Improving Out-of-Distribution Detection via Test-Time Augmentation

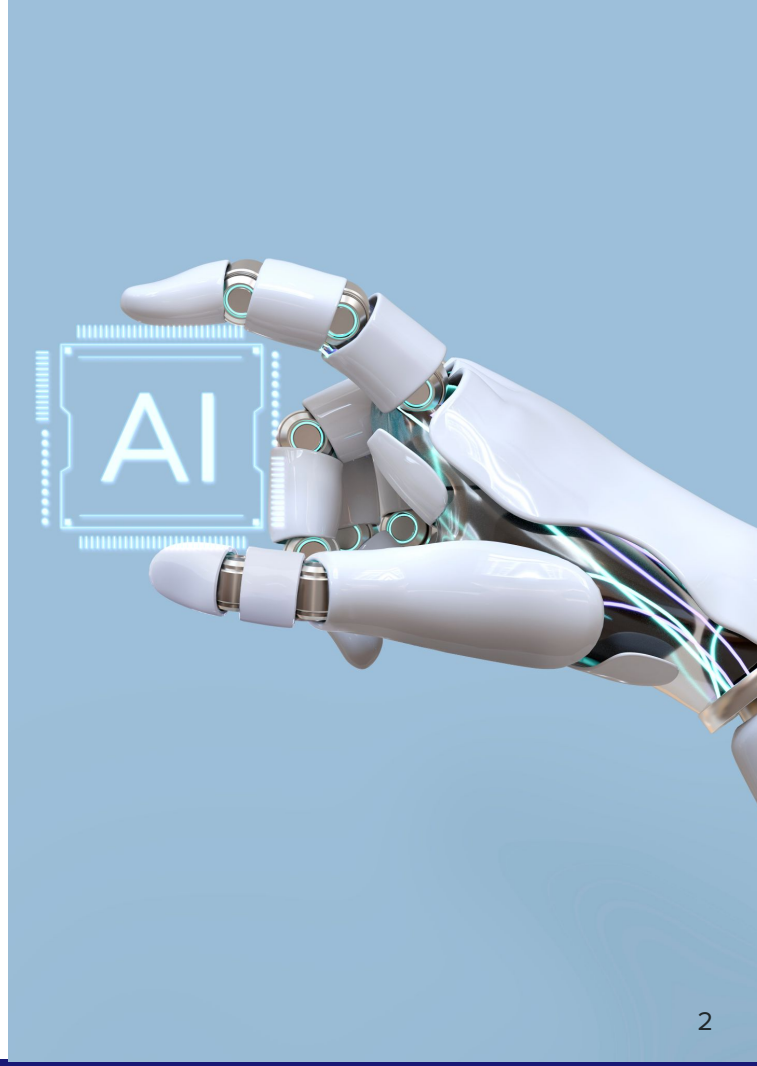
Dr. Ing. Imanol Allende
Nicholas Mc Guire
Javier del Campo
Dr. Carles Hernández

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

OpenTech

Context

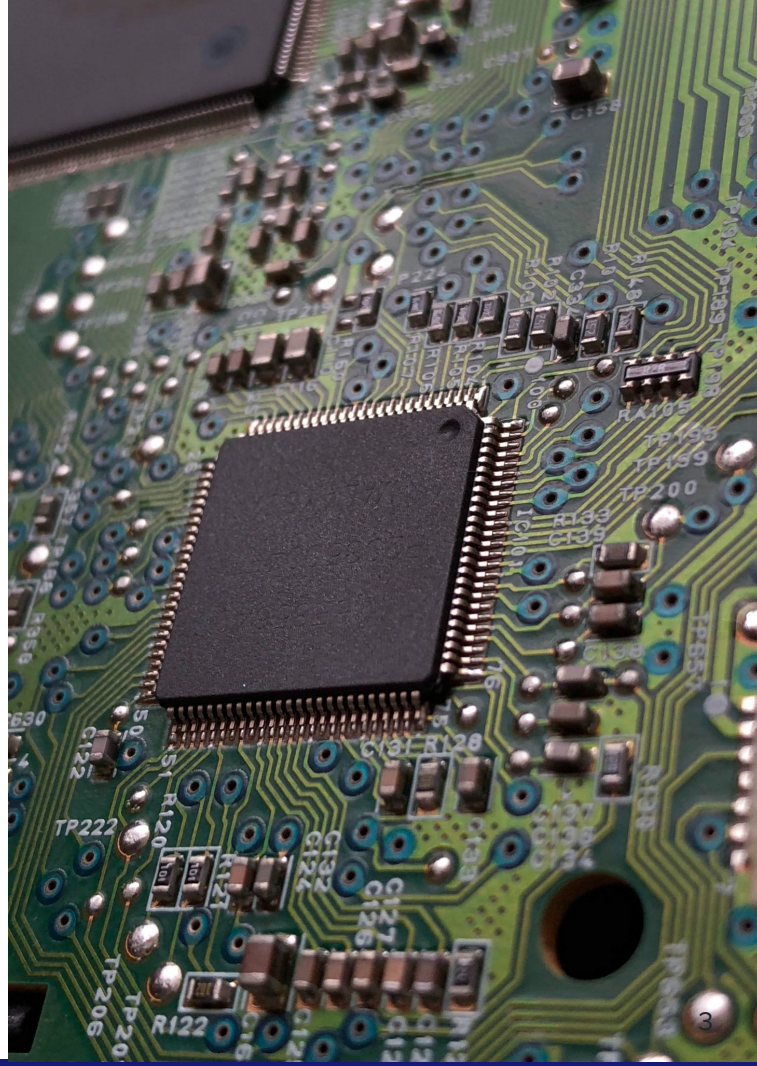
- **Next-generation** critical systems.
  - Deep Learning.
  - Security.
  - High-performance.



