Detecting the Undetectable

Robust Defense Strategies Against Audio Deepfakes

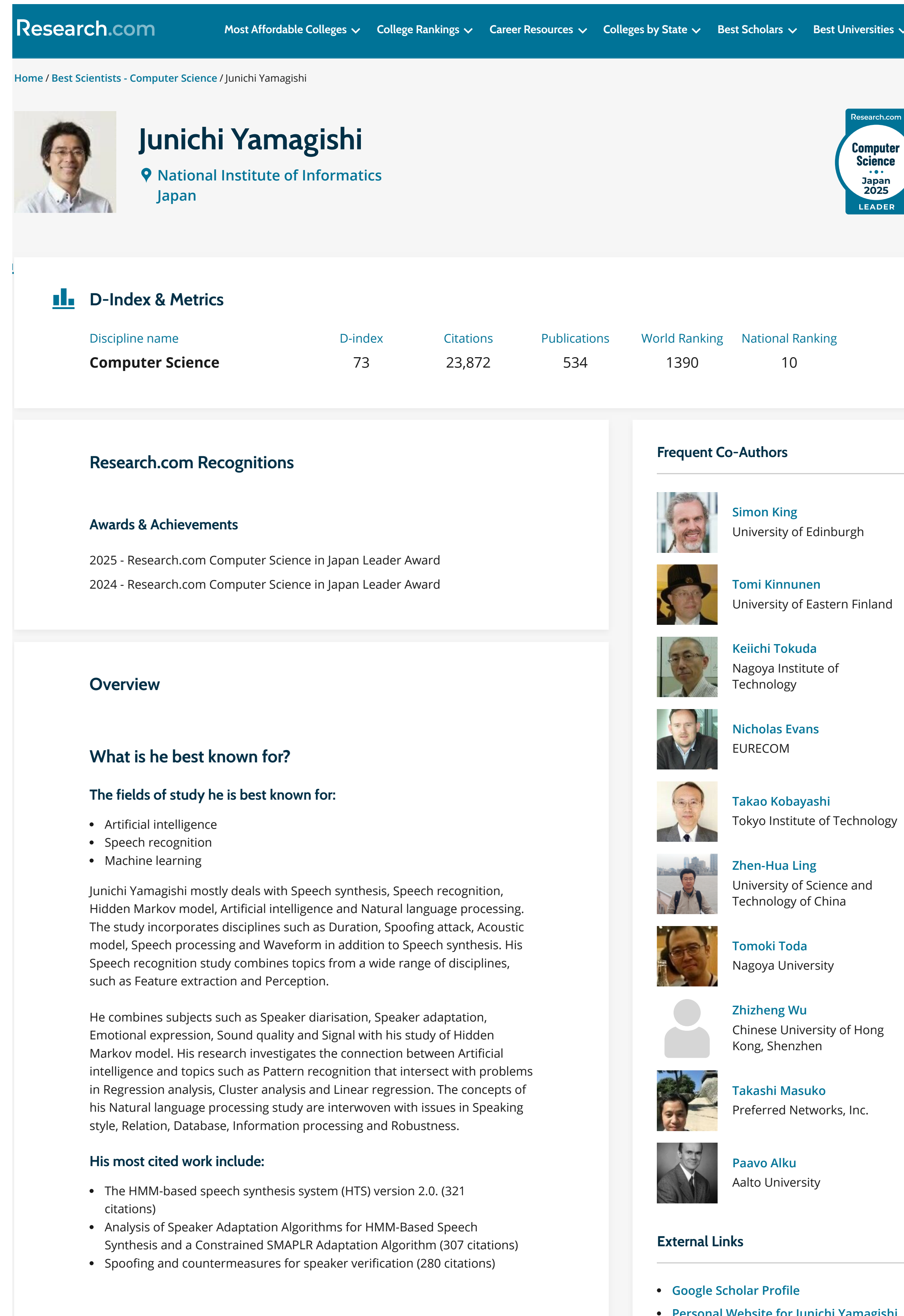
Prof. Junichi Yamagishi
National Institute of Informatics, Japan

Self introduction: Junichi Yamagishi

Major projects I worked on include:

- 2008-2011: **EMIME:** Speech translation using our own voice
- 2010-2013: **Listening Talker:** Improving the intelligibility of voice in noise
- 2011-2016: **VoiceBank:** Digital voice cloning technology for individuals with impaired speech
- 2012: **VCTK:** Voice Cloning Toolkit
- 2018: **MesoNet** for facial deepfake detection
- 2013-current: **ASVspoof:** audio anti-spoofing
- 2018-2023: **VoicePersonae**: Voice Protection and Privacy

The Yamagishi Lab at NII (yamagishilab.jp)



Research.com | Most Affordable Colleges | College Rankings | Career Resources | Colleges by State | Best Scholars | Best Universities

Home / Best Scientists - Computer Science / Junichi Yamagishi

Junichi Yamagishi

National Institute of Informatics
Japan

Research.com Computer Science Japan 2025 LEADER

D-Index & Metrics

Discipline name	D-index	Citations	Publications	World Ranking	National Ranking
Computer Science	73	23,872	534	1390	10

Research.com Recognitions

Awards & Achievements

2025 - Research.com Computer Science in Japan Leader Award

2024 - Research.com Computer Science in Japan Leader Award

Overview

What is he best known for?

The fields of study he is best known for:

- Artificial intelligence
- Speech recognition
- Machine learning

Junichi Yamagishi mostly deals with Speech synthesis, Speech recognition, Hidden Markov model, Artificial intelligence and Natural language processing. The study incorporates disciplines such as Duration, Spoofing attack, Acoustic model, Speech processing and Waveform in addition to Speech synthesis. His Speech recognition study combines topics from a wide range of disciplines, such as Feature extraction and Perception.

He combines subjects such as Speaker diarisation, Speaker adaptation, Emotional expression, Sound quality and Signal with his study of Hidden Markov model. His research investigates the connection between Artificial intelligence and topics such as Pattern recognition that intersect with problems in Regression analysis, Cluster analysis and Linear regression. The concepts of his Natural language processing study are interwoven with issues in Speaking style, Relation, Database, Information processing and Robustness.

His most cited work include:

- The HMM-based speech synthesis system (HTS) version 2.0. (321 citations)
- Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm (307 citations)
- Spoofing and countermeasures for speaker verification (280 citations)

Frequent Co-Authors

- Simon King — University of Edinburgh
- Tomi Kinnunen — University of Eastern Finland
- Keiichi Tokuda — Nagoya Institute of Technology
- Nicholas Evans — EURECOM
- Takao Kobayashi — Tokyo Institute of Technology
- Zhen-Hua Ling — University of Science and Technology of China
- Tomoki Toda — Nagoya University
- Zhizheng Wu — Chinese University of Hong Kong, Shenzhen
- Takashi Masuko — Preferred Networks, Inc.
- Paavo Alku — Aalto University

External Links

- Google Scholar Profile
- Personal Website for Junichi Yamagishi

Brand-new papers to be introduced in this talk

(1) Wanying Ge, Xin Wang, Xuechen Liu, Junichi Yamagishi, “**Post-training for Deepfake Speech Detection**” IEEE ASRU 2025

(2) Xuechen Liu, Xin Wang, Junichi Yamagishi, “**Frustratingly Easy Zero-Day Audio DeepFake Detection via Retrieval Augmentation** and Profile Matching,” Submitted to IEEE ICASSP 2026

(3) Xin Wang, Wanying Ge, Junichi Yamagishi, “**Towards Data Drift Monitoring for Speech Deepfake Detection in the context of MLOps**” Submitted to IEEE ICASSP 2026

(4) Yoshihiko Furuhashi, Xin Wang, Junichi Yamagishi, Huy Nguyen, Isao Echizen, “**Exploring Active Data Selection Strategies for Continuous Training in Deepfake Detection**” 23rd International Conference of the Biometrics Special Interest Group 2024

(5) Wanying Ge, Xin Wang, Junichi Yamagishi, “**FakeMark: Deepfake Speech Attribution With Watermarked Artifacts**” Arxiv 2025

Agenda of the talk

- **Background:**
  Why is deepfake detection such a challenging task?
- **Part 1:**
  Robust detection of unknown deepfake audio generation methods
- **Part 2:**
  Machine Learning Operations (MLOPs) of deepfake detection
- **Part 3:**
  Collective approach to passive and proactive deepfake defense

Background:

Why is deepfake detection such a challenging task?

Two types of deepfake detectors

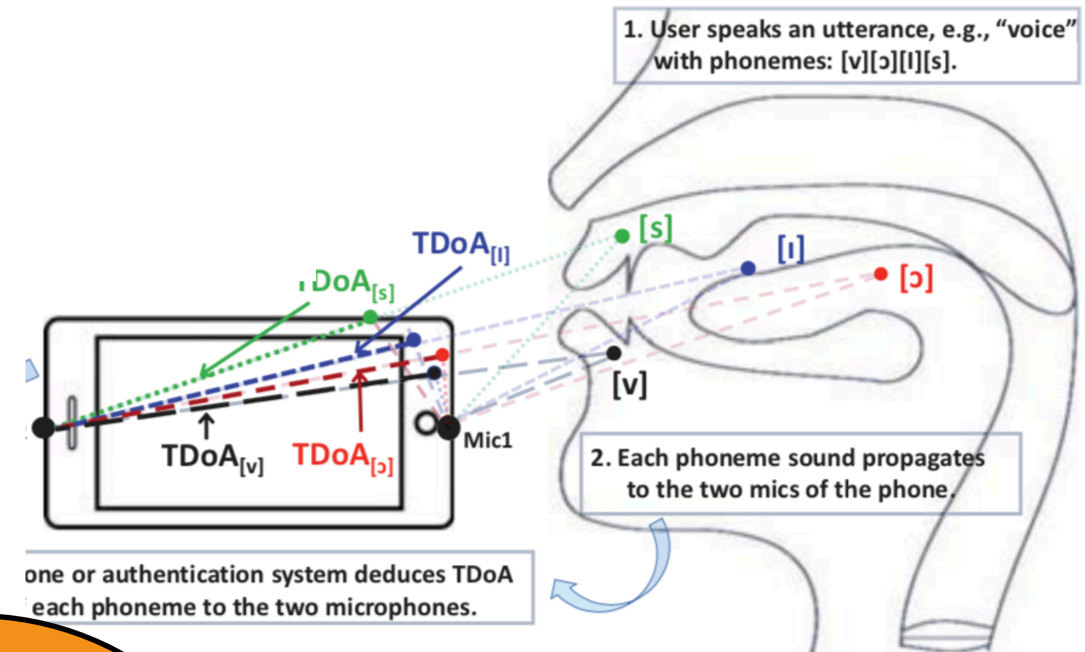
- What is the detector learning?

