Trustworthy AI in Healthcare

Niamh St John Lynch, PhD Researcher, p/t Lecturer of Software Validation and AI in Healthcare with DkIT

*Research made possible through **Research Ireland** under Collaboration Partnership between **DkIT** and **UCD**, Ireland*

Supervised by Dr. Róisín Loughran, Dr. Martin Mc Hugh,

Prof. Fergal McCaffery, DkIT

25+ Year's Industry Experience IT, Pharma and Medical Devices (incl. ex-QA/RA Director and ex-Notified Body as expert Software reviewer of Active Medical Devices)

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

Taighde Éireann Research Ireland

School of Informatics & Creative Arts

Agenda

- Introduction and evolution of AI in Healthcare incl. AI risks
- My Research Approach
- What is Trustworthy AI?
- A Proposal Framework for Trustworthy AIeMD in Healthcare
- Standards and SoTA
- Measurement and Metrics for AIeMD
- *3rd Party Pre-Trained AI Models; evaluation for ethical use*
- *Simplification of MDR/IVDR and impact on AI Act 2024/1689*

Time Permitting

Introduction to AI in Healthcare

Definitions and Growth of AI in Healthcare

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

Artificial Intelligence (AI) — systems performing tasks requiring human intelligence; includes symbolic (i.e., rules based) and statistical methods.

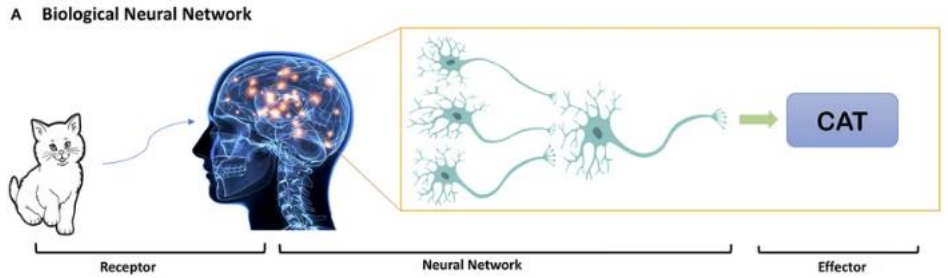
Machine Learning (ML) — subset of AI where algorithms learn patterns from data; includes supervised, unsupervised, reinforcement learning

Deep Learning (DL) — ML with multi-layer neural networks (CNNs, RNNs, Transformers) for perception and language and more!

Note — 'AI' in devices typically refers to ML/DL models trained on clinical data or rules embedded in device logic or a hybrid of these.

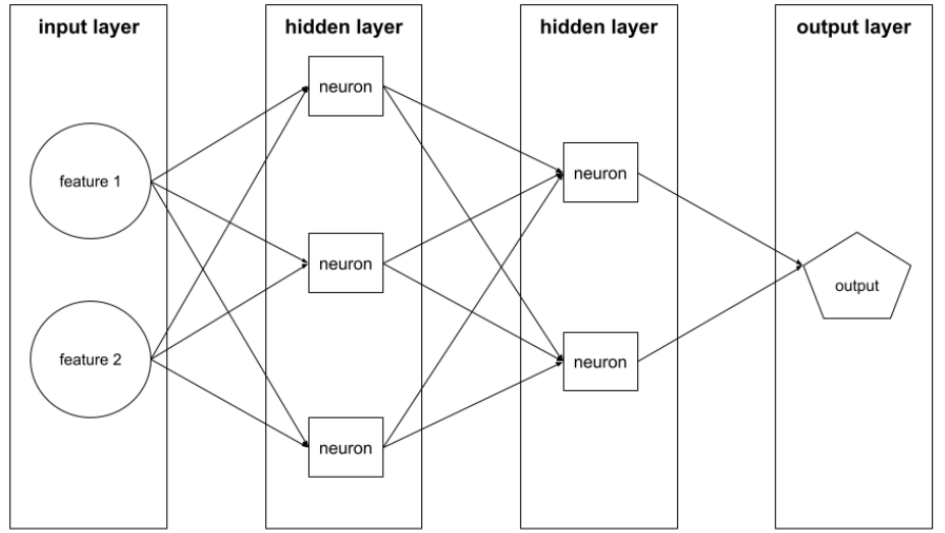
Generative LLM, LIMs and Multimodal models are on the rise in development.

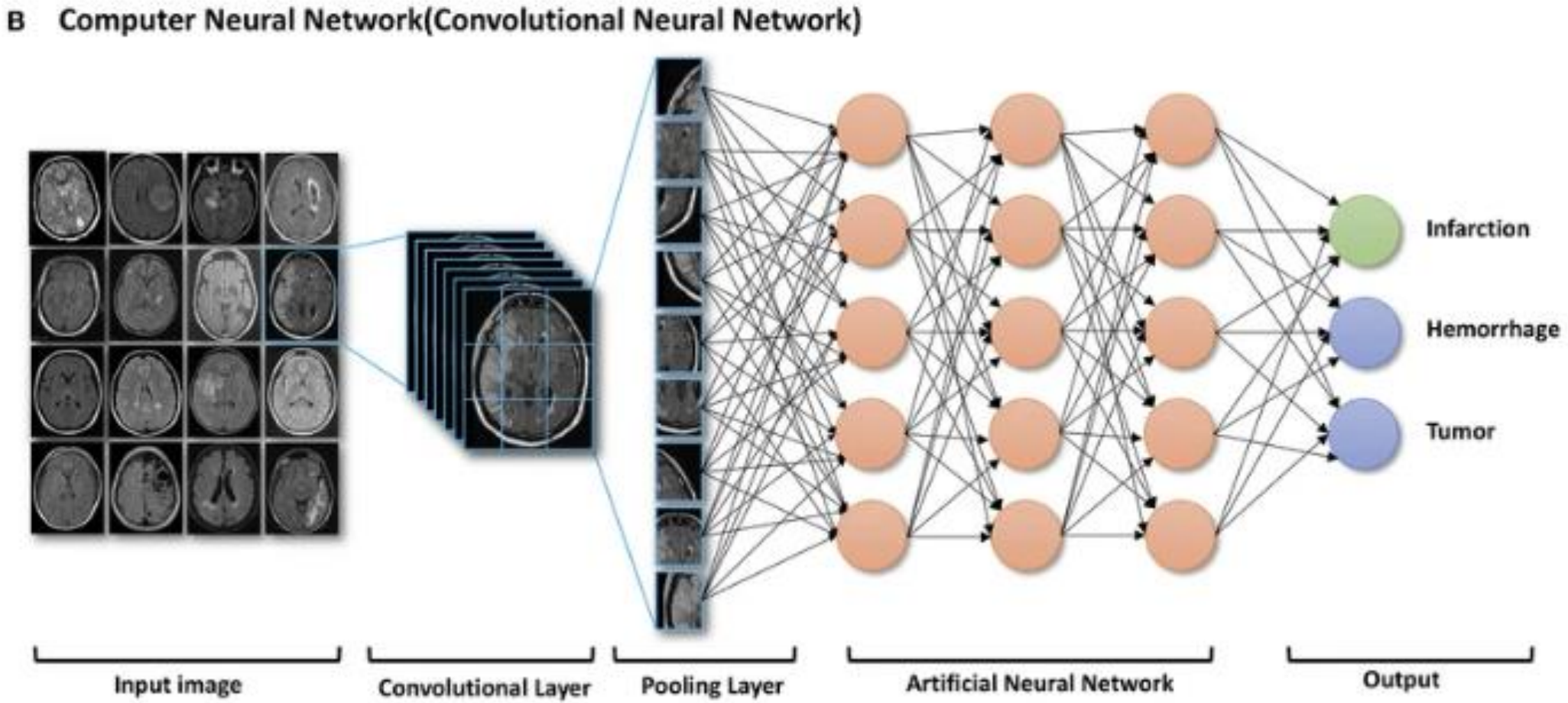
A typical Neural Network



- A set of **neurons** in a **neural network**. Three common types of layers are as follows:
- The **input layer**, which provides values for all the **features**.
- One or more **hidden layers**, which find nonlinear relationships between the features and the label.
- The **output layer**, which provides the prediction.

