2nd World Conference on eXplainable Artificial Intelligence

Can I trust my anomaly detection system? A case study based on eXplainable AI.

UNIVERSITÀ DI TORINO

Elvio G. Amparore
elviogilberto.amparore@unito.it

Muhammad Rashid
muhammad.rashid@unito.it

RuleX

Enrico Ferrari
enrico.ferrari@rulex.ai

Damiano Verda
damiano.verda@rulex.ai

Topics of the presentation

- Case study of Anomaly Detection in Industrial Quality Control System
- Use of Generative AI for Anomaly Detection
- Transparency & Trustworthiness in Anomaly Detection Systems
  - Review an explainable AD system architecture* that combines VAE-GAN models with the LIME and SHAP explanation methods.
  - Quantify the AD system efficacy using anomaly scores
  - Use XAI methods to determine if anomalies are indeed detected for the right reason, improving the framework of Ravi et al*.

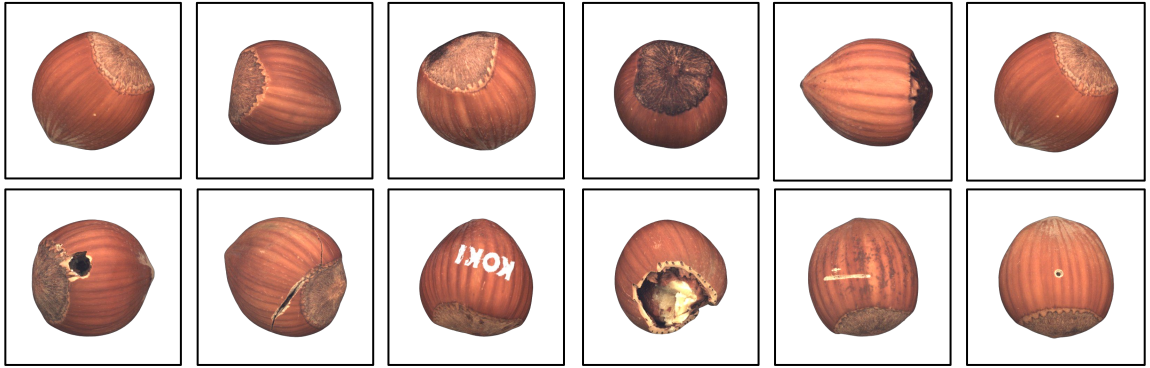
* Ravi, A., Yu, X., Santelices, I., Karray, F., & Fidan, B. (2021, August). **General frameworks for anomaly detection explainability: comparative study**. In *2021 IEEE International Conference on Autonomous Systems (ICAS)* (pp. 1-5). IEEE.

Defining Defective & Non-defective

Consider an industrial quality control system use case.

- Non defective products are common and easy to capture and describe
- Defective products are rare and unpredictable

→ setup for **anomaly detection**.



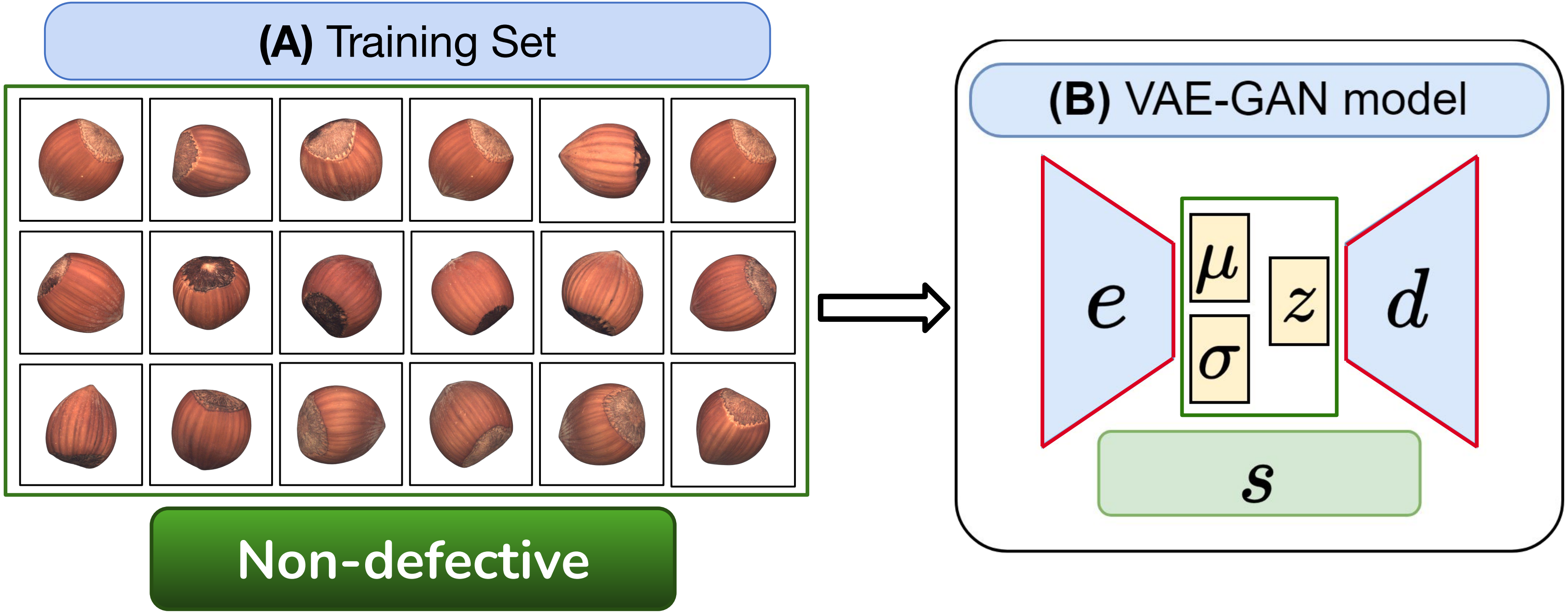
Non-defective:

Defective:

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., & Steger, C. (2021). The **MVTec** anomaly detection dataset: **a comprehensive real-world dataset for unsupervised anomaly detection**. *International Journal of Computer Vision*.

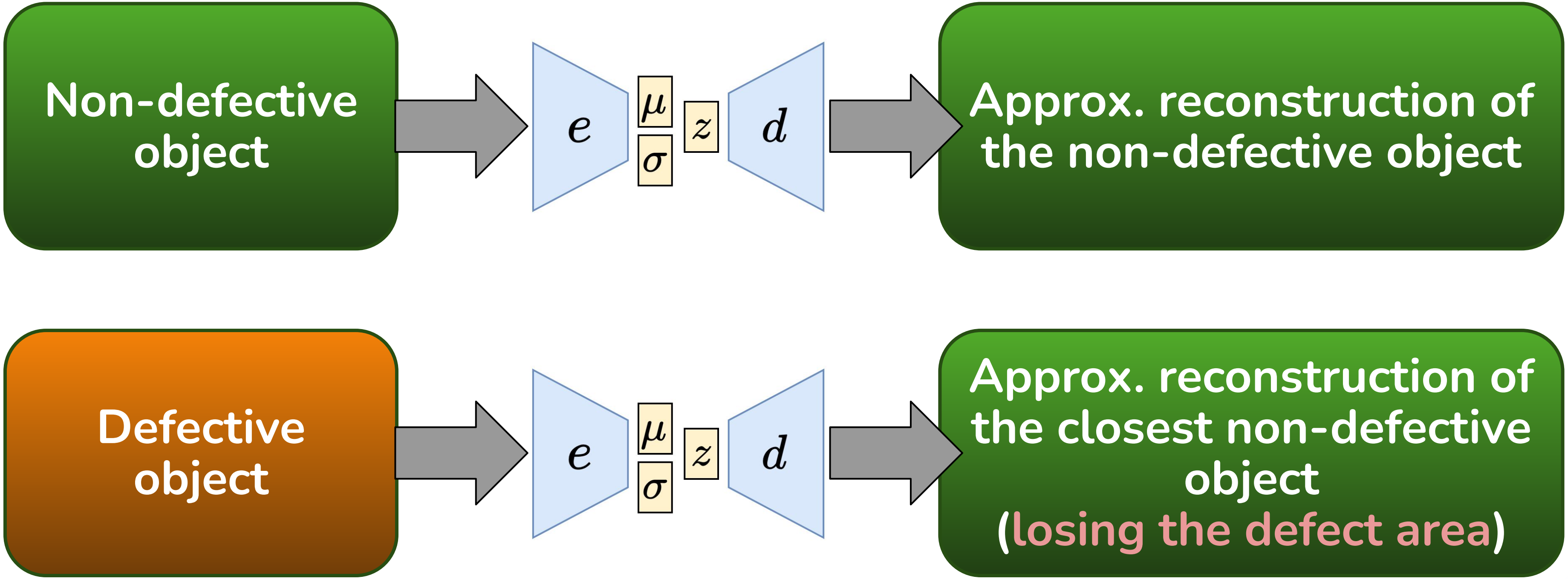
VAE-GAN model for non-defective instances

