#1

Standardization, bias and data quality – focus on trustworthy, clinically relevant AI

Enrico Giampieri

University of Bologna – Surgical and Medical Sciences Dept.

AI in Healthcare

Categorization of AI for healthcare

- WHO guidance document on ethics and governance of AI for health classifies AI health applications in:
  - **Healthcare** (diagnosis and prediction-based diagnosis, risk identification, decision support systems)
  - **Biomedical research** (drug repositioning, genomic medicine, big data exploitation)
  - **Health system management** (administrative workflows, logistics)
  - **Public health and public health surveillance** (monitoring of disease outbreaks, pandemic preparedness, health promotion)
- AI use requires balancing
  - Benefit for the patients, the healthcare system, society
  - Costs
  - Potential risks
- To be really used it needs:
  - Low absolute risk
  - Great risk/benefit ratio
  - High quality
  - **For these estimates to be believable!**

AI risks

Subtitle here

- What can go wrong?
  - Privacy breach
  - Worst clinical outcome
  - False improvements
  - System inefficiencies
- What are the components?
  - Data
  - Algorithms
  - Processes
- How can we trust our systems
  - Legitimacy
  - Morality
  - Robustness

- How can it go wrong?
  - missing data due to failed harmonization
  - data missingness due to pathway of care
  - distortion due to poor data quality
  - limitation in data exchange
  - Specification drift
  - data leakage
  - prediction bias
  - Human bias in the loop
  - error obfuscation
  - Lack of transparency
  - Conflict of interest

What do we need for trustworthy AI?

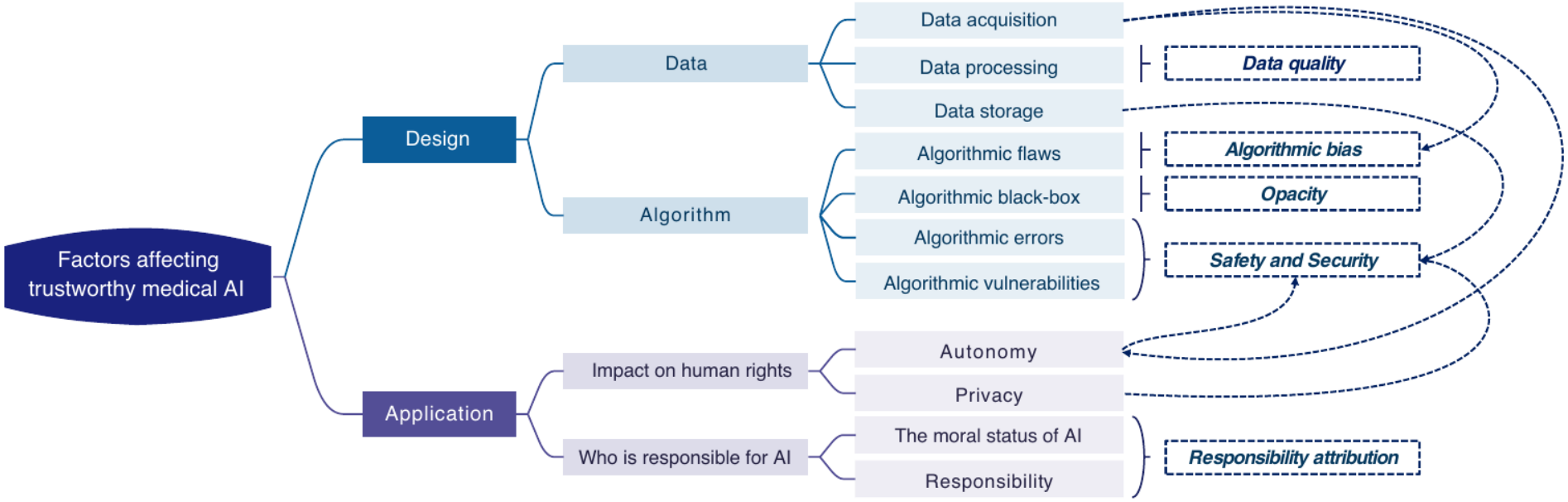


Fig. 1 Factors affecting trustworthy medical AI

From: ethics and governance of trustworthy medical artificial intelligence

Models are complicated

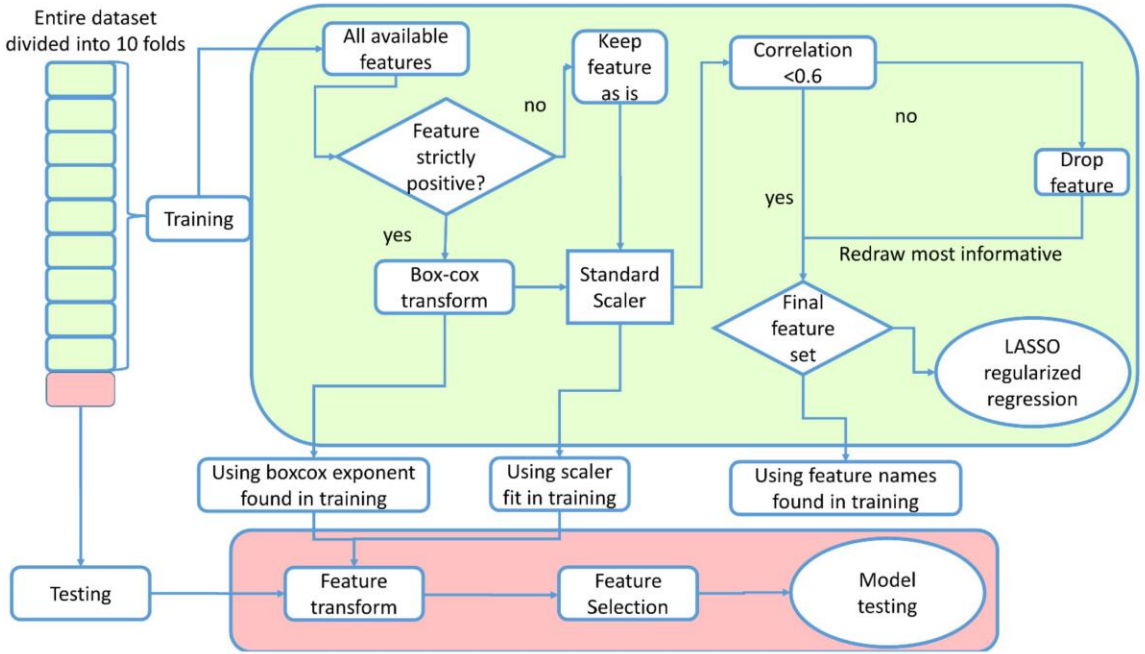


Figure 1. Representation of the 10-fold cross-validation approach used for the training and testing of a single feature group classifier. The input family was selected before entering the green box.

