speech, or, 'spoof' (1) for spoofing speech

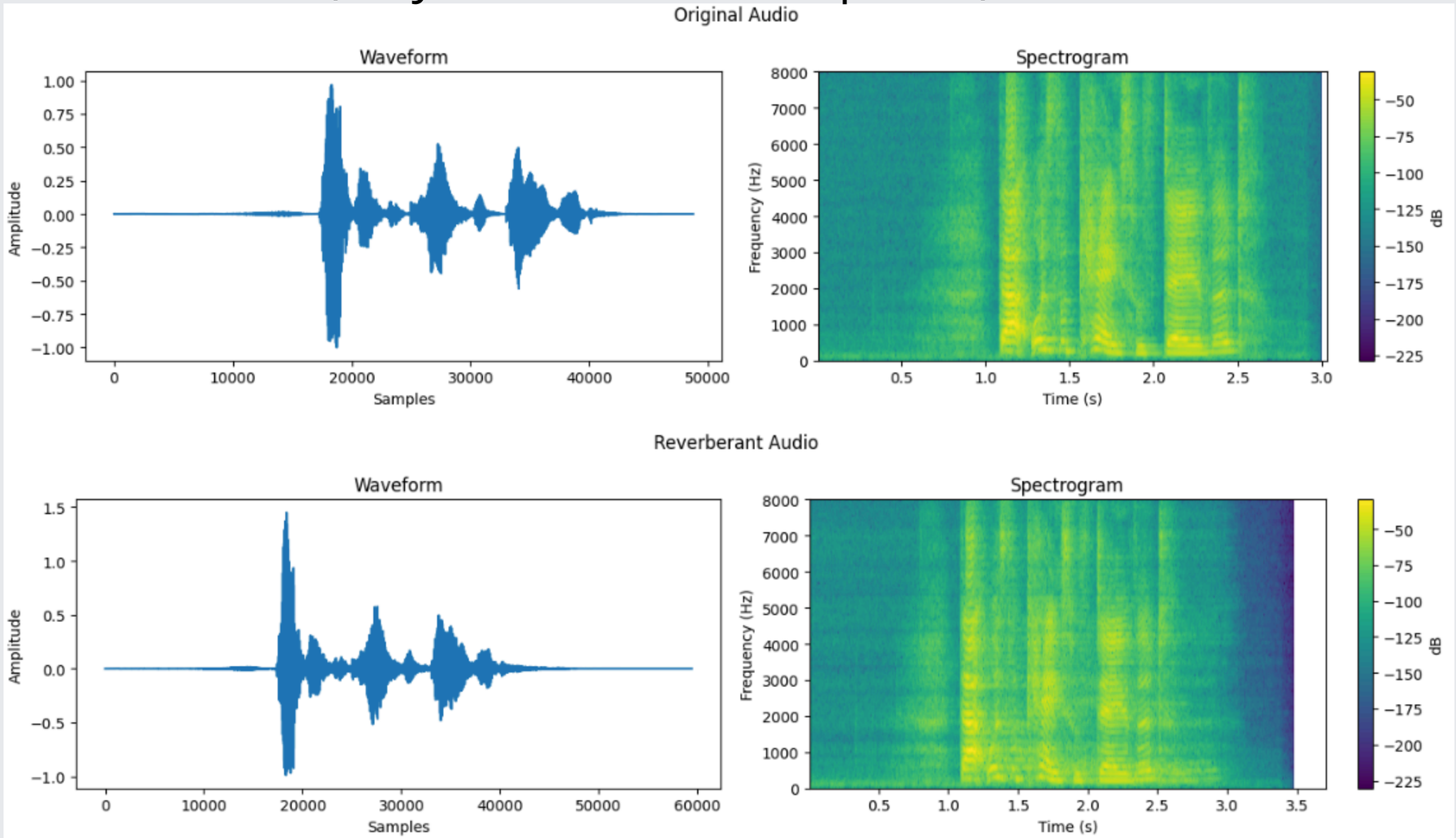
Metadata - PA

	speaker_id	audio_file_name	environment_id	attack_id	label	path	tag
0	PA_0079	PA_T_0000001	aaa	-	0	/kaggle/input/asvpoof-2019-dataset/PA/PA/ASVspoof2019_PA_train/flac/PA_T_0000001.flac	PA
1	PA_0079	PA_T_0000002	aaa	-	0	/kaggle/input/asvpoof-2019-dataset/PA/PA/ASVspoof2019_PA_train/flac/PA_T_0000002.flac	PA
2	PA_0079	PA_T_0000003	aaa	-	0	/kaggle/input/asvpoof-2019-dataset/PA/PA/ASVspoof2019_PA_train/flac/PA_T_0000003.flac	PA
3	PA_0079	PA_T_0000004	aaa	-	0	/kaggle/input/asvpoof-2019-dataset/PA/PA/ASVspoof2019_PA_train/flac/PA_T_0000004.flac	PA
4	PA_0079	PA_T_0000005	aaa	-	0	/kaggle/input/asvpoof-2019-dataset/PA/PA/ASVspoof2019_PA_train/flac/PA_T_0000005.flac	PA