Training the Variational Auto-Encoder



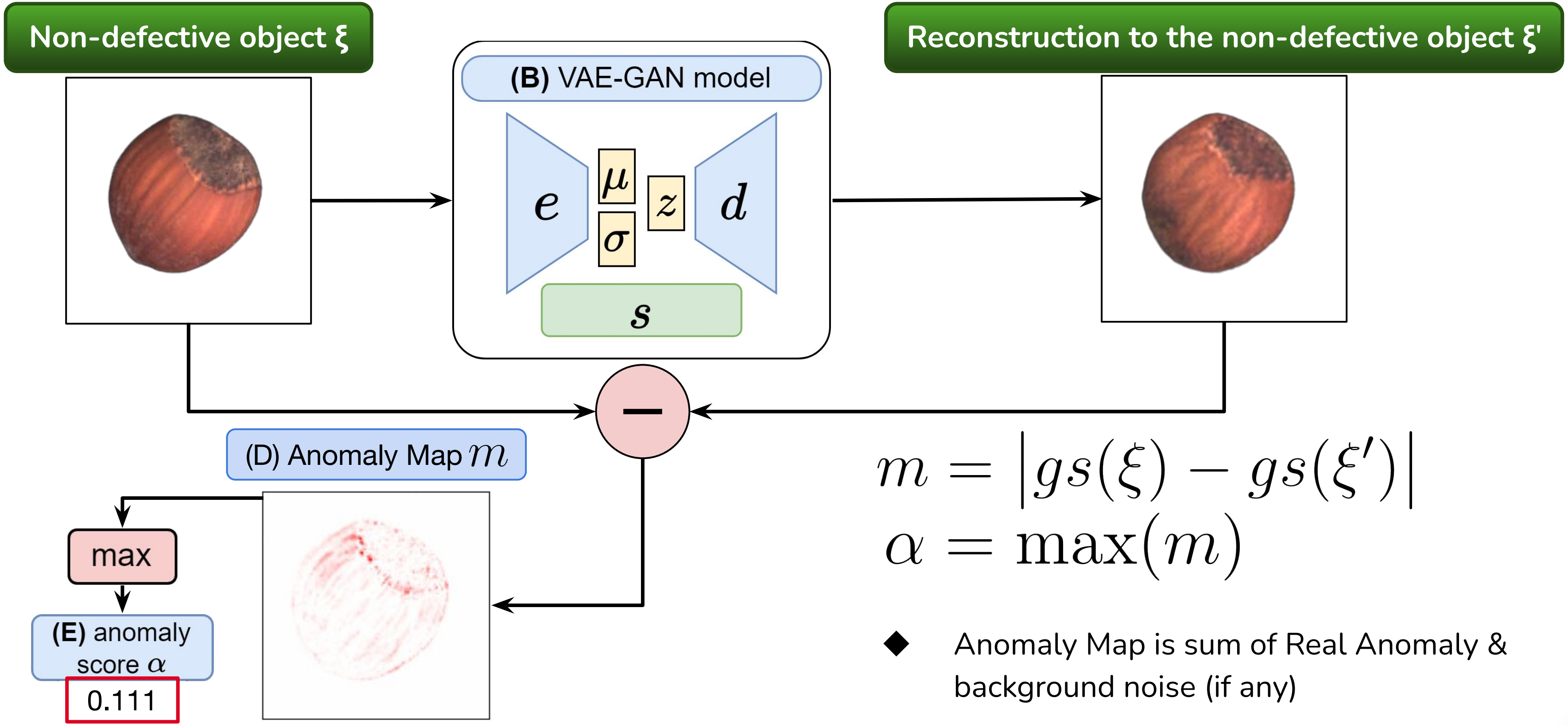
Anomaly detection with VAE-GAN

Anomaly Detection with Variational Auto-Encoder


Non-defective object
e
μ
σ
z
d
Approx. reconstruction of the non-defective object
Defective object
e
μ
σ
z
d
Approx. reconstruction of the closest non-defective object
(losing the defect area)