Context

- **Next-generation** critical systems.
  - Deep Learning.
  - Security.
  - High-performance.

- **AI** in **safety-related** systems.
  - Enable reliable real-world decision-making.
  - Dealing with uncertainty.



Context

- **Next-generation** critical systems.
  - Deep Learning.
  - Security.
  - High-performance.

- **AI** in **safety-related** systems.
  - Enable reliable real-world decision-making.
  - Dealing with uncertainty.

- **Need to deal with unknown scenarios**
  - Out-of-Distribution (OOD) Detection

Context

SafeComp 2024 Position Paper:

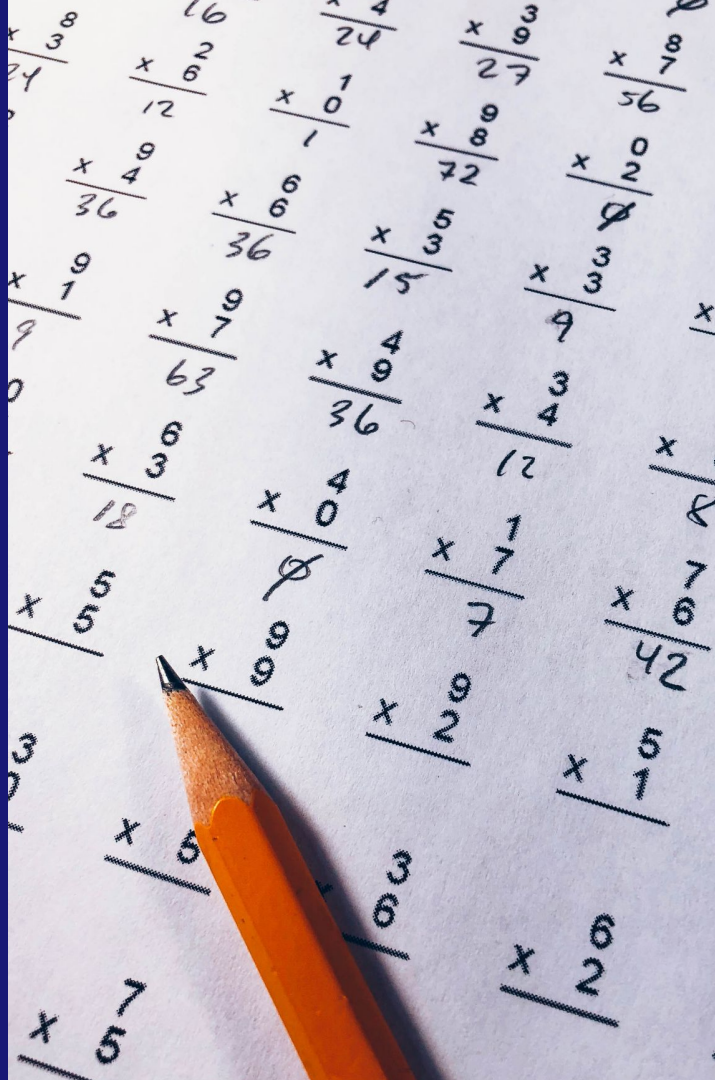
Herds of Dumb Models: A New Approach Towards Reliable and Safe AI

