AI Standards – Application & Interaction in Assessments

WSAI - 9th and 10th October 2024 - Amsterdam

Dilara Gumusbas: *AI Technical Specialist*
Simone Vazzoler: *AI Technical Specialist*

Agenda

1. ***The EU Artificial Intelligence Act (AIA)***
   - *Key requirements of the Act for High-Risk AI Systems*
   - *Compliance with the AIA*

2. ***State of the Art (SOTA), Harmonisation & Presumption of Conformity***
   - *Definitions*
   - *Key AI standards used when conducting AI assessments*

3. ***Case study : Biometric Identification System in Electronic Health Records (EHRs)***
   - *Observations*
   - *Critical Issues – Identification & Analysis*

Learning Objectives

By the end of this session, you will be able to:

1. ***Describe** key requirements of the AIA*
2. ***Explain** the concept of State of the Art (SOTA) harmonisation, and the presumption of conformity using standards*
3. ***List** key standards considered when conducting AI assessments*
4. ***Apply** knowledge gained to use SOTA to access against articles of the AIA*

The EU Artificial Intelligence Act (AIA)

	Article 9 *Risk Management System*	***Article 10*** *Data & Data Governance*	***Article 11*** *Technical Documentation*	***Article 12*** *Record Keeping*
What is it about?	*Risk assessment, evaluation, mitigation, control*	*Labeling, preprocessing, bias, training/validation/ testing*	*Intended purpose, model design, software used*	*Recording of events, logs, post-market monitoring*
Why is it important?	*Reduce the risk for the end user*	*Bad data implies bad model*	*Process quality*	*Traceability and risk management*

The EU Artificial Intelligence Act (AIA)

	Article 13 *Transparency and Provision of Information to Deployers*	***Article 14*** *Human Oversight*	***Article 15*** *Accuracy, Robustness, and Cybersecurity*	***Article 17*** *Quality Management System*
What is it about?	*Transparency, intended purpose, instruction for use*	*Control, safety*	*Accuracy, robustness, privacy*	*Compliance, quality control, quality assurance*
Why is it important?	*Transparency and interpretability for the end user*	*AI systems must be overseen by a human to minimize risks to user's safety*	*The model should perform well in every condition and protect the user's data*	*To ensure the AI system is compliant with EU AIA*

State of the Art (SOTA), Harmonisation & Presumption of Conformity

EU AIA → *Harmonisation* ← *AI Standards*

=

Presumption of Conformity

State of the Art (SOTA), Harmonisation & Presumption of Conformity

There is a correspondence between the EU AI Act Articles and ISO standards

EU AIA Articles	*ISO Standards*
9: Risk Management System	*23894, 42001, 5338*
10: Data & Data Governance	*4213, 24027, 24029-1, 5259-1, 2, 3, 4, 5*
11: Technical Documentation	*23894, 42001*
12: Record Keeping	*23894*
13: Transparency and Provision of Information to Deployers	*24028, 23984*
14: Human Oversight	*23894, 42001*
15: Accuracy, Robustness and Cybersecurity	*4213, 24029-1*
17: Quality Management System	*42001, 24029-1, 23894, 5259-3,4*

Case Study: Biometric Identification System

- *Access to Electronic Health Records (EHRs)*
- *Secure Personally Identifiable Information (PII)*
- *Utilizes Convolutional Neural Networks (CNNs)*

Observations

Dataset

- *Collected*
  - *by a data scientist*
  - *using X brand camera*
  - *within one month*
- *Consists of*
  - *1300 samples from 300 people (290 men and 10 women)*
  - *minimum of 3 samples per person*
- *No details provided on*
  - *gender, race or age, etc*
  - *poses taken*

Model

- *Convolutional Neural Network (CNN) based pre-trained network to be finetuned with collected dataset*
- *Accuracy used as evaluation metric*
- *No additional evaluation metric usage reported*
- *Inference time reported as average*
- *No details about*
  - *adversarial attack testing or modelling*
  - *false positive (FP) rates*
  - *acceptable inference time rates in real-time*

