POLITECNICO MILANO 1863

Leveraging Mixture of Experts for Improved Speech Deepfake Detection

Viola Negroni – *PhD Student*

Image and Sound Processing Lab – ISPL
Department of Electronics, Information e Bioengineering

20.10.2025

Image and Sound Processing Lab – *Forensics Division*

STEFANO TUBARO
Full Professor

PAOLO BESTAGINI
Associate Professor

SARA MANDELLI
Assistant Professor
Image/Video Forensics

ISPL
Image and Sound Processing Lab

DANIELE UGO LEONZIO
Post-Doc Researcher
Geophysics

VIOLA NEGRONI
PhD Student
Audio Forensics

GIOVANNI AFFATATO
PhD Student
Image/Video Forensics

WENDY WANG
Research Assistant
Geophysics

JUAN CAMILO ALBARRACIN
Research Assistant
Geophysics

Audio “Team”

VIOLA NEGRONI
PhD Student

- **Curiosity-driven**, this got me quite an unusual background
- Now almost at the end of my **2nd year** as a **PhD student**
- Specialized in **audio forensics**, focus on Speech Deepfake Detection
- **10 papers** published, 3 under review
  - *Analyzing the Impact of Splicing Artifacts in Partially Fake Speech Signals*, ASVspoof Workshop, *2024*
  - *Source Verification for Speech Deepfakes*, Interspeech, *2025*

Presentation Outline

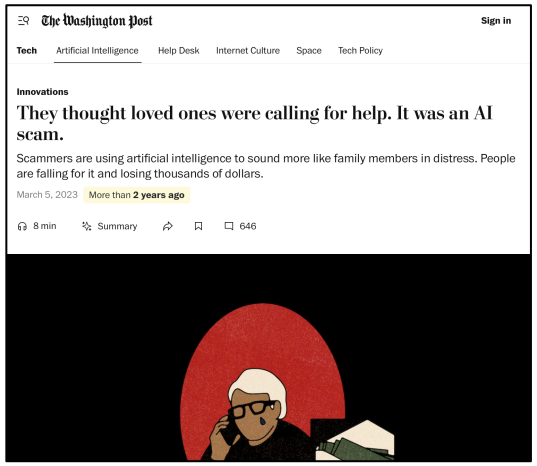
1. **Motivations**
2. **Background on MoEs**
3. **«Leveraging Mixture of Experts for Improved Speech Deepfake Detection »**
    a. Rationale
    b. Proposed Systems
    c. Evaluation Setup
    d. Experimental Results
4. **« Attention-based Mixture of Experts for Robust Speech Deepfake Detection »**
    a. Context: SAFE Challenge 2025
    b. Paradigm Shift: MELE vs MILE
    c. Evaluation Setup
    d. Experimental Results
5. **Outcomes and Limitations**
6. **Future directions on MoEs**

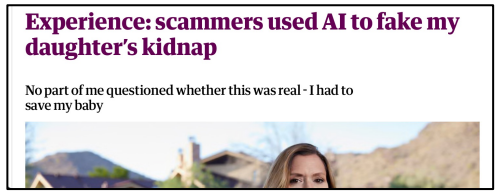
Motivations

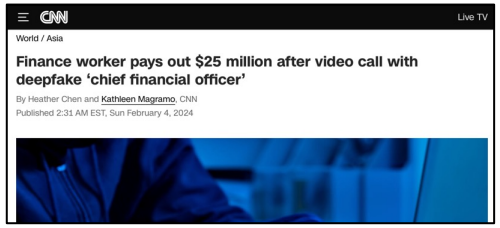
🎙️ **TTS and VC tools** are getting better and easier to use

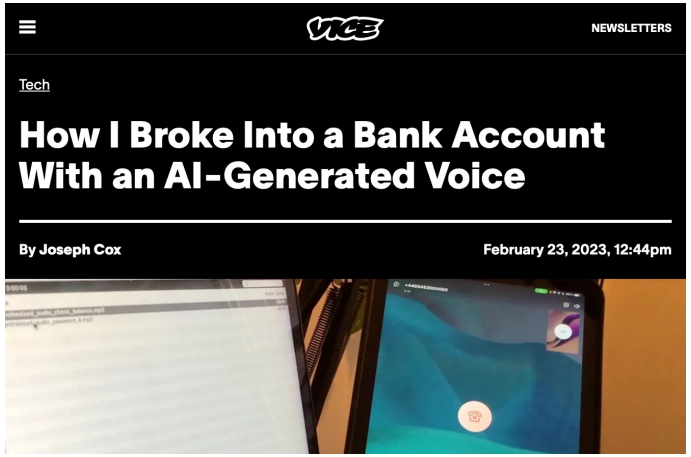
🕵️ **Fraud, scams and other crimes** that rely on generated audio have skyrocketed

🏛️ Fake speech detectors work on academic datasets but **underperform in the wild**









Motivations

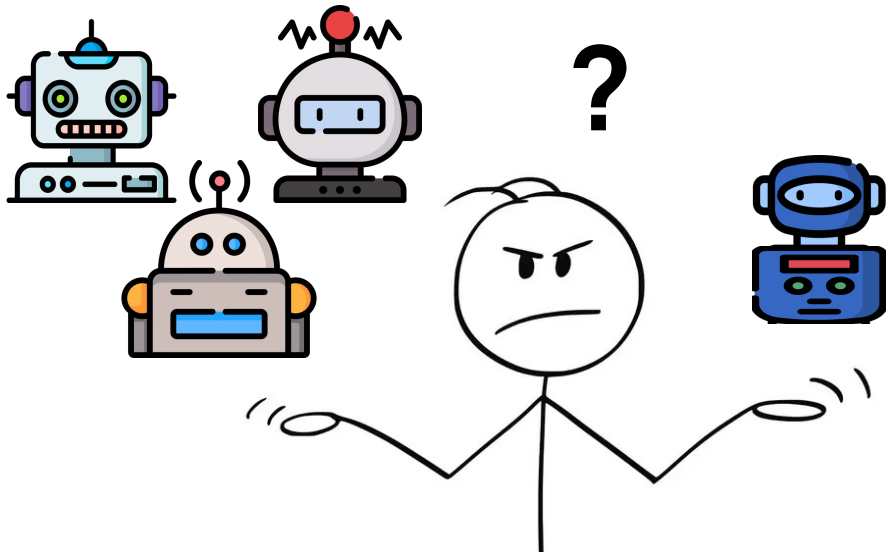
🏛️ Fake speech detectors work on academic datasets but **underperform in the wild**

Why?