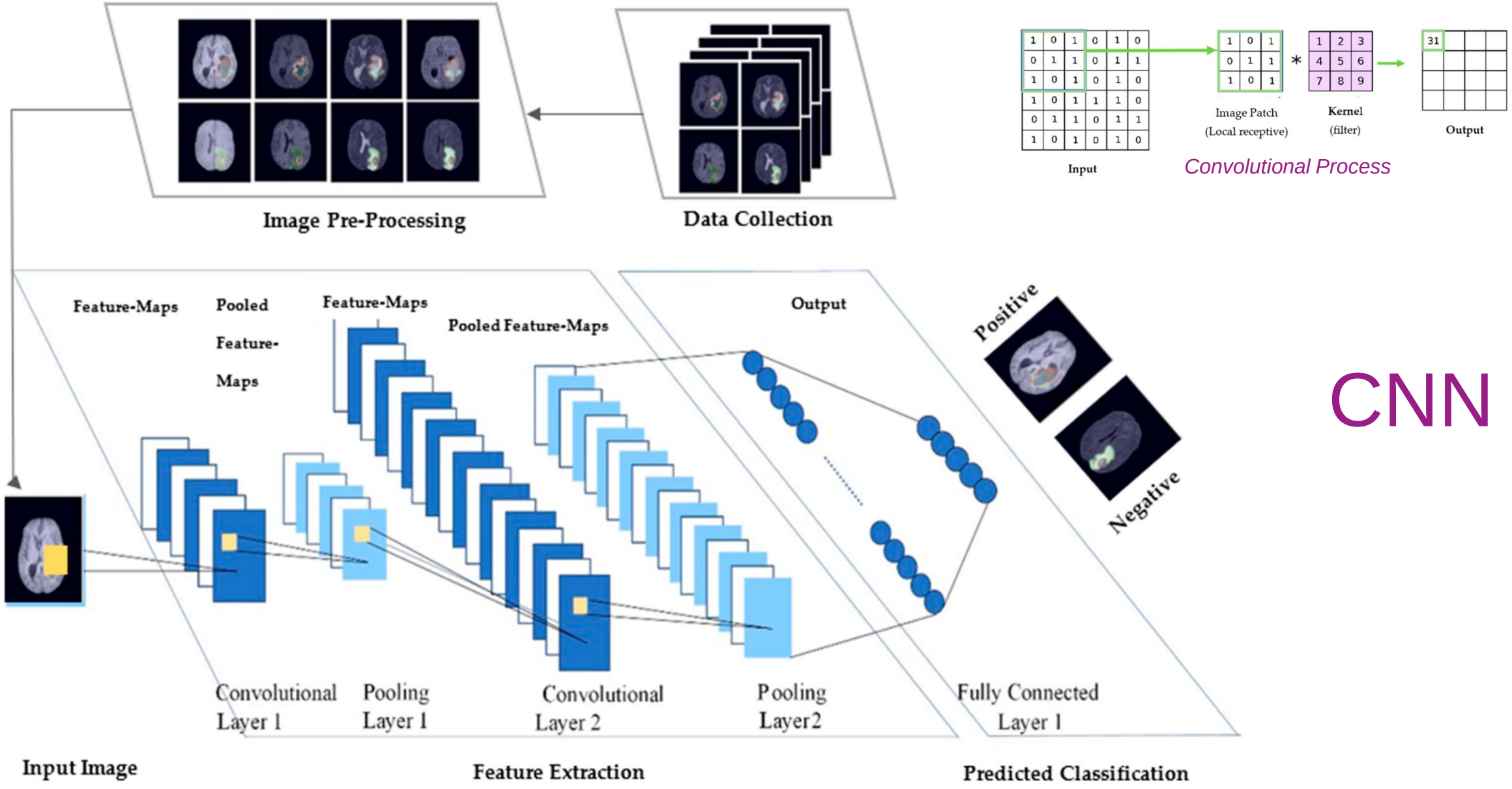
For example, the following illustration shows a neural network with one input layer, two hidden layers, and one output layer.





Example of components of Biologic Neural Network (A) and Computer Neural Network (B). Reprinted with permission from Zaharchuk et al. (15). Copyright American Journal of Neuroradiology.

https://www.researchgate.net/publication/335165437_Applications_of_Deep_Learning_to_Neuro-Imaging_Techniques



CNN

Salehi, A.W.; Khan, S.; Gupta, G.; Alabduallah, B.I.; Almjally, A.; Alsolai, H.; Siddiqui, T.; Mellit, A. A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability* **2023**, *15*, 5930. https://doi.org/10.3390/su15075930

Summary of Architectures used in Medical Use Cases

Architecture	Key Innovation	Typical Uses
LeNet	Early CNN blueprint	Simple vision tasks
AlexNet	ReLU, GPU training	General image classification
VGG	Deep, simple stacks	Feature extraction
Inception	Parallel multi-scale convs	Efficient large-scale models
ResNet	Skip connections	Deep architectures, medical imaging
DenseNet	Dense connectivity	Efficient feature reuse
MobileNet	Lightweight CNN	Mobile/edge AI
EfficientNet	Compound scaling	High-accuracy efficient models
U-Net	Encoder-decoder for segmentation	CT/MRI segmentation
3D CNNs	3D convolutions	Volumetric imaging

Supervised

- Learn from labeled examples
- Used for diagnosis, classification, risk prediction
- Requires curated labels; common in medical imaging

Unsupervised

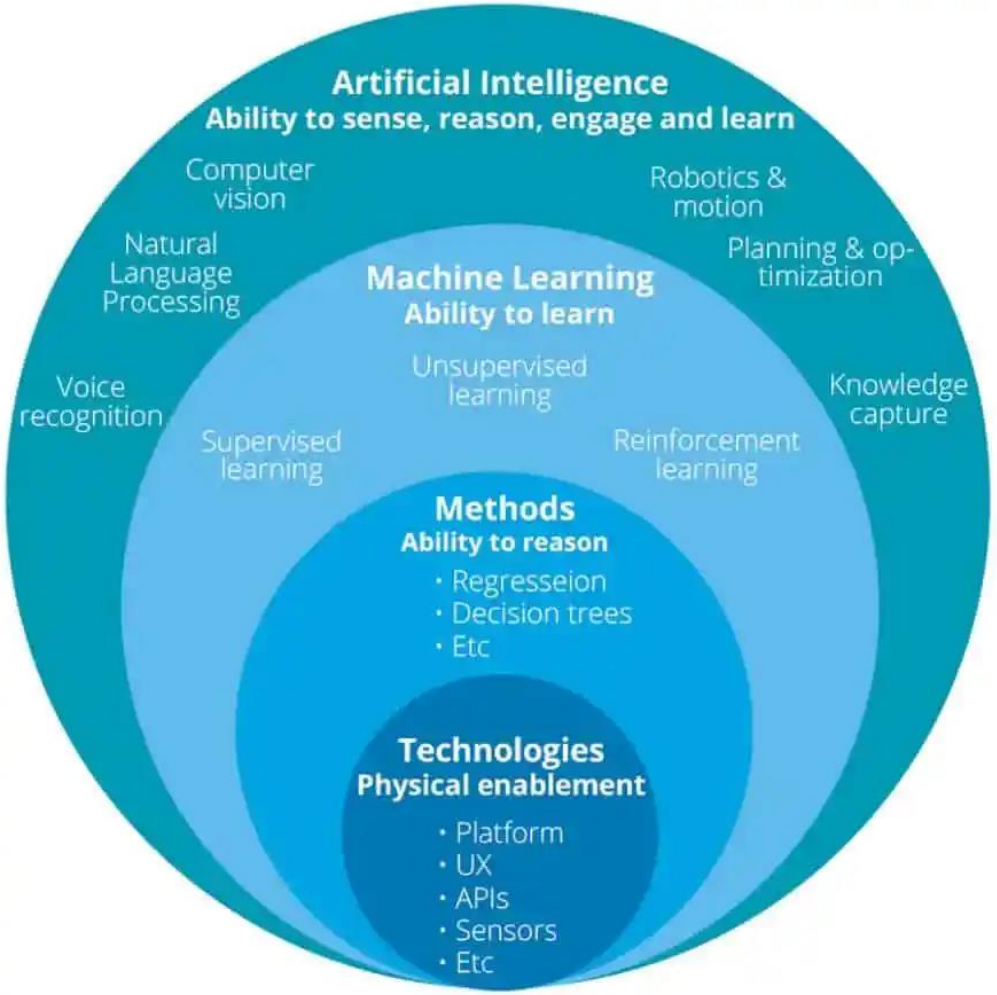
- Detect structure without labels
- Used for clustering, anomaly detection, representation learning
- Useful when labels are scarce

Reinforcement

- Learn policies via reward signals
- Applied in control systems and treatment optimization research
- Requires simulation and safe exploration

Learning Paradigms: Supervised, Unsupervised, Reinforcement

How each approach learns, where it's applied in healthcare, and practical constraints



DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

FDA Accepted AI/ML Devices at May 2024

1950 — **Turing: Computing Machinery and Intelligence**
Alan Turing publishes the seminal paper defining machine intelligence

1956 — **Dartmouth workshop**
Term **"Artificial Intelligence"** coined at the Dartmouth workshop

1957–1960s — **Perceptron & symbolic systems**
Early **perceptron**, symbolic AI and expert systems (MYCIN)

1986 — **Backpropagation revival**
Backpropagation re-popularized; neural networks regain research focus

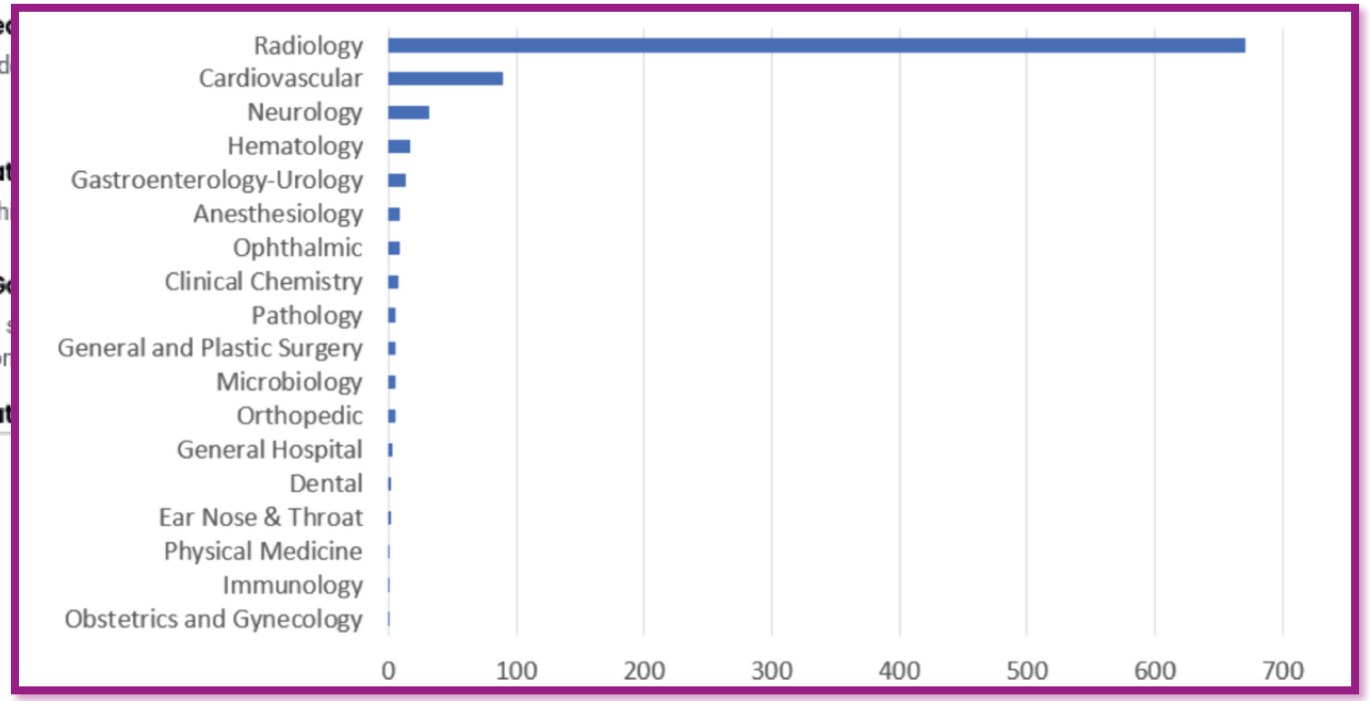
1997 — **Deep Blue vs. Kasparov**
Deep Blue defeats Kasparov demonstrating symbolic search strength

2006–2012 — **Deep le**
Modern d
vision

2011 — **IBM Wat**
Watson h

2016–2020 — **AlphaG**
AlphaGo
predictio

2020–2024 — **Foundat**



Dating back to

- 1995 (n=1),
- 1997 (n=1),
- 1998 (n=1),
- 2008 (n-5),

- 2018 (n=64)

- 2023 (n=221)

1,357 FDA-authorised AI/ML medical devices as of Feb 20, 2026 – not an exhaustive list!