Documentation

- *No post-market monitoring*

Critical Issues

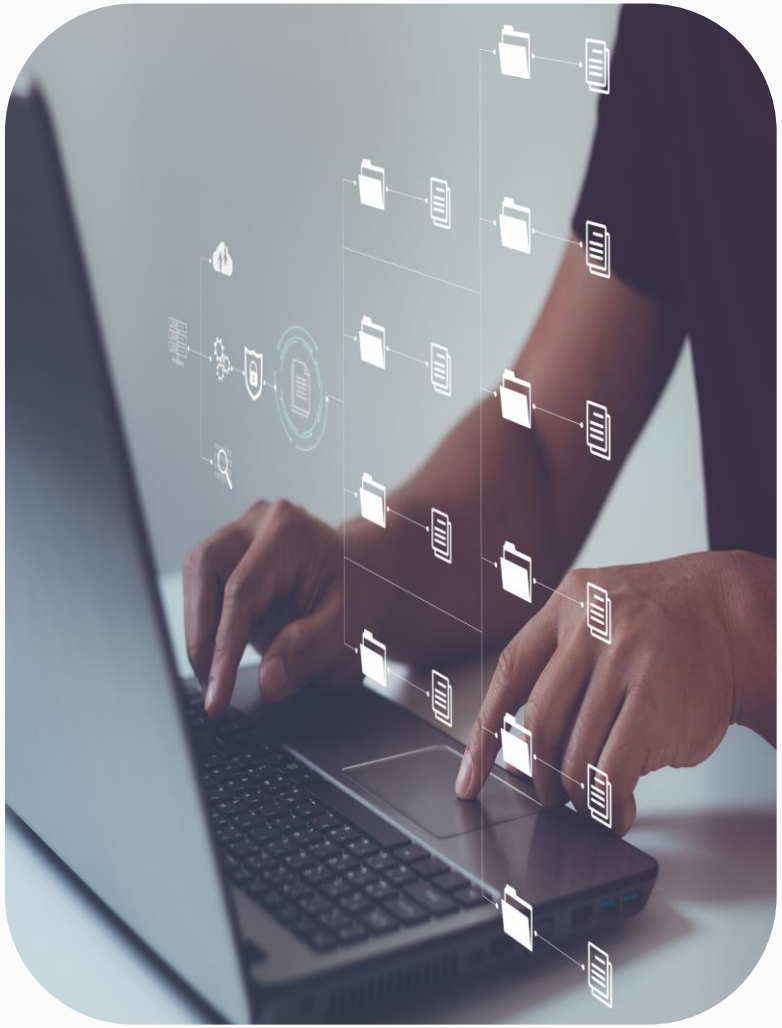
1. *Data is possibly imbalanced and biased (men vs women)*
   ***ISO 24027** Bias in AI systems and AI aided decision making*

2. *Is the dataset diverse enough?*
   ***ISO 5259** Data quality for analytics and machine learning (ML)*

3. *No evidence of testing for robustness and adversarial attacks*
   ***ISO 24029** Assessment of the robustness of neural networks*

4. *Is accuracy the right metric?*
   ***ISO 4213** Assessment of machine learning classification performance*

5. *No post-market monitoring*
   ***ISO 23894** Guidance on risk management*



Critical Issue 1: Data is possibly imbalanced and biased (men/women)

***ISO 24027** – Bias in AI systems and AI-aided decision-making*

Human Factor	***Favouritism***	***Indirect Bias***	***Fairness and Balance***	***Fairness Metrics***	***Inherited Bias***

ISO 24027

***EU AI Act – Article 10:** Data and Data Governance*

Critical Issue 1: Data is possibly imbalanced and biased (men/women)

***ISO 24027** – Bias in AI systems and AI-aided decision-making*

Human Factor	***Favouritism***	***Indirect Bias***	***Fairness and Balance***	***Fairness Metrics***	***Inherited Bias***

ISO 24027 clauses 6.3 Data bias and 6.3.2.1.3 Coverage bias:
"Coverage bias occurs when a population represented in a dataset does not match the population that the ML model is making predictions about."