- SPEAKER_ID: PA_****, a 4-digit speaker id
- AUDIO_FILE_NAME: PA_****, name of the file audio
- ENVIRONMENT_ID: a triplet (S,R,D_s), which take one letter in the set {a,b,c} as categorical value
- ATTACK_ID: a duple (D_a,Q), which take one letter in the set {A,B,C} as categorical value
- LABEL: 'bonafide' (0) for genuine speech, or, 'spoof' (1) for spoofing speech

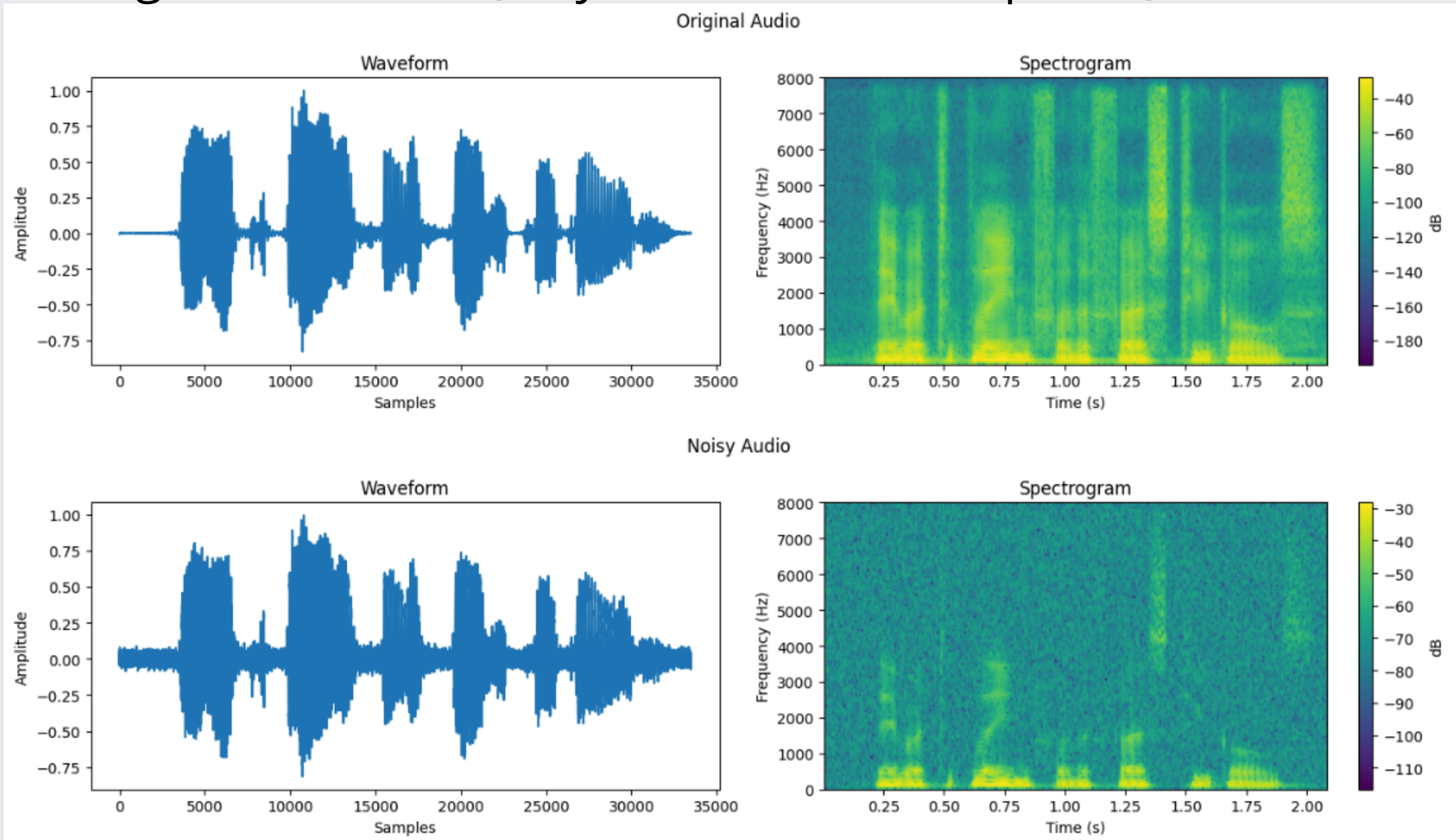
Augmentation

- reverberation (only for LA and add wp <0.3) → NOT IMPLEMENTED



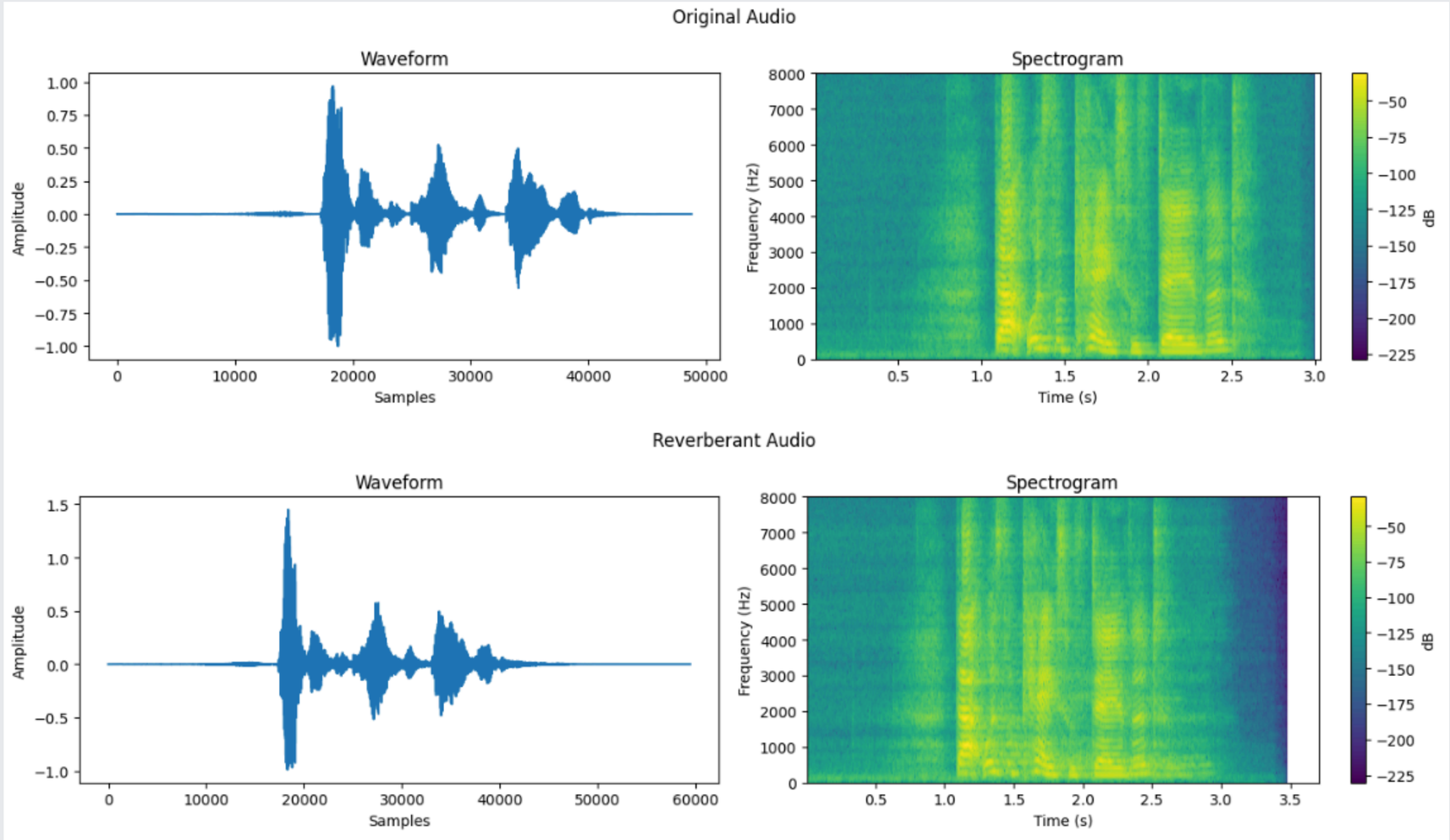
Augmentation

- white gaussian noise (only for LA and add wp <0.2)