Motivations

The Generalization Problem

- Existing speech deepfake detectors lack **generalization** capabilities on *unknown* data
  a. Unseen synthetic speech generators
  b. Out of domain genuine data *
- With the increasing frequency of **new generative models** releases, it is essential for detectors to adapt efficiently

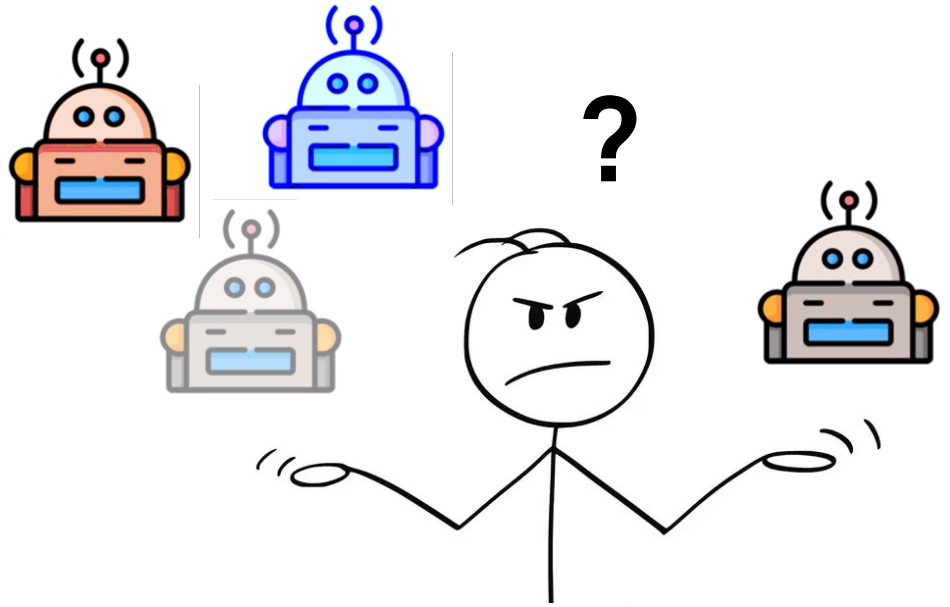


* Let's not forget about *shortcut learning*!

Motivations

The Robustness Problem

- Existing speech deepfake detectors lack **robustness** in mismatched conditions
    a. Post-processed data
    b. Degraded data
- With the increasing **spread** of synthetic content **online**, it is essential for detectors to adapt efficiently

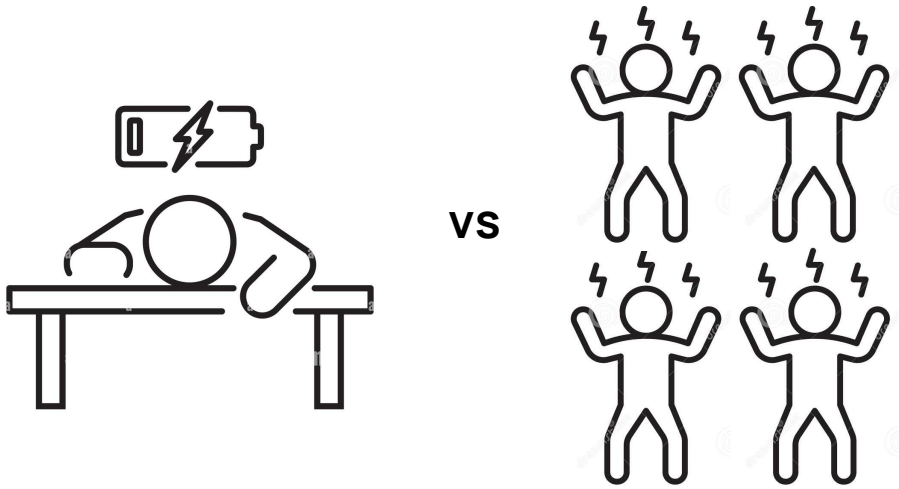


⚠ *Robustness ≠ Generalization !*

Motivations

- Practical and conceptual **limitations** of current solutions.
  - **Fine tuning** leads to catastrophic forgetting.
  - **Re-training** and **joint-training** are not scalable.

- To tackle this, we can leverage **Mixture of Experts** frameworks!



Background on MoEs

Useful relevant literature

📚 **Adaptive Mixtures of Local Experts**
R. A. Jacobs, M. I. Jordan, S. J. Nowla, G. E. Hinton
Neural Computation (1991)

📚 **Mixture of Experts: A Literature Survey**
S. Masoudnia, R. Ebrahimpour
Artificial Intelligence Review (2014)

📚 **Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer**
N. Shazeer , A. Mirhoseini , K. Maziarz , A. Davis , Q. Le , G. Hinton, J. Dean
ICLR (2017)

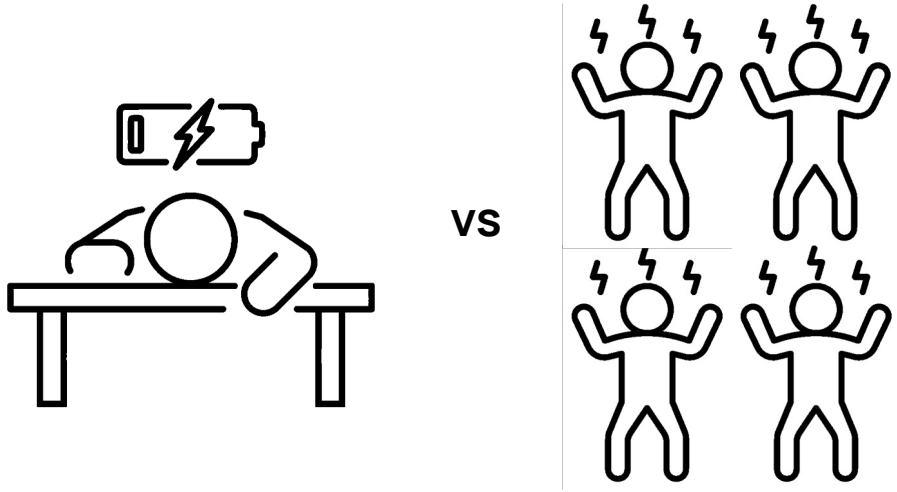
Background on MoEs

Core elements

- A MoE is a hierarchical ensemble of specialized **experts** controlled by a **gating function**

- **EXPERT**
  A model **specialized in** a specific **domain or representation**, trained to detect certain patterns.