ASVspoof
Automatic Speaker Verification and Spoofing Countermeasures Challenge

ADD
Audio Deep Synthesis Detection

Artifacts
(speech community)

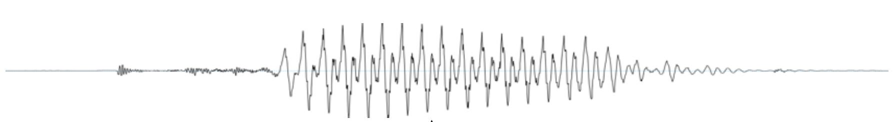
Liveness evidence
(Security community)



Linghan Zhang, Sheng Tan, Jie Yang, Yingying Chen, VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones 23rd ACM Conference on Computer and Communications Security (CCS 2016) Vienna, Austria, October 2016

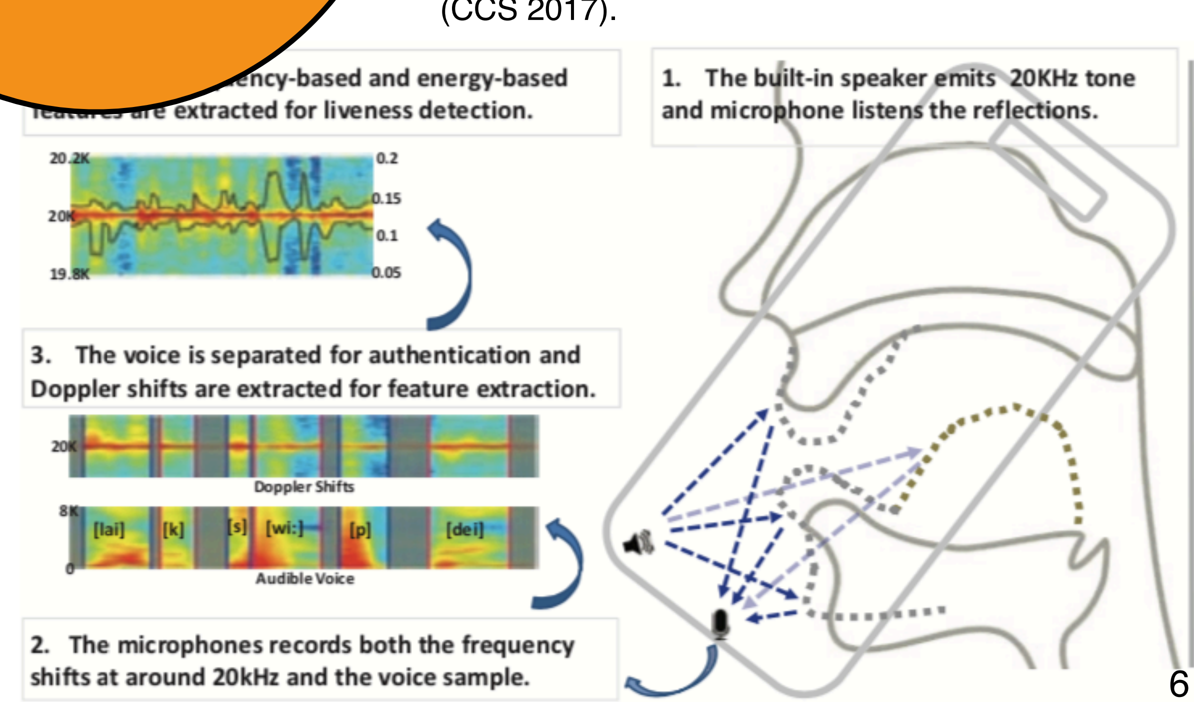
Linghan Zhang, Sheng Tan, Jie Yang. "Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication". 24th ACM Conference on Computer and Communication Security (CCS 2017).

Artifacts: **audible or non-audible differences between real and generated waveforms**

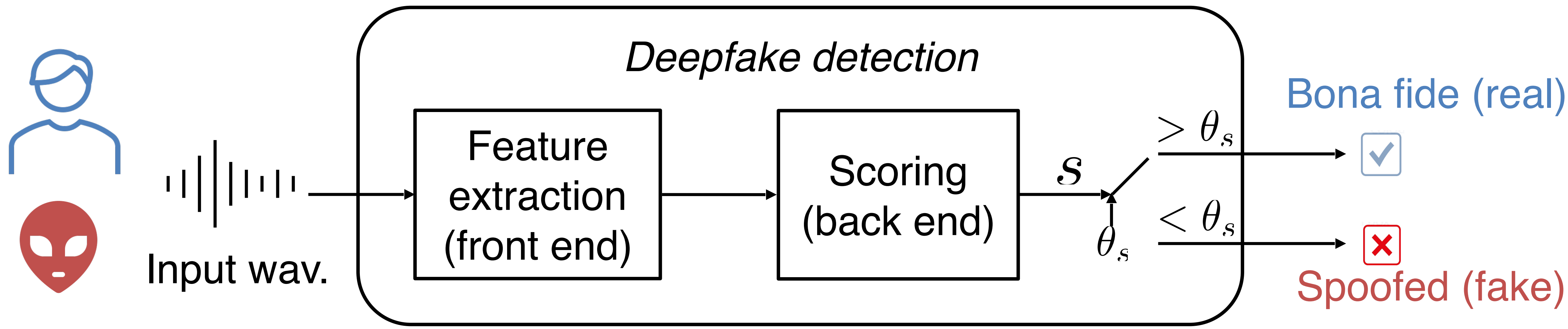


Approximation

Neural vocoder models

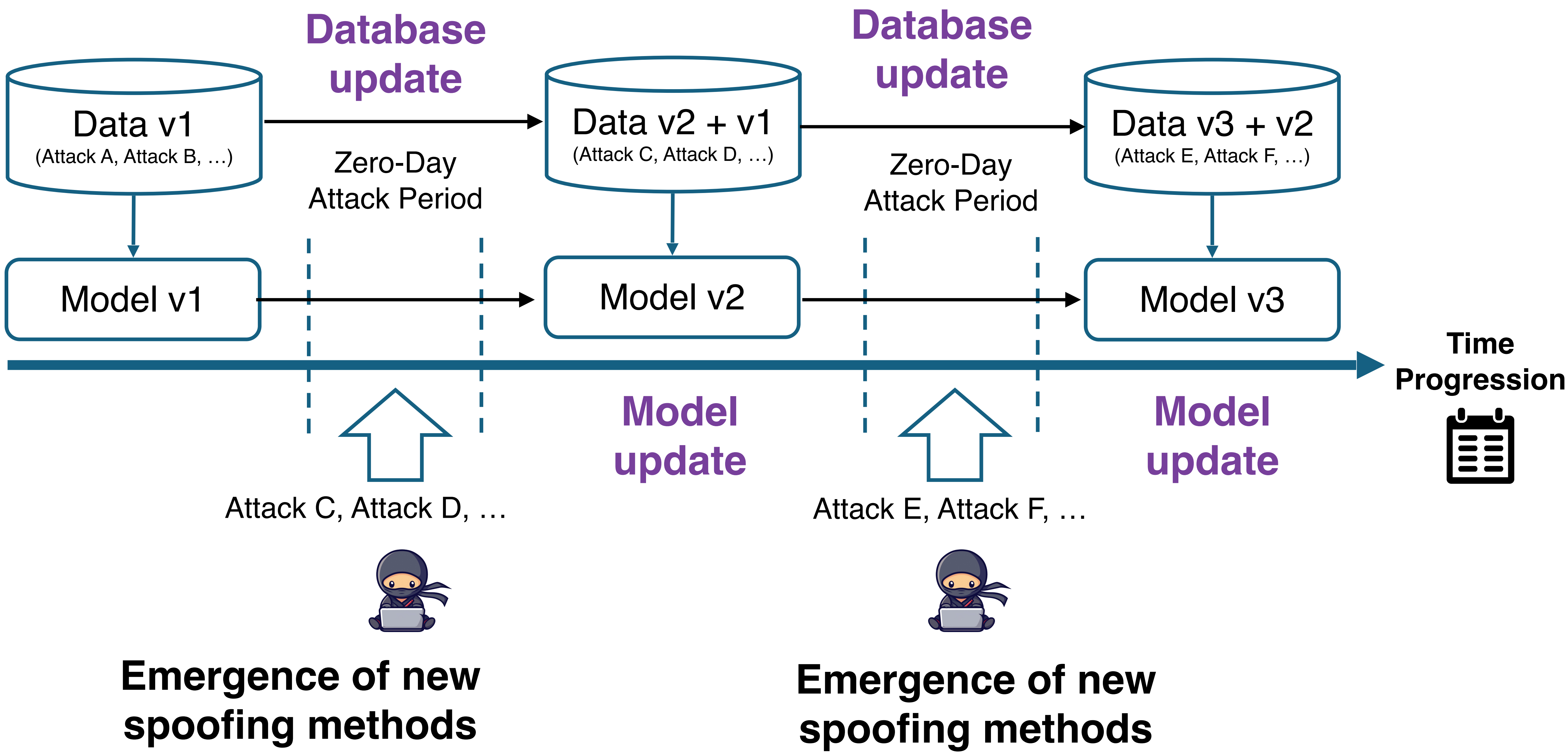


Deepfake detection isn't an ordinal binary classification task



- **Challenges of Deepfake Detection**
  - New spoofing methods and their distinct unseen artifacts in spoofed data make deepfake detection challenging
- **Domain Shift in Test Data**
  - Substantial domain shifts caused by the new spoofing methods necessitate robust generalization to handle unseen methods

Importance of database and model updates

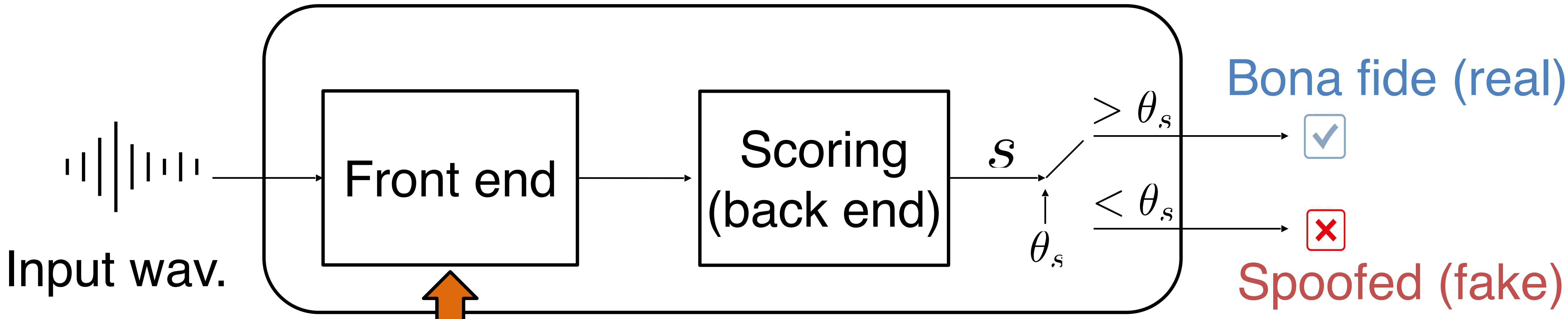


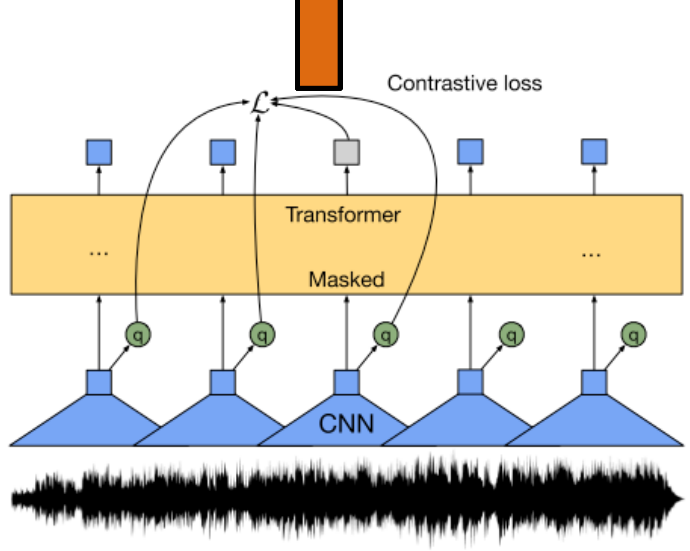
High update frequency raises costs, while low update frequency extends the zero-day attack period