Nicholas Mc Guire*†, Imanol Allende*†, Carles Hernández ‡
*OpenTech EDV Research GmbH, Bullendorf, Austria
† Open Source Automation Development Lab (OSADL), Germany
‡ Universitat Politécnica de Valéncia, Spain

https://www.safecomp2024.unifi.it/upload/sub/other/position%20papers/SAFECOMP2024_paper_108.pdf

Statement of the Problem

1. **Dealing with uncertainty**
2. **Out-of-Distribution Detection**



Why Out-of-Distribution Detection Matters?

- Safety-critical domain seeks to deploy AI model.
- **AI models must reliably identify** when they encounter **unfamiliar data to prevent catastrophic failures.**
- **Out-of-Distribution (OOD)** inputs **deviate from training distribution**.
- **OOD** can cause **unpredictable model behaviour**.
- We need the capacity of saying: I do not know.

Out-of-Distribution detectors

- **Detectors** derive a **scoring function** from a trained network.
- The idea is that **In-Distribution (ID) and Out-of-Distribution (OOD)** samples exhibit a **statistically distinct scores**.
- Different approaches: parametric vs non-parameters, last-layer vs penultimate-layer

OOD Detectors limitations

- **Existing methods struggle** with subtle distribution shifts and complex real-world noise.

- **Robust OOD detection is essential** for functional safety in AI systems

Test-Time Augmentation Background

- **Traditional Data Augmentation**:
  - Applied during training.
  - Diversifies training data.
  - Boost prediction accuracy.
  - Improves model generalization.

- **Test-Time Augmentation:**
  - Applied during inference.
  - Multiple versions of input.
  - Aggregates predictions.
  - Captures model uncertainty.

TTA traditionally improves prediction accuracy, but we hypothesize that the uncertainty signal it captures can be repurposed for OOD detection.

Our Contribution

- Examine state-of-the-art **OOD detectors** with **TTA.**
- **Not proposing new** OOD detector.
- Using TTA as a **modular plug-in.**

Based on psychologist

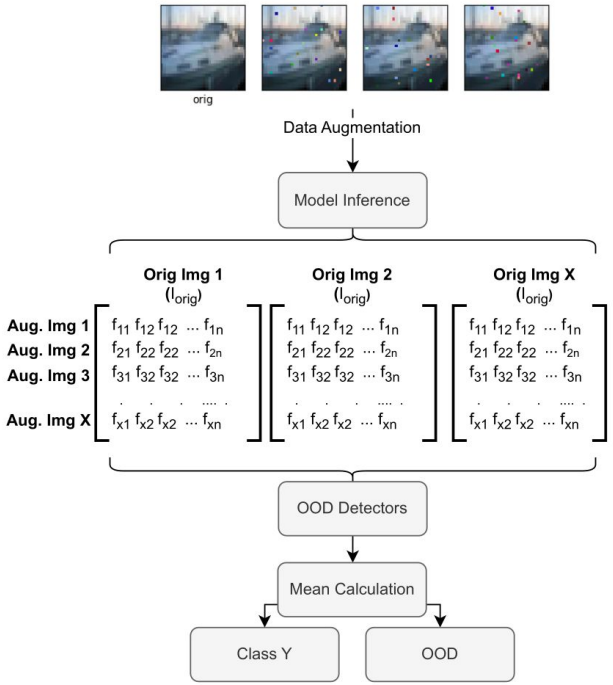
- Psychological tests use **varied phrasing of similar questions** to enhance accuracy, completeness, and reliability.
- Our proposed method **evaluates the consistency of detector results** when subjected to sets of transformed inputs.
- Ensures that **responses** are **not influenced** by the **interpretation** of a **single question** (or input image, in our case).
- From a functional safety point of view, **diversity of inputs** is particularly **valuable** as it **reduces the likelihood of common mode failures**.

TTA + OOD

- As psychologists ask the **same question, rephrased differently**,
- We evaluate the model with the **same original image** but **replicated and transformed differently** (i.e., Test-Time Augmentation (TTA)).
- Input transformations cannot be either **too small**, as this would only result in an **increase of resource usage without any benefit**,
- Or **too much**, as this would **only create a significant reduction of the classification capabilities** of the model.