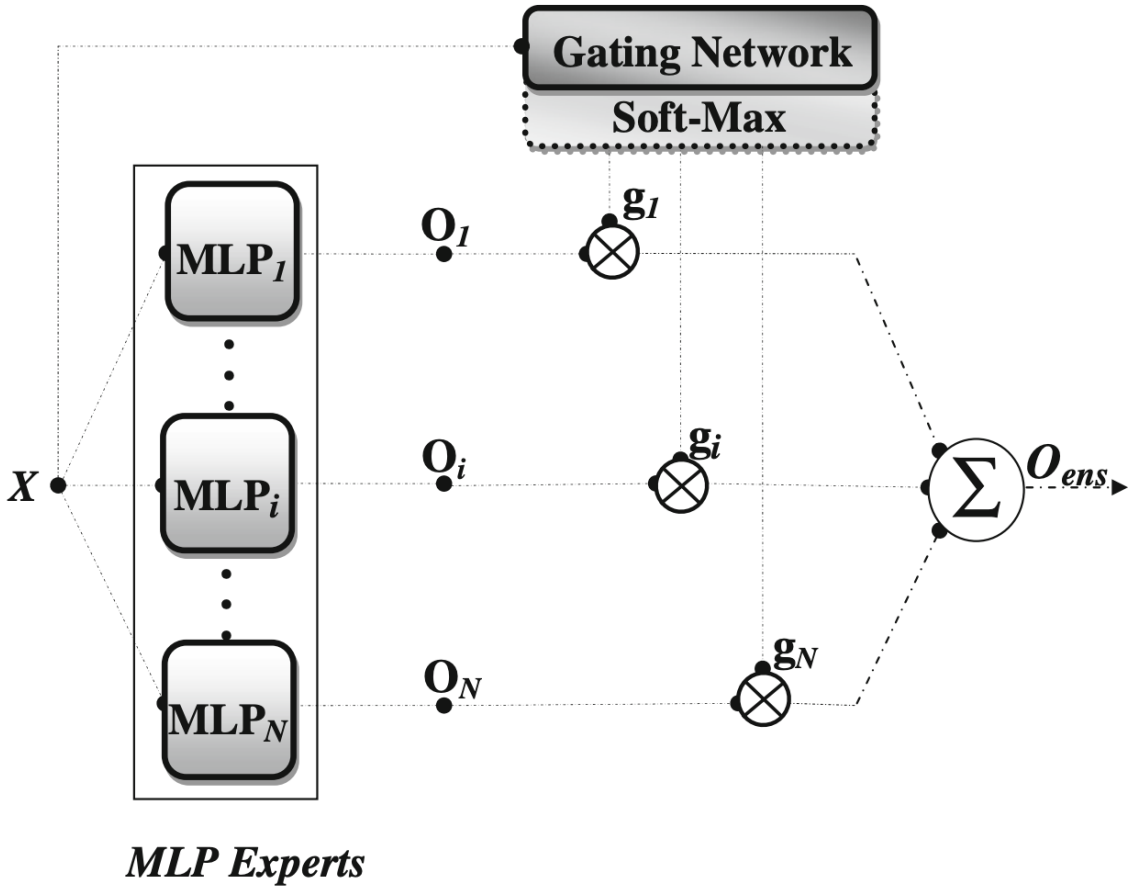
- **GATING NETWORK**
  A controller model that analyzes the input and **dynamically determines the contribution** of each expert to the output



Background on MoEs

Flexibility as key

- Training data can be partitioned among models **deterministically or stochastically**
- Experts and the gating network can be trained **jointly or sequentially**
- Expert selection at inference can be **hard** (few) **or soft** (weighted)
- MoEs act as “smart” ensembles, using a **divide-and-conquer** strategy to maximize performance



Prototypical MoE framework during inference with soft expert selection.

Background on MoEs

MoEs for Speech Deepfake Detection

- **EXPERT**
  A synthetic speech detector
  - Traditional: ResNets, LCNNs
  - SincNet-based: RawNet2, AASIST
  - SSL: wav2vec 2.0, XLS-R
- **GATING NETWORK**
  A shallower custom neural network
  - MLP
  - Transformer Layer

🧩 Leverage each model's strengths and compensate for their weaknesses

🧠 Remember a wide range of different deepfake types simultaneously

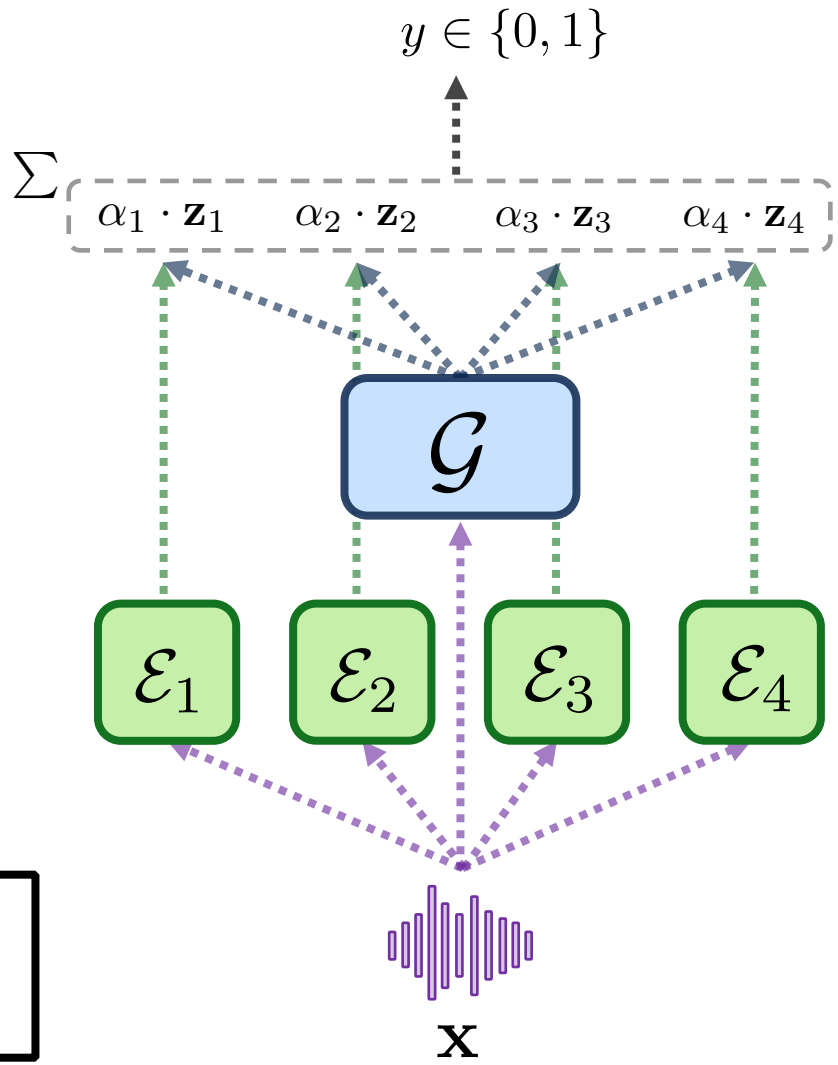
Leveraging Mixture of Experts for Improved Speech Deepfake Detection