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

Typical AI Risk Failures

AI Failure Risk	Example Clinical Scenario	What Can Go Wrong
False Negative (missed finding)	AI misses a small pneumothorax on ER trauma CT	Delayed recognition → respiratory deterioration; patient may not receive urgent chest tube placement.
False Positive (incorrect flag)	AI flags a benign lung nodule as suspicious	Unnecessary follow-up CT, invasive testing, radiation exposure, patient anxiety.
Data Bias / Poor Generalization	Model trained mostly on adult data used on paediatric chest X-rays	Misclassification due to anatomical differences → incorrect triage or missed pathology.
Dataset Shift / Model Drift	Hospital upgrades to a new CT scanner with different reconstruction algorithms	AI performs worse because image characteristics differ → increased error rate.
Out-of-Distribution Imaging	Motion-blurred MRI or metal-artifact-heavy CT is analyzed by AI	AI misreads pathology or fails to detect key findings due to unfamiliar image artifacts.
Automation Bias (over-trusting AI)	Radiologist assumes AI's "no PE detected" output is correct on CT pulmonary angiography	Human reviewer may miss an embolus if they defer too heavily to AI judgment.
Low Explainability	AI labels an area as "possible stroke" on non-contrast head CT with no explanation	Clinician unsure if model is hallucinating → wasted time validating or misprioritizing care.
Failure on Rare or Atypical Presentations	Patient has an uncommon congenital heart defect not represented in training data	AI missegments anatomy → incorrect measurements or misleading diagnostic cues.
Workflow / Integration Issues	AI triage alert fails to appear in PACS due to interface error	Critical, time-sensitive findings (e.g., intracranial hemorrhage) are not escalated.
User Interface (UI) Misinterpretation	AI displays a "confidence score" that clinicians assume reflects diagnostic certainty	Over- or under-weighting AI advice → misprioritization of patient cases.
Security / Cyber Vulnerability	Ransomware attack corrupts CT images before AI evaluation	Corrupted images mislead AI → unsafe outputs or system downtime delays diagnosis.

Risk Mitigation Strategies

AI Failure Risk	Mitigation Strategies
False Negative (missed finding)	• Mandatory human review of all AI-negative high-risk studies • • Use AI as second-reader, not replacement • Regular audit of FN rates
False Positive (incorrect flag)	• Radiologist final authority on interpretation • Threshold tuning to reduce false alarms • Implement double-reading for flagged cases
Data Bias / Poor Generalization	• Diverse, representative training datasets • Local validation before deployment • Continuous performance monitoring by demographic group
Dataset Shift / Model Drift	• Re-validation after equipment changes • Periodic retraining or recalibration • Drift monitoring systems with alerts
Out-of-Distribution Inputs	• Automated OOD detection (flag degraded images) • Image quality checks before AI analysis • Radiologist prioritization of artifact-heavy scans
Automation Bias (over-trusting AI)	• Training on cognitive bias awareness • Require independent human interpretation • Interface design that discourages blind acceptance

AI Failure Risk	Mitigation Strategies
Low Explainability	• Prefer models with saliency maps or interpretable outputs • Educate clinicians on how to interpret AI confidence • Provide structured reasoning summaries where possible
Failure on Rare / Atypical Cases	• Human override capability • Specialist review workflows for rare disease populations • Continuous enrichment of datasets with rare cases
Workflow / Integration Issues	• Robust IT QA processes • Direct alert redundancy (SMS, dashboard, EHR flag) • Regular integration testing
UI / Usability Problems	• UI standardization and clear labelling • Clinician training on meaning of scores • Color-coding and tiered alert design
Security / Cyber Risks	• Regular cybersecurity audits • Encrypted data pipelines • Fail-safe mode to block AI use during suspicious activity

Regulatory Context — Standards, Documentation, Evidence for AI Devices

What do regulators expect and how manufacturers must document, test, monitor, and govern AI-enabled devices

Regulatory expectations — technical file, clinical evidence, risk management, post-market surveillance aligned with FDA/region-specific guidance

Standards & frameworks — apply ISO 13485, IEC 62304, IEC 62366 and relevant AI assurance frameworks

Documentation needs — datasets, labeling protocols, model architecture, validation results, monitoring plan, update governance

Regulatory trend — FDA approvals growing (>1000 AI-enabled devices by 2025);
emphasis on real-world monitoring

My Research Approach

Trustworthy AI-enabled Medical Devices in Healthcare

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

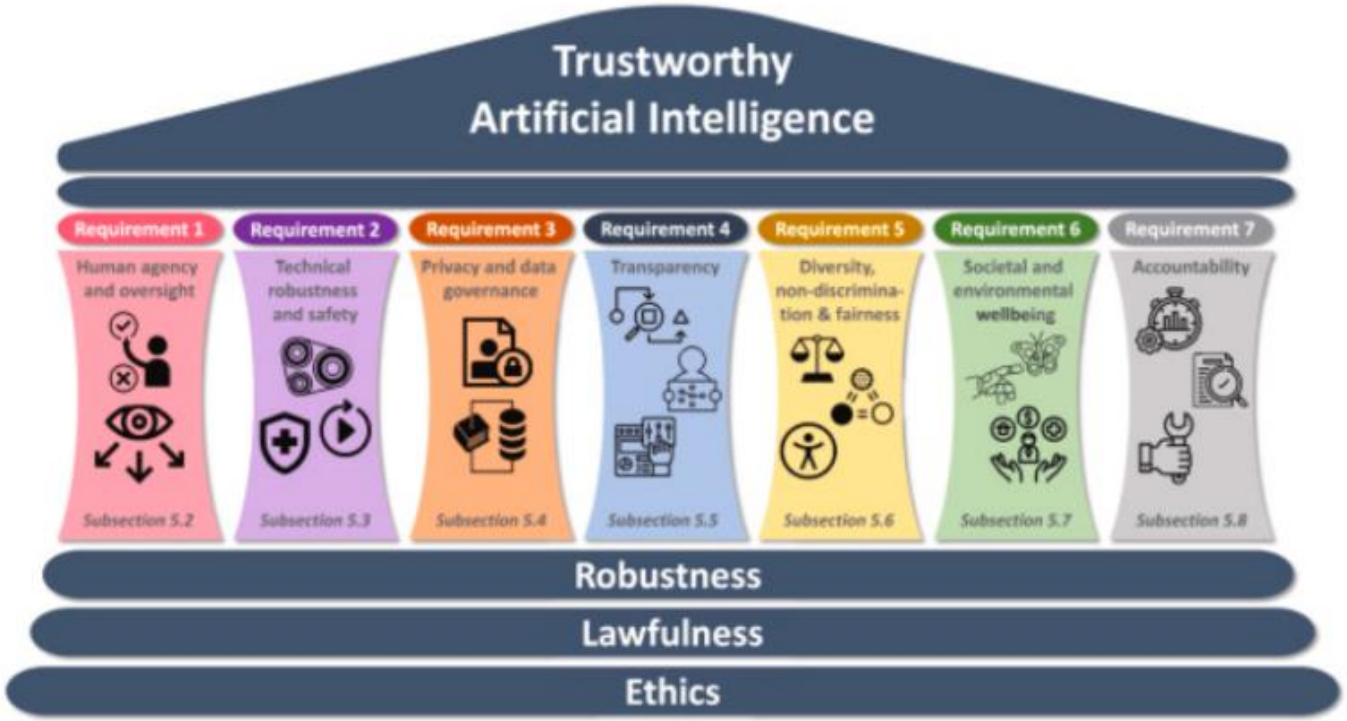
School of Informatics & Creative Arts

My Research Design

Research Philosophy	Post-positivist objectivism	Constructivist	Pragmatism	PhD Resesarch Phase
Source of Truth	Sourced in Theory	Researcher's domain knowledge / understanding	Problem Centred	Planning Phase 1
Methodology Approach	Reductionism, Empirical Observation & Measurement (e.g. hypothesis)	Social and Historical Construction / Multiple Participant Views	Pluralistic / Consequences of Actions '*whatever works*'	
Research Aim	Prove or disprove existing theory	Theory Generation	Suitable for real-world practice	
Suitable Method according to Philosophy	Quantitative	Qualitative	Mixed Methods (case study; research design, etc.)	
Output from PhD Research	TwAI Framework & CMP Tool	TwAI Framework	TwAI Framework & CMP Tool for use in industry	Implementation Phase 2
Methods used in this research		Semi-Qualitative Interviews - Ongoing	Expert Reviews in Real-World Practice - Ongoing	
	Quantitative Survey - Pending		Research Design Case-Study in Industry - Pending	Validation Phase 3

What is Trustworthy AI?

And why do we need it?