Data are complicated

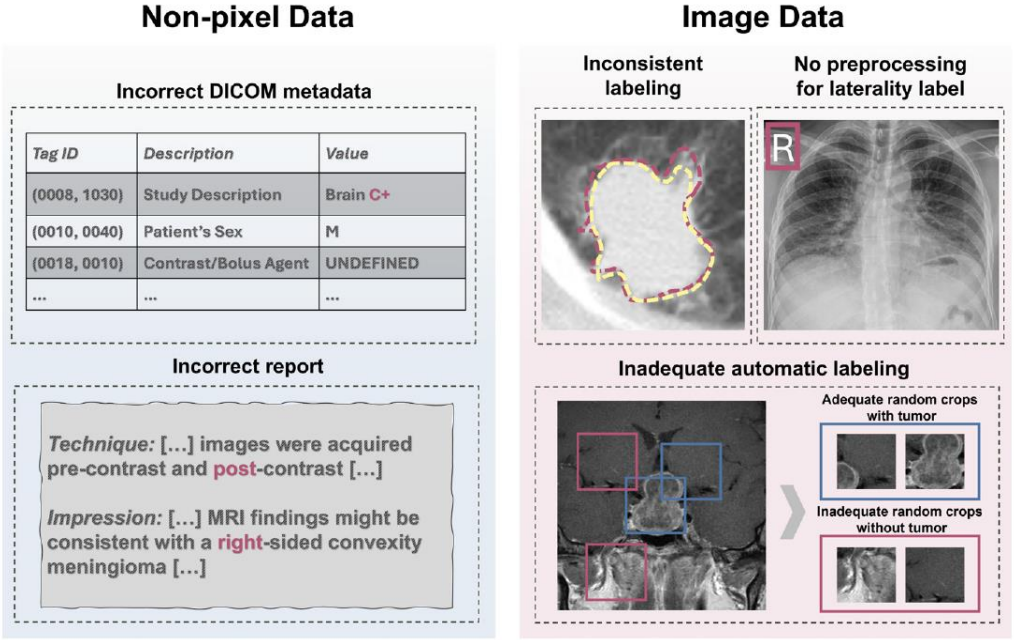


Figure 6. Potential and practical bias sources relevant to medical imaging artificial intelligence based on data type (i.e., non-pixel and image data). Radiological images belong to chest computed tomography (upper left panel), chest X-ray (upper right panel), and pituitary magnetic resonance imaging (lower panel).

From: bias in artificial intelligence for medical imaging - fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects

European Data Strategy

The GDPR was only the beginning

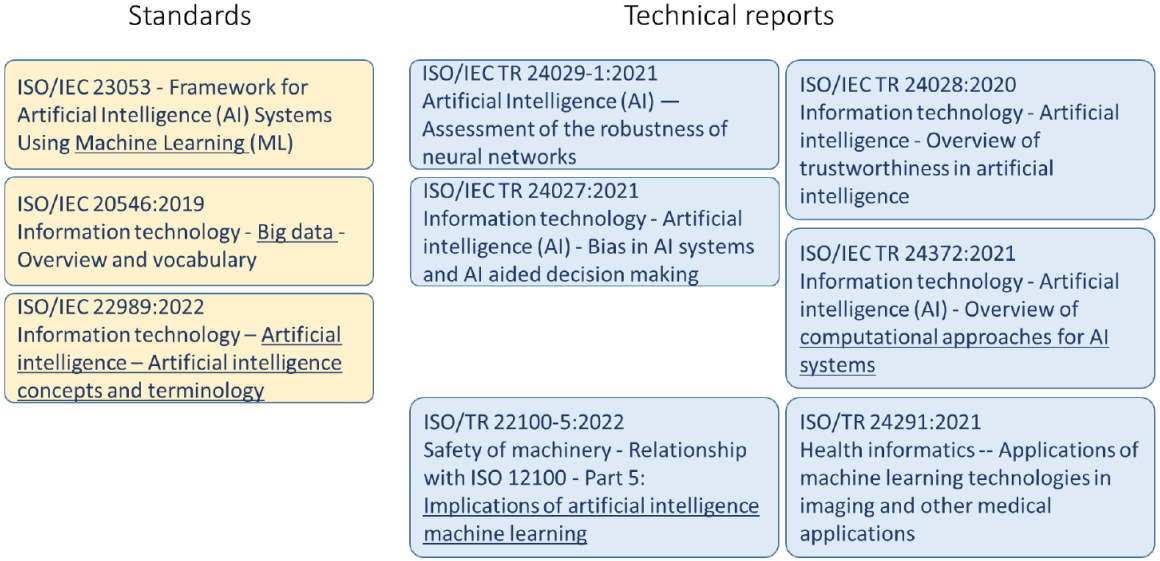
- **Data Act**
- In effect from 12 September 2025
- How to exchange data
  - Between private and public
  - With extra European countries
- **Might require an update to the informed consents!**

- **AI Act**
- In effects from 2 August 2026
- Risk based approach
- Other principles
  - Transparency
  - Reliability
  - Minimizing data
  - Responsibility

Standard frameworks for trustworthy AI

The ISO/IEC approach

Figure 4 Relevant ISO/IEC publications (standards and technical reports) regarding both horizontal aspects of AI (e.g. robustness, bias, machine learning = ML) and health specific aspects (i.e. ML applications for imaging and other medical applications).