Anomaly detection with generative AI

Pipeline with a Variational Auto-Encoder

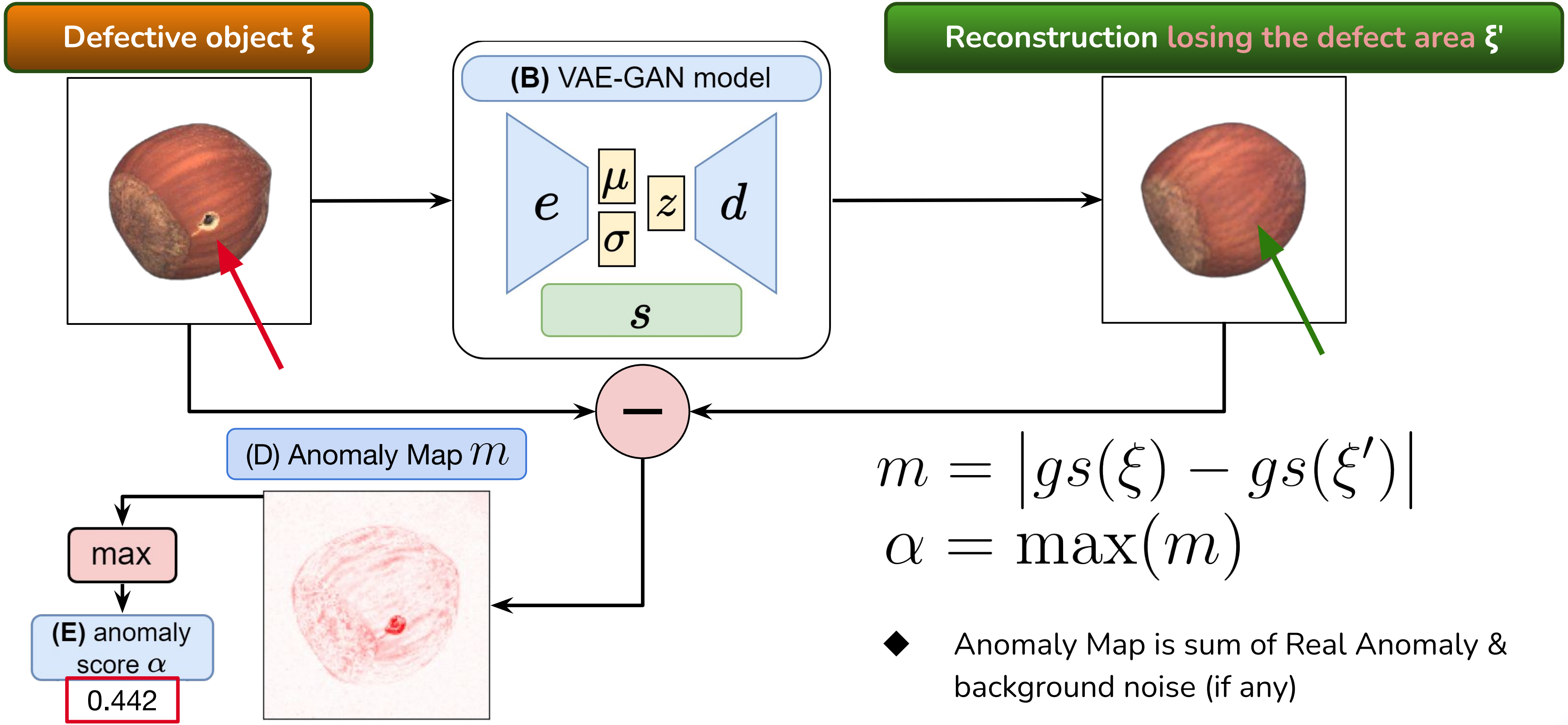


$$m = \left| gs(\xi) - gs(\xi') \right|$$
$$\alpha = \max(m)$$

- Anomaly Map is sum of Real Anomaly & background noise (if any)

Anomaly detection with generative AI

Pipeline with a Variational Auto-Encoder



$$m = \left| gs(\xi) - gs(\xi') \right|$$
$$\alpha = \max(m)$$

- Anomaly Map is sum of Real Anomaly & background noise (if any)

Anomaly detection threshold

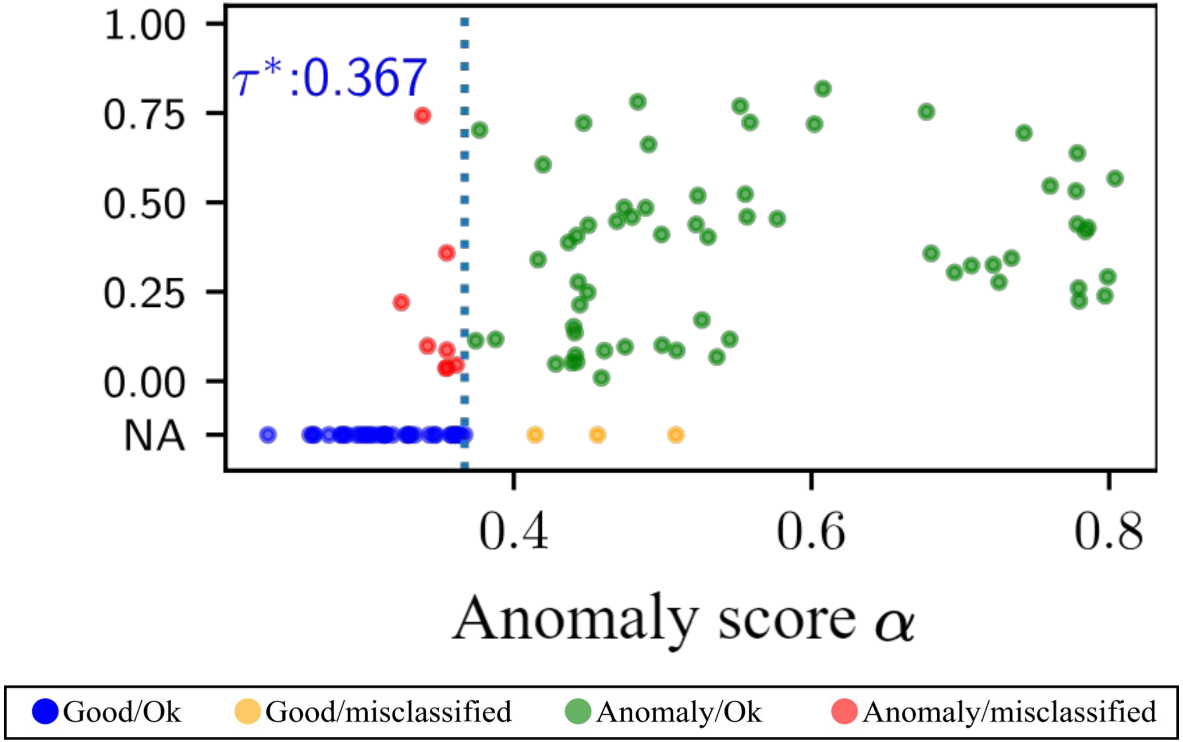
Anomalous: if $\alpha \geq \tau*$

Finding the optimal threshold τ^* means solving an optimization problem.

$$\tau^* = \underset{\tau}{\operatorname{argmax}} \sqrt{\mathrm{TPR}(\tau) \times (1 - \mathrm{FPR}(\tau))}$$

True Positive Rate	: Anomalous as anomalous
False Positive Rate	: Normal as anomalous