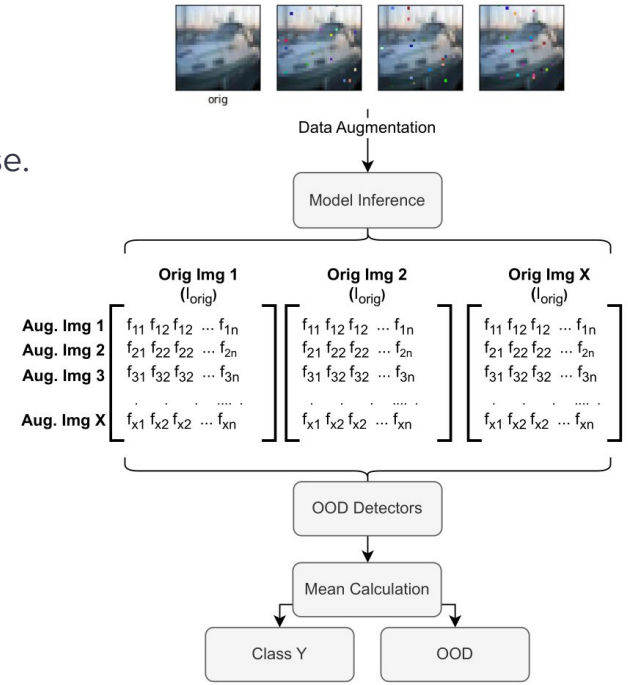
Proposed Approach

1. Image Replication
2. Random Transformations
3. Model Inference
4. OOD Detection
5. Mean Calculation



Drawbacks

- Mean calculation as first step to validate the approach.
- Increase in processing time and hardware resources use.
- No timing and performance analysis.



Experiment

1. **SOTA OOD Detectors**
2. **Test Time Augmentation**
3. **Different setups**

Experimental Setup

- **OOD Detectors:**
  - Energy-based detection
  - NNGuide (Nearest Neighbor Guidance)
  - MSP (Maximum Softmax Probability)
  - ViM (Virtual-logit Matching)
  - MaxLogit
  - SSD (Self-Supervised Detection)
  - Mahalanobis distance
  - KNN (K-Nearest Neighbors)

- **Metrics**
  - FPR95: False positive rate at 95% TPR
  - AUROC: Area under ROC curve

- **Model Architectures:**
  - ResNet18/50
  - Vision Transformer (ViT-B16)
  - RegNet-Y-16GF
  - MobileNet-V2

Experimental Setup

- **Model:** ResNet18
- **ID Datasets:** CIFAR-10
- **OOD Datasets:**
  - Street View House Numbers (SVHN)
  - Describable Textures Dataset (DTD)
  - Places365
  - iSUN
  - Large-scale Scene UNderstanding (LSUN)
  - Cifar-100
- **Augmentation techniques:**
  - Random Crop
  - Random Horizontal Flip
  - Random Rotation

Results: CIFAR-10 Experiments

- Consistent Improvements Across Detectors
- TTA with 16 replicas shows significant performance gains over baseline (no augmentation)

Table 1: FPR95 values for different detectors, datasets and number of transformed images (i.e., no augmentation vs 8 and 16). Results obtained with ResNet18 model trained with Cifar-10 dataset.

	SVHN			iSUN			LSUN			Textures			Places365		
Detectors	No Aug	8	16	No Aug	8	16	No Aug	8	16	No Aug	8	16	No Aug	8	16
Energy	12.73	10.18	9.78	28.69	24.17	23.23	16.43	11.43	10.70	28.56	18.83	15.92	20.82	3.96	1.88
NNguide	15.32	12.23	11.56	27.04	21.62	21.11	18.09	12.37	11.31	27.64	18.39	15.32	21.81	4.82	2.16
ViM	32.39	22.27	21.01	49.95	40.27	38.72	37.75	27.87	25.23	44.72	29.91	25.55	42.33	16.53	9.78
MSP	14.12	10.90	10.38	30.89	25.88	24.63	17.98	12.53	11.70	29.77	19.17	16.21	22.49	4.32	1.99
MaxLogit	13.72	8.80	8.59	75.55	68.72	69.43	43.60	39.69	39.23	42.82	25.62	23.76	46.60	20.00	15.56
SSD	15.47	13.73	13.56	94.38	93.68	93.50	49.25	58.13	58.26	38.35	29.96	28.03	53.24	33.97	30.68
Mahalanobis	13.40	11.72	11.54	90.12	88.39	88.58	44.22	50.27	50.10	33.87	24.47	22.25	46.32	23.80	19.48
KNN	22.79	18.51	18.03	33.96	29.83	28.91	28.61	24.01	22.91	35.78	26.38	22.77	31.60	11.52	6.63