Critical Issue 2: Is the dataset diverse enough?

ISO 5259 – *Data quality for analytics and machine learning (ML)*

Human Factor	***Completeness***	***Sensor Element***	***Time Factor***	***Privacy / User Consent***

ISO 5259

EU AI Act – Article 10: *Data and Data Governance*

Critical Issue 2: Is the dataset diverse enough?

***ISO 5259 –** Data quality for analytics and machine learning (ML)*

Human Factor	***Completeness***	***Sensor Element***	***Time Factor***	***Privacy / User Consent***

ISO 5259 series clauses 6.5.3 Effectiveness, 6.5.4 Balance and 6.5.5 Diversity (ISO 5259-2)
In particular, 6.5.5 Diversity: "If all or most data records in a dataset are alike, an ML model trained from that dataset can have the risk of overfitting and consequently being less generalizable."

Critical Issue 3: No evidence of testing for robustness and adversarial attacks

ISO 24029 – *Assessment of the robustness of neural networks*

Input Perturbations

Adversarial Attacks

Consistency / Performance Over Time

ISO 24029

EU AI Act – Article 15: *Accuracy, Robustness and Cybersecurity*

Critical Issue 3: No evidence of testing for robustness and adversarial attacks

ISO 24029 – *Assessment of the robustness of neural networks*

Input Perturbations

Adversarial Attacks

Consistency / Performance Over Time

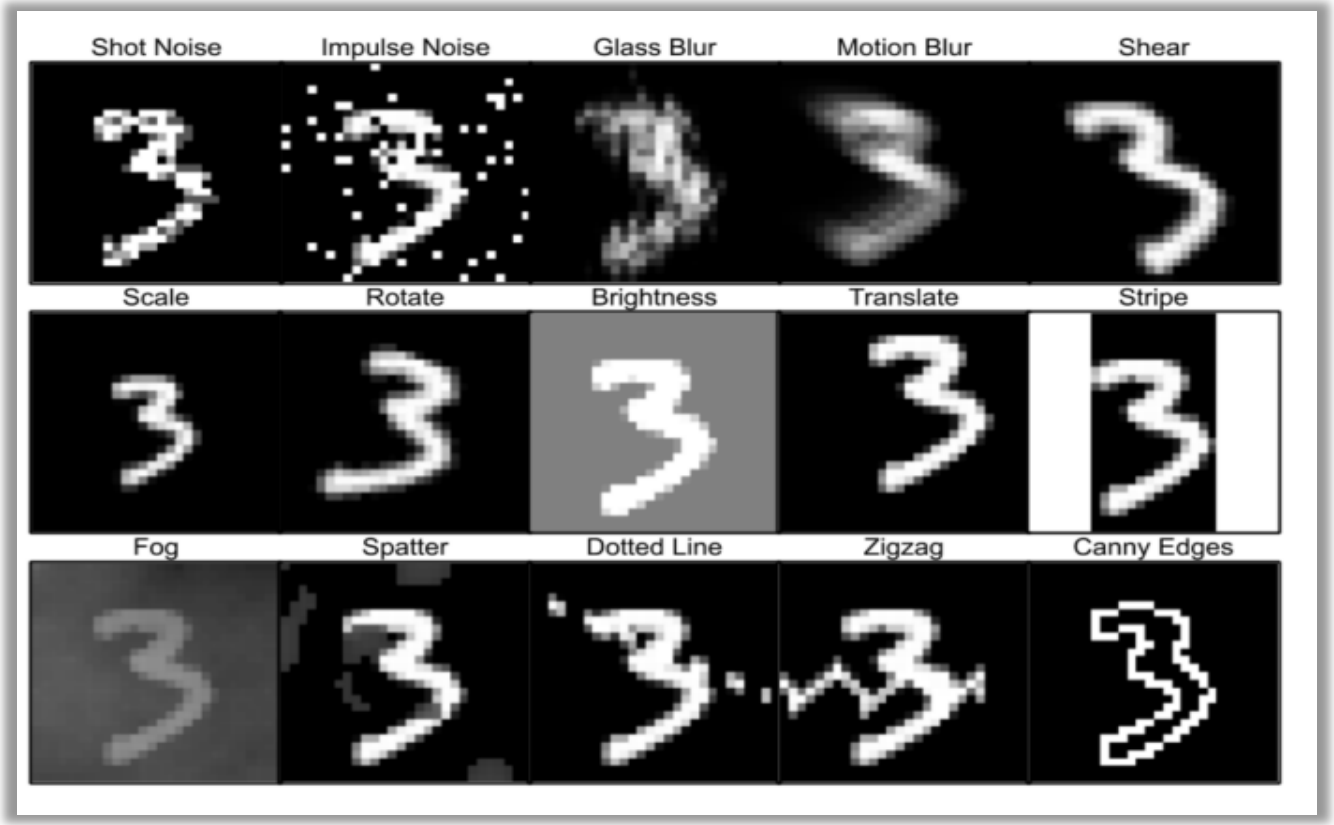
ISO 24029 clauses 4.1.1 Robustness concept