EU Ethical Guideline 2019

Ethical Foundations and Trust in AI-eMD

Trust as a complex psychological construct extending to AI through computational trust theories

- Trust involves dynamic perceptions influenced by context and stakes
- In healthcare AI, debates focus on whether AI can and should be trusted given high-impact decisions

1

Ethical principles underpinning trust in AIeMD

- Respect for autonomy ensures patient control and informed decisions
- Fairness addresses bias mitigation and equitable outcomes
- Transparency promotes clarity on AI decision processes

2

Limitations of current AI ethics guidelines in healthcare contexts

Existing frameworks provide high-level guidance but lack healthcare-specific operational details

3

Our research focus: Integrating ethical values early in AIeMD development

- Emphasizes embedding trustworthiness beyond conceptual understanding
- Supports practical, ethical design to build reliable AI-enabled medical devices

4

AIeMD Standards: A Multidisciplinary Literature Review

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

In the beginning….

- Conducted a systematic search screening over 3,900 studies focused on trustworthy AI in healthcare
- Selected 43 studies meeting strict inclusion criteria for detailed qualitative analysis
- Analyzed regulatory frameworks from the EU, FDA, and ISO standards to ensure compliance
- Performed meta-synthesis to identify core factors for trustworthy AI (TwAI)
- Validated findings through expert qualitative review to ground the framework in real-world expertise
- Ensured the TwAI framework is evidence-based and aligned with regulatory standards

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

Semi-Qualitative Interview (SQI)

- *No* regulatory requirements
- Continue to comply with IEC 62304
- Waiting for AI Standards
- Not clear how to measure (metrics)
- Cross-validation used; not accepted by FDA for clinical proof
- Heavy reliance on “locked” ML
- Not clear on risks (drift, degradation)
- Development continues without documented procedures
- *Greatest Challenge*: lack of data
- Post-market surveillance uses traditional methods, complaints, etc.
- Not clear how to incorporate Ethics or Explainability (XAI)

Company	AI Type	Treatment Type	Country
Siemens-Healthineers	ML-Locked	Radiology	Pakistan
MD101	ML/LLM - Locked	Radiology & support of other developers (LLMs)	France
Blackford Analysis (AI Partner)	ML	Software support Partner for delivering AI results	Scotland
Varian Medical Systems Imaging Laboratory GmbH	ML-Locked	Radiography imaging for specialist Cancer Treatment	Switzerland
GE Healthcare	ML-Locked	Patient Monitoring (Cardiac Health)	Europe
ASCO Group	ML-Locked	Radiography imaging	Bucks, UK
BlueBridge Technologies	ML-Locked	Respiratory	Ireland
QAIR	DL-Locked ML-adaptive	Respiratory imaging; robotic AI surgery; laparoscopic surgery; genomic tumour profiling.	The Netherlands
Elekta	ML-Locked	Various	The Netherlands
Communitech	ML-Locked, ML-Adaptive / LLM	Various with students doing LLMs in basements.	Canada
Brain Lab	ML-Locked/ Adaptive	Radiology & Image-based Surgery	Germany

Bridging AI Regulatory & Ethical Gaps

Insights from 15 AI Experts Highlight Urgent Needs in Healthcare Quality Systems

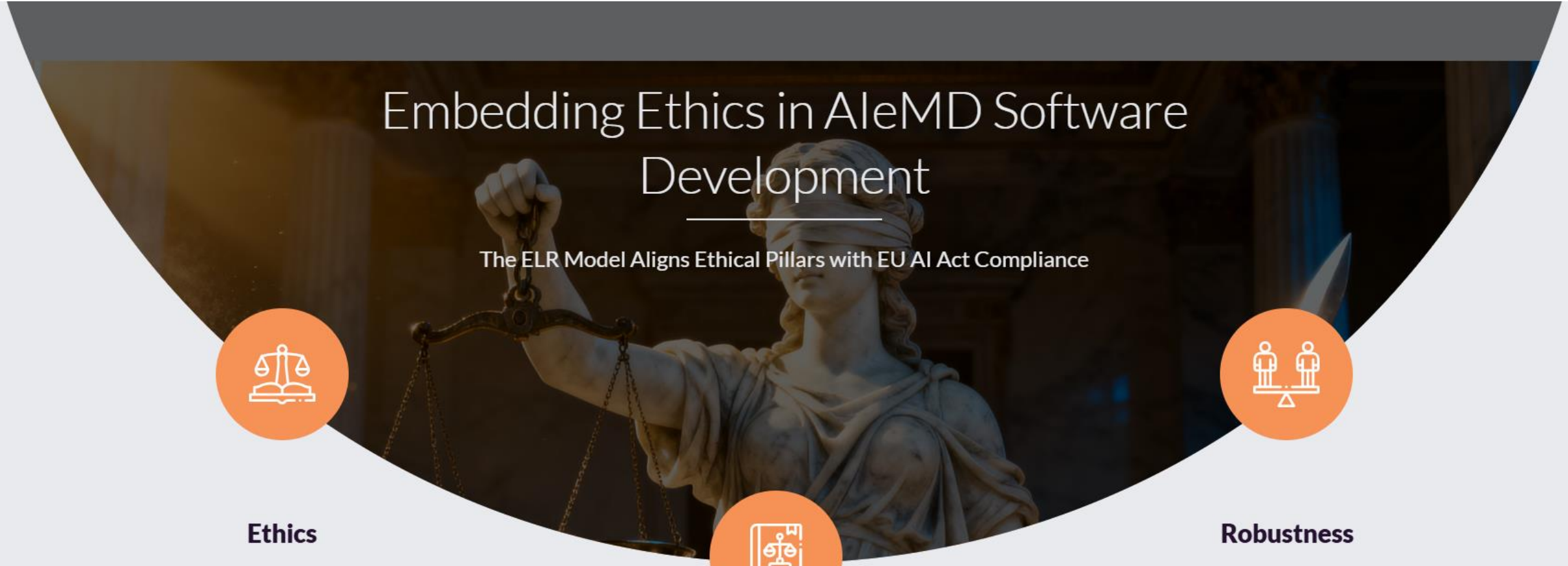
Widespread lack of understanding of AI-specific regulatory and ethical requirements among healthcare organizations

Most quality management systems remain outdated, relying on IEC 62304 without AI-specific updates

Experts emphasize urgent need for clearer guidance on ethical values and protection of fundamental rights

Comprehensive performance metrics are required to accurately assess AI system effectiveness

Current post-market surveillance is limited to traditional complaint handling, inadequate for AI's dynamic nature



Embedding Ethics in AIeMD Software Development

The ELR Model Aligns Ethical Pillars with EU AI Act Compliance

Ethics

Ensures respect for patient fundamental rights and integrates ethical risk assessments throughout AIeMD development.

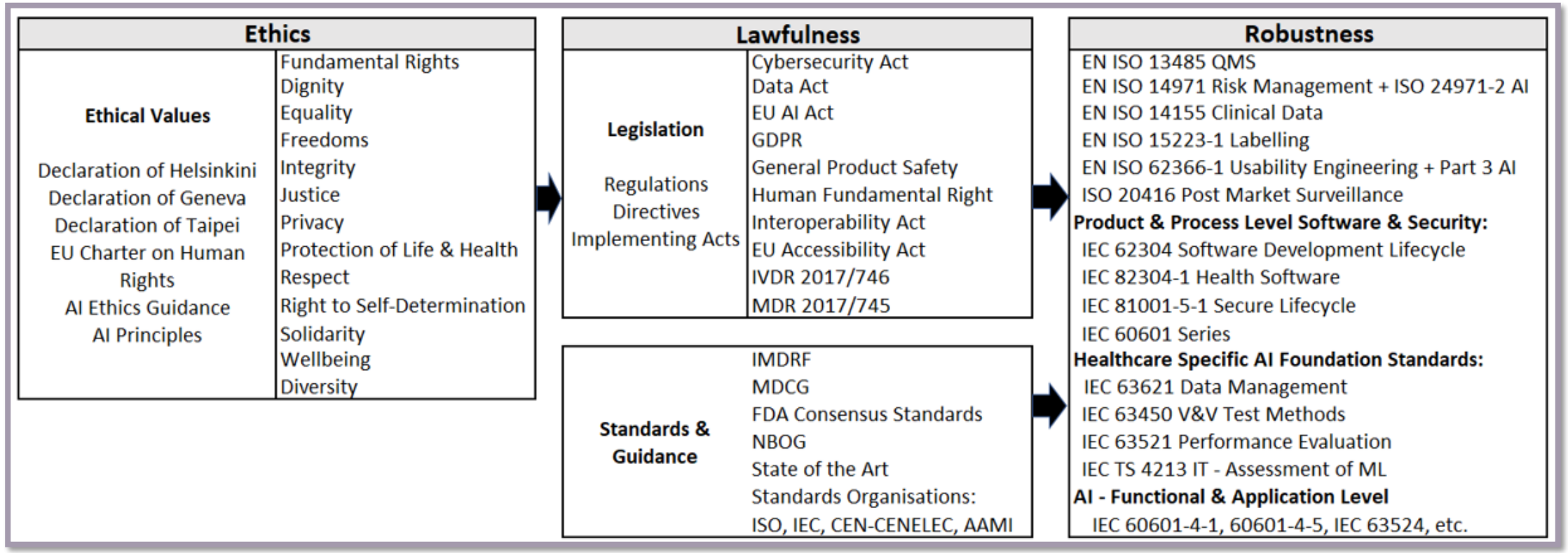
Lawfulness

Guarantees compliance with the EU AI Act and legal requirements governing medical device software