Results: CIFAR-10 Experiments

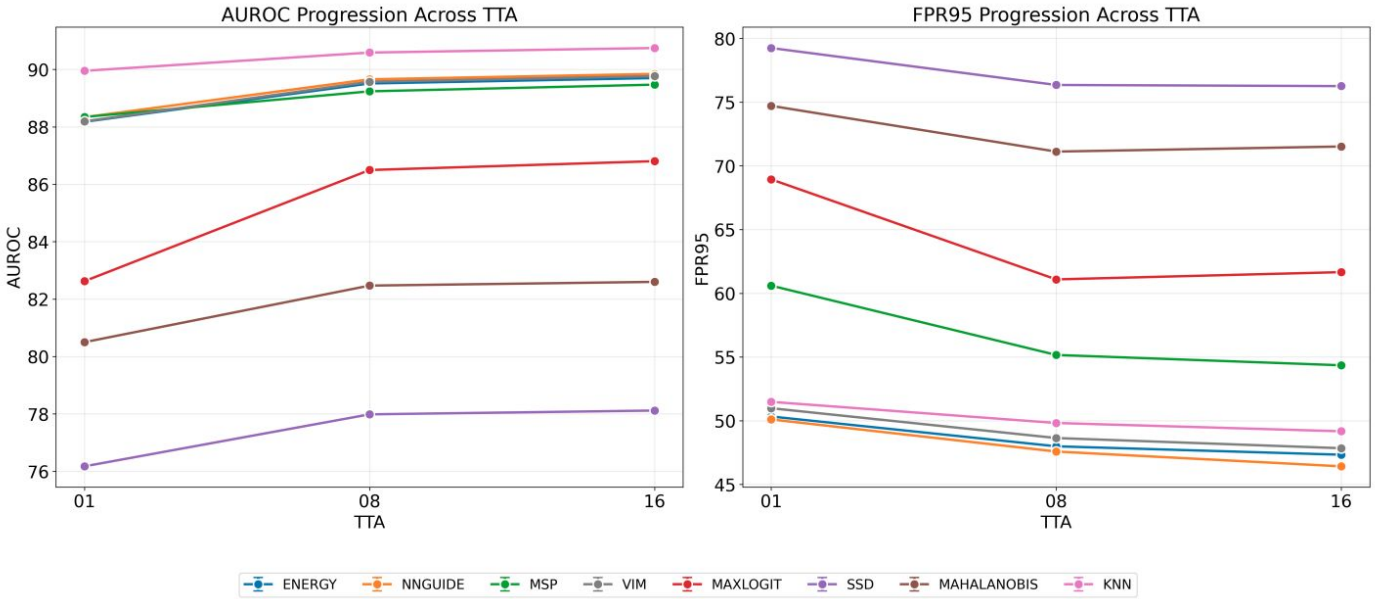
- Consistent Improvements Across Detectors
- TTA with 16 replicas shows significant performance gains over baseline (no augmentation)

Table 2: AUROC values for different detectors, datasets and number of transformed images (i.e., no augmentation vs 8 and 16). Results obtained with ResNet18 model trained with Cifar-10 dataset.

	SVHN			iSUN			LSUN			Textures			Places365		
Detectors	No Aug	08	16	No Aug	08	16	No Aug	08	16	No Aug	08	16	No Aug	08	16
Energy	97.54	98.02	98.11	94.47	95.57	95.69	96.91	97.81	97.90	94.52	96.76	96.98	96.16	98.40	98.57
NNguide	97.02	97.66	97.76	94.85	96.00	96.12	96.68	97.70	97.80	94.76	96.90	97.11	96.01	98.26	98.44
ViM	95.39	96.38	96.58	92.29	93.29	93.49	94.64	95.75	95.99	92.99	95.73	96.12	93.94	97.13	97.53
MSP	97.40	97.96	98.05	94.34	95.48	95.60	96.76	97.72	97.81	94.44	96.79	97.01	96.02	98.40	98.57
MaxLogit	97.42	98.34	98.47	83.52	87.84	88.20	92.11	94.37	94.68	91.76	96.29	96.71	91.30	96.84	97.29
SSD	96.99	97.32	97.42	69.02	71.37	71.50	88.62	87.97	88.17	91.99	95.02	95.50	88.86	95.19	95.85
Mahalanobis	97.53	97.84	97.92	77.92	80.61	80.75	91.66	91.81	91.98	93.70	96.29	96.66	91.71	96.58	97.05
KNN	96.32	96.89	97.00	94.26	95.11	95.24	95.57	96.23	96.35	94.51	96.16	96.40	95.03	97.27	97.50