From: data quality requirements for inclusive non biased and trustworthy AI

Bias in AI

Sources of distortion

The strength of the chain and the links

- Data imbalance and distortion
  - Data quality issues
  - variability over time or across sites,
  - information uncertainty
- Algorithms
- Optimization criteria
- Statistical assumptions
- Human procedures

Algorithmic bias

human-induced bias and data-induced bias

- Data-induced bias refers to the bias when the training data is not accurate, representative or sufficient
  - Even if the data is accurate and representative, the result will be meaningless if there is a problem in the data annotation process or inconsistent data collection
    - Es: gold standard for diagnosis of pulmonary nodules is a biopsy, but not every patient will have one
- Human-induced bias written by the developers
  - individuals are always influenced by
    - their own moral perceptions
    - relevant interests
  - Can be is intentional or unintentional
  - likely to be reinforced and amplified with the accumulation of data and iterations of algorithms.
- deep learning is a “black box”, it is opaque and uninterpretable, which makes the biases difficult to be detected
  - Es: Melanoma detection for underrepresented and more complex phototypes
  - Es: Using health care costs (rather than disease) to represent the level of health need
    - Affect underprivileged areas and categories

Dataset biases

Are you looking at reality?

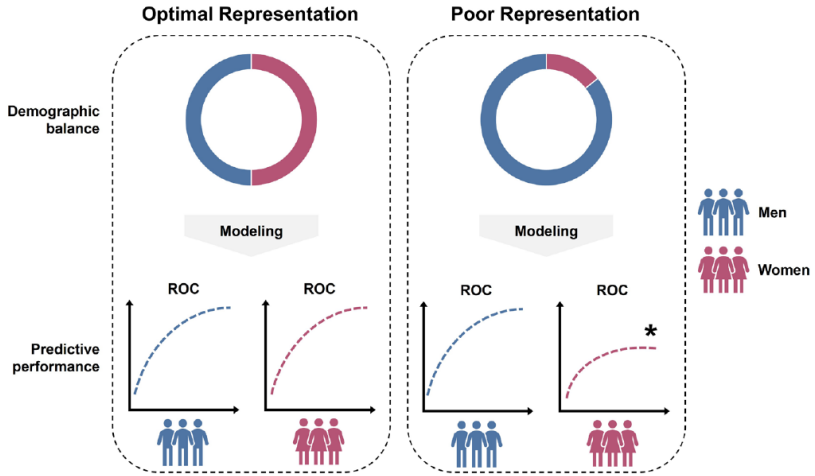


Figure 4. Over-simplified illustration of optimal and poor representation of subgroups, such as gender in this case, and their effect (*) in subsequent modeling. ROC, receiver operating characteristics.

Treatment A had higher success rates for both small and large kidney stones when looking at each individually.

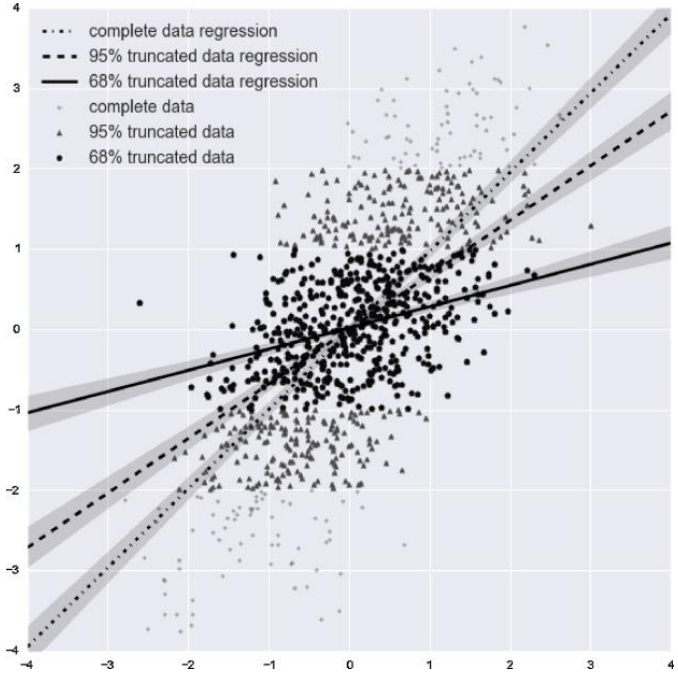
	Treatment A	Treatment B
Small Kidney Stones	93% (81/87)	87% (234/270)
Large Kidney Stones	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

Treatment B had a higher success rate when looking at all sizes at once.

From: bias in artificial intelligence for medical imaging - fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects

Wrong statistical assumptions

Defining what right means is not easy



Human limitations

You can't think about things you don't have words for

Puppy

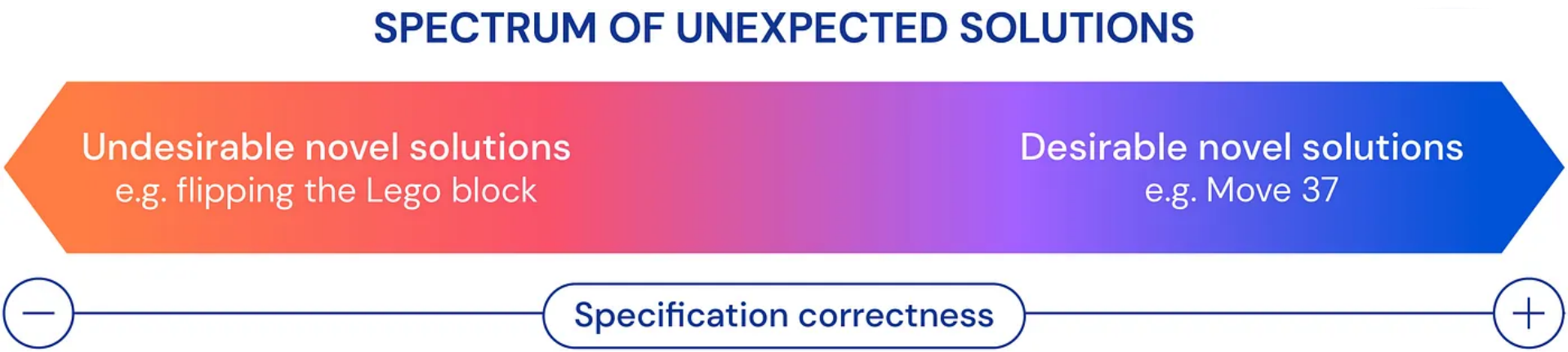
or

muffin?

source: boredpanda.com

Specification gaming

My model is smarter than me



From: Specification gaming: the flip side of AI ingenuity

Specification gaming

My model is smarter than me

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.