Part 1:

Robust detection of unknown deepfake audio generation methods

Basic structure of deepfake audio detectors





Pre-trained speech SSL models:
- wav2vec 2.0 model
- HuBERT
- AudioMAE

- Robustness can be improved by introducing self-supervised learning (**SSL) models**—such as wav2vec 2.0 or HuBERT—that are pre-trained on large amounts of natural speech waveforms as feature extraction models instead of using spectral features [1-3]

[1] Xin Wang, Junichi Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures" Speaker and Language Recognition Workshop (Odyssey 2022), June 2022
[2] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,"Speaker and Language Recognition Workshop (Odyssey 2022), June 2022
[3] Ruibo Fu, Xiaopeng Wang, Zhengqi Wen, Jianhua Tao, Yuankun Xie, Zhiyong Wang, Chunyu Qiang, Xuefei Liu, Cunhang Fan, Chenxing Li, Guanjun Li, "RPRA-ADD: Forgery Trace Enhancement-Driven Audio Deepfake Detection" 2025

Post-training for deepfake detection [4]

	Pre-training	Post-training [4]
Training criterion	Masked language style or masked auto-encoder style	**Discriminative objective to distinguish natural speech from other types of speech**
Data	Natural speech	Natural speech + various generated speech
Purpose	Feature extraction	Feature extraction

AntiDeepfake

- Various post-trained SSL models (HuBERT, wav2vec, MMS, and XLS-R) using 74,000 hours of speech that contains over 100 languages

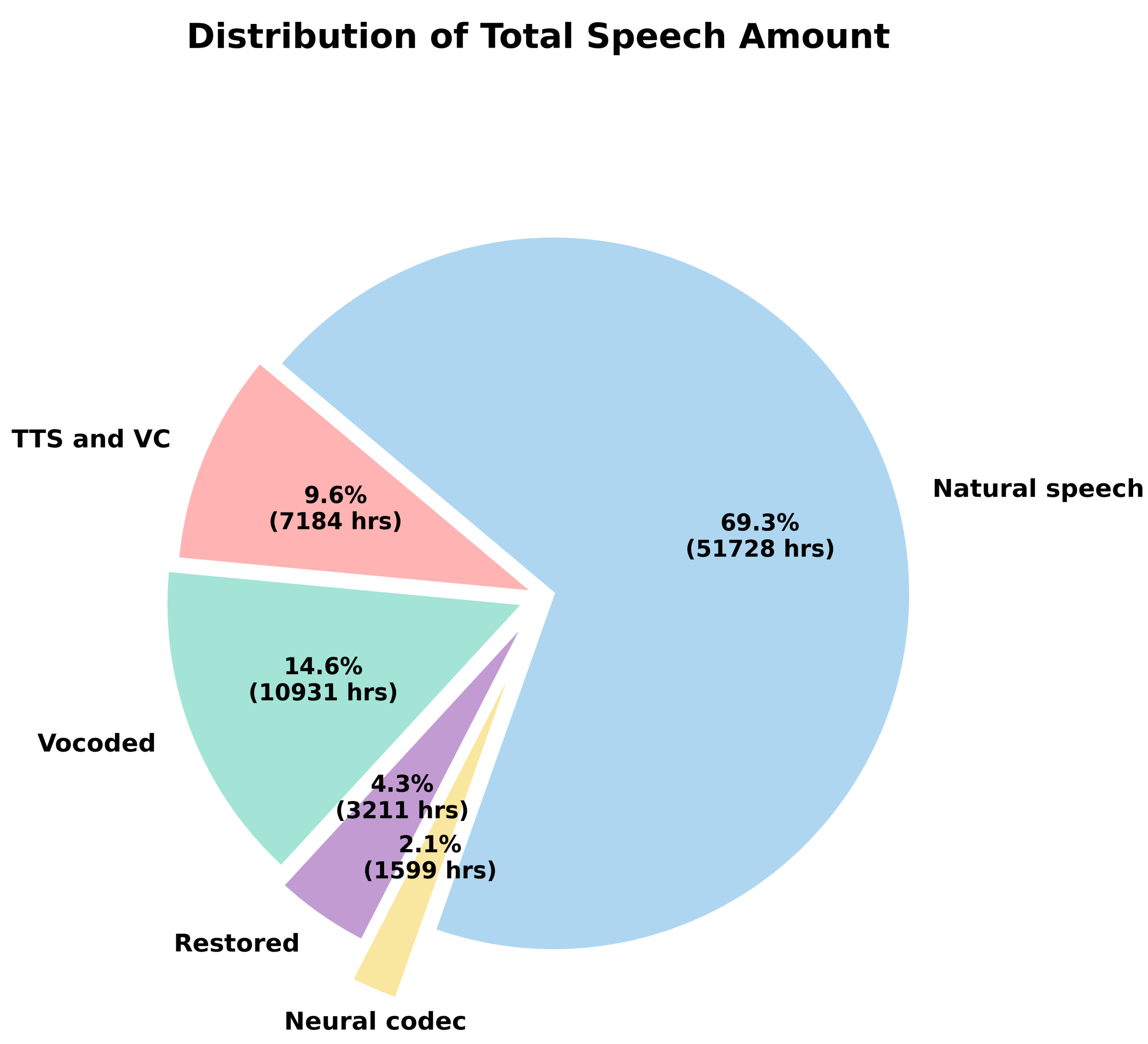
GitHub https://github.com/nii-yamagishilab/AntiDeepfake

https://huggingface.co/nii-yamagishilab

[4] Wanying Ge, Xin Wang, Xuechen Liu, Junichi Yamagishi, "Post-training for Deepfake Speech Detection," IEEE ASRU 2025, Dec. 2025

Large-scale multi-lingual post-training dataset

Dataset	Language	Genuine (hrs)	Fake (hrs)	Attack
TTS and VC				
ASVspoof2019-LA [23]	en	11.85	97.80	TTS, VC
ASVspoof2021-LA [24]	en	16.40	116.10	TTS, VC
ASVspoof2021-DF [24]	en	20.73	487.00	TTS, VC
ASVspoof5 [25]	en	413.49	1808.48	TTS, VC
CFAD [26]	zh	171.25	224.55	TTS
DECRO [27]	en, zh	35.18	102.44	TTS, VC
DFADD [28]	en	41.62	66.01	TTS
Diffuse or Confuse [29]	en	0	231.66	TTS
DiffSSD [30]	en	0	139.73	TTS
DSD [31]	en, ja, ko	100.98	60.23	TTS, VC
HABLA [32]	es	35.56	87.83	TTS, TTS-VC
MLAAD [33]	38 languages	0	377.96	TTS
SpoofCeleb [34]	Multilingual	173.00	1916.2	TTS
VoiceMOS [35]	en	0	448.44	TTS
Vocoded speech				
CVoiceFake [36]	en, fr, de, it, zh	315.14	1561.16	Vocoded
LibriTTS [37]	en	585.83	0	–
LibriTTS-Vocoded	en	0	2345.14	Vocoded
LJSpeech [38]	en	23.92	0	–
VoxCeleb2 [39]	Multilingual	1179.62	0	–
VoxCeleb2-Vocoded	Multilingual	0	4721.46	Vocoded
WaveFake [40]	en, ja	0	198.65	Vocoded
Restored speech				
FLEURS [41]	102 languages	1388.97	0	–
FLEURS-R [42]	102 languages	0	1238.83	Restored & vocoded
LibriTTS-R [43]	en	0	583.15	Restored & vocoded
Neural codec speech				
Codecfake [44]	en, zh	129.66	808.32	Neural codec
CodecFake [45]	en	0	660.92	Neural codec
Additional genuine speech				
AISHELL3 [46]	zh	85.62	0	–
CNCeleb2 [47]	zh	1084.34	0	–
MLS [48]	8 languages	50558.11	0	–
Train Set	Over 100 languages	56.37 k	18.28 k	–



Equal Error Rate results on various test sets **under zero-shot evaluation**

with RawBoost without RawBoost

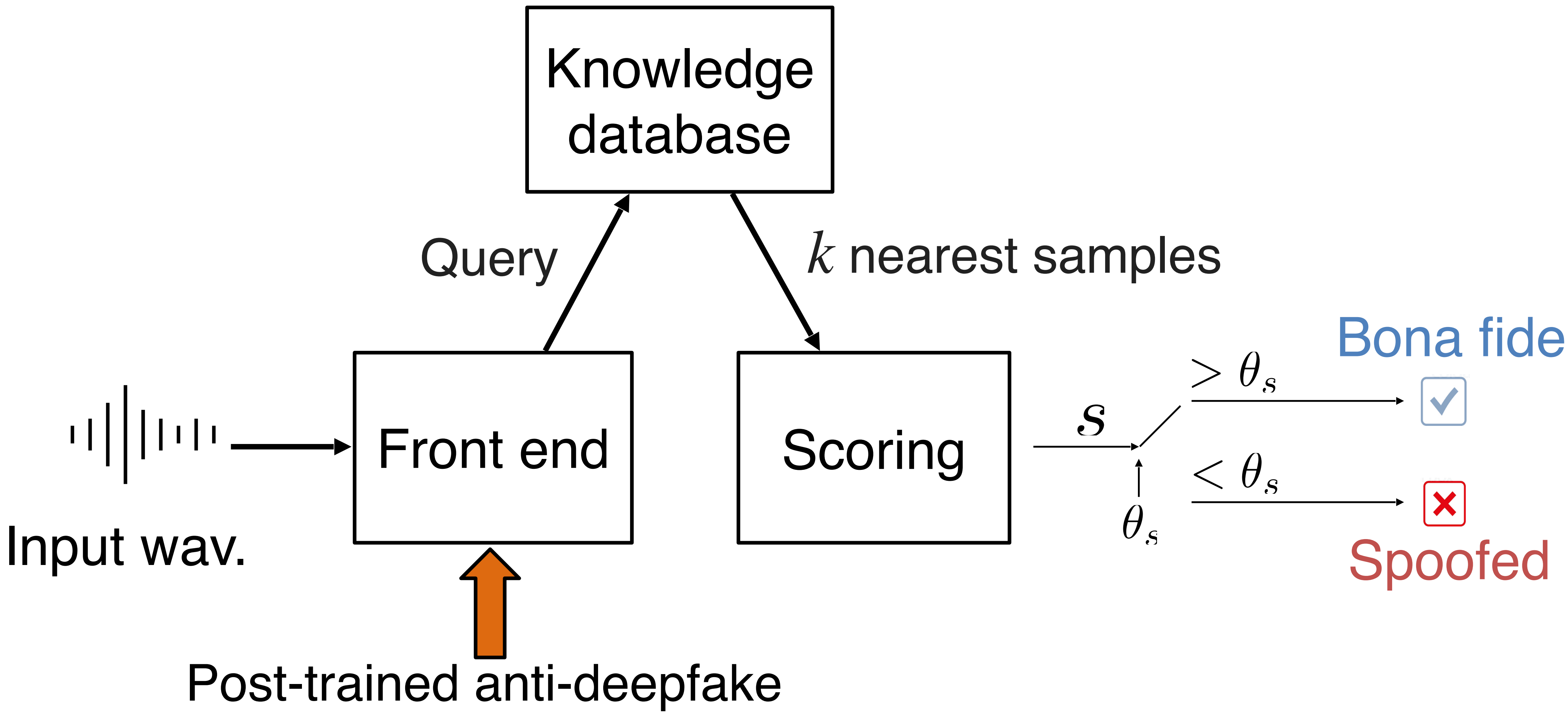
	Model ID	# of params.	ADD 2023	DEEP-VOICE	FakeOrReal		In-the-Wild
			Track-1.2-R2-Test	Segmented Full Set	original-Test	norm-Test	Full Set
Pre-training + Post-training (AntiDeepfake)	HuBERT-XL	964 M	18.90 / 35.34	5.67 / 14.87	2.49 / 3.67	3.17 / 15.52	5.23 / 17.99
	W2V-Small	95 M	13.02 / 19.41	9.80 / 16.22	21.94 / 1.05	17.85 / 6.47	4.24 / 4.65
	W2V-Large	317 M	13.25 / 12.67	4.53 / 5.01	0.63 / 0.80	**0.97** / 1.44	1.91 / 2.25
	MMS-300M	317 M	7.93 / 11.22	**2.27** / 3.04	1.35 / **0.46**	5.92 / 2.71	2.90 / 2.00
	MMS-1B	965 M	9.06 / 9.46	2.56 / **2.27**	1.22 / 0.89	1.73 / 1.10	1.82 / 1.86
	XLS-R-1B	965 M	5.39 / 6.58	2.52 / 2.96	5.74 / 3.16	12.14 / 10.91	1.35 / 1.36
	XLS-R-2B	2.2 B	**4.67** / 6.84	2.30 / 2.63	2.62 / 1.18	1.65 / 1.73	**1.23** / 1.31
Zero-shot evaluation results in the literature	XLSR-Mamba [52]	319 M	19.36	-	6.71	-	6.70
	Resemble AI [53]	2.1 B	6.11	-	1.36	-	3.94
	SpeechFake [2]	317 M	-	-	4.88	-	2.01
	Wav2Vec + VIB [31]	-	-	-	-	3.93	1.99
	UniSpeech-SAT [53], [54]	96 M	28.21	-	1.06	-	15.05
	XLS-R + SLS [55]	340 M	21.10	-	5.08	-	7.45
	XLSR-Conformer + TCM [56]	319 M	22.74	-	10.69	-	7.79
	AdaLAM & f-InfoED [57]	-	-	-	-	-	8.36
	P3 [20], [58]	317 M	-	-	-	-	-
	AASIST [20], [59]	0.3 M	32.47	-	21.64	-	43.01
	RawNet2 [20], [60]	18 M	64.55	-	65.68	-	49.19