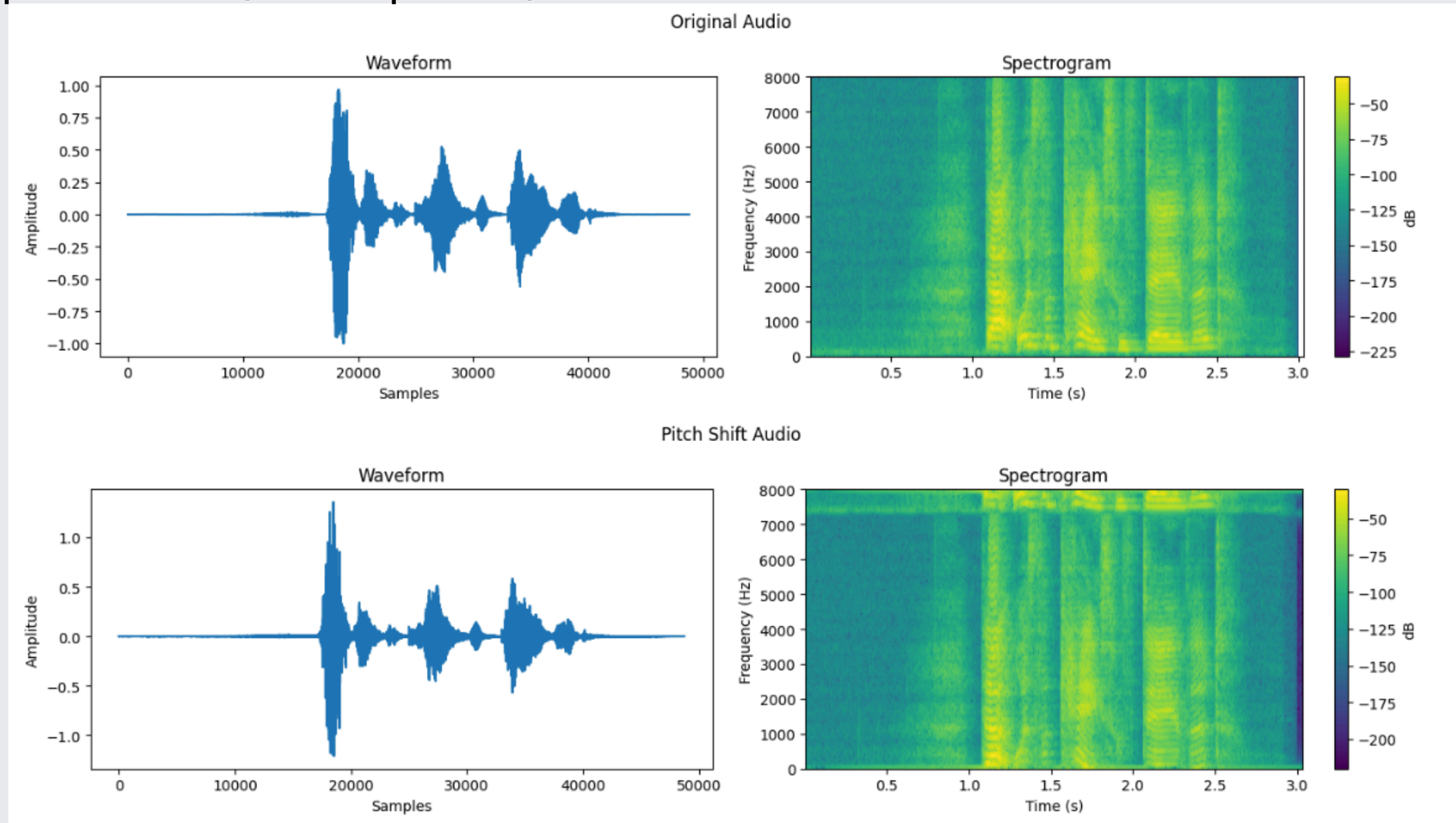
Augmentation

- time-strech (add wp <0.2)