Specification gaming

My model is smarter than me

Coyote

Wolf

Specification gaming

My model is smarter than me

It also muddies the origin of certain data sets. This can mean that researchers miss important features that skew the training of their models. Many unwittingly used a data set that contained chest scans of children who did not have covid as their examples of what non-covid cases looked like. But as a result, the AIs learned to identify kids, not covid.

Driggs's group trained its own model using a data set that contained a mix of scans taken when patients were lying down and standing up. Because patients scanned while lying down were more likely to be seriously ill, the AI learned wrongly to predict serious covid risk from a person's position.

In yet other cases, some AIs were found to be picking up on the text font that certain hospitals used to label the scans. As a result, fonts from hospitals with more serious caseloads became predictors of covid risk.

Specification gaming

My model is smarter than me

He and his colleagues had one such problem in their their study with rulers. When dermatologists are looking at a lesion that they think might be a tumor, they'll break out a ruler—the type you might have used in grade school—to take an accurate measurement of its size. Dermatologists tend to do this only for lesions that are a cause for concern. So in the set of biopsy images, if an image had a ruler in it, the algorithm was more likely to call a tumor malignant, because the presence of a ruler correlated with an increased likelihood a lesion was cancerous. Unfortunately, as Novoa emphasizes, the algorithm doesn't know why that correlation makes sense, so it could easily misinterpret a random ruler sighting as grounds to diagnose cancer.

Explainability and Reliability

How to be wrong right

A.S. Albahri et al. *Information Fusion 96 (2023) 156–191*

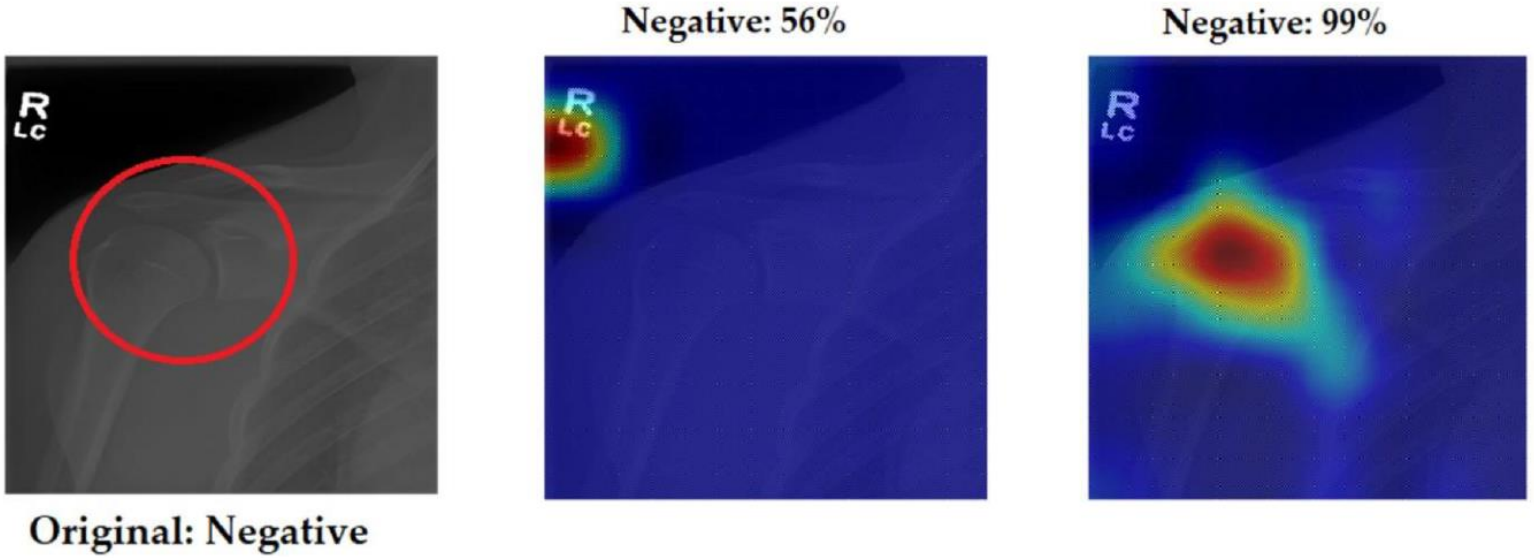


Fig. 17. Grad-CAM visualisation of two DL models.

Explainability and Reliability

How to be wrong right

- Challenges:
  - Complexity of AI Models (black boxes)
  - Trade-off Between Accuracy and Explainability
  - Human-AI Interaction (knowledge barriers)

Data Quality and access

Data Harmonization

merging different datasets together

- We need shared:
  - Common data models
  - Acquisition procedures
  - Analisys target

- Challenges in data harmonization
  - Heterogeneous Data Sources
  - Data Privacy and Security
  - Bias in Multimodal Data
  - Subjectivity in data interpretation
  - Complex logistics of data collection
  - Balancing patient needs with accuracy and costs

Dataset documentation

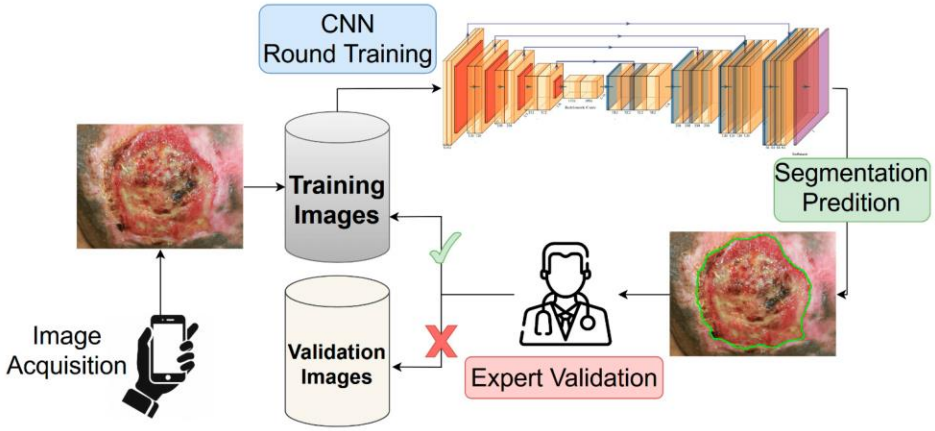
Article: Datasheets for Datasets

- FAIRification of the data is not enough
- Topics to be disclosed:
  - Motivation
  - Composition
  - Collection Process
  - Preprocessing/cleaning/labeling
  - Uses
  - Distribution
  - Maintenance