The internal representations of the large post-trained SSL models (e.g. 1B, 2B) are effective for detecting audio generated using previously unseen generation methods

[4] Wanying Ge, Xin Wang, Xuechen Liu, Junichi Yamagishi, "Post-training for Deepfake Speech Detection," IEEE ASRU 2025, Dec. 2025

Further improvement using the retrieval of the knowledge source [5]

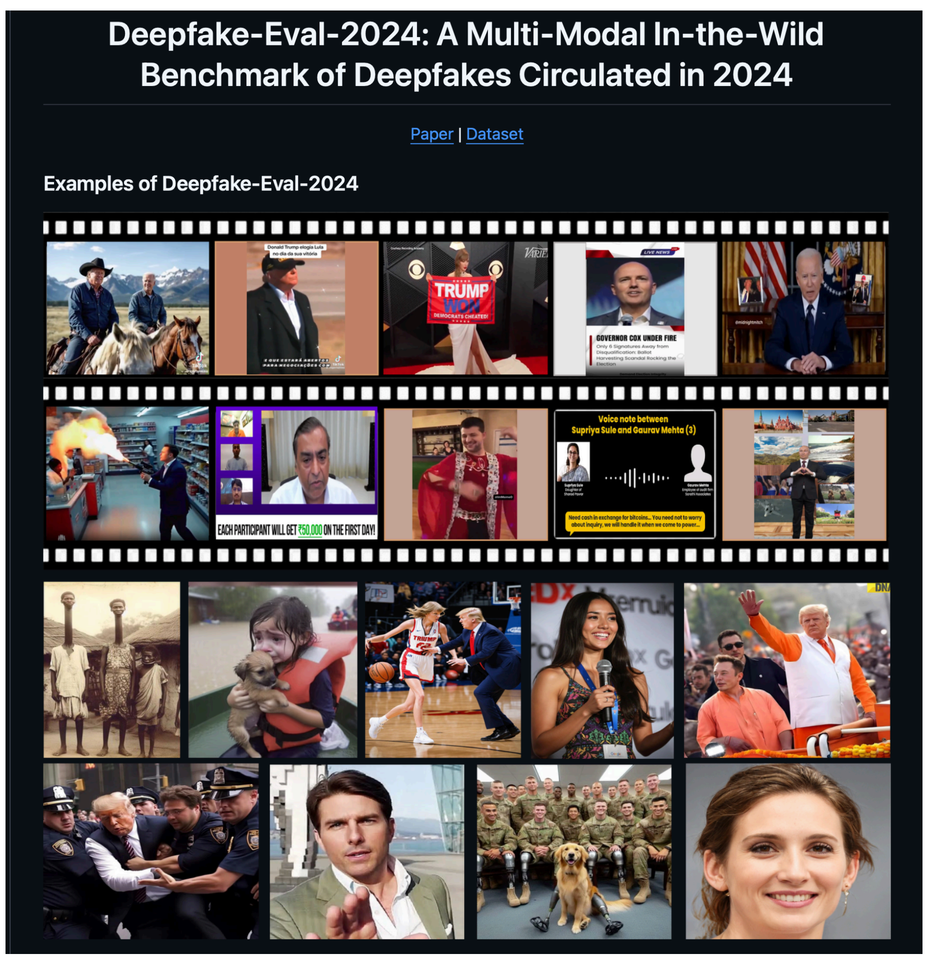
- Improving detection accuracy further without additional training
- **Retrieve the knowledge source using the post-trained SSL embedding**
- Use k nearest samples in the knowledge database for inference (k-NN)

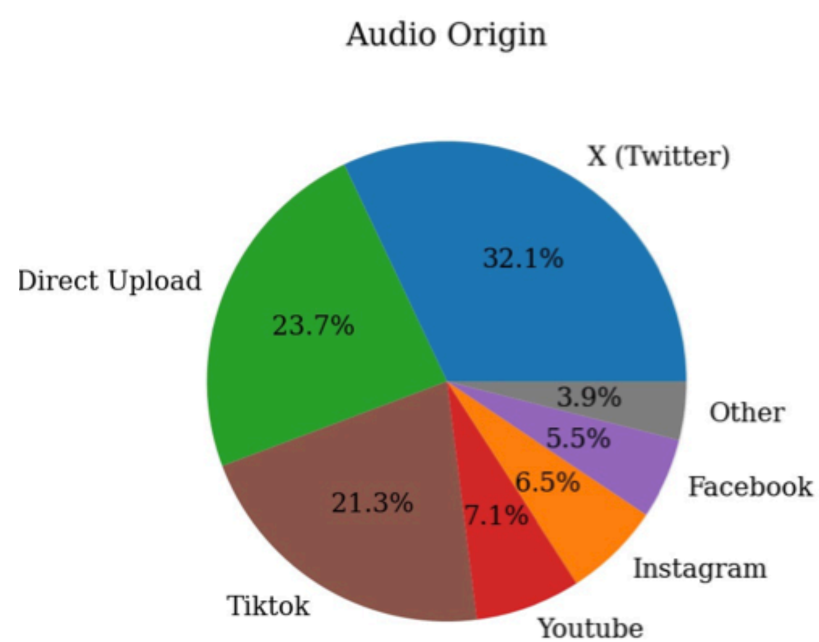


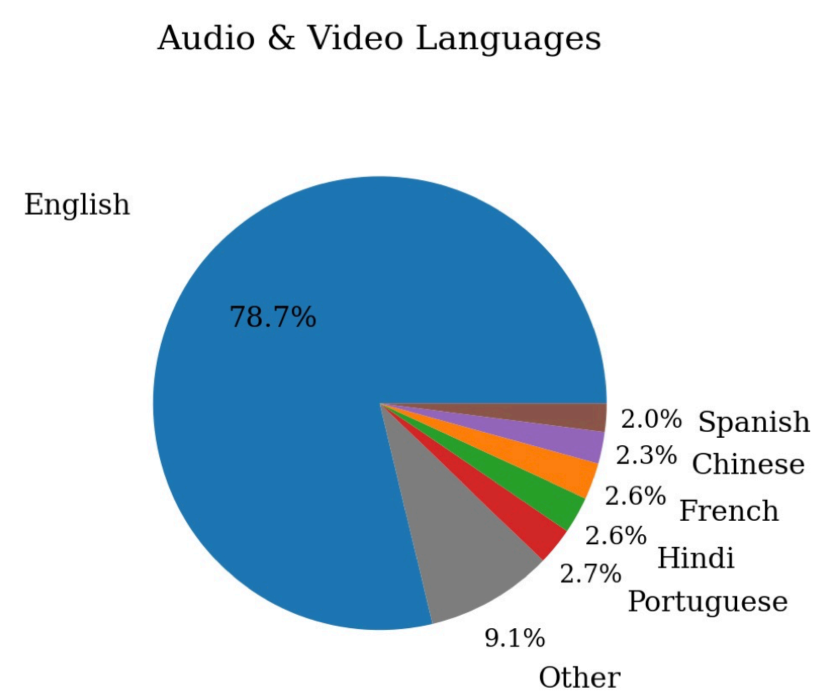
[5] Xuechen Liu, Xin Wang, Junichi Yamagishi, "Frustratingly Easy Zero-Day Audio DeepFake Detection via Retrieval Augmentation and Profile Matching," Submitted to ICASSP 2026

Audio deepfakes on social media platforms

- **Evaluation Using Deefake-Eval-2024 [6]**
  - Deefake-Eval-2024 is a dataset constructed by the nonprofit organization *TrueMedia*, released in 2024, that collects deepfake content spread on social media
  - Zero-shot and fine-tuning scenarios have been tested
  - The train set was used for either fine-tuning or knowledge resource







[6] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, Oren Etzioni “Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024”

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	**AASIST [7]**	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	**AASIST [7]**	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024"

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, Nicholas Evans AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks, ICASSP 2022

[8] Xin Wang, Junichi Yamagishi "Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?" ICASSP 2024

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	**AASIST [7]**	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	**AASIST [7]**	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

Zeroshot condition

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

Zeroshot condition

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

Fine-tuning condition

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

Fine-tuning condition

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

W/o and w/ knowledge source

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

W/o and w/ knowledge source

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in “Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi “Can large-sc
supervised front end?” ICASSP 2024

Post-training improves the zero-shot performance

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

Post-training improves the zero-shot performance

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	**AASIST [7]**	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	$\boxed{27.76}$
ZS + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	**AASIST [7]**	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	$\boxed{\mathbf{12.52}}$
FT + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

Fine-tuning improves the performance further
(but causes overfitting to a specific dataset)

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	**AASIST [7]**	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	**AASIST [7]**	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	**AntiDeepfake (XLS-R-2B) + k-NN**	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2024: A

[7] Jee-weon Jung, Hee-Soo Heo, Hemlata Ta
AASIST: Audio Anti-Spoofing using Integrated

[8] Xin Wang, Junichi Yamagishi "Can large-sc
supervised front end?" ICASSP 2024

Fine-tuning improves the performance further
(but causes overfitting to a specific dataset)