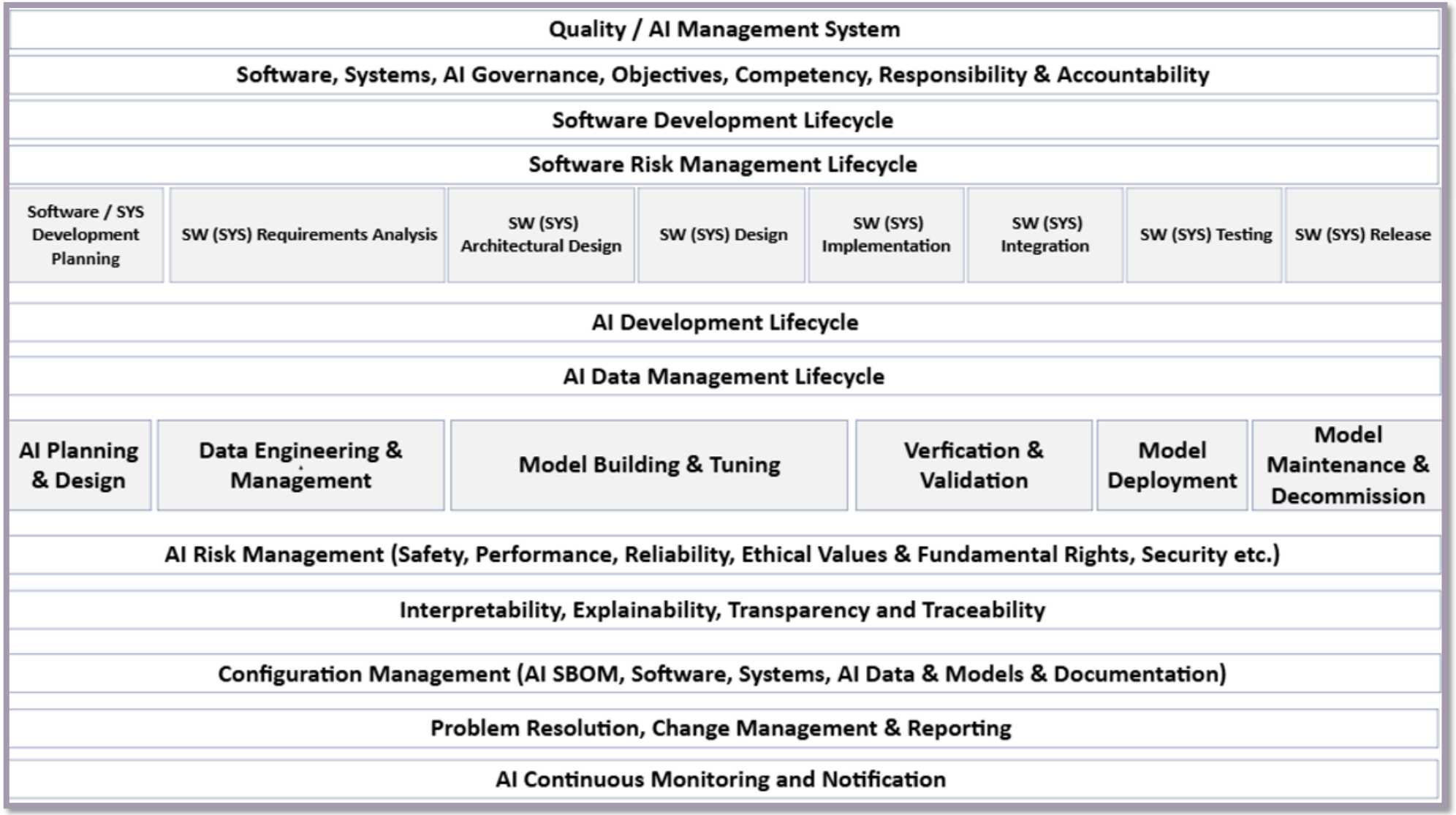
Robustness

Promotes reliable, secure, and resilient AI algorithms within the software development lifecycle.

Ethics, Lawfulness and Robustness Model

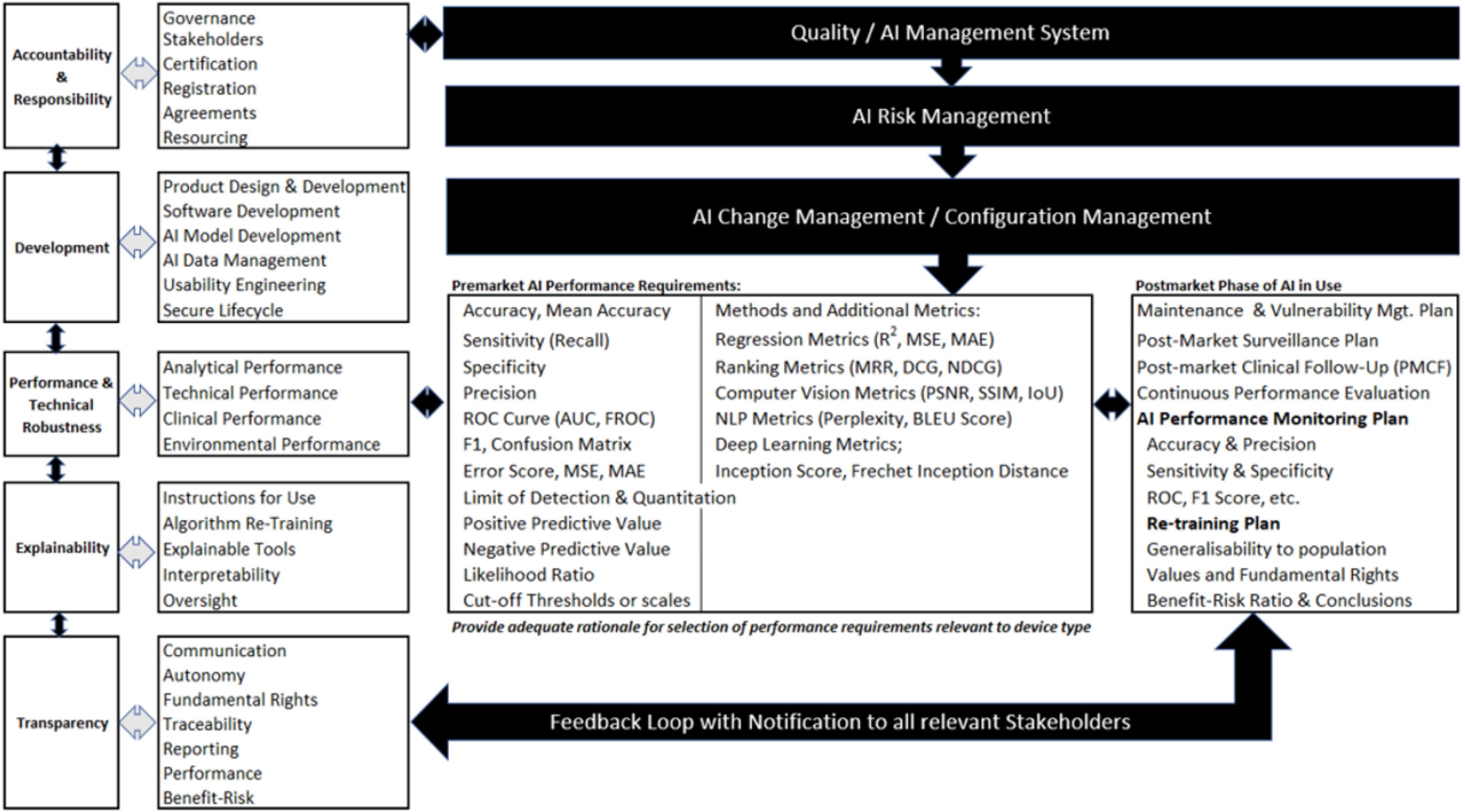


https://eprints.dkit.ie/id/eprint/935



DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts


Accountability & Responsibility
Governance
Stakeholders
Certification
Registration
Agreements
Resourcing
Development
Product Design & Development
Software Development
AI Model Development
AI Data Management
Usability Engineering
Secure Lifecycle
Performance & Technical Robustness
Analytical Performance
Technical Performance
Clinical Performance
Environmental Performance
Explainability
Instructions for Use
Algorithm Re-Training
Explainable Tools
Interpretability
Oversight
Transparency
Communication
Autonomy
Fundamental Rights
Traceability
Reporting
Performance
Benefit-Risk
Quality / AI Management System
AI Risk Management
AI Change Management / Configuration Management
Premarket AI Performance Requirements:
Accuracy, Mean Accuracy
Sensitivity (Recall)
Specificity
Precision
ROC Curve (AUC, FROC)
F1, Confusion Matrix
Error Score, MSE, MAE
Limit of Detection & Quantitation
Positive Predictive Value
Negative Predictive Value
Likelihood Ratio
Cut-off Thresholds or scales
Methods and Additional Metrics:
Regression Metrics (R^2, MSE, MAE)
Ranking Metrics (MRR, DCG, NDCG)
Computer Vision Metrics (PSNR, SSIM, IoU)
NLP Metrics (Perplexity, BLEU Score)
Deep Learning Metrics;
Inception Score, Frechet Inception Distance
Provide adequate rationale for selection of performance requirements relevant to device type
Postmarket Phase of AI in Use
Maintenance & Vulnerability Mgt. Plan
Post-Market Surveillance Plan
Post-market Clinical Follow-Up (PMCF)
Continuous Performance Evaluation
AI Performance Monitoring Plan
Accuracy & Precision
Sensitivity & Specificity
ROC, F1 Score, etc.
Re-training Plan
Generalisability to population
Values and Fundamental Rights
Benefit-Risk Ratio & Conclusions
Feedback Loop with Notification to all relevant Stakeholders

Unified TwAI Framework for AIeMD Excellence

Integrating ELR, AIDL, and AICM to enhance ethical standards, lifecycle management, and performance under EU AI Act and MDR compliance.

ELR: Ethical Lifecycle Requirements

Defines **ethical values** and lifecycle steps to embed integrity and transparency throughout AIeMD development.

1

AIDL: AI Development Lifecycle

2

Specifies a **top-down design** guiding manufacturers to update QMS and SDLC for AI-specific demands, boosting safety and traceability.

AICM: AI Compliance Model

Focuses on **measurable performance attributes** ensuring regulatory alignment with the EU AI Act and MDR, reducing defects and improving outcomes.