Data preprocessing

Garbage In – Garbage Out

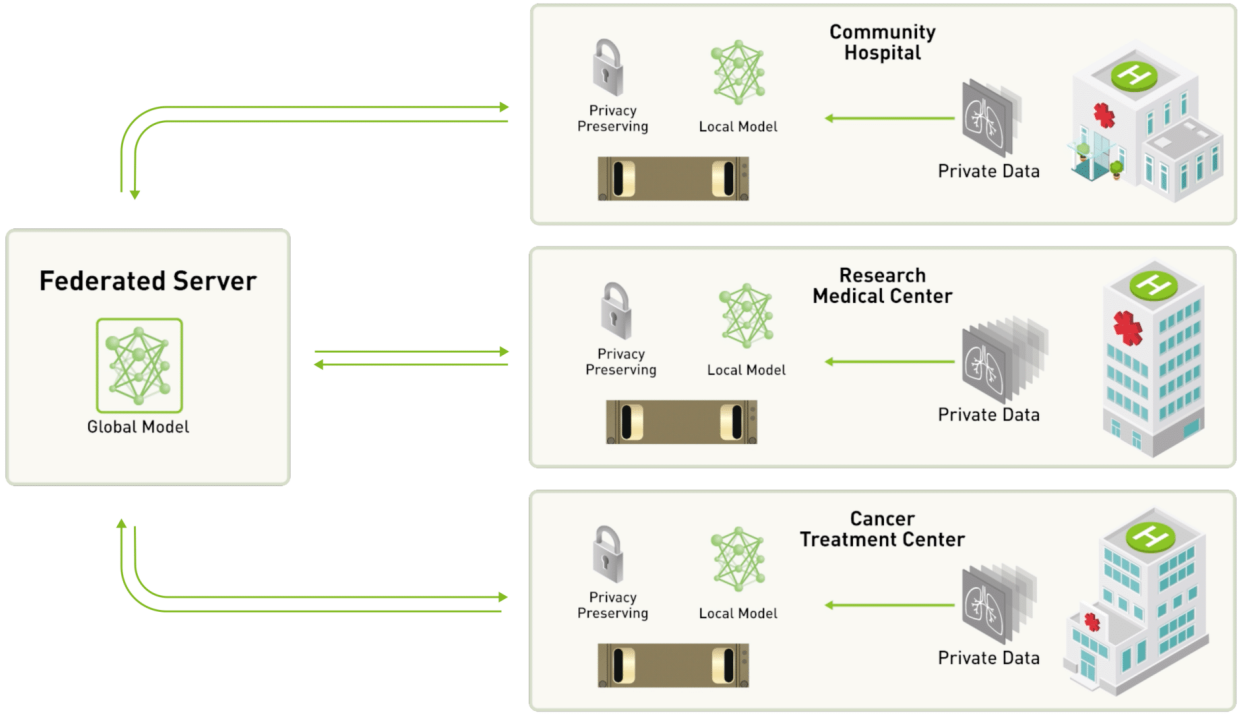
- Good data > big data
  - Practice makes ~~perfect~~ **permanent**
- Getting **a lot** of data is **easy**
- Getting **good** data is **hard**
- **Even defining good data is hard**
  - **Who does the annotation?**
  - **The agreement problem**



Data Access

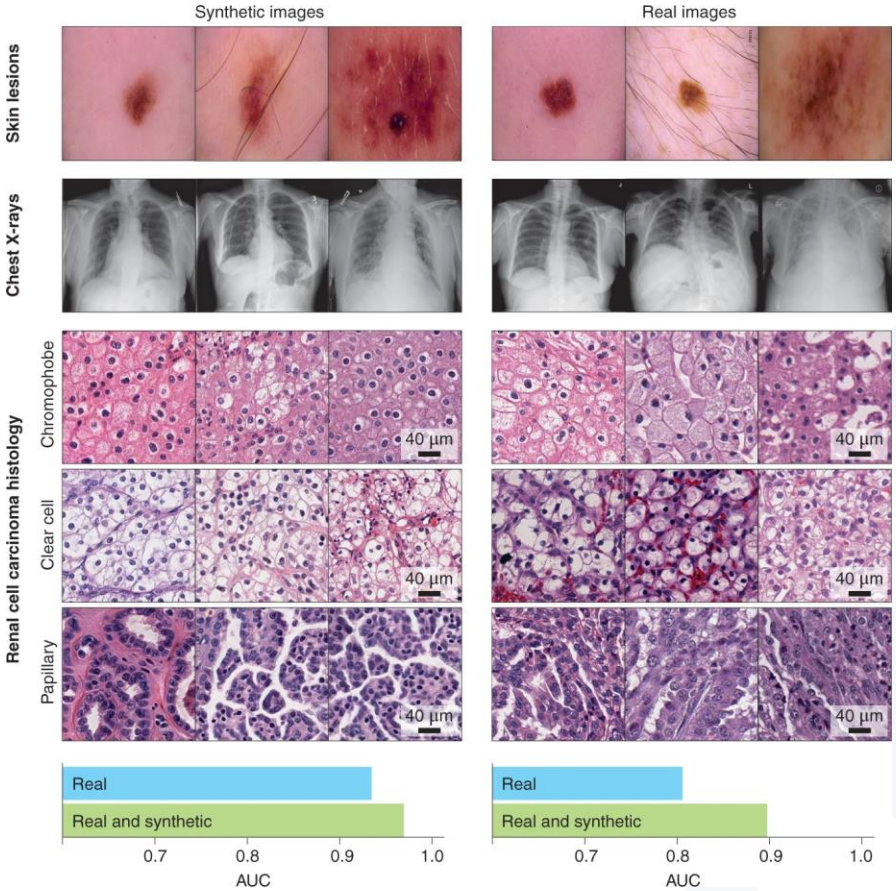
Legally, of course

- Single center datasets
- Shared data vaults
- Federated learning
- Synthetic data