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2

[7] Jee-weon Jung, Hee-Soo Heo, Hem
AASIST: Audio Anti-Spoofing using Inte

[8] Xin Wang, Junichi Yamagishi "Can l
supervised front end?" ICASSP 2024

The use of a knowledge resource is also a good choice (to avoid a model overfitting to a specific dataset)

Results on the TrueMedia database: Audio

		Accuracy	AUC	F1	EER (%)
Zero-shot (ZS)	AASIST [7]	0.42*	0.43*	0.39*	55.22*
	NII's P3 (XLSR-large) + MLP [8]	0.36*	0.58*	0.53*	43.00*
	AntiDeepfake (XLS-R-2B) +MLP	0.75	0.80	0.83	27.76
ZS + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.87**	**0.90**	**0.84**	**14.78**
Fine-tuned (FT)	AASIST [7]	0.84*	0.91*	0.78*	16.99
	NII's P3 (XLSR-large) + MLP [8]	0.86*	0.92*	0.81*	15.38
	AntiDeepfake (XLS-R-2B) +MLP	0.88	**0.93**	0.83	**12.52**
FT + knowledge source	AntiDeepfake (XLS-R-2B) + k-NN	**0.89**	0.91	**0.84**	12.86

* Numbers reported in "Deepfake-Eval-2

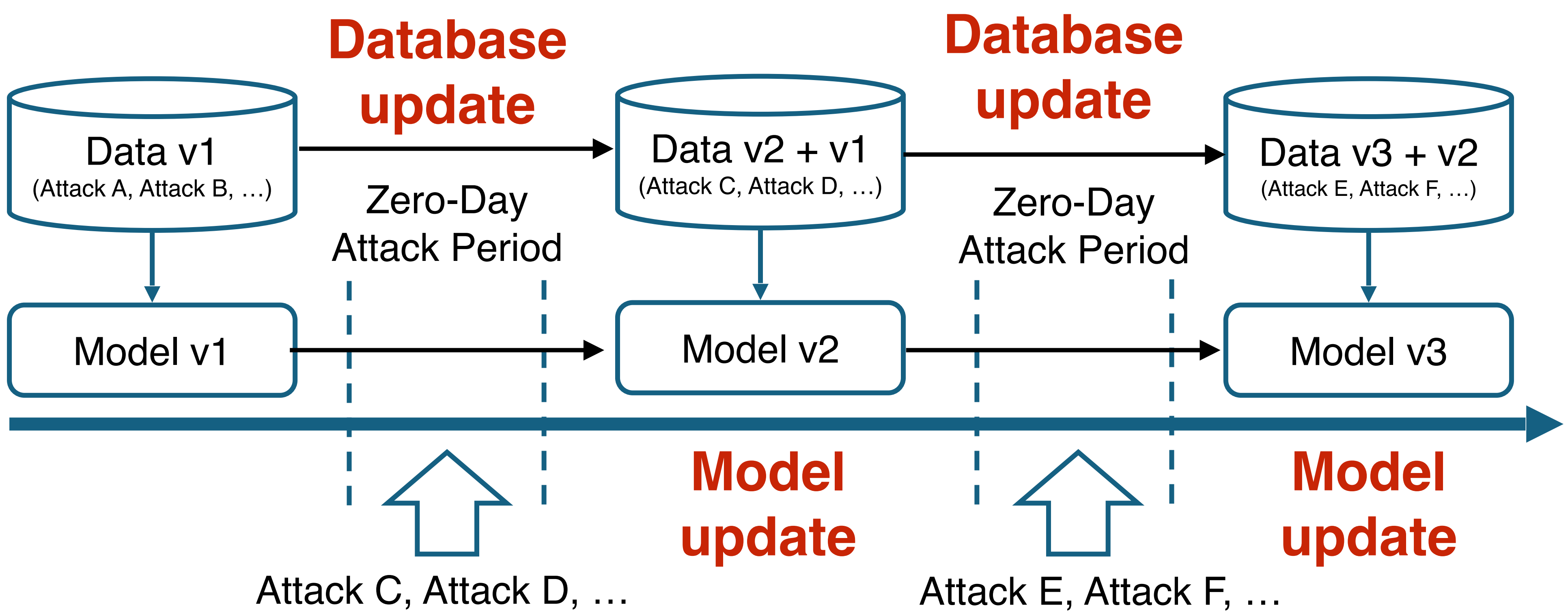
[7] Jee-weon Jung, Hee-Soo Heo, Hem
AASIST: Audio Anti-Spoofing using Inte

[8] Xin Wang, Junichi Yamagishi "Can l
supervised front end?" ICASSP 2024

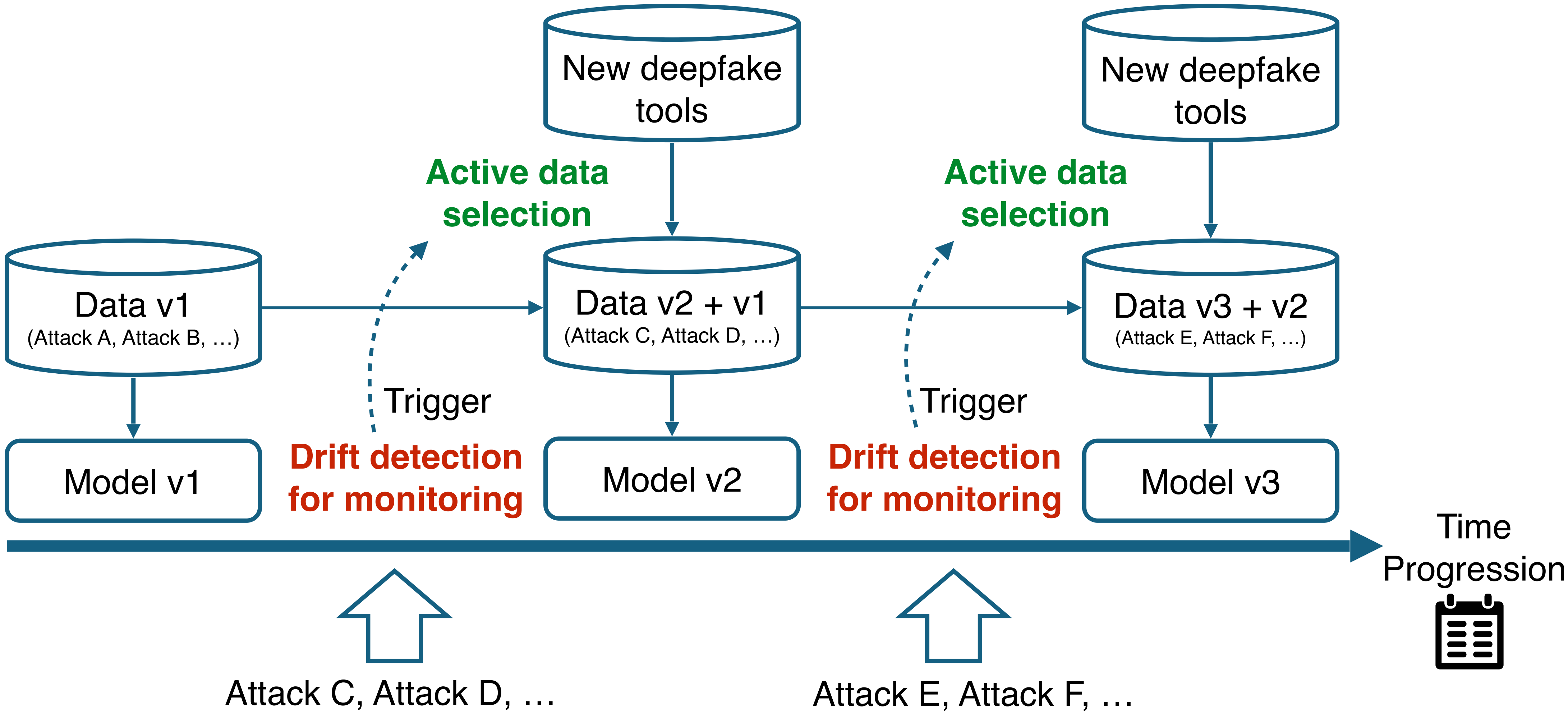
The use of a knowledge resource is also a good choice (to avoid a model overfitting to a specific dataset)

Part 2:

Machine Learning Operations **(MLOPs)** of deepfake detection



Four RQs for regular model and database updates



RQ1: Do the new deepfake attacks (Attacks C, D, etc.) exhibit unseen artifacts that significantly differ from the previously seen attack methods (Attack A and B)?