Results: CIFAR-10 Experiments

The model achieves: 94.24% accuracy (no augmentation), 94.76% (8 TTAs) and 94.82% (16 TTAs)



Experimental Setup

- **Model:** ResNet50
- **ID Datasets:** ImageNet
- **OOD Datasets:**
  - Street View House Numbers (SVHN)
  - Describable Textures Dataset (DTD)
  - Places365
  - SUN
  - iNaturalist
- **Augmentation techniques:**
  - Resize
  - Random Crop
  - Random Horizontal Flip
  - Random Rotation

Results: ImageNet Experiments

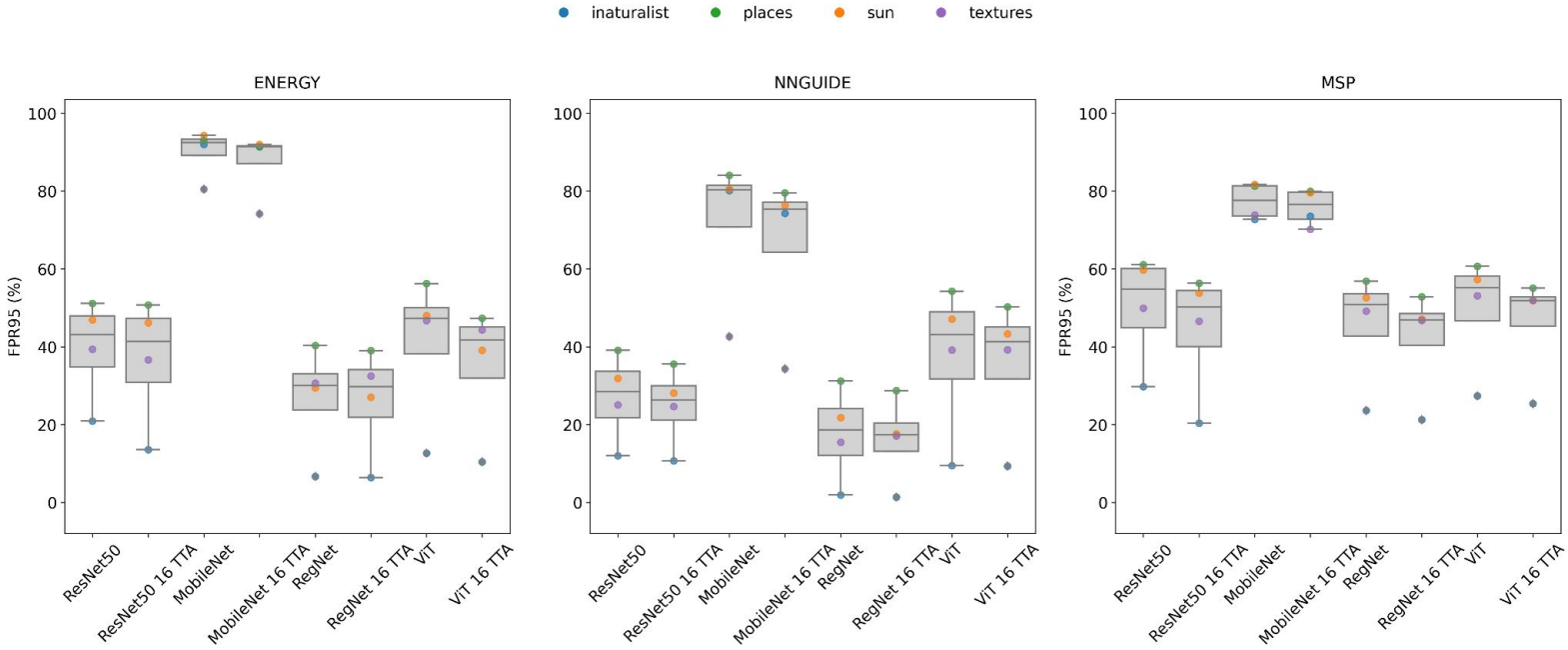
- Consistent Improvements Across Detectors
- TTA with 16 replicas shows significant performance gains over baseline (no augmentation)

Table 3: Mean FPR95 and AUROC values across all datasets.