Anomaly detection threshold



Contribution of XAI in Anomaly Detection

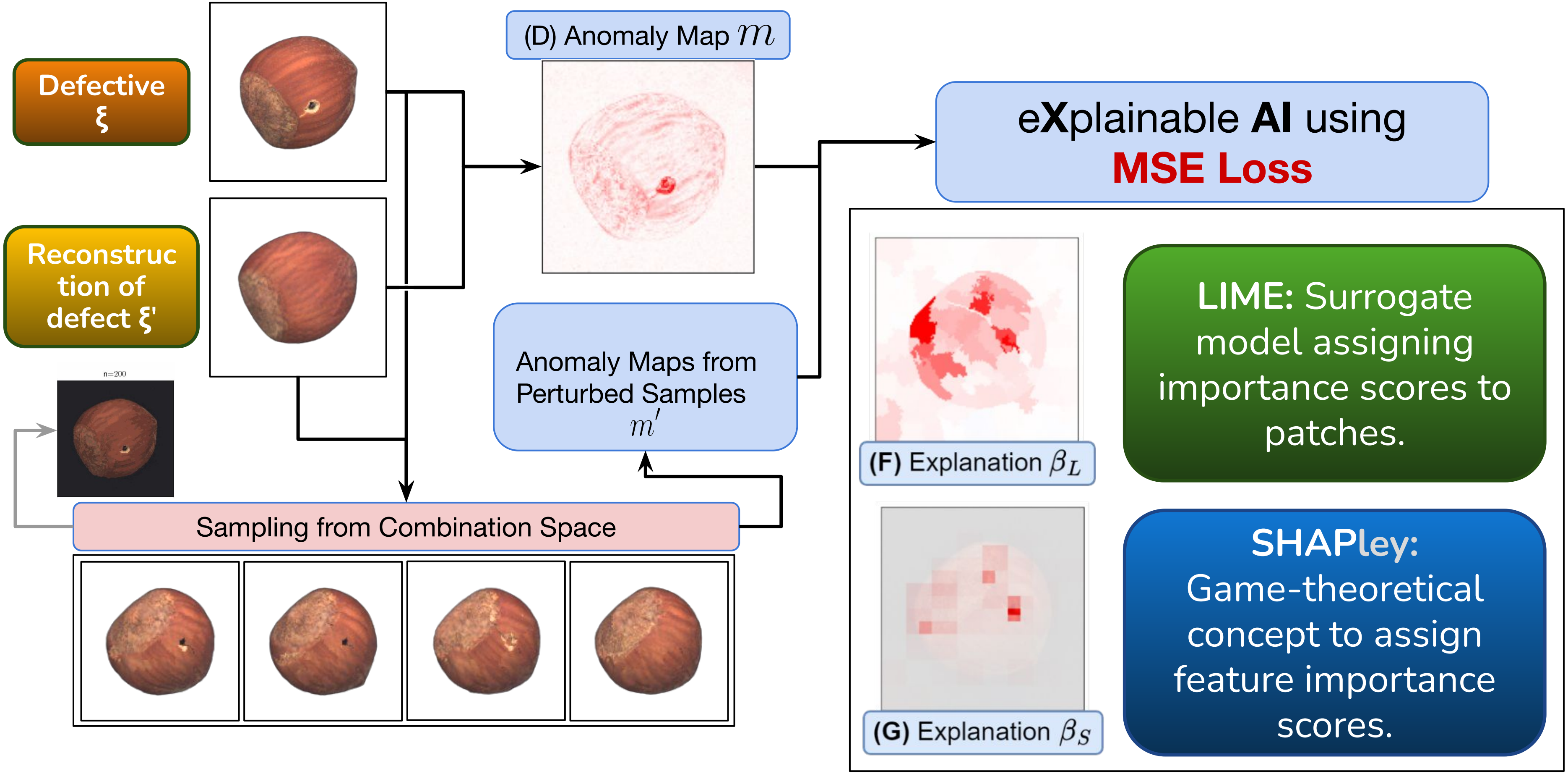
- We have the anomaly map + a detection threshold.
  Is it enough for explaining anomaly?
  → Of course not.

- Problem: anomaly map α is
  the sum of **reconstruction error** (noise) + **anomaly** (if any)
  → Need a way to:
  - separate the **anomaly** from the **noise**;
  - and to localize the region of the **anomaly**.

- More precise information

- Localization of anomaly

- Is the **anomaly** a real anomaly?



Defective ξ

Reconstruction of defect ξ'

n=200

(D) Anomaly Map m

eXplainable **AI** using **MSE Loss**

Anomaly Maps from Perturbed Samples m'