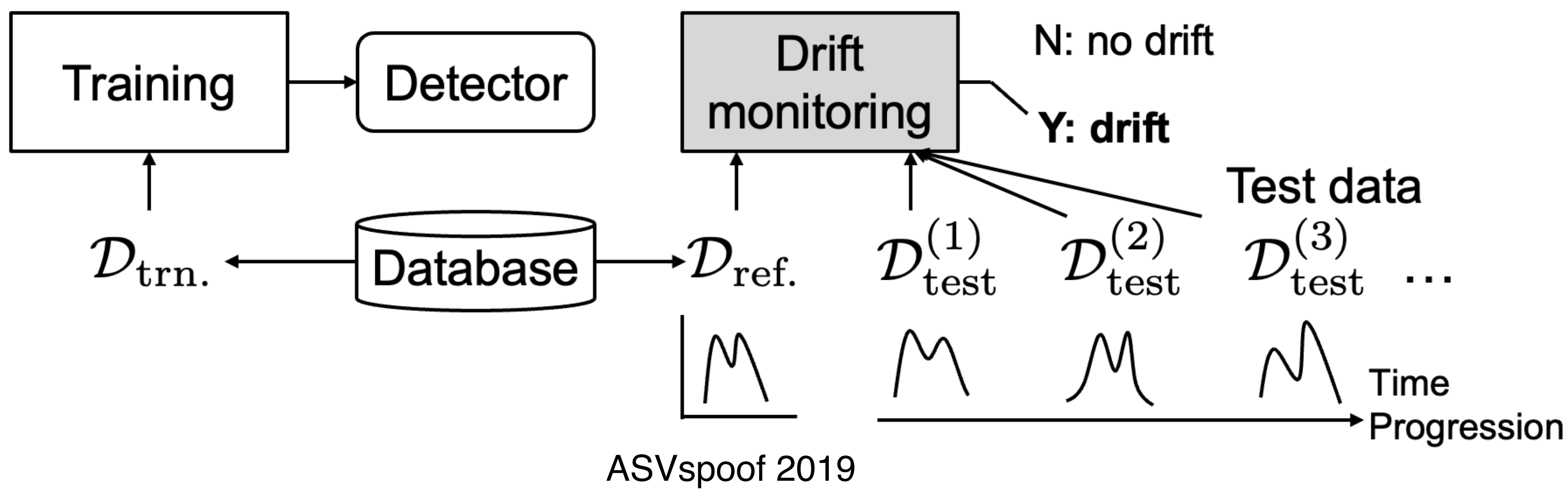
Drift detection for monitoring

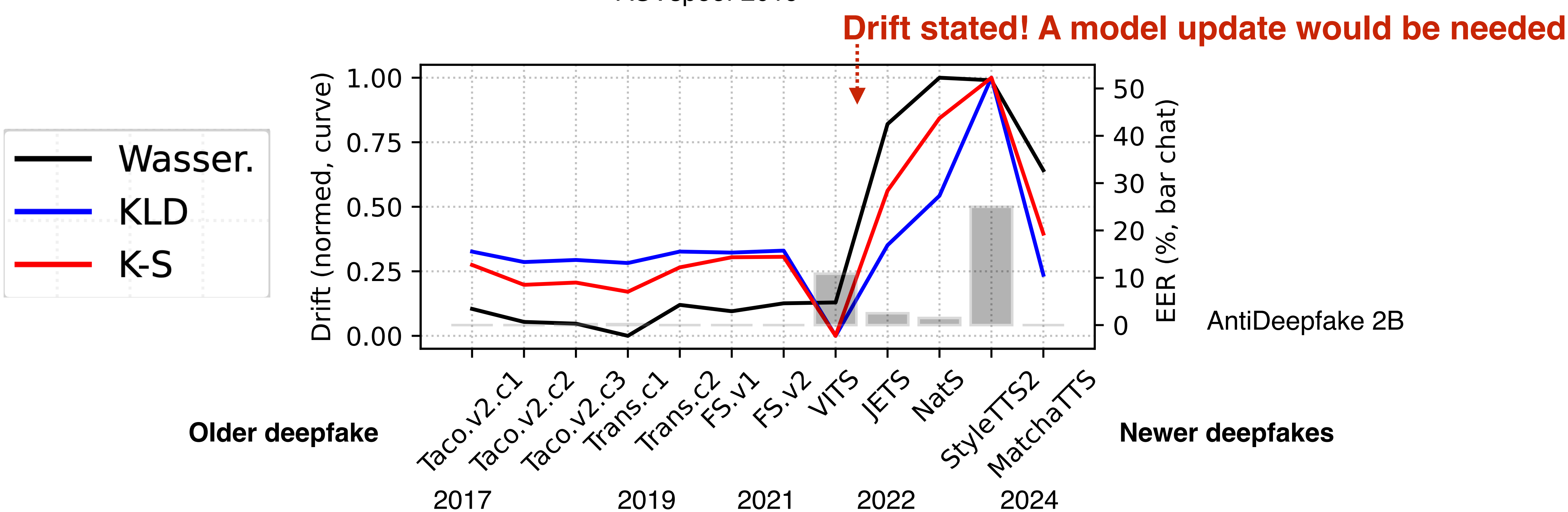
RQ2: Which of the new attacks (Attacks C, D, etc.) should be incorporated into the training dataset?

Automatic selection of “useful” unseen deepfake attacks

RQ1: Drift (Change Point) Detection in Deepfakes [9]

- Observe the test data at each time step during operation and compare it to the fixed reference data to develop the model
- Identify the points where the distance to the reference data significantly increases **(Note: this is NOT the distance to human speech)**

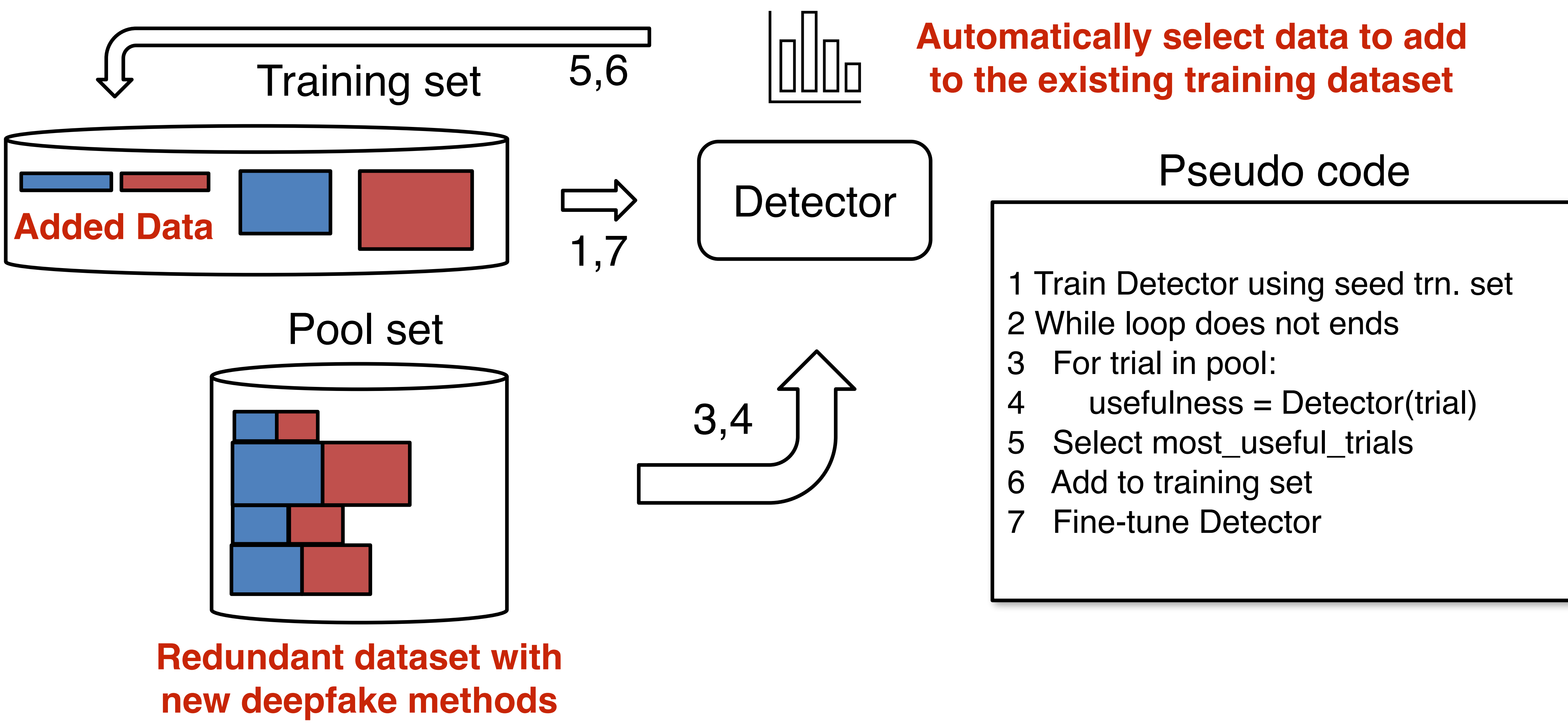




[9] Xin Wang, Wanying Ge, Junichi Yamagishi, "Towards Data Drift Monitoring for Speech Deepfake Detection in the context of MLOps" Submitted to ICASSP 2026

RQ2: Active data addition based on confidence scores [10]

- While new deepfake methods are proposed almost daily, **many of them share similarities with existing techniques in terms of artifacts**
  - It is unnecessary to include all new deepfake methods in the training dataset for detectors if the initial database is sufficiently rich
- **Automatically select samples with low detection confidence as new *additional* training data for the model update [8]**

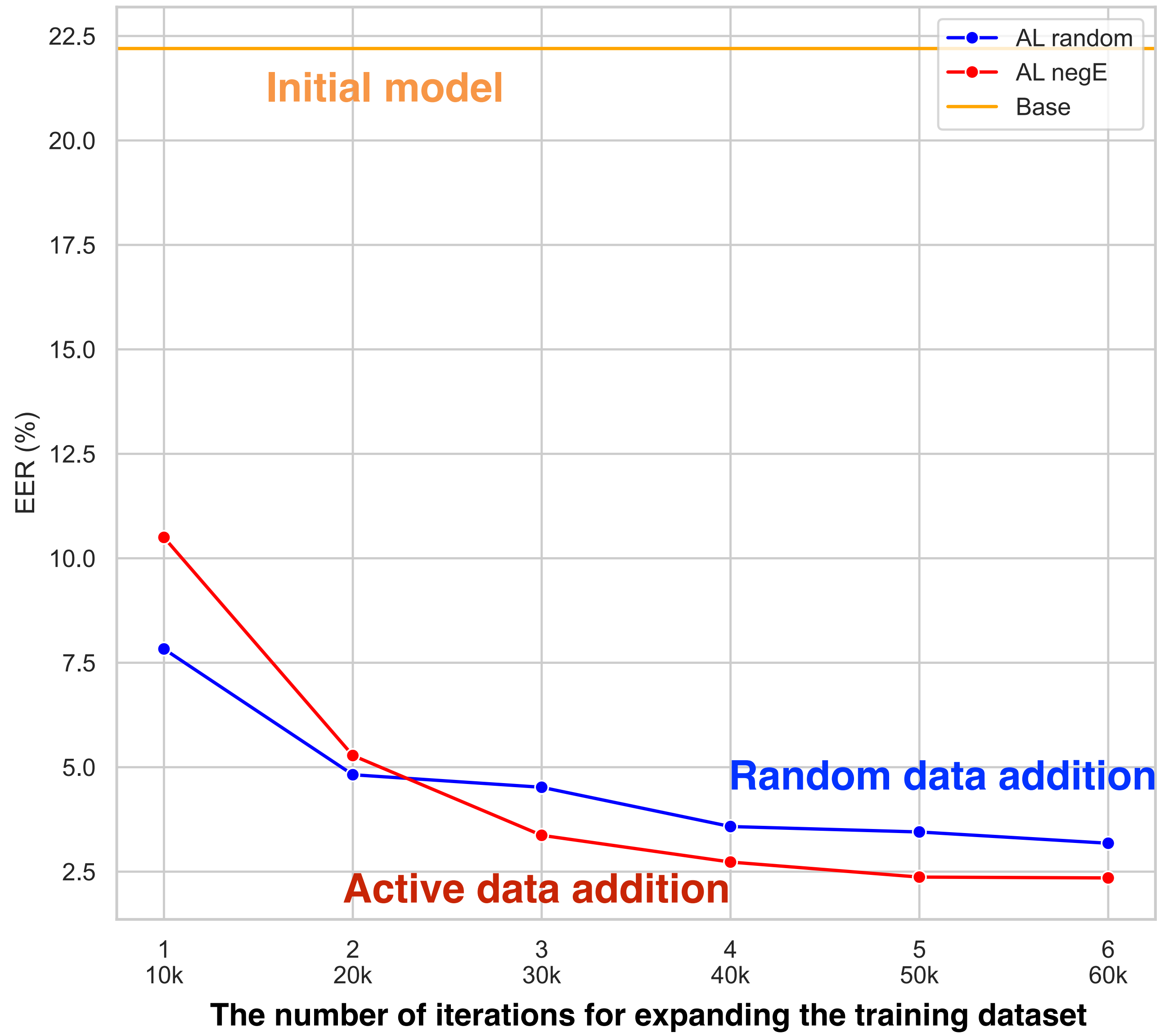


[10] Xin Wang, Junich Yamagishi, "Investigating Active-learning-based Training Data Selection for Speech Spoofing Countermeasure" the 2022 IEEE Spoken Language Technology Workshop (SLT 2022) Jan

RQ2: Active training data addition for facial deepfake detection [11]

- **Model**: EfficientNet V2 pre-trained on ImageNet 21k.
- **The initial dataset**: ForgeryNet dataset
- **# of additions**: When expanding the training dataset, select 10,000 images each time from the pool set
- **Test set:** combinations of multiple test sets below