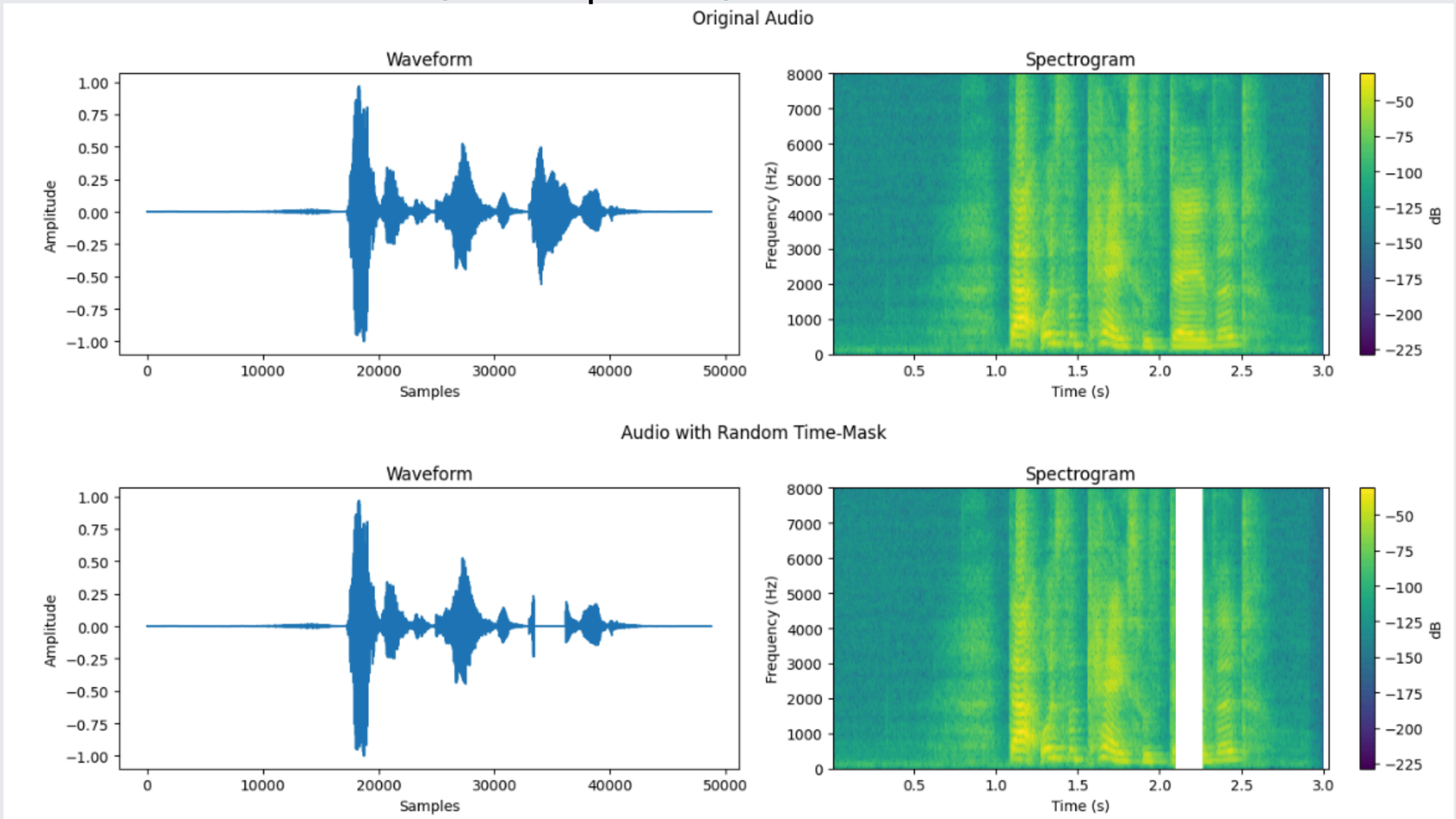
Augmentation

- pitch shift (add wp <0.2)