Synthetic data

Generative AI for the rescue



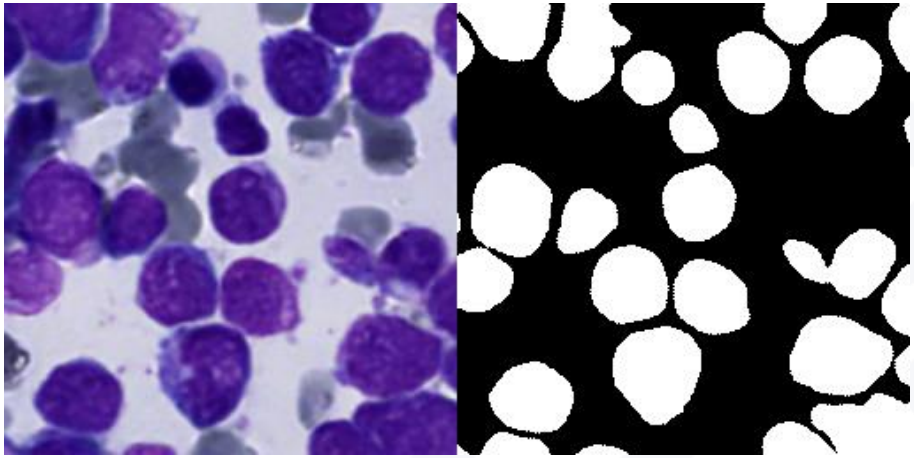
From: Synthetic data in machine learning for medicine and healthcare

A practical example

Segmentation of cells in AML cytological data

The first step of many

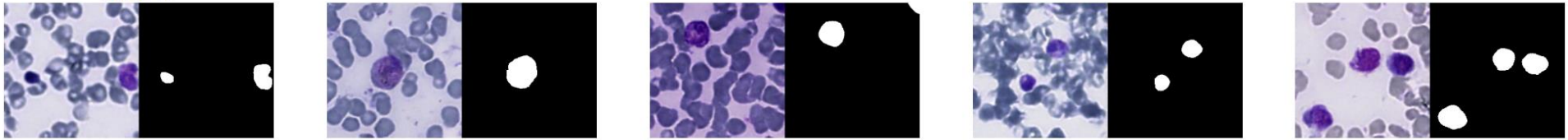
- Preparing the data
  - Access, preprocess, anonymize
- Getting the segmentations right
  - Millions of patches
- Identifying the model to generate the images
  - Time-quality tradeoff
  - assumptions
- Making it behave