Database	Type	Initial	AL Pool	Val.	Test
Starter master set					
ForgeryNet [He21]	Real	163,200		1,000	1,000
ForgeryNet [He21]	Fake	163,200		1,000	1,000
Pool set					
FF++ [Ro19]	Real		40,000	1,000	1,000
FF++ (5 types) [Ro19]	Fake		40,000	1,000	1,000
Google DFD [DG19]	Real		40,000	1,000	1,000
Google DFD [DG19]	Fake		40,000	1,000	1,000
VoxCeleb [CNZ18]	Real		40,000	1,000	1,000
YouTube DF [Ku20]	Fake		40,000	1,000	1,000
KoDF [Kw21]	Real		40,000	1,000	1,000
KoDF [Kw21]	Fake		40,000	1,000	1,000
FFHQ [KLA19]	Real		40,000	1,000	1,000
Stable Diffusion 2.1 [Ro22]	Fake		40,000	1,000	1,000

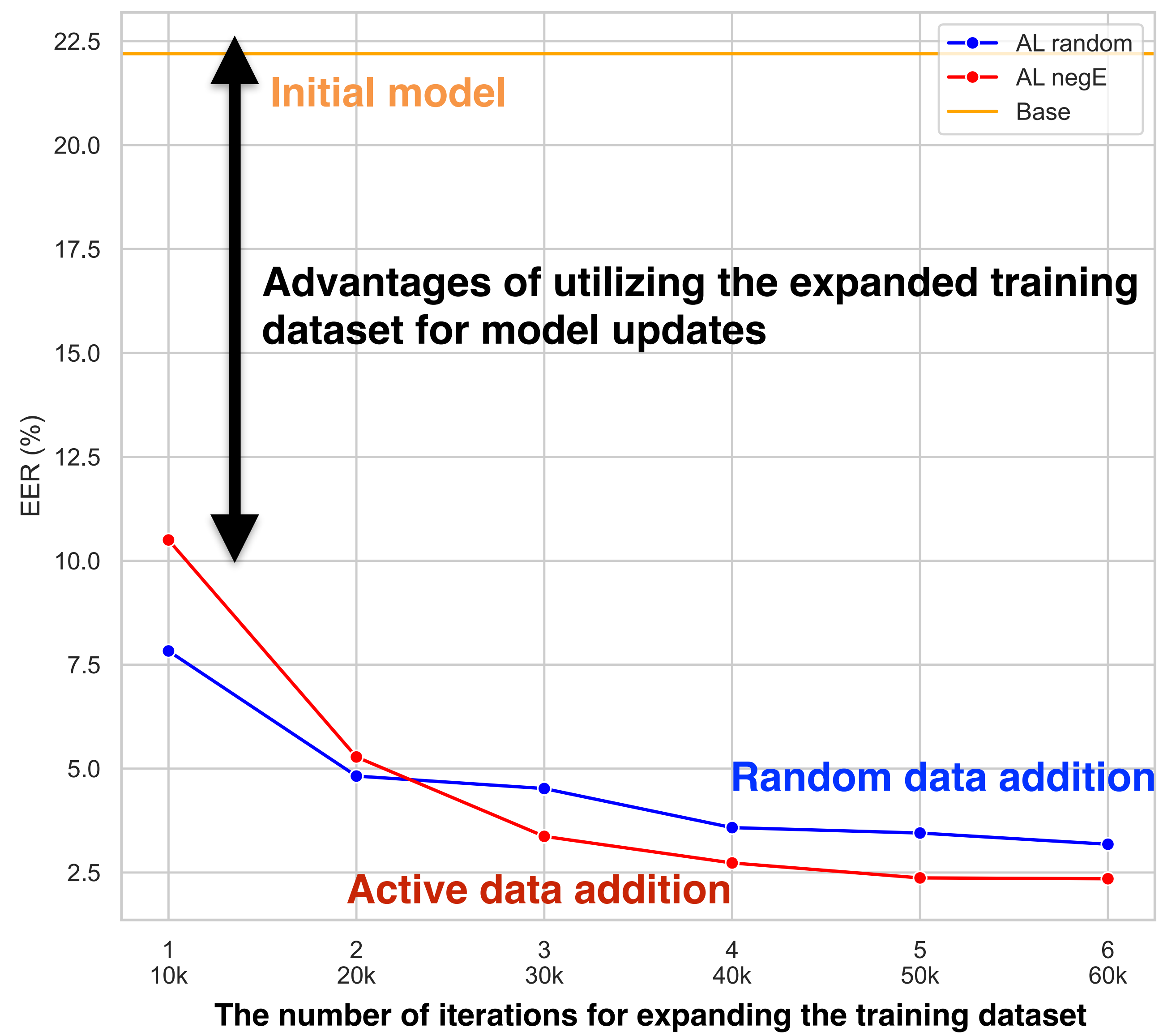


[11] Yoshihiko Furuhashi, Xin Wang, Junichi Yamagishi, Huy Nguyen, Isao Echizen, “Exploring Active Data Selection Strategies for Continuous Training in Deepfake Detection” 23rd International Conference of the Biometrics Special Interest Group 2024

RQ2: Active training data addition for facial deepfake detection [11]

- **Model**: EfficientNet V2 pre-trained on ImageNet 21k.
- **The initial dataset**: ForgeryNet dataset
- **# of additions**: When expanding the training dataset, select 10,000 images each time from the pool set
- **Test set:** combinations of multiple test sets below

Database	Type	Initial	AL Pool	Val.	Test
Starter master set					
ForgeryNet [He21]	Real	163,200		1,000	1,000
ForgeryNet [He21]	Fake	163,200		1,000	1,000
Pool set					
FF++ [Ro19]	Real		40,000	1,000	1,000
FF++ (5 types) [Ro19]	Fake		40,000	1,000	1,000
Google DFD [DG19]	Real		40,000	1,000	1,000
Google DFD [DG19]	Fake		40,000	1,000	1,000
VoxCeleb [CNZ18]	Real		40,000	1,000	1,000
YouTube DF [Ku20]	Fake		40,000	1,000	1,000
KoDF [Kw21]	Real		40,000	1,000	1,000
KoDF [Kw21]	Fake		40,000	1,000	1,000
FFHQ [KLA19]	Real		40,000	1,000	1,000
Stable Diffusion 2.1 [Ro22]	Fake		40,000	1,000	1,000

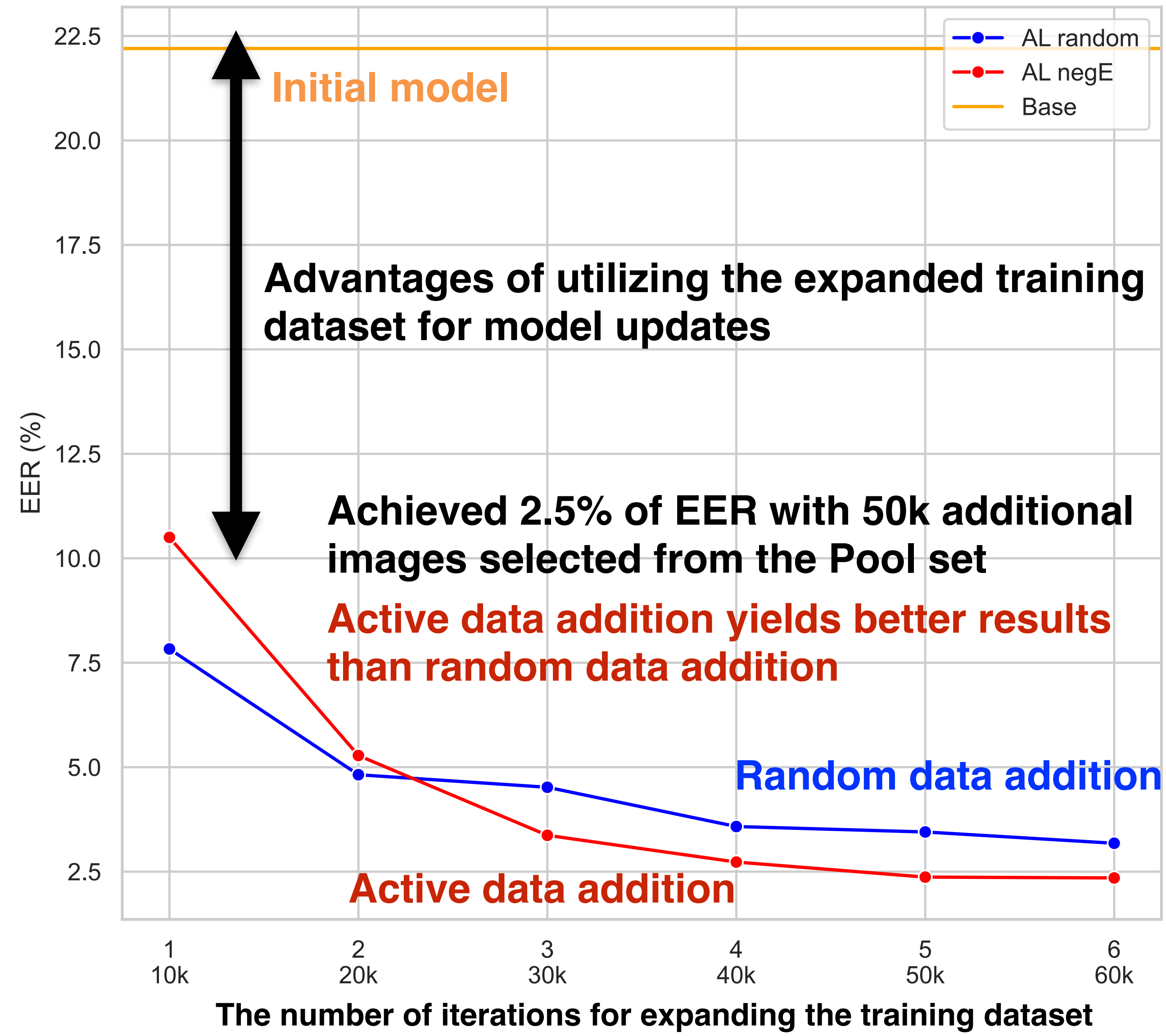


[11] Yoshihiko Furuhashi, Xin Wang, Junichi Yamagishi, Huy Nguyen, Isao Echizen, “Exploring Active Data Selection Strategies for Continuous Training in Deepfake Detection” 23rd International Conference of the Biometrics Special Interest Group 2024

RQ2: Active training data addition for facial deepfake detection [11]

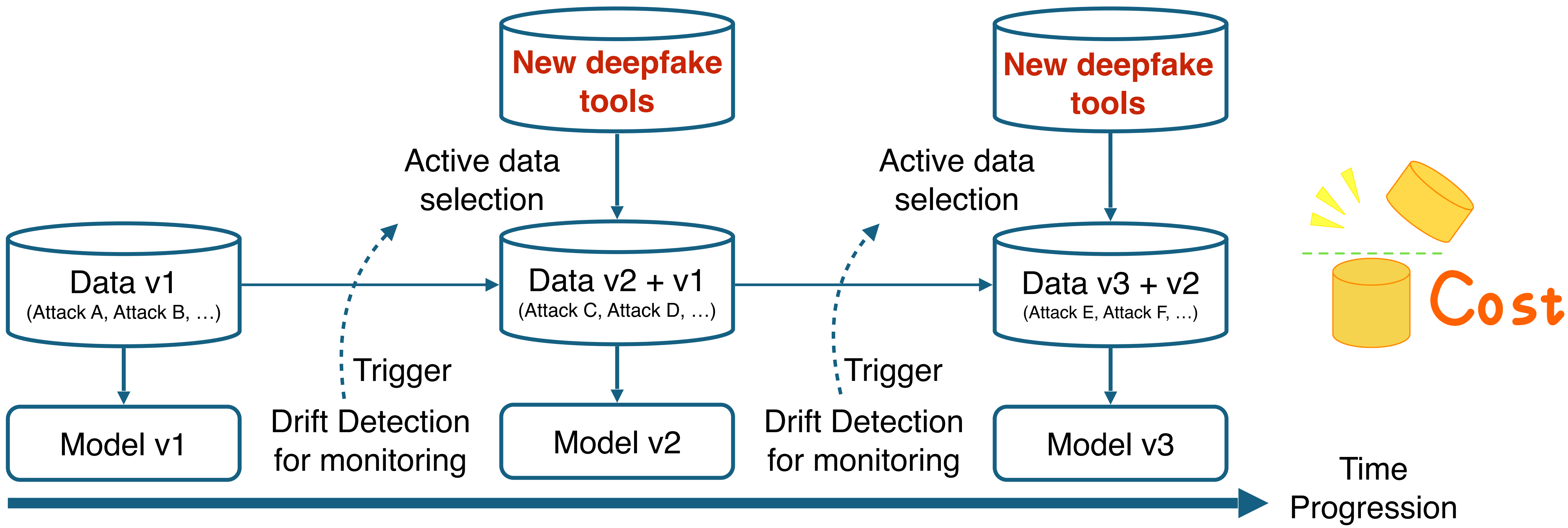
- **Model**: EfficientNet V2 pre-trained on ImageNet 21k.
- **The initial dataset**: ForgeryNet dataset
- **# of additions**: When expanding the training dataset, select 10,000 images each time from the pool set
- **Test set:** combinations of multiple test sets below