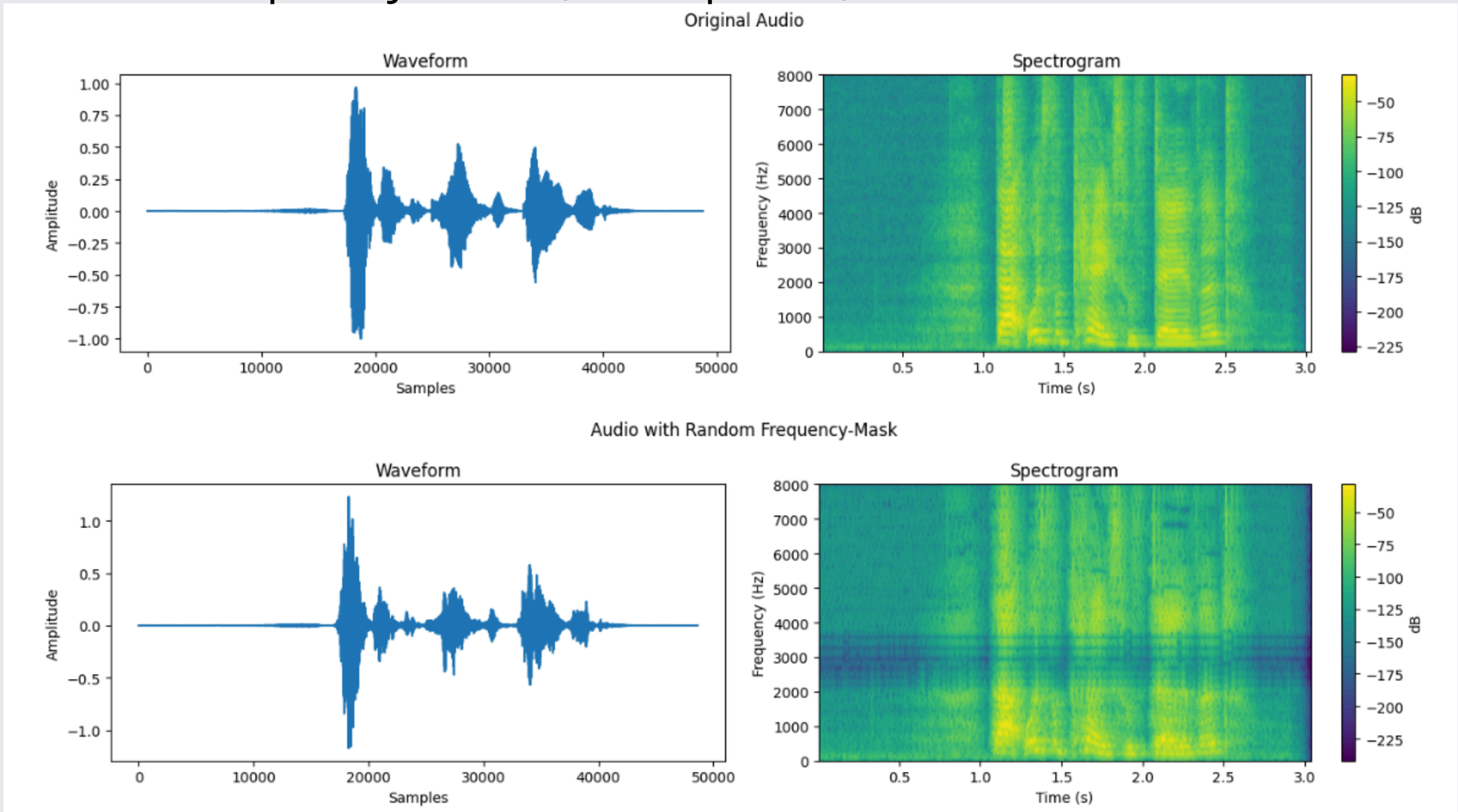
Augmentation

- random time mask (add wp <0.2)



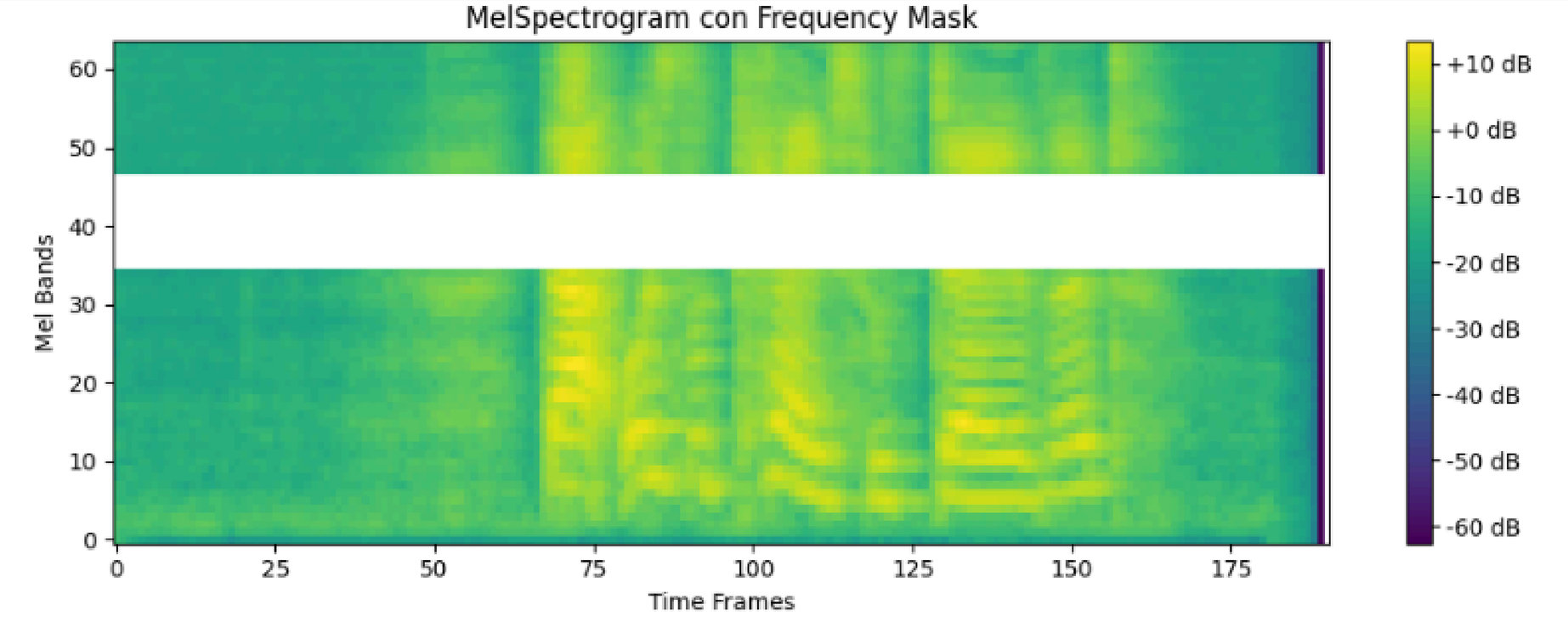
Augmentation

- random frequency mask (add wp <0.2)