Sampling from Combination Space

(F) Explanation β_L

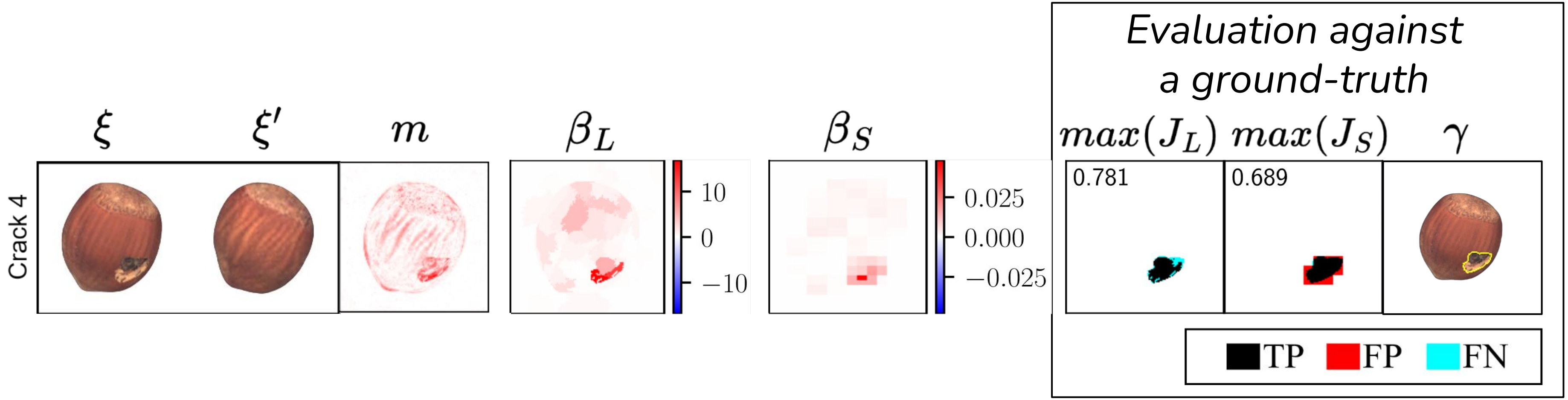
(G) Explanation β_S

LIME: Surrogate model assigning importance scores to patches.

SHAPley: Game-theoretical concept to assign feature importance scores.