"Robustness properties demonstrates the degree to which the system performs with atypical data as opposed to the data expected in typical operations."

Critical Issue 3: No evidence of testing for robustness and adversarial attacks

***ISO 24029 –** Assessment of the robustness of neural networks*

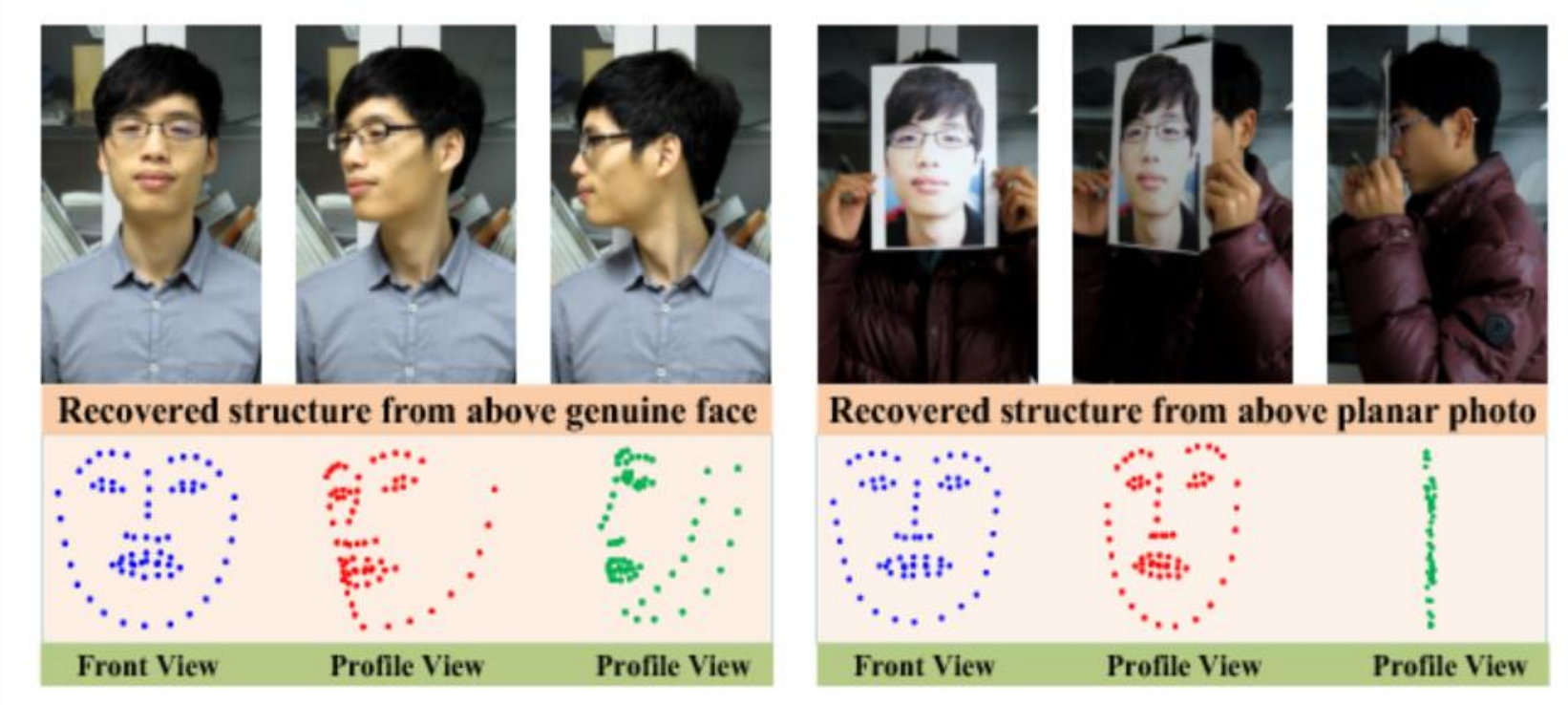
Input Perturbations



Critical Issue 3: No evidence of testing for robustness and adversarial attacks

ISO 24029 – Assessment of the robustness of neural networks

Adversarial Attacks



REF: [PDF] Face liveness detection using 3D structure recovered from a single camera | Semantic Scholar

Critical Issue 4: Is accuracy the right metric?

ISO 4213 – *Assessment of machine learning classification performance*

Evaluation Metric Correctness

Overfitting Issue

Real-Time Performance

Human Factor

ISO 4213

EU AI Act – Article 15: *Accuracy, Robustness and Cybersecurity*

Critical Issue 4: Is accuracy the right metric?

ISO 4213 – *Assessment of machine learning classification performance*

Evaluation Metric Correctness

Overfitting Issue

Real-Time Performance

Human Factor

ISO 4213 clauses 6. Statistical measures of performance, 6.2.3 Accuracy and 5.3.13 Appropriate baselines

6.2.3: "Accuracy should not be used to express comparative performance across models unless classes are known to be reasonably balanced"

5.3.13: "A baseline method can be necessary as a basis of comparison for machine learning classification performance."

Critical Issue 5: No post-market monitoring

ISO 23894 - *Guidance on risk management*

Unauthorised Use	*Monitoring / Post Market Surveillance*	*Misconfiguration*	*Insider Threats*	*Cybersecurity Attacks*

ISO 23894

EU AI Act – Article 61: *Post-Market monitoring by providers and post-market monitoring plan for high-risk AI systems*

Critical Issue 5: No post-market monitoring

ISO 23894 - *Guidance on risk management*

Unauthorised Use	***Monitoring / Post Market Surveillance***	*Misconfiguration*	*Insider Threats*	*Cybersecurity Attacks*

ISO 23894 clauses 6.4 Risk assessment and 6.6 Monitoring and review

"AI risks should be identified, quantified or qualitatively described and prioritized against risk criteria and objectives relevant to the organization."

Learning Objectives

Now we have reached the end of the session, you should be able to:

1. ***Describe** key requirements of the AIA*
2. ***Explain** the concept of State of the Art (SOTA) harmonisation, and the presumption of conformity using standards*
3. ***List** key standards considered when conducting AI assessments*
4. ***Apply** knowledge gained to use SOTA to access against articles of the AIA*

Thank you