V. Negroni, D. Salvi, A. I. Mezza, P. Bestagini, S. Tubaro – IEEE ICASSP 2025

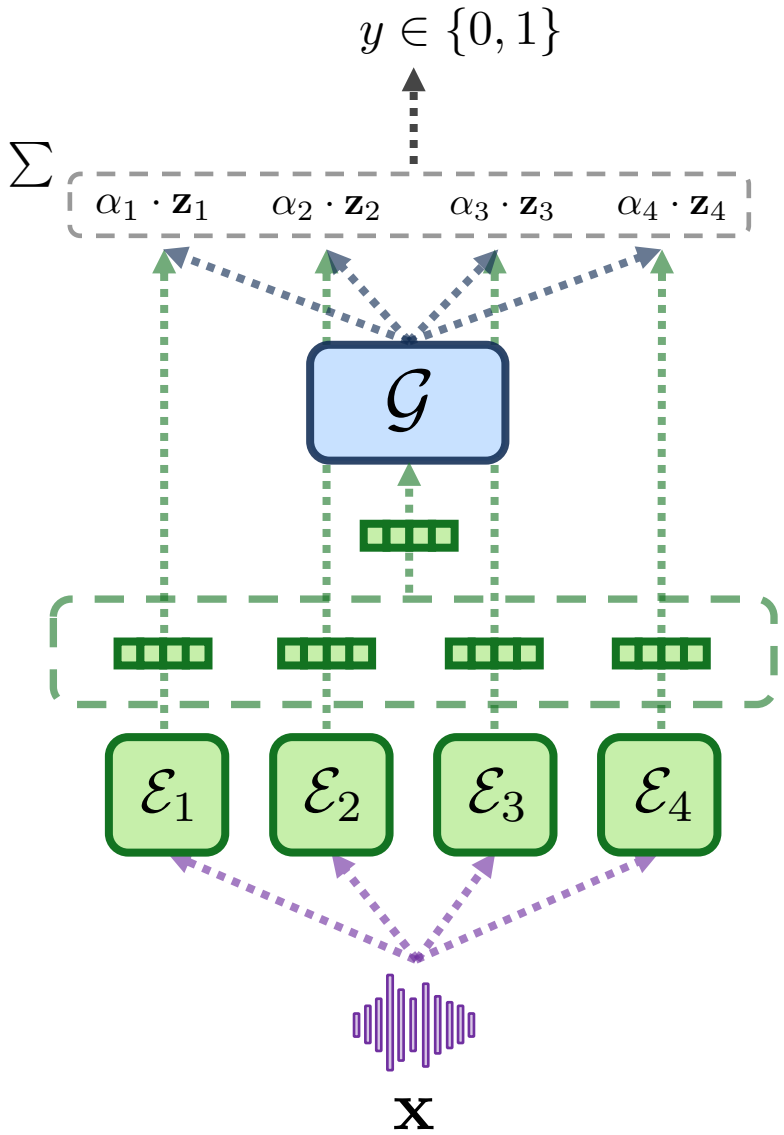
Rationale

- A priori (**deterministic**) partitioning of the input data: domain experts
  - Experts are **first pre-trained** in their domain, **then trained jointly** with the gating network
- Inference: **soft** expert **selection**
  - Expert weights α are calculated by applying a softmax function to the logits produced by the gating network.
  - Each weight can be interpreted as the **percentage of involvement** of each expert in the decision for a given input.

Proposed Systems



Standard MoE



Enhanced MoE

Evaluation Setup

CONSIDERED DATASETS

KNOWN DOMAINS:

1. ASVspoof 2019
2. Fake-Or-Real
3. ADD 2022
4. In-The-Wild

UNKNOWN DOMAINS:

1. Purdue dataset
2. TIMIT-TTS

ARCHITECTURES

1. EXPERTS: LCNN
2. GATING NETWORK:
   - FC Layer
   - Dropout
   - BN
   - LeakyReLU

BASELINES

1. Average Ensemble
2. Joint Training

Experimental Results

Detection Results

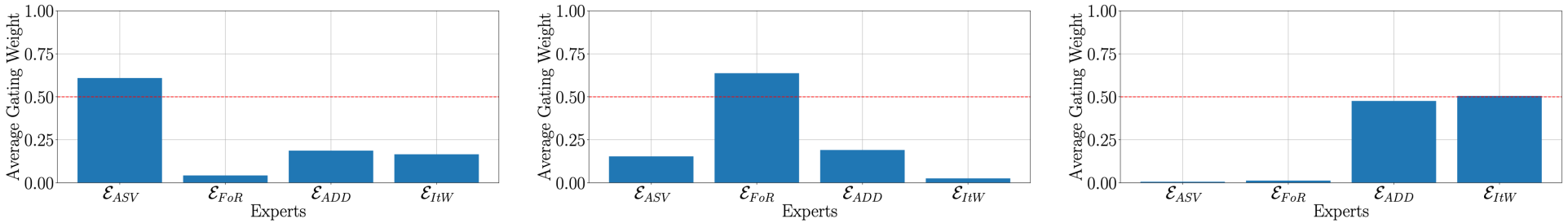
- MoE models, particularly the **Enhanced MoE**, outperform all other approaches.
- **Individual** experts' results reveal extremely **poor** generalization.
- **Effectiveness** on both **known and unseen** datasets.
- **Unusual** discrepancy between known and unknown performances due to a challenging known dataset.

