https://tinyurl.com/4vj26tv3

Appraising medical AI literature and model

Evidence-Based Medicine

August 8th, 2024

Sira Sriswasdi, PhD

- Research Affairs, Faculty of Medicine, Chulalongkorn University
- Computational Molecular Biology Group (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Appraising AI literature

Our assumptions for today

- There is no obvious red flag, such as
  - Predatory journal
  - Questionable authors
- The clinical aspect (research question and study design) makes sense
- **We will focus on the AI aspect**
  - Dataset
  - Model
  - Evaluation
  - Interpretation

Technical aspects of an AI literature

Data

- Sample size
- Inclusion/exclusion
- Input pre-processing
- Input definition
- Label definition

Model

- Model type
- Model complexity
- Explainability
- (Special techniques)

Evaluation

- External or internal
- Performance metrics
- Ablation
- Benchmarking
- Explainability

AI’s quality depends on data quality

How many samples are enough?

- Sample size should be “enough” to capture the diversity of the data

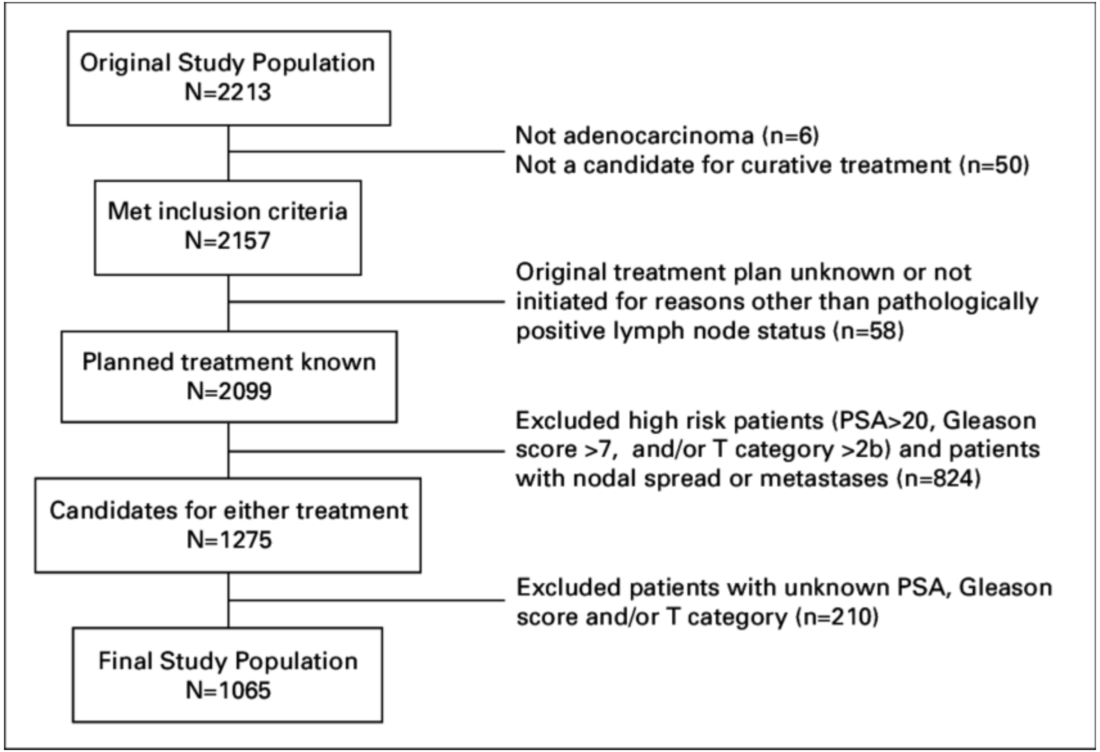
- When small sample size may be ok
  - 300 patients + tabular clinical data + logistic regression/random forest
  - Gene expression data from 20 tumors and 20 adjacent normal tissues

- When small sample size is likely not ok
  - Gut microbiome from 39 children
  - 100 CT images
  - 200 patients + artificial neural network model with 10k parameters

Data diversity and pre-processing

- Inclusion and exclusion criteria can introduce bias or simplify the problem

- Data preprocessing, such as imputation and outlier removal, also has the same effect

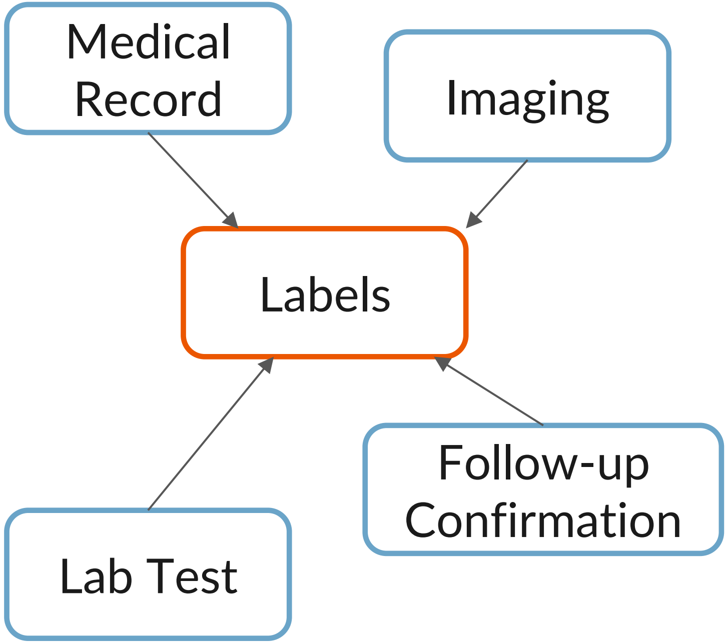


How were the input and output labels defined?