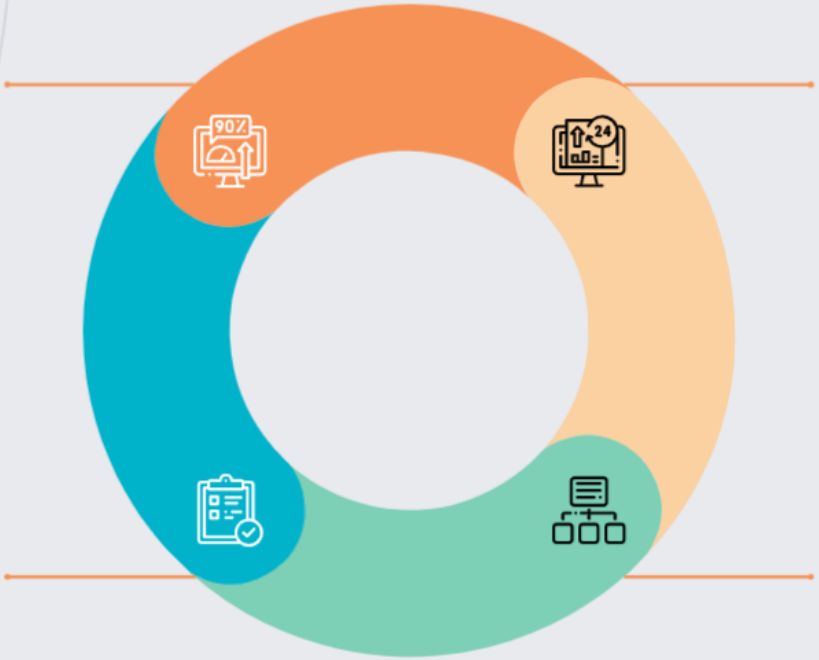
3

AI Change Management (AICM) Model & Performance Measures

Ensuring AIeMD Safety, Performance, and Regulatory Compliance through Robust Metrics and Continuous Oversight

Performance Metrics Selection

Define critical premarket AI performance metrics—accuracy, sensitivity, specificity, ROC curves—tailored to device type. Justify metric choice as single accuracy measures are insufficient.



Continuous Post-Market Monitoring

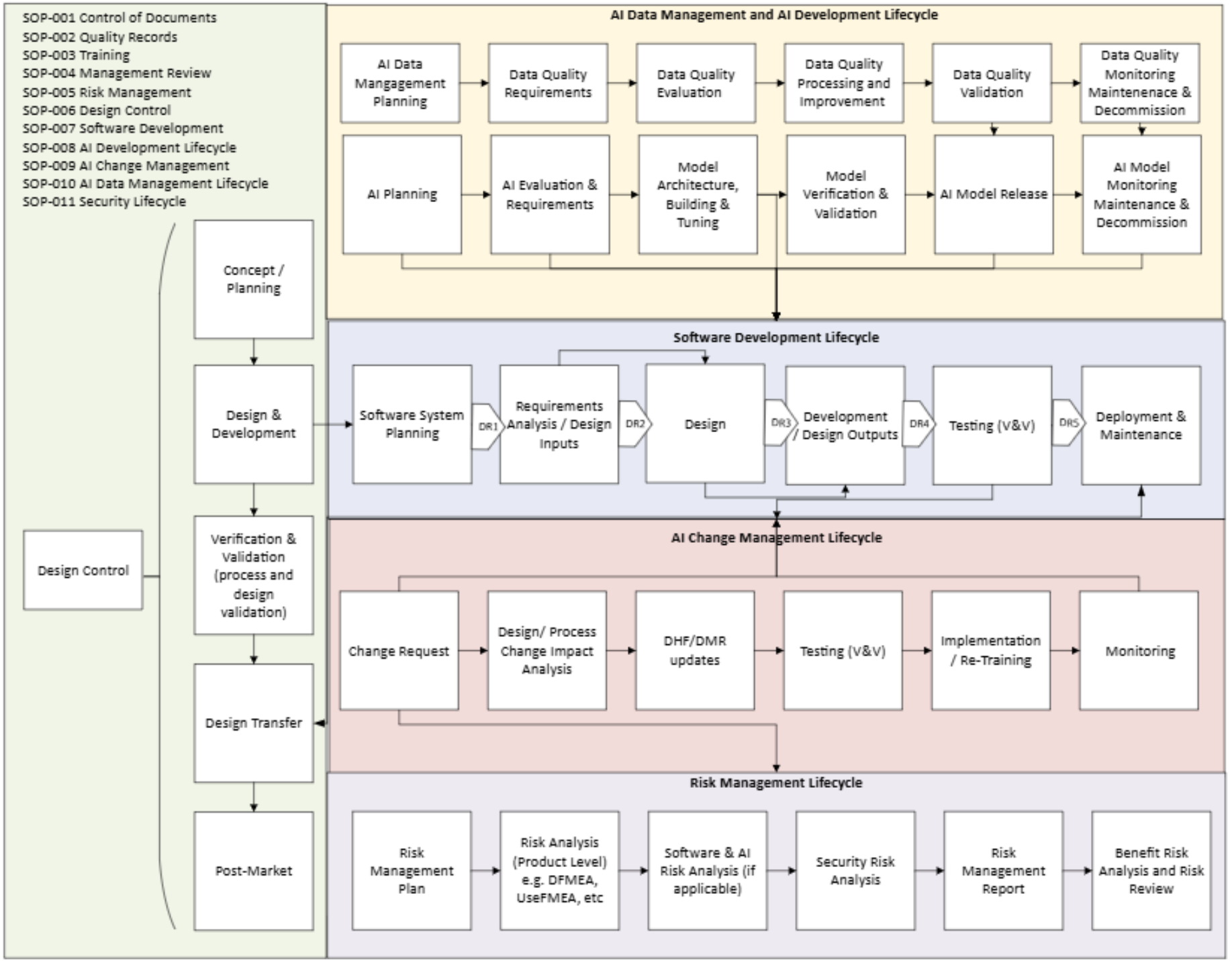
Implement ongoing monitoring of AI algorithms post-deployment to ensure sustained safety, reliability, and performance under real-world conditions.

Regulatory Compliance Requirements

Align operations with regulatory standards by integrating performance metrics and monitoring outcomes into formal compliance and safety documentation.

Change Management & Transparency

Manage iterative updates for both locked and adaptive AI algorithms, maintaining transparency to support regulatory compliance and build stakeholder trust.



Integrated AIDL Framework is copyright of St John Lynch, N. (2025) & DkIT as part of PhD Research

School of Informatics & Creative Arts

Standards

State of the Art (SoTA)

Harmonised Standards and State of the Art (SoTA) in Healthcare

International Electrotechnical Commission

[Check out this link for work programme]

Base Level State of the Art /Harmonised Standards
- EN ISO 13485 QMS
- EN ISO 14971 Risk Management
- EN ISO 14155 Clinical Data
- EN ISO 15223-1 Labelling
- EN ISO 62366-1 Usability Engineering
- ISO 20416 Post Market Surveillance

Software & Security:
- IEC 62304 Software Development Lifecycle
- IEC 82304-1 Health Software
- IEC 81001-5-1 Secure Lifecycle

Electrical Medical Equpment
- IEC 60601 Series

Artificial Intelligence - Foundation Level
- IEC 63450 AI-Verification & Validation Test Methods
- IEC 63521 AI-Performance Evaluation
- IEC TS 4213 IT - AI - Assessment of ML class. perf.

AI - Functional Level
- IEC TR 60601-4-1
- IEC 60601-4-5
- Etc. ....

AI - Application Level
- IEC 63542, etc.

Additional AI Guidance
- IEC 25024
- IEC 42001
- IEC 5338
- IEC 8183
- IEC 5259-2
- IEC 5259-4
- AAMI TIR 34971

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

IEC 63450 Test Methods for V&V

- Test Strategy & Planning
- Data Quality
- Bias and Fairness including some mitigation techniques
- Methods for Algorithm Modelling & Selection (e.g. cross-validation)
- Performance Characteristics
- Transparency Requirements
- Explainability (LIME, SHAP)

62/520/CD Committee Draft
Circulated 16 Aug 2024.
Closing date for comments 01 Nov 2024.

IEC 63521 - Performance Evaluation

TITLE OF PROPOSAL:

Machine Learning-enabled Medical Device – Performance Evaluation Process

SCOPE

(AS DEFINED IN ISO/IEC DIRECTIVES, PART 2, 14):

This document defines a standardized performance evaluation process for Machine Learning-enabled Medical Devices (MLMD). The set of processes, activities, and tasks described in this document establishes a common framework for MLMD performance evaluation.

To achieve this purpose, this document builds on established terms and concepts from IMDRF and medical device standards, while taking into account relevant AI-specific standards. One of the foundations for this document is IMDRF N41, Software as a Medical Device (SaMD): Clinical Evaluation. While IMDRF N41 has SaMD in its scope, whereas MLMD can also be SiMD, the concepts described in IMDRF N41 can be used for the purpose of this document and therefore beyond SaMD.

TARGET DATE(S)	FOR FIRST CD:	2024-11-30	FOR PUBLICATION:	2028-12-31

Robustness Properties using Statistical Methods

- Stability
- Sensitivity
- Relevance
- Reachability
- ….

Measurement & Metrics

ISO/IEC DIS 24029-3 and others

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

How We Measure Models — Accuracy, Sensitivity, Specificity, AUC, and Bias

Key metrics, clinical implications, and equity considerations for AI in medical devices

Accuracy — (TP+TN)/Total: overall correct predictions

Sensitivity (Recall) — TP/(TP+FN): detects positives; critical for screening

Specificity — TN/(TN+FP): avoids false alarms; reduces workflow burden

AUC / ROC — Area under curve: discrimination across thresholds

Bias & Equity — Performance gaps (%): measure across subgroups; often under-measured (17% always measured)

Clinical implication — High **sensitivity** can increase false positives; balance per use case

Equity note — Health equity metrics rarely *always* measured: **17%** always measured in surveys

Accuracy – is it always the best metric?