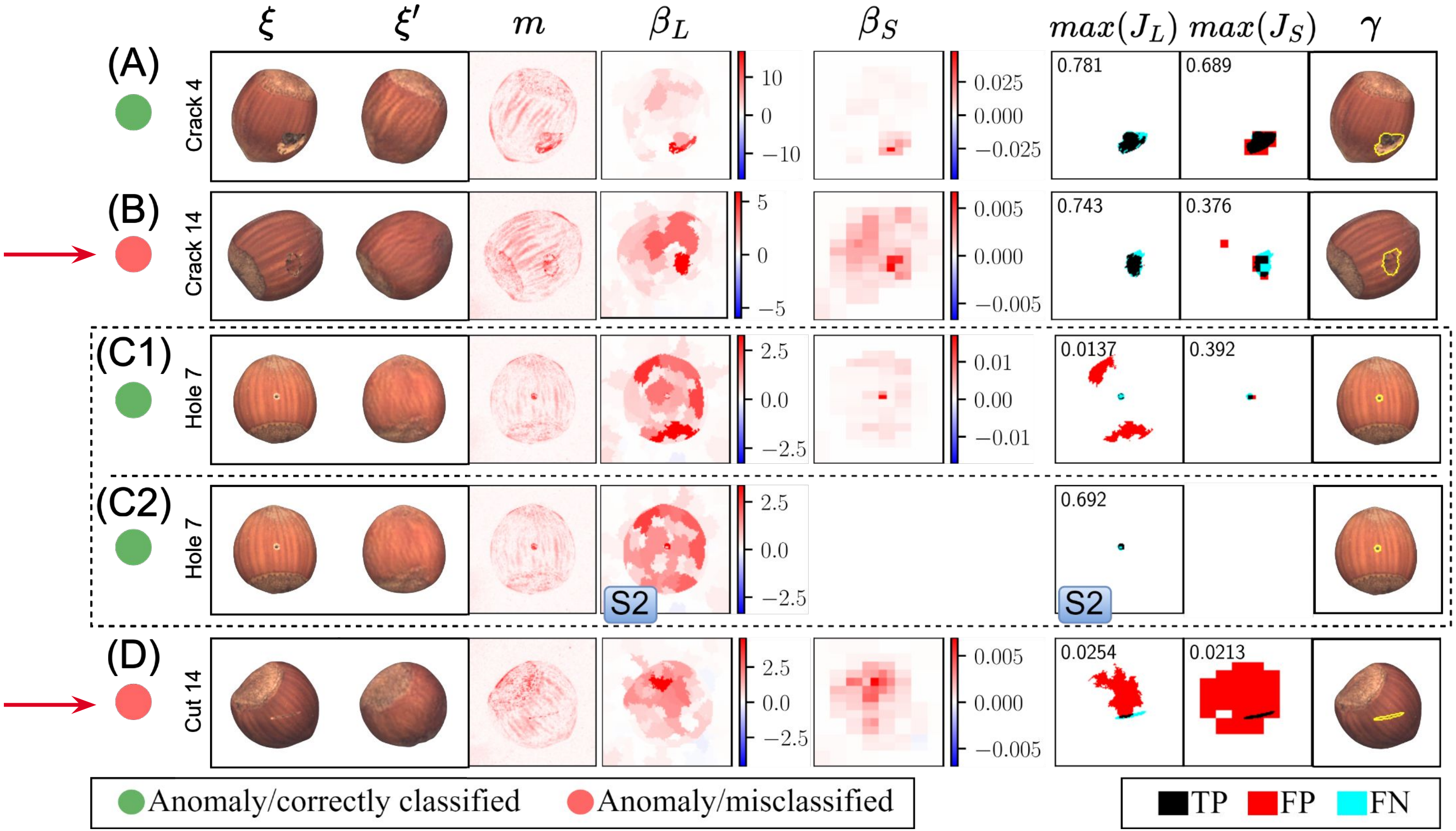
Anomaly region localization

Explaining anomaly detection with XAI

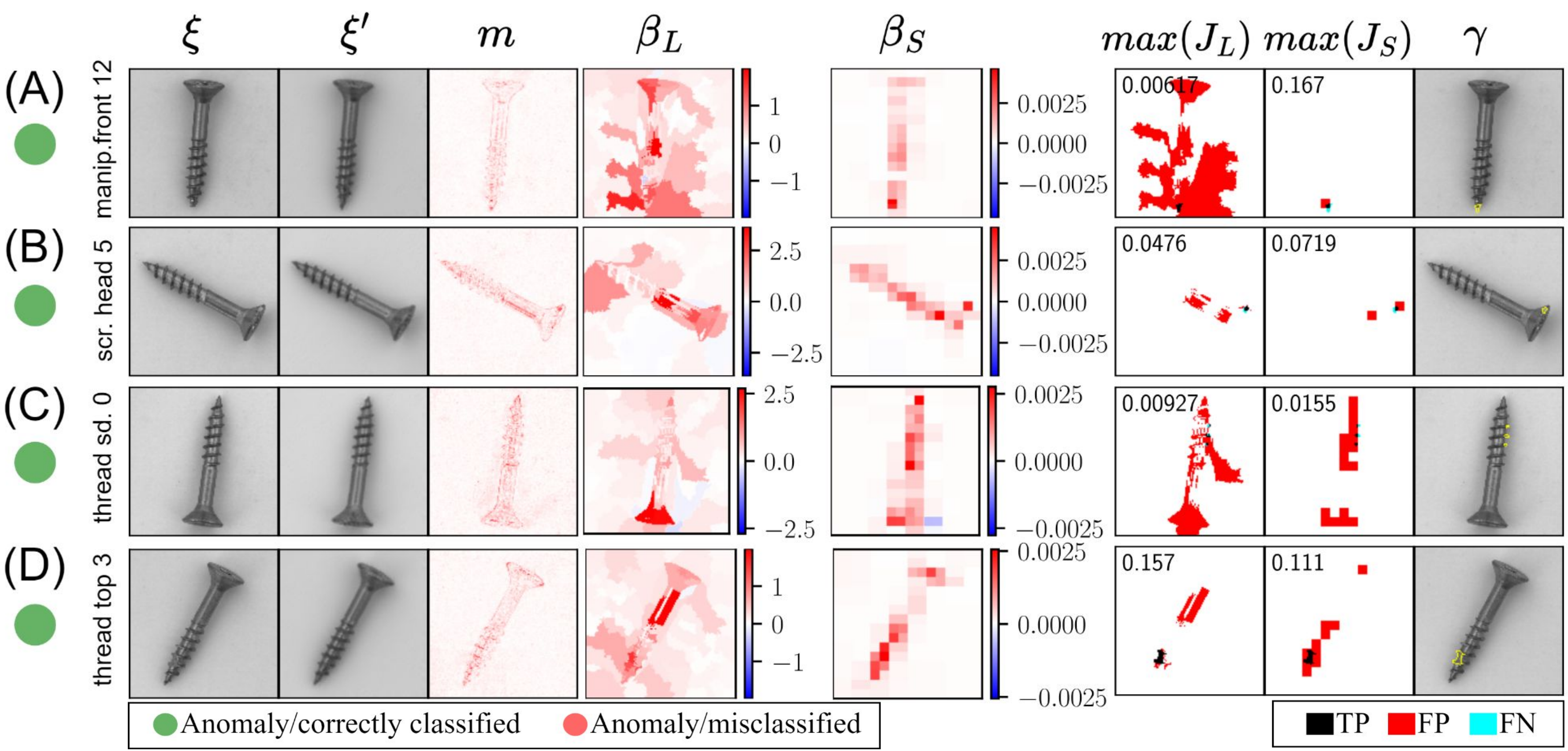


Evaluation against a ground-truth

Explaining Anomalies

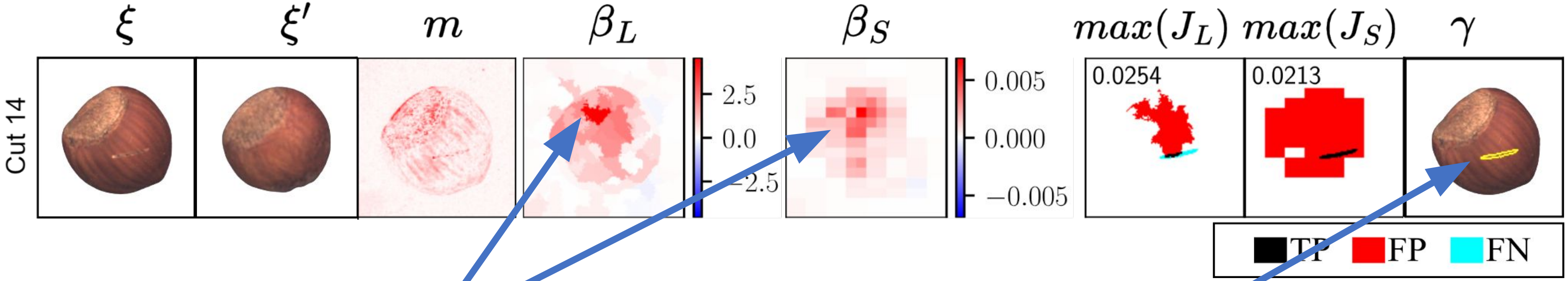


Explaining Anomalies



Revealing model misbehaviours

Samples may be classified as anomalous for the wrong reason, and only XAI can reveal such behaviour.



model detect anomaly from a badly reconstructed region...