Detector	No Augmentation		4 TTA		16 TTA	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Energy	39.58	90.41	39.08	90.64	36.76	91.10
NNGuide	27.00	92.68	25.69	92.82	24.74	93.05
ViM	50.12	86.87	46.31	88.00	44.27	88.55
MSP	42.28	90.04	41.42	90.27	39.63	90.71
MaxLogit	30.30	92.59	32.24	92.42	29.90	92.84
SSD	41.73	91.00	44.60	90.56	42.29	91.14
Mahalanobis	46.39	90.18	48.17	89.70	45.59	90.25
KNN	43.29	89.63	46.01	89.97	45.32	90.17

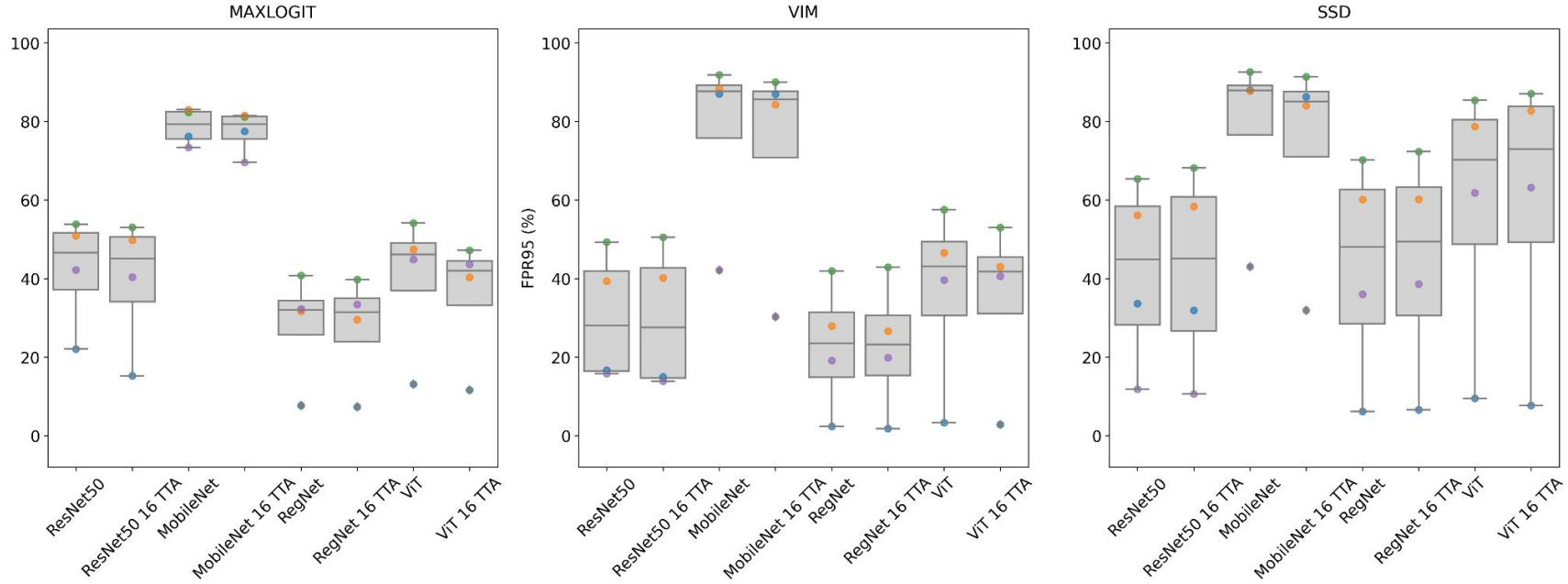
Other Model Architectures

- Consistent Improvements Across Detectors
- TTA with 16 replicas shows significant performance gains over no augmentation



Other Model Architectures

- Consistent Improvements Across Detectors
- TTA with 16 replicas shows significant performance gains over no augmentation



Conclusions

1. **Summary**
2. **Critical review**
3. **Future lines**

Conclusions: Summary

- Validated Across
  - 8 state-of-the-art detectors
  - 5 model architectures
  - Diverse training datasets and OOD datasets
  - Near and far OOD scenarios

- Performance gains
  - Consistent FPR95 reductions
  - Improved AUROC scores
  - Maintained classification accuracy
  - Statistical significance confirmed

Conclusions: Critical review

— Mean calculation is too basic.