Accuracy = (TP + TN) / (TP + TN + FP + FN)

the proportion of correct predictions (both true positives and true negatives out of all predictions).

A perfect model would have zero false positives and zero false negatives and therefore an accuracy of 1.0, or 100%.

Because it incorporates all four outcomes from the confusion matrix (TP, FP, TN, FN), given a balanced dataset, with similar numbers of examples in both classes, accuracy can serve as a coarse-grained measure of model quality.

However, when the dataset is imbalanced, or where one kind of mistake (FN or FP) is more costly than the other, which is the case in most real-world applications, it's better to optimize for one of the other metrics instead.

For heavily imbalanced datasets, where one class appears very rarely, say 1% of the time, a model that predicts negative 100% of the time would score 99% on accuracy, despite being useless.

ISO/IEC TS 4213:2022

	A	B	C
Accuracy	91,74	88,27	29,15
Binary Accuracy	95,97	86,46	89,40
Precision	70,92	95,79	15,01
Recall	91,74	88,27	29,15
Specificity	96,38	74,66	92,24
F_1	80,00	91,88	19,82

6.2.3 Accuracy

Accuracy should not be used to express comparative performance across mod
known to be reasonably balanced.

6.2.4 Precision, recall and specificity

As precision increases, more true positives are detected, but false negatives are not accounted for. Precision of a class is calculated as:

$$p = \frac{T_P}{T_P + F_P}$$

As recall increases, more true positives are detected, but false positives are not accounted for. Recall of a class is calculated as:

$$r = \frac{T_P}{T_P + F_N}$$

As specificity increases, more true negatives are detected, but false negatives are not accounted for. Specificity of a class is calculated as:

$$s = \frac{T_N}{T_N + F_P}$$

The **true positive rate (TPR)**, or the proportion of all actual positives that were classified correctly as positives, is also known as **recall**.

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

A hypothetical perfect model would have zero false negatives and therefore a recall (TPR) of 1.0, which is to say, a 100% detection rate.

In an *imbalanced dataset* where the number of actual positives is very low, recall is a more meaningful metric than accuracy because it measures the ability of the model to correctly identify all positive instances.

Another name for recall is **probability of detection**

F1 Score – harmonic mean

Harmonic mean (a kind of average) of precision and recall

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}$$

This metric balances the importance of precision and recall, **and is preferable to accuracy for class-imbalanced datasets**.
When precision and recall both have perfect scores of 1.0, F1 will also have a perfect score of 1.0. More broadly, when precision and recall are close in value, F1 will be close to their value.
When precision and recall are far apart, F1 will be similar to whichever metric is worse.

Confusion Matrix

Device Result	True Positive	True Negative	Post-test Risk	Likelihood Ratio
Positive	A	D	$\frac{A}{A+D}$	$\frac{A / (A + B + C)}{D / (D + E + F)}$
Negative	B	E	$\frac{B}{B+E}$	$\frac{B / (A + B + C)}{E / (D + E + F)}$
Ungradable	C	F	$\frac{C}{C+F}$	$\frac{C / (A + B + C)}{F / (D + E + F)}$

Worst Case Scenario	Sensitivity	Specificity	Pre-test Risk	1
	$= \frac{A}{A+B+C}$	$= \frac{E}{D+E+F}$	$= \frac{A+B+C}{A+B+C+D+E+F}$	

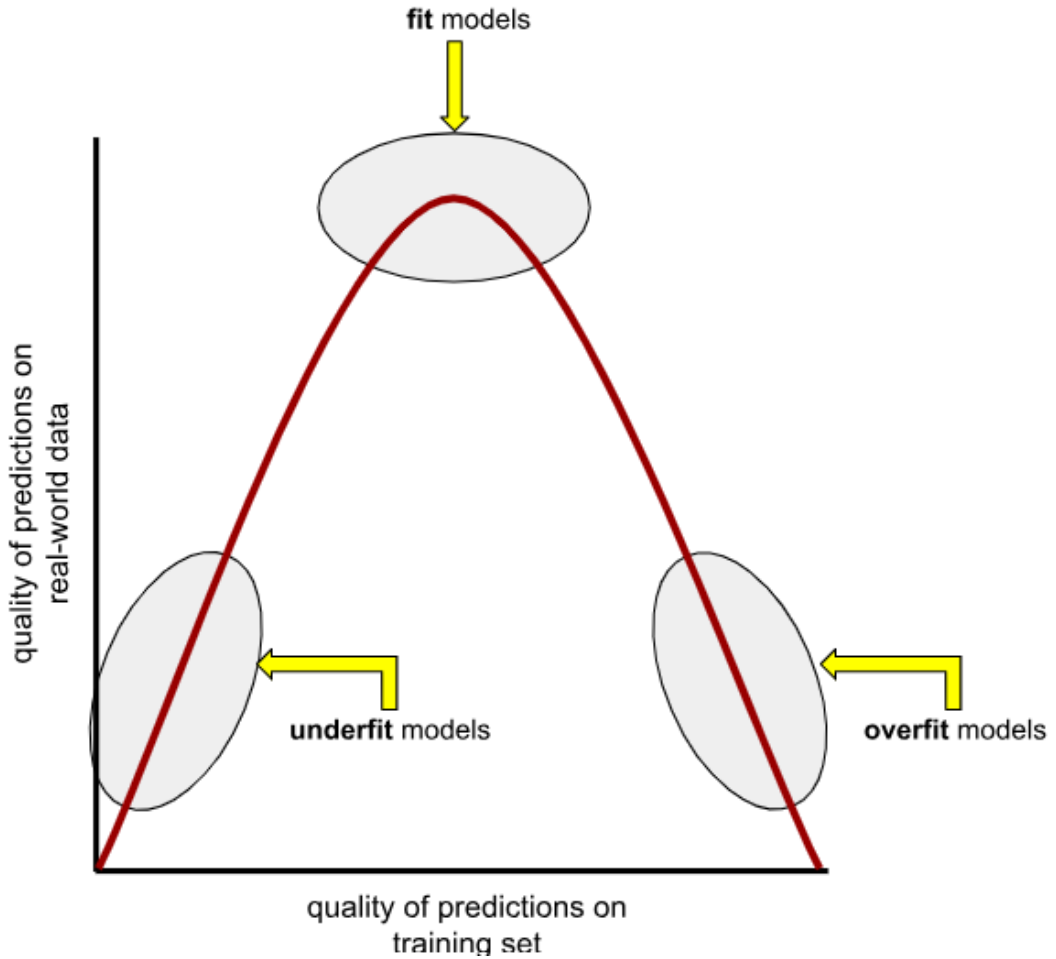
Guideline for Performance Metrics

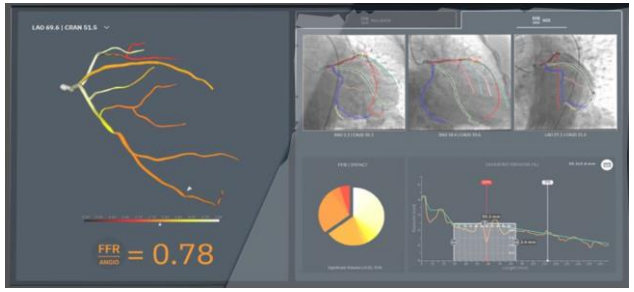
Metric	Guidance
Accuracy	• Indicator only of model training progress/convergence for balanced datasets • Use in combination with other suitable metrics • Understand use in imbalanced datasets and provide rationale for use and risk assess
Recall (True positive rate)	• Use when false negatives are more expensive than false positives
False positive rate	• Use when false positives are more expensive than false negatives
Precision	• Use when it is important for positive predictions to be accurate (clinical diagnosis of disease)

Overfitting – a common problem

Overfitting means creating a model that matches (*memorizes*) the **training set** so closely that the model fails to make correct predictions on new data.

Performs well in the lab but is worthless in the real world.





Explanation of Bias

Understanding the types and sources of bias that impact MLMD safety and performance

Ref. ISO/IEC TR 24971-2

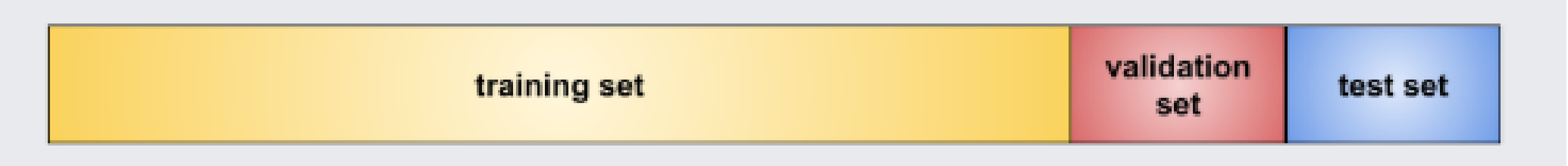
Selection bias: missing data, sample bias, coverage bias, or restricted access to patient data.

Interpretation bias: systematic difference in treatment of certain patient groups by different individuals.

Bias by confounding variables: causing the MLMD to believe in false cause-and-effect relationships.

Experimenter's bias: training continued until output agrees with trainer's pre-existing beliefs.