	D_{ASV}		D_{FoR}		D_{ADD}		D_{ItW}		D_{PUR}		D_{TIM}		**Known**		**All**	
	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑
$\mathcal{E}_{\text{ASV}}$	**8.80**	**97.24**	23.10	86.33	47.48	53.26	43.56	59.07	33.45	72.29	11.40	94.40	30.74	73.98	27.97	77.10
$\mathcal{E}_{\text{FoR}}$	29.11	78.03	14.75	92.25	46.84	57.90	19.09	88.15	23.21	83.36	8.37	95.62	27.45	79.08	23.56	82.55
$\mathcal{E}_{\text{ADD}}$	55.62	36.39	**2.34**	61.78	35.60	69.62	63.68	32.43	24.39	47.71	98.84	0.07	39.31	50.06	46.75	52.93
$\mathcal{E}_{\text{ItW}}$	24.23	81.76	19.30	89.72	43.55	59.31	0.38	**99.89**	15.73	91.34	7.21	97.62	21.87	82.67	18.40	86.61
Ensemble	13.43	93.80	10.29	97.03	38.54	67.29	11.20	95.96	21.98	82.77	29.30	75.49	18.37	88.52	20.79	85.39
Joint Training	9.29	96.57	3.80	98.56	30.86	72.23	0.98	99.73	12.38	84.66	3.26	94.26	11.23	91.77	10.10	91.00
MoE (proposed)	9.45	96.17	2.74	**99.43**	30.87	76.04	0.55	99.83	**2.53**	**94.52**	6.98	**97.73**	**10.90**	**92.87**	**8.85**	**93.95**

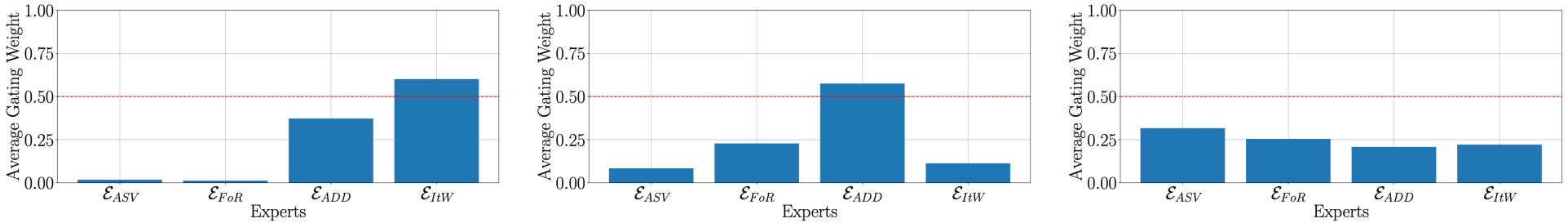
Experimental Results

Gating Network Analysis

- Analyze the **average gating weights** across various datasets
- Assess each **expert's contribution** to predictions on each domain
- Valuable insights into the **similarity between different domains**.



(a) ASVspoof 2019 (D_{ASV}) (b) FakeOrReal (D_{FoR}) (c) ADD 2022 (D_{ADD})



(d) In-the-Wild (D_{ItW}) (e) Purdue speech dataset (D_{PUR}) (f) TIMIT-TTS (D_{TIM})

Experimental Results

Contributions so far

🚀 Feasibility

MoE frameworks can be easily employed within the speech deepfake detection scenario

🌍 Performance

MoEs can significantly improve **generalization** without introducing scalability issues

🔍 Interpretability

MoEs can provide insights into domain similarities and differences

Attention-based Mixture of Experts for Robust Speech Deepfake Detection

V. Negroni, D. Salvi, A. I. Mezza, P. Bestagini, S. Tubaro – IEEE WIFS 2025

Context: SAFE Challenge 2025

Overview

Organization

3 different tasks, 90 days of time
10 teams, >900 submissions

Blind Evaluation Protocol

No data is released
Models & sources kept secret

Evaluation

Submit the detector with a *binary* output
Evaluated on a *Public* and a *Private* split
Metrics: Bal. Acc., TPR, TNR