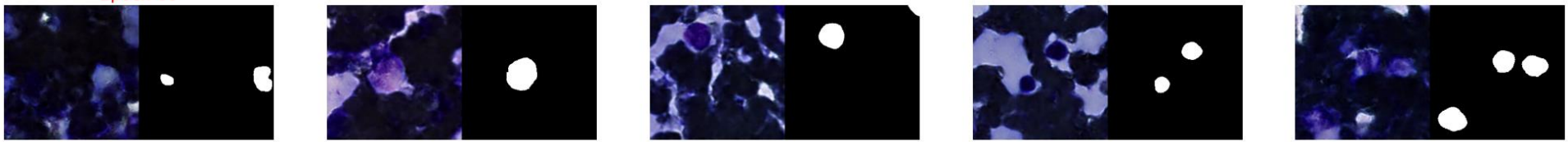
Mode oscillation problem

The model sometimes can't decide what is best

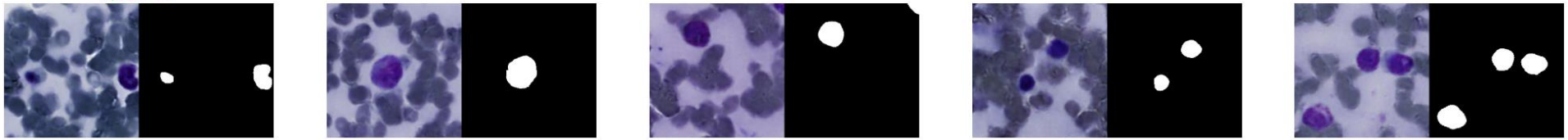
Epoch 55



Epoch 60

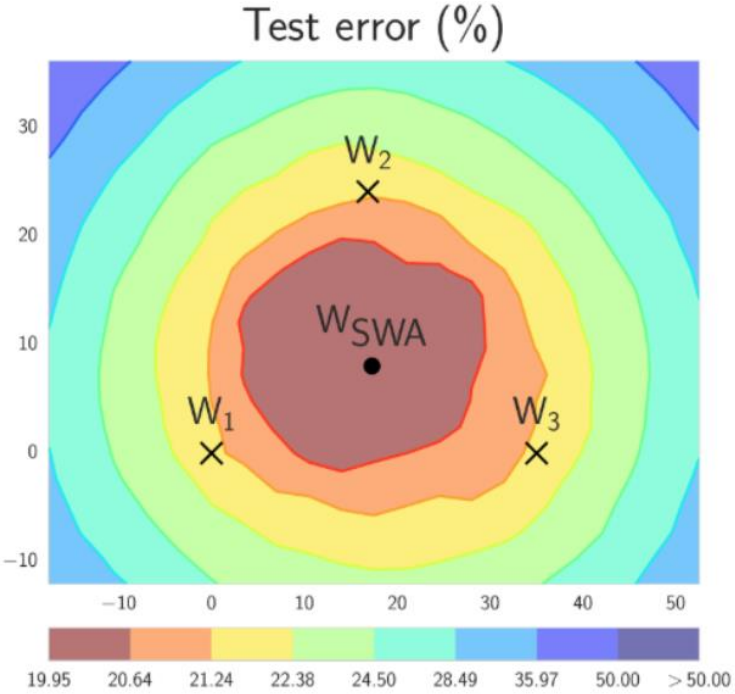


Epoch 65



How to solve the issue

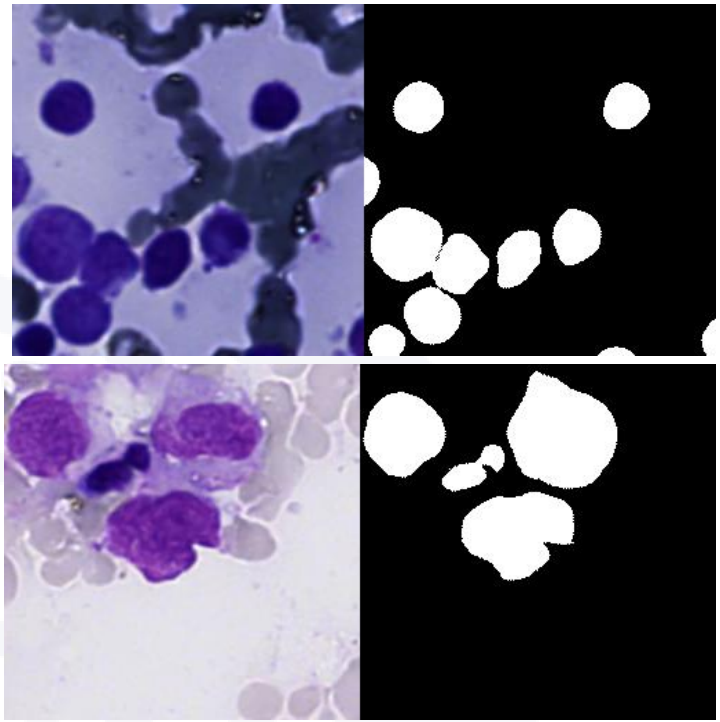
Mode averaging

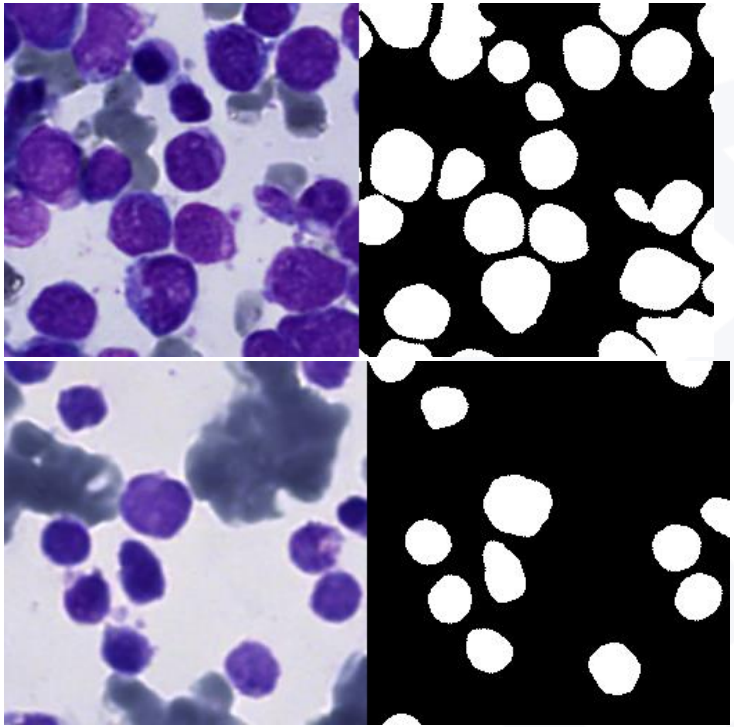


From: Averaging Weights Leads to Wider Optima and Better Generalization

Some Results

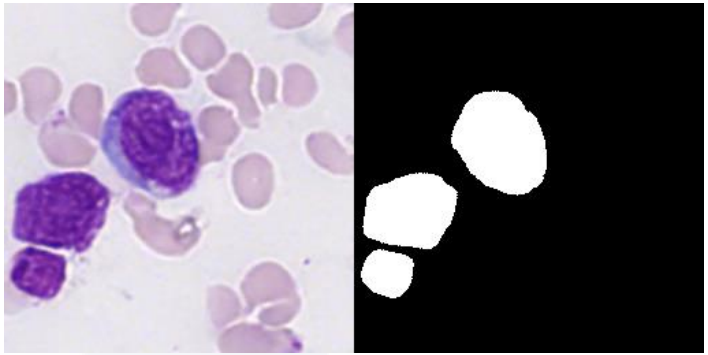
AML

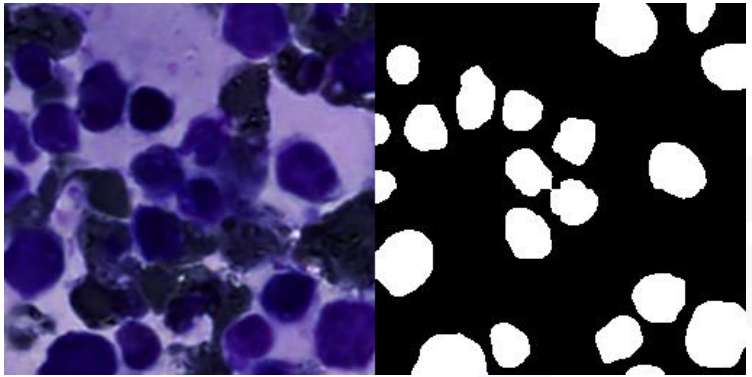


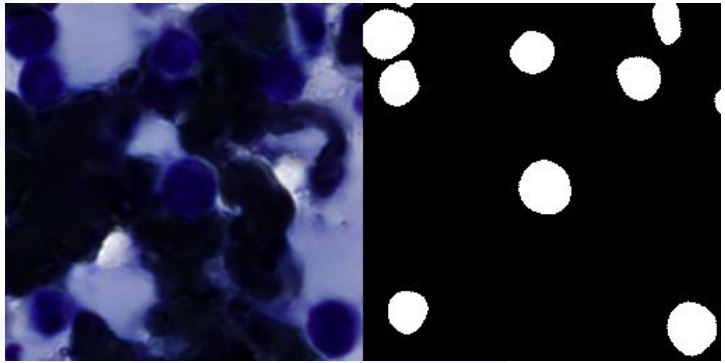


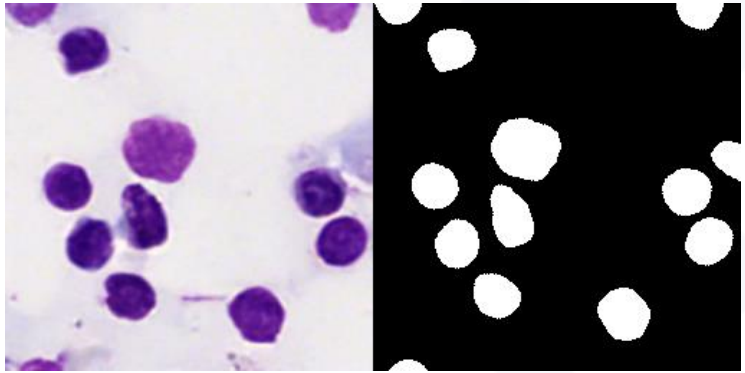
Some Results

Control



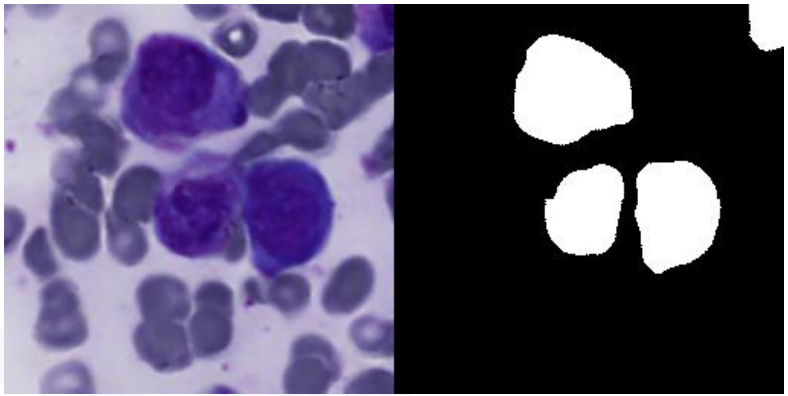


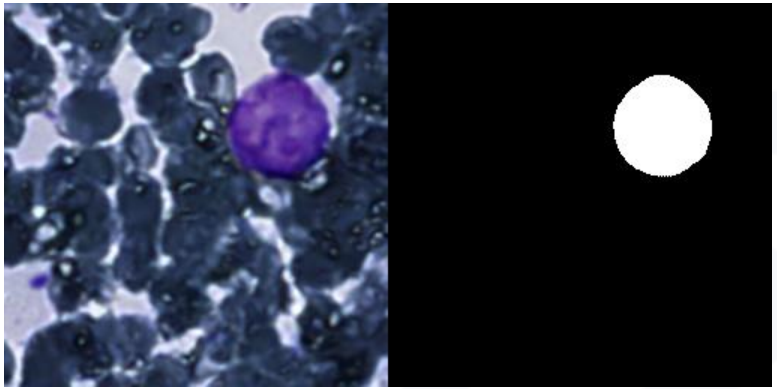


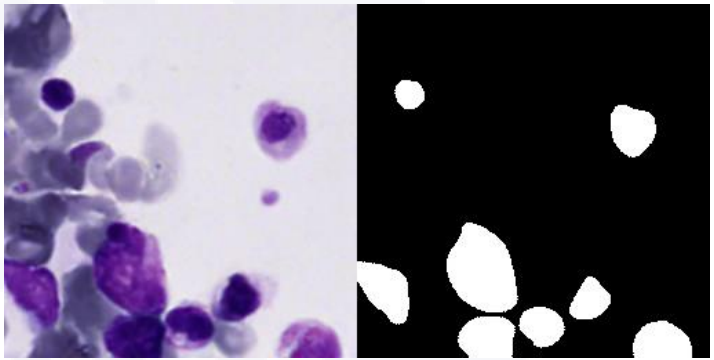


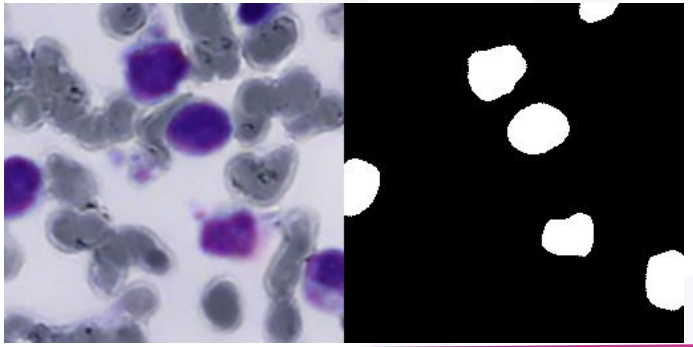
Some Results

MDS









Closing Thoughts

Current Challenges

- Improving Data Fusion Techniques
- Developing Clear Standards for Trustworthy AI
- Balancing Explainability and Performance
- Enhancing Bias Detection and Mitigation
- Engaging Clinicians in AI Development

Closing Thoughts

Where are we headed, and how can we get there?

- There will be an increase in the “busywork” needed
  - It’s going to be challenging
  - It’s going to be necessary
- “The difference between science and messing around is writing stuff down”
  - Data documentation
  - Accessible model explainability
  - Shared metrics for bias and errors
- We’ll need a lot of coordination
  - What do we want to achieve?
  - How we measure if we are there?
  - What can go wrong?
  - How to describe and collect data?
  - How to model it?

Thanks!

Any questions?

GenoMed4All & ERN-EuroBloodNet

Educational Program on AI in Hematology for an expert audience

Follow us!

genomed4all.eu

@genomed4all

/genomed4all

eurobloodnet.eu

@ERNEuroBloodNet

/ERNEuroBloodNet

Acknowledgements

European Reference Network
for rare or low prevalence complex diseases
Network Hematological Diseases (ERN EuroBloodNet)

Co-funded by the European Union

This project is supported by the European Reference Network on Rare Haematological Diseases (ERN-EuroBloodNet)-Project ID No 101085717. ERN-EuroBloodNet is partly co-funded by the European Union within the framework of the Fourth EU Health Programme.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

GENOMED4ALL

Funded by the European Union

GenoMed4All has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017549.