...but the anomaly is a small cut region in a different location.

Conclusions

- XAI methods are relevant in finding the **true drivers** behind AI systems using techniques like classification and/or anomaly detection.
- Case study based on **reconstruction error maps** generated from **VAE-GAN** models.
- **Multiple XAI** techniques to separate the **reconstruction error** (noise) from the **anomaly** (if any).
- A sample may be detected as **anomalous** for the **wrong reasons**, yet this misbehaviour may not be detectable from the information provided by the anomaly detection system alone → Role of XAI!

Can I trust my anomaly detection system? A case study based on eXplainable AI.

https://github.com/rashidrao-pk/anomaly_detection_trust_case_study

Thank you! – Questions?

UNIVERSITÀ DI TORINO

Elvio G. Amparore
elviogilberto.amparore@unito.it

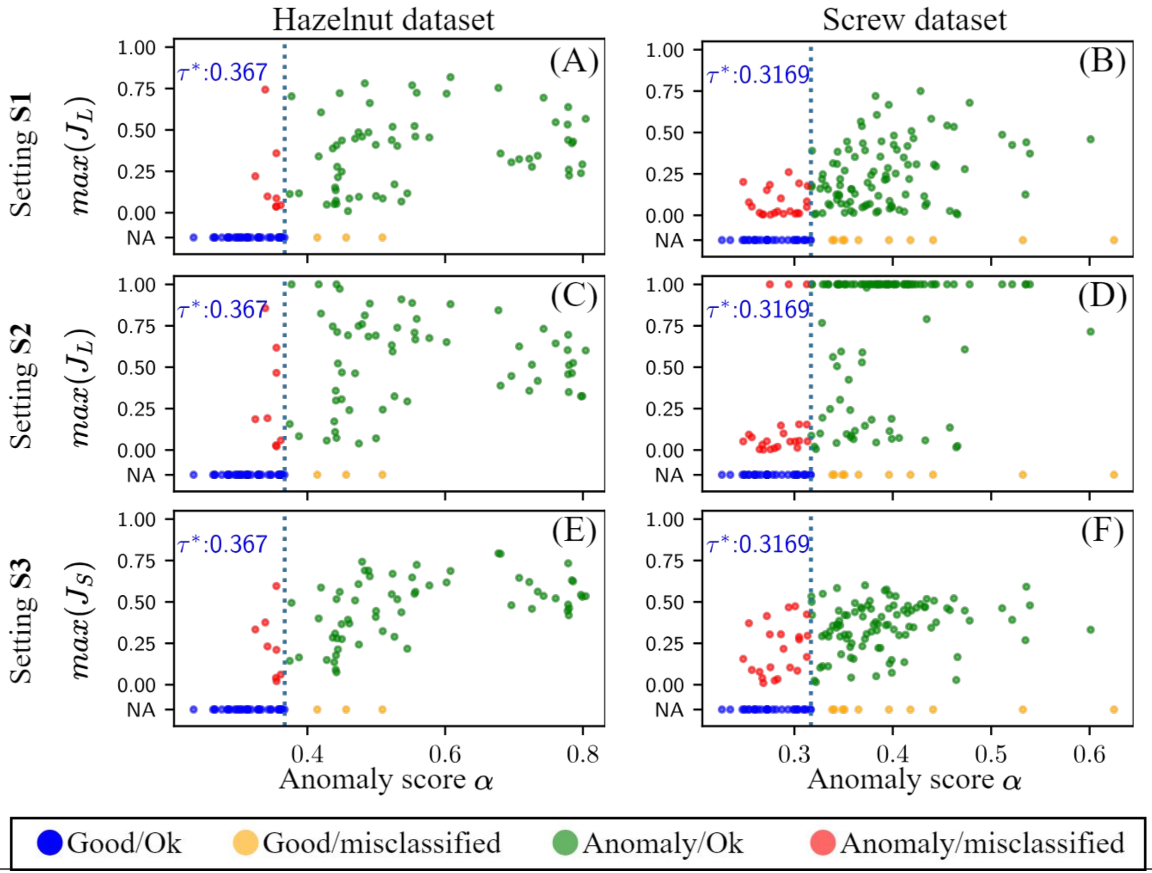
Muhammad Rashid
muhammad.rashid@unito.it

RuleX

Enrico Ferrari
enrico.ferrari@rulex.ai

Damiano Verda
damiano.verda@rulex.ai

Supplementary Material


Hazelnut dataset
Screw dataset
Setting S1
Setting S2
Setting S3
$max(J_L)$
$max(J_S)$
(A)
(B)
(C)
(D)
(E)
(F)
τ^*:0.367
τ^*:0.3169
NA
Anomaly score α
● Good/Ok
● Good/misclassified
● Anomaly/Ok
● Anomaly/misclassified