🏁 *Let's engage with challenges more!*

Context: SAFE Challenge 2025

Tasks Structure

🤖 **Task 1: Detection of Generated Voice Audio**

21 real sources - 13 TTS models

🎛️ **Task 2: Detection of Post-processed Generated Audio**

Test how performance is affected by post-processing techniques

🧺 **Task 3: Detection of Laundered Generated Audio**

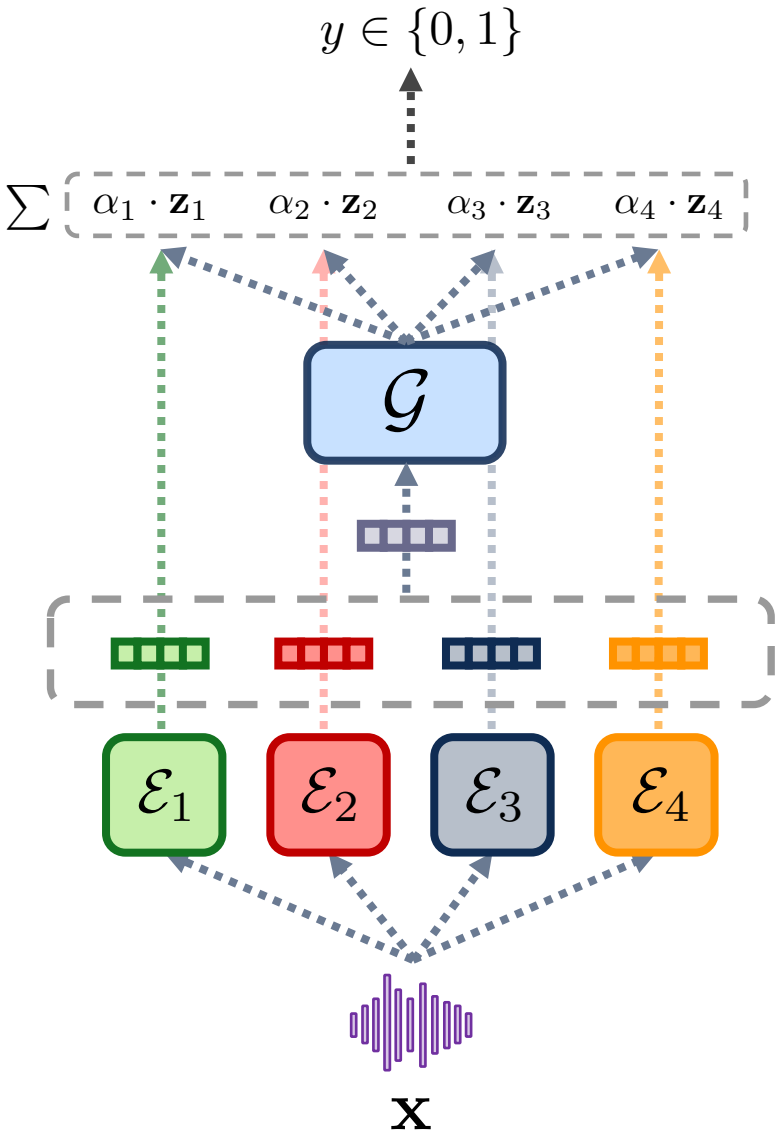
Synthetic audio is laundered to evade detection

Paradigm Shift: MELE vs MILE

Previously...

- Each expert was **trained on a different dataset**, specializing it in a specific domain
- Gating network **dynamically weights** experts at inference
  - ➡️ Strong generalization
  - ➡️ Interpretability

🤔 Can we boost performance even more?



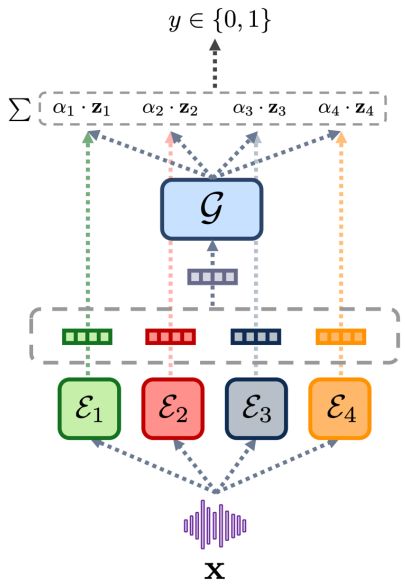
Paradigm Shift: MELE vs MILE

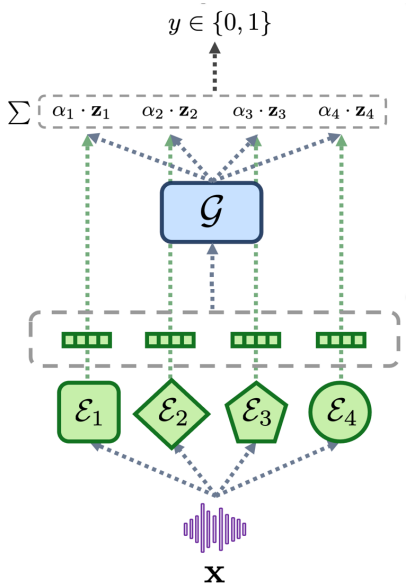
- Given **strong generalization capabilities** of MoEs, we used it as foundation for our detector for the SAFE challenge

- **Key changes**:
  - 1️⃣ **Enhanced Gating Network**
    - Transformer-based encoder
    - Captures temporal & contextual dependencies across experts
  - 2️⃣ **MELE → MILE**
    - From identical, domain-specific experts (MELE)
    - To diverse architectures trained on the same data (MILE)