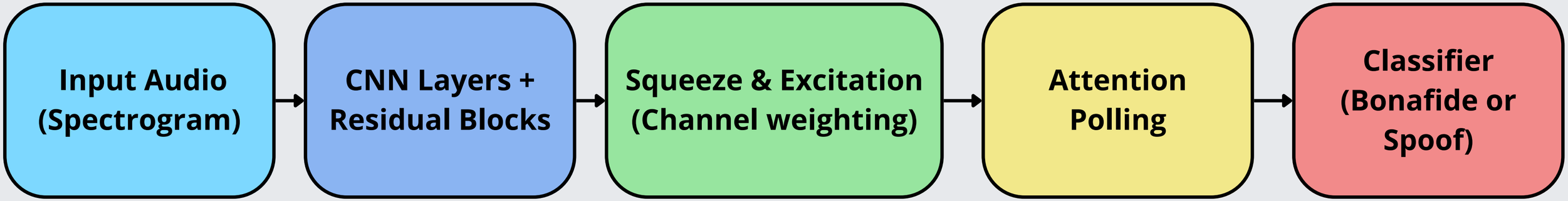


Augmentation

- random frequency mask (add wp <0.2)



Model



- Input (Spectrogram) → Audio as time–frequency image.
- CNN + Residual → Extract local patterns, stable training.
- Squeeze & Excitation → Emphasize key frequency bands.
- Attention Pooling → Focus on most relevant regions.
- Classifier (MLP) → Decision: Bonafide or Spoof.

Metrics

- Accuracy $$\frac{TP + TN}{TP + TN + FP + FN}$$
- Precision $$\frac{TP}{TP + FP}$$
- Recall $$\frac{TP}{TP + FN}$$
- F1 score $$\frac{2 \times Precision \times Recall}{Precision + Recall}$$
- Equal Error Rate (EER) EER = FPR where FPR = FNR