Database	Type	Initial	AL Pool	Val.	Test
Starter master set					
ForgeryNet [He21]	Real	163,200		1,000	1,000
ForgeryNet [He21]	Fake	163,200		1,000	1,000
Pool set					
FF++ [Ro19]	Real		40,000	1,000	1,000
FF++ (5 types) [Ro19]	Fake		40,000	1,000	1,000
Google DFD [DG19]	Real		40,000	1,000	1,000
Google DFD [DG19]	Fake		40,000	1,000	1,000
VoxCeleb [CNZ18]	Real		40,000	1,000	1,000
YouTube DF [Ku20]	Fake		40,000	1,000	1,000
KoDF [Kw21]	Real		40,000	1,000	1,000
KoDF [Kw21]	Fake		40,000	1,000	1,000
FFHQ [KLA19]	Real		40,000	1,000	1,000
Stable Diffusion 2.1 [Ro22]	Fake		40,000	1,000	1,000



[11] Yoshihiko Furuhashi, Xin Wang, Junichi Yamagishi, Huy Nguyen, Isao Echizen, “Exploring Active Data Selection Strategies for Continuous Training in Deepfake Detection” 23rd International Conference of the Biometrics Special Interest Group 2024

Unsolved two RQs for MLOps of deepfake detection



RQ3: How can we swiftly and automatically identify new deepfake tools being used by the general public and reliably curate new deepfake data?

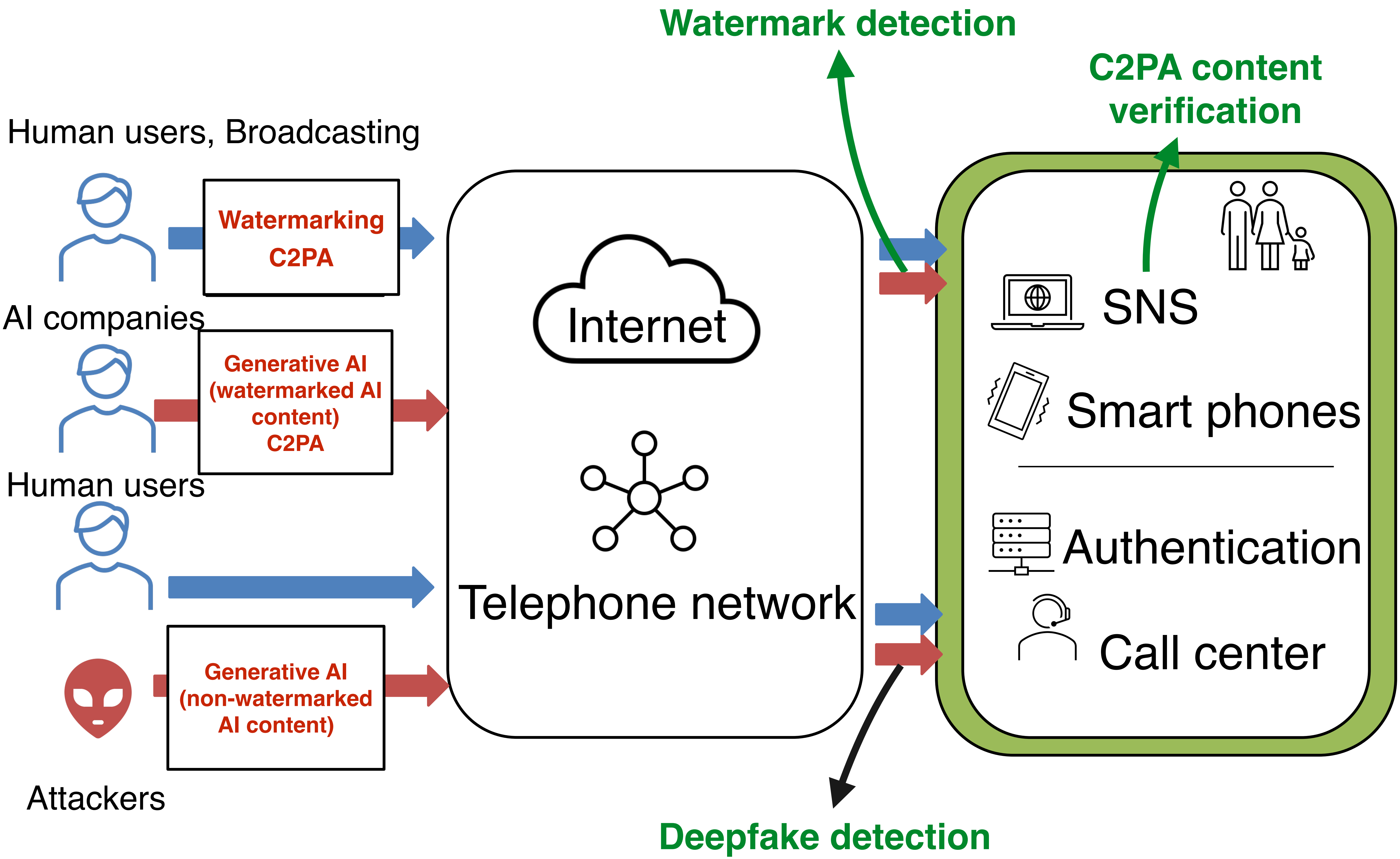
Deepfake voices on social media are different from any scientific benchmark databases, including ASVspoof databases [12]

RQ4: How can we reduce the costs associated with data collection, database updates, and model updates (thereby enabling us to increase the frequency of these updates)?

[12] David Combei, Adriana Stan, Dan Oneata, Nicolas Müller, Horia Cucu, “Unmasking real-world audio deepfakes: A data-centric approach,” Proc. Interspeech 2025

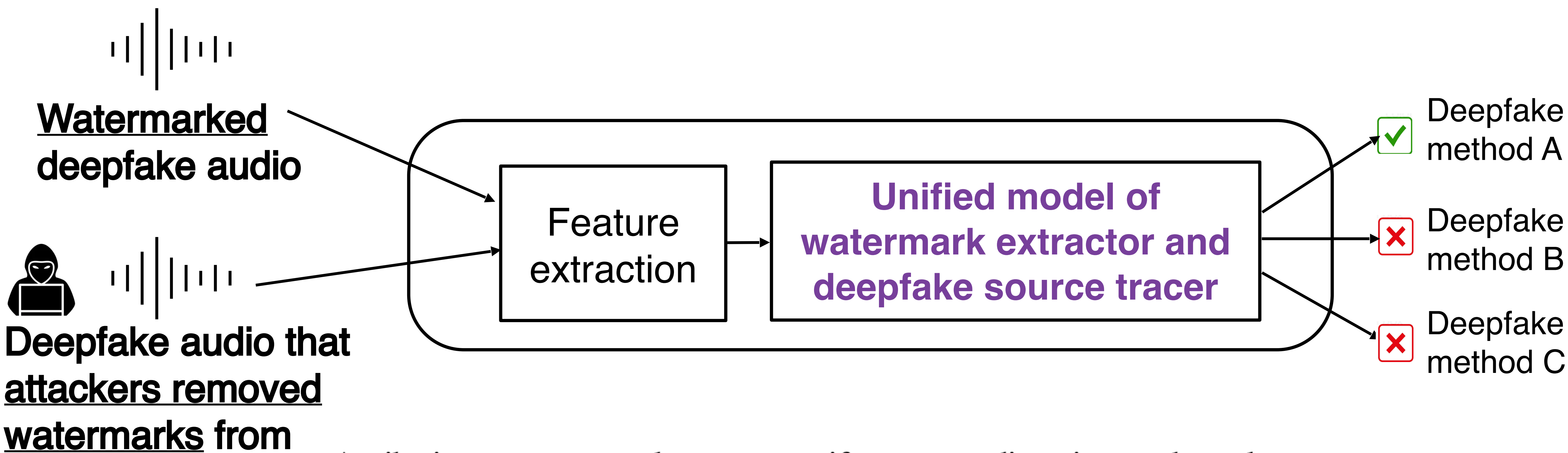
Part 3:
Collective approach to passive and proactive deepfake defense

Passive and proactive deepfake defense



Mixture of neural watermark and source tracing models [13]

- If extracting the watermark is impossible, use the source tracer based on acoustic features. If the watermark exists, extract its information



Attribution accuracy results on seen artifacts across distortions and attacks.

	Distortion \ System	Proposed Method		Watermarking Baselines		Classifier Baselines	
		FakeMark^{A}	**FakeMark**^{T}	**AudioSeal** [14]	**Timbre** [15]	**MMS-300M**	**ResNet34**
	None	1.00	1.00	1.00	1.00	1.00	0.97
Removal Attack	Overwriting	0.99	0.95	0.68	0.55	0.95	0.75
	Averaging	0.98	0.99	0.79	1.00	1.00	0.96

Experiments using the MLAAD v5 test set

[13] Wanying Ge, Xin Wang, Junichi Yamagishi, “FakeMark: Deepfake Speech Attribution With Watermarked Artifacts” Arxiv, 2025

[14] AudioSeal: Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, Hady Elsahar, "Proactive Detection of Voice Cloning with Localized Watermarking” ICML 2024

[15] Timbre: Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, Nenghai Yu, "Detecting Voice Cloning Attacks via Timbre Watermarking” NDDS 2024

Agenda of the talk and future topics

- **Background:**
  Why is deepfake detection such a challenging task?
- **Part 1:**
  Robust detection of unknown deepfake audio generation methods
- **Part 2:**
  Machine Learning Operations (MLOPs) of deepfake detection
- **Part 3:**
  Collective approach to passive and proactive deepfake defense

- **Important topics that I couldn’t cover today include**
  - explainability
  - adversarial attacks against deepfake detection
  - combination of misinformation detection and deepfake detection

Acknowledgement

- This study is based on results obtained from a project, *JPNP22007, commissioned by the New Energy and Industrial Technology Development Organization (NEDO)*
- This study is partially supported by *JST AIP Acceleration Research (JPMJCR24U3) and JST PRESTO (JPMJPR23P9)*
- This study was carried out using the TSUBAME4.0 supercomputer at the Institute of Science Tokyo

Q & A