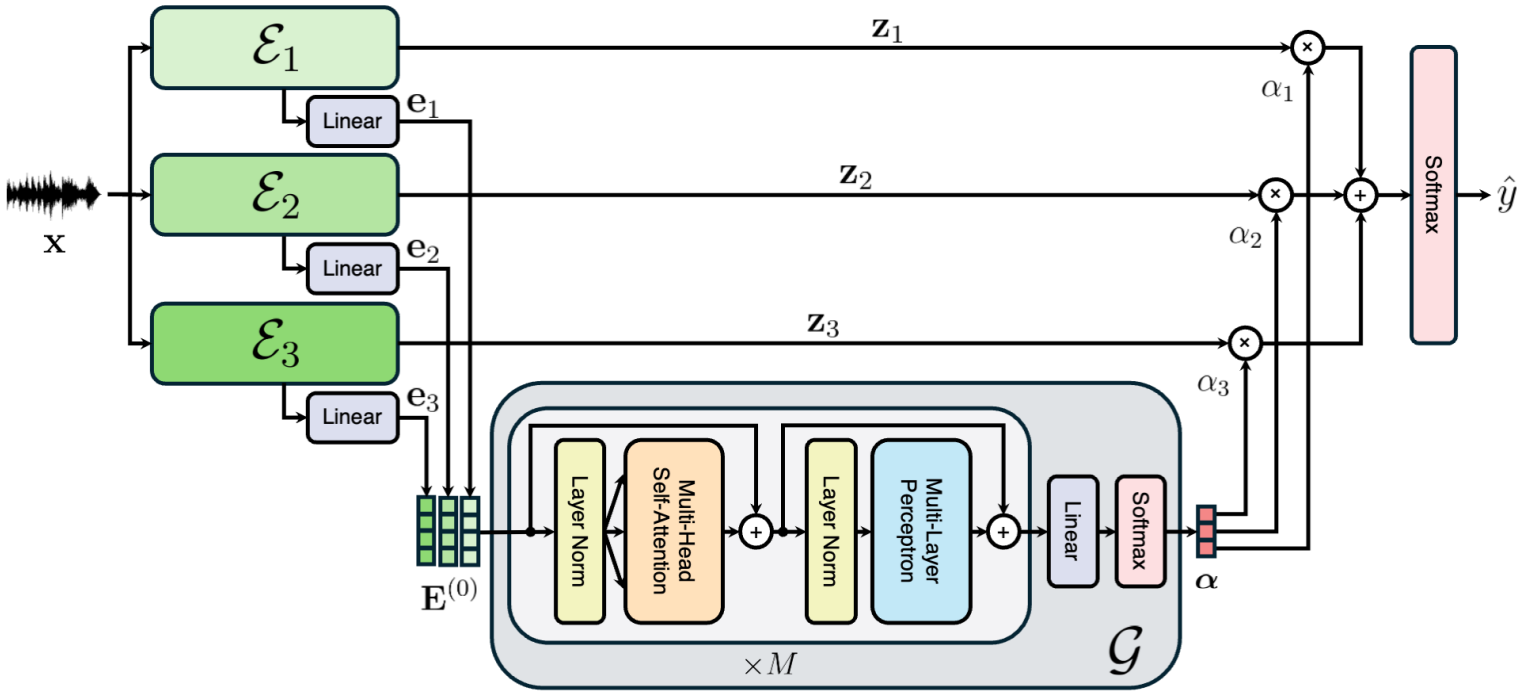
Paradigm Shift: MELE vs MILE

- In our ICASSP paper we considered a **Mixture of Explicitly Localized Experts (MELE)**
  - Identical expert architectures, each trained on **different domains** (i.e., different datasets)
- In this work we adopt a **Mixture of Implicitly Localized Experts (MILE)**
  - **Different expert architectures**, all trained on the same data
- We consider **3 experts**, each with its own architecture, and a **Transformer-based encoder for cross-expert attention** as a gating network





	MELE Mixture of Explicitly Localized Experts ICASSP 2025	MILE Mixture of Implicitly Localized Experts WIFS 2025
EXPERTS' ARCHITECTURE	**SAME**	**DIFFERENT**
EXPERTS' TRAINING SET	**DIFFERENT**	**SAME**



- **Implicit Expert Localization**: *the system stochastically partitions the data domain by leveraging the architectural biases of all its distinct deepfake detectors*
- The **gating network** is as a **transformer encoder** with **2 layers**

Evaluation Setup

Challenge Setup

TRAINING DATASETS

1. ASVspoof 2019
2. Fake-Or-Real
3. MLAAD
4. In-The-Wild
5. DiffSSD
6. LibriSpeech
7. LJspeech
8. VCTK
9. Mozilla CommonVoice

EXPERTS

1. LCNN + MelSpec
2. ResNet + MelSpec
3. ResNet + LogSpec

DATA AUGMENTATION

SpecAugment
Noise injection (RawBoost)

Paper Setup

TRAINING DATASETS

1. ASVspoof 2019
2. Fake-Or-Real
3. MLAAD
4. In-The-Wild

TEST DATASETS

1. ASVspoof 2021
2. Purdue Dataset
3. ASVspoof 5

EXPERTS

1. LCNN + MelSpec
2. ResNet + MelSpec
3. ResNet + LogSpec

DATA AUGMENTATION

N/A

Experimental Results

Public Evaluation Set – Task 1

Model	Generated Acc.	Pristine Acc.	Balanced Acc.
MoE (Proposed)	0.859	**0.933**	**0.896**
LCNN + MelSpec	**0.864**	0.799	0.832
ResNet + MelSpec	0.836	0.866	0.851
ResNet + LogSpec	0.837	0.916	0.877

Experimental Results

Public + Private Evaluation Set – Task 1

REAL DATA

	ISPL	VIPER-PURDUE	ISSF	ANON-PEKING	JAIST-HIS	DMF
Mandarin Podcast 1	**0.90**	0.87	0.87	0.89	0.68	0.61
Fleurs German	**0.90**	0.86	0.58	0.69	0.72	0.83
VSP Semi-professional	**0.90**	0.80	0.87	0.89	0.71	0.65
youtube phonecall	**0.86**	0.78	0.84	0.84	0.73	0.64
VSP Documentary	**0.87**	0.84	0.87	0.81	0.66	0.69
Arabic Speech Corpus	0.62	0.38	0.39	0.59	**0.68**	0.43
High Quality Podcasts	**0.85**	0.74	0.83	0.71	0.71	0.52
Japanese Shortwave	0.90	0.65	0.48	0.84	0.68	**0.92**
Conference	**0.88**	0.79	0.86	0.63	0.63	0.48
English Podcast	**0.89**	0.78	0.86	0.78	0.73	0.48
Fleurs English	**0.90**	0.84	0.71	0.56	0.73	0.89
Dipco	0.88	0.87	0.84	0.54	0.41	**0.90**
Digitized Cassette	**0.90**	0.87	0.69	0.89	0.73	0.87
Librivox	0.86	0.83	0.69	**0.86**	0.73	0.65
Old Radio	**0.90**	0.76	0.83	0.65	0.72	0.51
phone home	**0.89**	0.69	0.53	0.69	0.73	0.83
Russian Audiobook	**0.89**	0.56	0.52	0.48	0.46	0.53
Mandarin Podcast 2	0.90	0.87	0.86	0.77	0.73	**0.91**
VSP Home Mic	0.88	0.83	0.87	**0.88**	0.73	0.59
Radio Drama	**0.89**	0.82	0.81	0.78	0.73	0.54
VSP Professional	**0.89**	0.76	0.84	0.79	0.73	0.67
accuracy	**0.87**	0.77	0.75	0.74	0.68	0.67

FAKE DATA

	ISPL	VIPER-PURDUE	ISSF	ANON-PEKING	JAIST-HIS	DMF
elevenlabs	**0.97**	0.88	0.64	0.82	0.58	0.71
fish	**0.94**	0.79	0.81	0.81	0.91	0.62
hierspeech	0.76	0.62	**0.87**	0.62	0.86	0.63
kokoro	**0.98**	0.90	0.87	0.84	0.80	0.73
parler	**0.97**	0.74	0.86	0.80	0.48	0.66
seamless	0.93	0.90	0.76	0.86	**0.94**	0.75
style	0.82	0.71	**0.86**	0.46	0.46	0.75
cartesia	**0.91**	0.84	0.47	0.81	0.67	0.72
edge	0.68	0.83	0.82	**0.85**	0.73	0.73
f5	**0.90**	0.67	0.63	0.72	0.50	0.56
metavoice	**0.88**	0.80	0.58	0.76	0.56	0.71
openai	0.86	0.85	0.87	0.75	**0.92**	0.72
zonos	**0.71**	0.48	0.65	0.56	0.45	0.52
accuracy	**0.87**	0.77	0.75	0.74	0.68	0.67