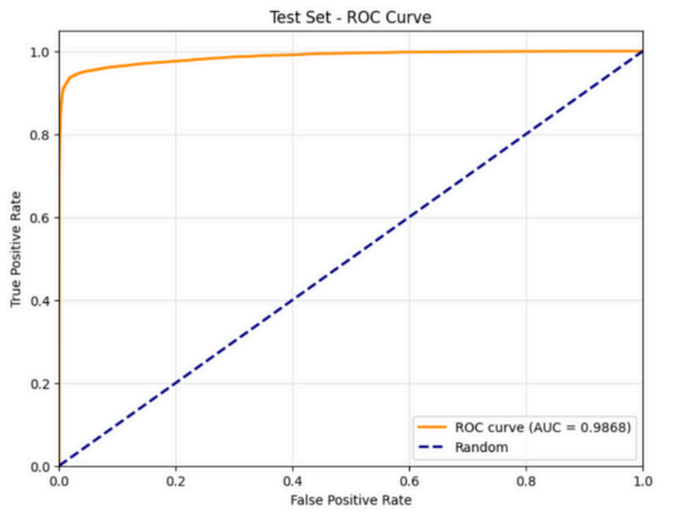
UNIVERSITÀ DEGLI STUDI DI PADOVA

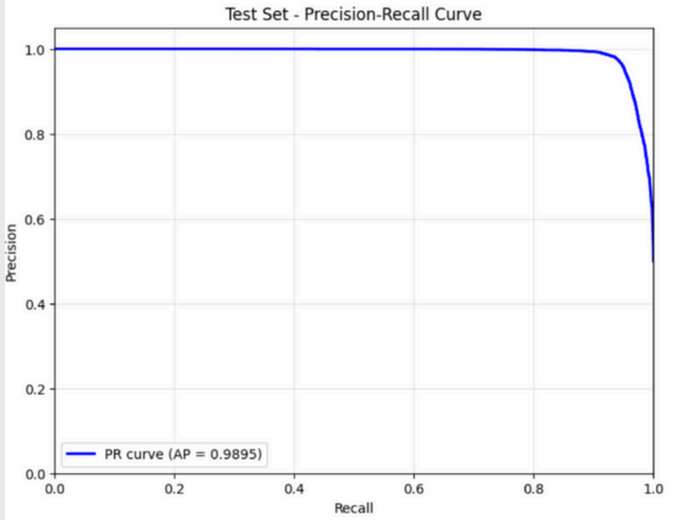
Results

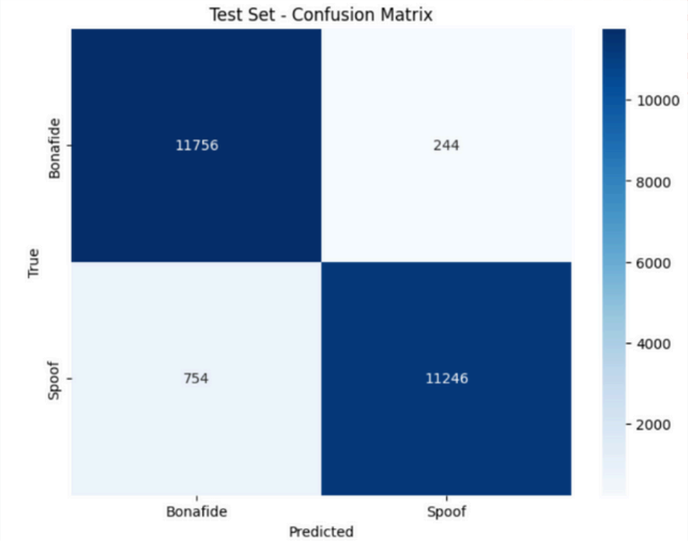
The model reaches the following results:
Epoch 24/50 | Train Loss: 0.3004 | Val Loss: 0.0508 | Acc: 0.9824 |
Prec: 0.9901 | Rec: 0.9745 | F1: 0.9822

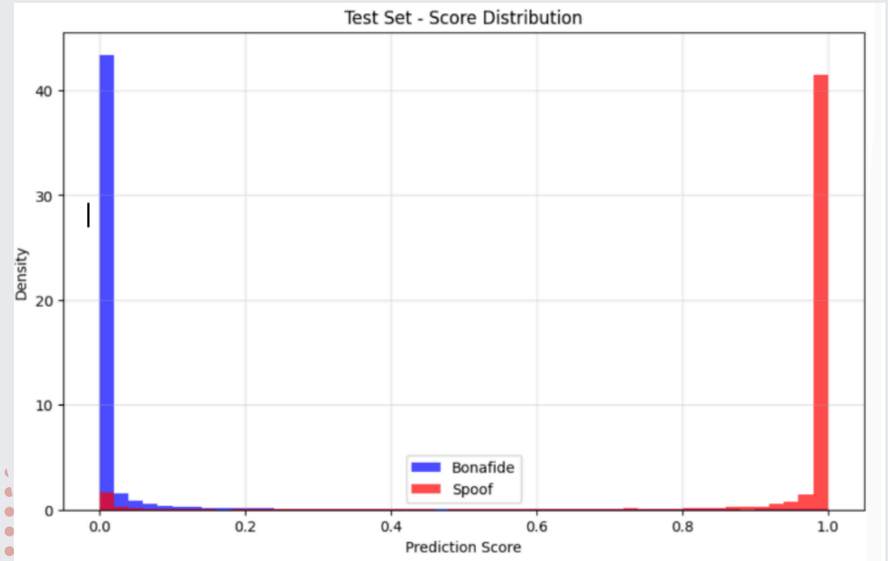
Then we use the test set and we obtaining:
Loss: 0.1771 | Accuracy: 0.9523 (95.23%) | Precision: 0.9832
Recall: 0.9203 | F1-Score: 0.9507 | AUC-ROC: 0.9837
EER: 4.62% | EER Threshold: 0.1437

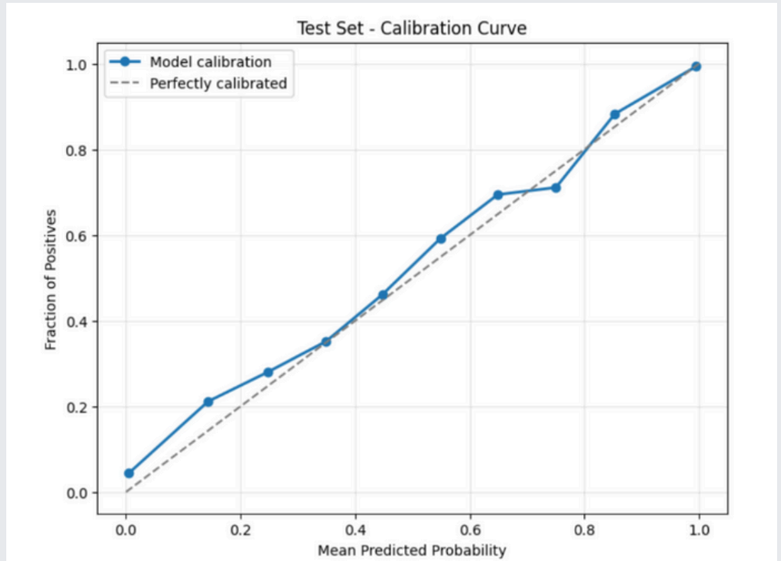
Results


Test Set - ROC Curve
True Positive Rate
False Positive Rate
ROC curve (AUC = 0.9868)
Random


Test Set - Precision-Recall Curve
Precision
Recall
PR curve (AP = 0.9895)


Test Set - Confusion Matrix
Bonafide
Spoof
True
Predicted
11756
244
754
11246


Test Set - Score Distribution
Density
Prediction Score
Bonafide
Spoof


Test Set - Calibration Curve
Model calibration
Perfectly calibrated
Fraction of Positives
Mean Predicted Probability