Data Management is an essential process



Independence in critical

3rd Party Pre-Trained AI Models

Evaluating Open / Closed Source

3rd Party Pre-Trained Models

LLM Medical Pre-trained Models

Table 1. LLM Medical AI/ML Pre-Trained Models

Model Name	Description	Location	Use Case
BioBERT	A pre-trained biomedical language representation model.	https://github.com/dmis-lab/biobert	Biomedical text mining, NLP tasks
BioGPT	A generative pre-trained transformer model for biomedical text generation.	https://github.com/microsoft/BioGPT	Biomedical text generation, NLP
BioMistral-7B	A Collection of Open-Source Pretrained Large Language Models for Medical Domains	https://huggingface.co/BioMistral/BioMistral-7B	Biomedical text mining, NLP tasks
BlueBERT	BERT-based model trained on PubMed abstracts and MIMIC-III clinical notes.	https://github.com/ncbi-nlp/bluebert	Biomedical research, clinical notes
ClinicalBERT	A Pretrained Model on a large corpus of 1.2B words of diverse diseases. Fine-tuned on 3 million patient records.	https://huggingface.co/medicalai/ClinicalBERT	Medical question-answering (QA) tasks
ClinicalBERT	BERT model fine-tuned on clinical notes.	https://github.com/EmilyAlsentzer/clinicalBERT	Clinical text analysis, EHR data
LLaVA-Med	A large language and vision model trained using a curriculum learning method for adapting LLaVA to the biomedical domain.	https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b	Medical question-answering (QA) tasks
SciBERT	A BERT-based model trained on scientific text.	https://github.com/allenai/scibert	Scientific literature analysis

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

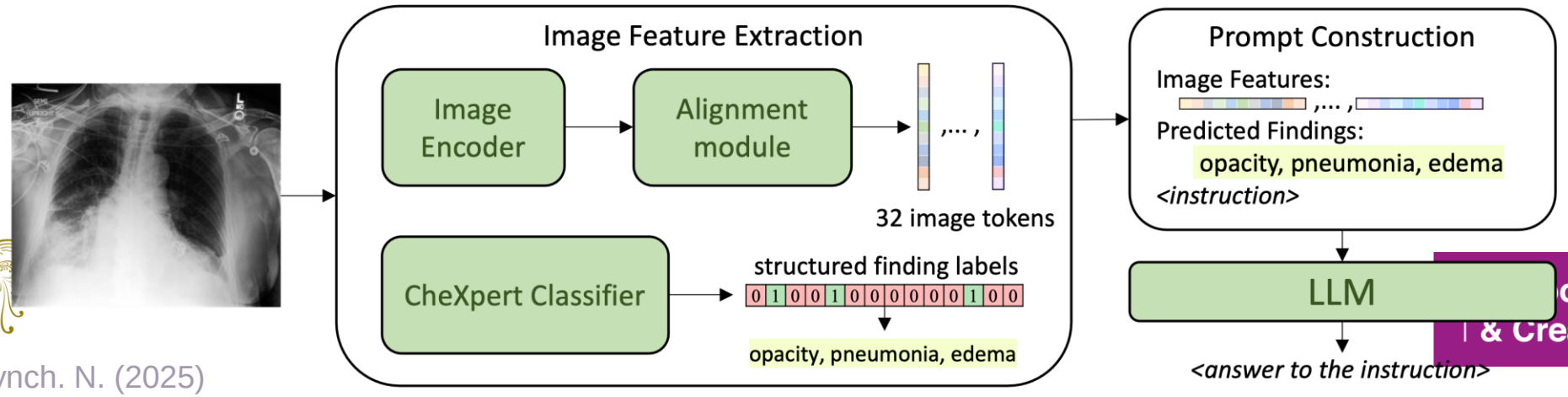
& Creative Arts

Pre-Trained Vision Models

(aka Foundation Models)

Table 2. Pre-Trained Vision Models

Model Name	Description	Link	Use Case
Vision Transformer (Open-source, (OS))	A deep learning model for classification & segmentation on imaging data.	https://github.com/google-research/vision_transformer	Radiology, Pathology
CheXbert (OS)	Automatic Labelers and Expert Annotations.	https://github.com/stanford-mlgroup/CheXbert	Radiology
UNet (OS)	A convolutional neural network for biomedical image segmentation.	https://github.com/zhixuhao/unet	Biomedical image segmentation
nnU-Net (OS)	A self-configuring method for biomedical image segmentation.	https://github.com/MIC-DKFZ/nnUNet	Biomedical image segmentation
AI for Breast Cancer (CS)	AI model for improving breast cancer screening accuracy	https://health.google/caregivers/mammography/	Breast cancer screening, diagnostics
AI for Diabetic Retinopathy (CS)	Detects diabetic retinopathy and diabetic macular edema	https://health.google/caregivers/arda/	Diabetic retinopathy screening, ophthalmology



What's the problem?

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

Table 4. Risks and Potential Mitigations to use of Pre-Trained Models in Medical Devices

Risk	Example	Mitigation
Security & Privacy	Data poisoning, privacy leakage, model inversion, model theft, member inference, exfiltration, backdoor attack, manipulation, exposed API and credentials, tampering, insecure components, unauthorized use of model/data.	Full assessment of pre-trained model selection and purpose with security and privacy controls in place. Model versioning, continuous monitoring, zero-trust for secure deployment and maintenance.
Bias & Inequity	Intrinsic Bias (Data, Spatial, Temporal, Collection Bias), Extrinsic Bias (Task-specific, Gen-AI biases), Algorithmic Bias (architecture, feature selection, loss function, metrics, training methods, competency, sampling, labelling, etc.	Risk Analysis and Adequate Controls; including understanding source of data used in pre-learning, re-sampling controls, data balancing, independence, augmentation, filtering; de-biasing, competency; standardisation, etc.
Transparency	Awareness of pre-training model details for assessment; labelling / documentation with adequate information available	Documentation availability for assessment and maintenance
Performance	Metric suitability, adequacy; model drift assessment, over-fitting, under-fitting, generalizability.	Refinement and evaluation for model generalizability
Technical Suitability	Architecture type, purpose, use case fit, data suitability to use case, etc.	Assessment for suitability including layers, nodes, loss function, data, sample size, representativeness of population / problem, parameters, hyperparameters, etc.

More AI-related risks; 3rd Party SOUP

Data Poisoning - Injecting corrupted or inauthentic data into datasets for training compromises the integrity of AI. The impact could be inaccurate diagnoses, which could cost lives and break down trust in the healthcare system.

Model Evasion - What if adversarial data became part of a learning model? That's the gist of this technique, and it could affect how medical devices with AI diagnose. They could be wrong, endangering patients and possessing

Model Inversion – stealing or exploiting an AI model and/or access sensitive information. Attackers can employ model inversion on predictive models for diagnosis, violating patient privacy and undermining trust.

Performance Drift - Predictive accuracy from new inputs "drifts" from the model's performance during training. In short, it impacts the model's accuracy, possibly resulting in erroneous diagnoses and unnecessary treatments.

Bias - When AI learning models learn from datasets that aren't diverse, it could limit what populations can safely use the devices for diagnoses. Misdiagnosis or failure to recognize conditions could lead to further mistrust from marginalized communities.

And more, data leakage, overfitting, etc.

No problem – Just evaluate

Table 5. Pre-Trained Model Checklist

Category	Consideration for selection and development[1]	Supporting Comment[2]
Transparency	Is the model under development utilising a 'pre-trained model' or 'foundational' model?	Pre-trained models refer to those used in the development of AI medical devices. They are not limited to models specifically developed for medical devices. Sources of pre-trained models are diverse and may include: • Medical device manufacturers • Third-party suppliers • Third-party service platforms • Open-source networks the types of pre-trained models, which may vary in modality and parameter scale. It includes, but is not limited to:
Security	Adversarial Security; what level of documentation is provided by 3rd party supplier. What additional measures are necessary if assuming zero trust principle?	The provider is encouraged to declare the model's adversarial security. If applicable, examples should be provided showing the types of adversarial attacks handled and the model's performance under such conditions.
Security	Privacy Protection; what level of documentation is provided by 3rd party supplier. What additional measures are necessary if assuming zero trust principle?	The provider should declare the privacy protection measures adopted by the model, meeting the following requirements: • Use appropriate techniques (e.g., differential privacy) to prevent leakage of training data, including distribution and individual data inference • Ensure protective measures are in place for data upload and storage operations generated by the model code
Evaluation Methods	General methods for Quality Compliance of Pre-Trained Models; What additional requirements/procedures are necessary to ensure pre-trained model is adequately evaluated for its intended purpose?	The quality evaluation of pre-trained models includes assessment of the model description, quality characteristics, and other relevant aspects. The model provider should submit the pre-trained model itself, its documentation, and other necessary materials for evaluation.

EU AI Act

Proposal Solution for Simplification of MDR/IVDR and impact to AI Act 2024/1689

Check out paper accepted by Journal of Medical Device Regulations in May 2026 and DkIT Stór for Open Source Papers

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

References

1. Lynch, N.S.J., Loughran, R., McHugh, M., McCaffrey, F. (2026). Trustworthy Artificial Intelligence in Healthcare: A Proposed Framework. In: Yilmaz, M., Clarke, P., Riel, A., Messnarz, R., Zelmenis, M., Buce, I.A. (eds) Systems, Software and Services Process Improvement. EuroSPI 2025. Communications in Computer and Information Science, vol 2657. Springer, Cham. https://doi.org/10.1007/978-3-032-04288-0_5
2. St John Lynch, N., Loughran, R., McHugh, M., McCaffrey, F. (2024). Artificial Intelligence-Enabled Medical Device Standards: A Multidisciplinary Literature Review. In: Yilmaz, M., Clarke, P., Riel, A., Messnarz, R., Greiner, C., Peisl, T. (eds) Systems, Software and Services Process Improvement. EuroSPI 2024. Communications in Computer and Information Science, vol 2179. Springer, Cham. https://doi.org/10.1007/978-3-031-71139-8_8
3. St John Lynch, N. et al (2025) Evaluating Pre-trained 3rd Party AI Models in Medical Device Software Development for Responsible and Ethical Use. In Press
4. St John Lynch, N. et al (2026) A Single Regulatory Framework for AI-Enabled Medical Devices: Implications of the EU Simplification Digital Package and Omnibus Proposal. Journal of Medical Device Regulations. May 2026. In Press

Open Source from DkIT Stór

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

School of Informatics & Creative Arts

Any Questions?

Thank You

DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

This publication has emanated from research conducted with the financial support of Research Ireland under Grant number 21/FFP-A/9255, a collaboration between DkIT and UCD. With thanks to Lead PI: Prof Fergal McCaffrey, DkIT.

Niamh St John Lynch
BSc MSc MScSED P.Dip MSc MIEI
PhD Candidate

Taighde Éireann
Research Ireland

School of Informatics & Creative Arts