Real Accuracy: 0.87
Fake Accuracy: 0.87

Experimental Results

Tasks 2 and 3

TASK 2 – Postprocessing

	ISPL	VIPER-PURDUE	ISSF	ANON-PEKING	JAIST-HIS	DMF
aac 16k	0.77	0.72	0.80	**0.82**	0.62	0.59
encodec	**0.92**	0.66	0.76	0.85	0.82	0.58
focalcodec	**0.95**	0.63	0.84	0.83	0.87	0.57
mp3-aac-mp3 16k	0.67	0.79	0.83	**0.83**	0.62	0.59
mp3-aac 16k	0.73	0.78	0.81	**0.82**	0.60	0.59
mp3 16k	**0.83**	0.78	0.79	0.78	0.60	0.58
mp3 VBR	**0.84**	0.74	0.79	0.76	0.68	0.58
noise	**0.70**	0.45	0.52	0.54	0.61	0.58
opus 16k	0.69	0.73	0.73	**0.83**	0.77	0.61
phone audio	0.71	0.61	**0.85**	0.79	0.64	0.56
pitch shift	0.85	0.70	0.81	0.77	**0.87**	0.60
resample down	0.77	0.63	**0.77**	0.76	0.67	0.57
resample up	**0.82**	0.61	0.76	0.76	0.69	0.57
semanticodec	0.83	0.67	0.74	0.85	**0.88**	0.58
snac	0.78	0.65	0.73	0.81	**0.87**	0.60
speech filter	0.76	0.61	**0.79**	0.67	0.67	0.45
time stretch	0.85	0.69	0.85	0.78	**0.88**	0.61
vorbis 16k	**0.86**	0.67	0.75	0.75	0.69	0.57
accuracy	**0.80**	0.67	0.77	0.78	0.72	0.58

TASK 3 – Laundering

	ISPL	VIPER-PURDUE	ISSF	ANON-PEKING	JAIST-HIS	DMF
car	**0.67**	0.50	0.51	0.53	0.60	0.57
played	0.62	0.54	0.67	0.54	**0.67**	0.54
played reverb car	0.58	0.69	0.67	0.49	**0.71**	0.35
reverb	**0.75**	0.46	0.58	0.70	0.64	0.58
accuracy	**0.66**	0.55	0.61	0.57	0.66	0.51

We obtained the **best performance** also in Tasks 2-3, even if our solution was not tailored for them

Experimental Results

Further Experiments and Validation

- Extended evaluation of the proposed architecture in the paper
- **Attention-based MILE MoE**, outperformed alternative architectures, showing stronger detection on both **in-domain** and **out-of-domain** data

	ASVspoof 2019	FakeOrReal	In-the-Wild	MLAAD	Purdue	ASVspoof 2021	ASVspoof 5	Known	Unknown	Overall
$\mathcal{E}_{\text{ASV19}}$	9.11	5.70	39.08	43.78	33.84	31.43	29.43	24.42	31.44	27.43
$\mathcal{E}_{\text{FOR}}$	27.90	6.05	20.44	46.49	11.53	34.29	35.54	25.22	27.97	26.40
$\mathcal{E}_{\text{ITW}}$	21.65	2.03	**0.40**	35.31	13.61	43.64	37.13	14.85	30.94	21.75
$\mathcal{E}_{\text{MLA}}$	27.97	61.84	23.91	1.56	60.77	60.91	30.35	28.82	50.67	38.18
MELE-Cat	7.37	2.69	1.17	1.24	**3.88**	31.82	**15.00**	3.12	16.31	8.77
MELE-Att	**7.25**	**1.90**	1.10	**1.02**	4.53	**29.87**	15.34	**2.82**	**15.75**	**8.36**

	ASVspoof 2019	FakeOrReal	In-the-Wild	MLAAD	Purdue	ASVspoof 2021	ASVspoof 5	Known	Unknown	Overall
MILE-Att	8.38	0.04	0.57	**0.04**	6.49	**17.27**	17.69	**2.26**	**14.15**	**7.35**
MELE-Att	**7.25**	1.90	1.10	1.02	**4.53**	29.87	**15.34**	2.82	15.75	8.36
MILE-Cat	8.86	**0.00**	**0.37**	0.07	10.68	18.18	18.72	2.32	18.09	9.08

Outcomes and Limitations

- MoEs hold more capacity and can be significantly more robust than individual detectors
- Different theorethical frameworks can offer equal valuable contribution to SDD
- Attention in the gating mechanism can make a difference

- MoEs' flexibility makes systematic ablations difficult
- We likely explored only ~1% of what's possible in SDD with MoEs
- We lost something nowadays we cannot afford to sacrifice anymore: *interpretability*

Future Directions on MoEs

🤖 Include **SSL** models in the recipe

🧩 Systematically evaluate expert **combinations and training** strategy

- Same **experts**, different **feature sets**
- Different experts, same feature set
- Expert **pre-training** vs no pre-training
- Shared **loss** with the gate or distinct loss functions

Future Directions on MoEs

- 👥 Systematic investigation of the **optimal number** of experts
  - Explore sequential addition
  - Monitor complexity vs. gain
- 🧠 Investigate potential in **continual learning** scenarios
  - Adaptive MoEs with experts pruning or merging
- 🚦 Explore novel ***hard*** **gating** mechanisms
  - Dynamic expert routing varying with computational budget
  - Confidence-based or context-based routing strategies

Future Directions on MoEs

🦖 Explore the potential of **MoE *layers*** for large heavy models

- Evaluate efficiency (e.g., sparse activation, routing constraints)
- Investigate routing patterns to gain interpretable insights

🔎 **Interpretability** of MoE Decisions

- Weight analysis can also be valuable **in a MILE** configuration!
- Carefully choose experts kind to later **draw cues** on what led to the final prediction based on their **characteristics**
- Deliberately design experts with known **inductive biases**

Feel free to contact me!

viola.negroni@polimi.it