— More sophisticated statistics needed.

— Performance impact.

— Result still fall short for functional safety.

Conclusions: Future lines

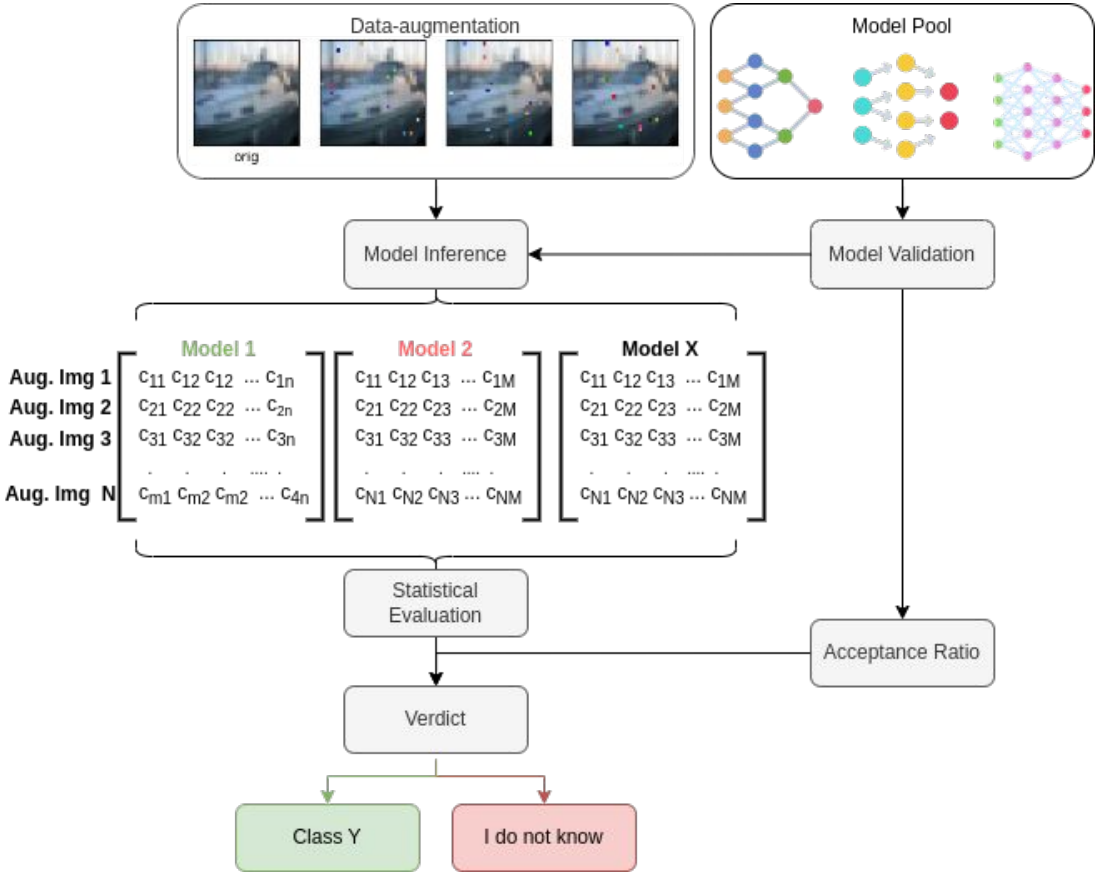
- **Voting architecture** of models.
- **Combine OOD detectors** and **models.**
- Aim is to **reduce the bias.**
- **Heterogeneous redundant** architecture.

Conclusions: Future lines

Herds of Dumb Models: A New Approach Towards Reliable and Safe AI

Nicholas Mc Guire*†, Imanol Allende*†, Carles Hernández ‡
*OpenTech EDV Research GmbH, Bullendorf, Austria
† Open Source Automation Development Lab (OSADL), Germany
‡ Universitat Politécnica de Valéncia, Spain

Conclusions: Future lines



Questions

Nicholas Mc Guire <safety@osadl.org>
Imanol Allende <imanol.allende@codethink.co.uk>
Carles Hernández <carherlu@upv.es>