- How trustworthy are the labels?
  - Gold standard test
  - Provisional diagnosis
  - Future outcome

- Are factors that define the output labels part of the input data? Data leak!

- When labels were made by applying a simple rule on input clinical data…



Examples of good input-output definitions

Input	Output
Present clinical data	**Future follow-up** confirmation or treatment response
Present clinical data	Diagnosis made by **clinicians' expertise** (no simple rules)
Easy-to-collect clinical data	**Gold standard diagnosis** from imaging and other medical devices

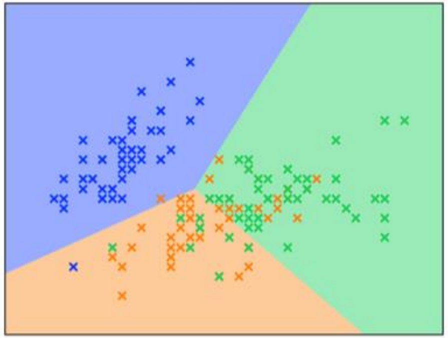
Questions to keep in mind regarding data

- **Q1**: Is the sample size large enough to reflect the diversity of patients?
  - Think about the number of clinical factors that can influence the disease
- **Q2**: Does the sample size match the choice of AI model and performance evaluation used?
- **Q3**: Are the inclusion/exclusion criteria and data cleaning justified?
- **Q4**: Is the definition of input and output labels appropriate?
  - Imagine yourself in place of the AI

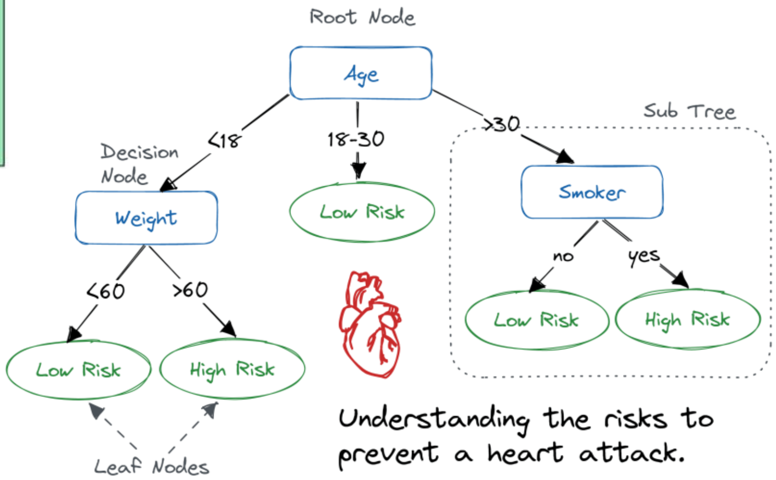
The right AI model for the right task

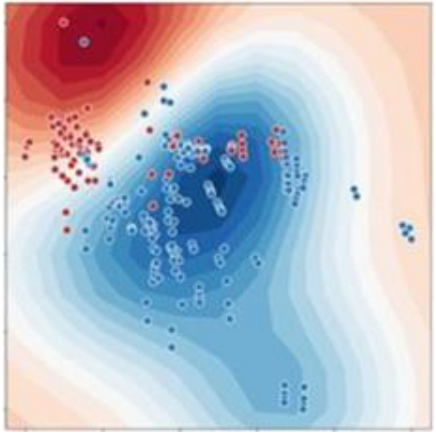
Broad types of classical AI models

Linear: output = (input$_1$ x w$_1$) + … + (input$_n$ x w$_n$)



Tree: Collection of binary decisions



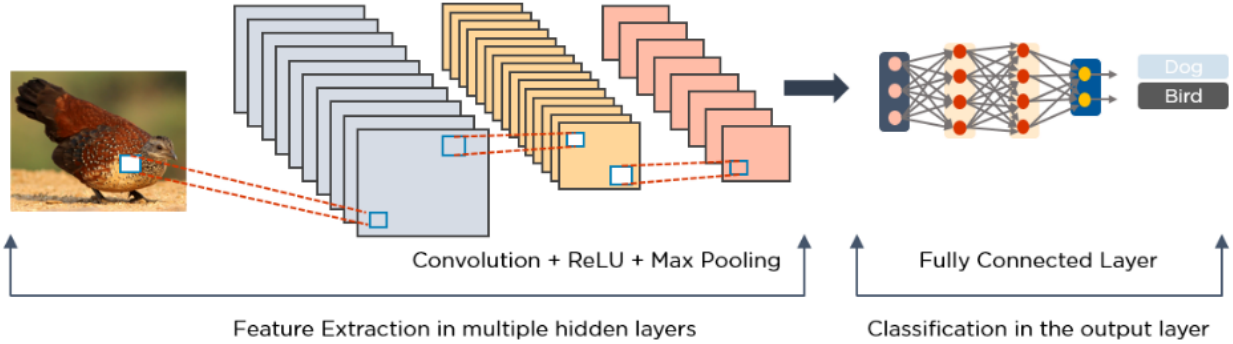


Neighbor-based: Predict new data points using labels of nearby data points

No model is strictly superior

- **Linear model** is good when risk scales with clinical measurements
  - Risk = clinical measurement x effect size
- **Tree model** is good for cutoff-based risk
  - Risk = (age > 65) and (blood cell count < 1,000)
- **Neighbor-based model** is good when you have a large cohort that can be used as reference
  - Diagnosis by referring to past similar cases

Artificial neural network (deep learing) models

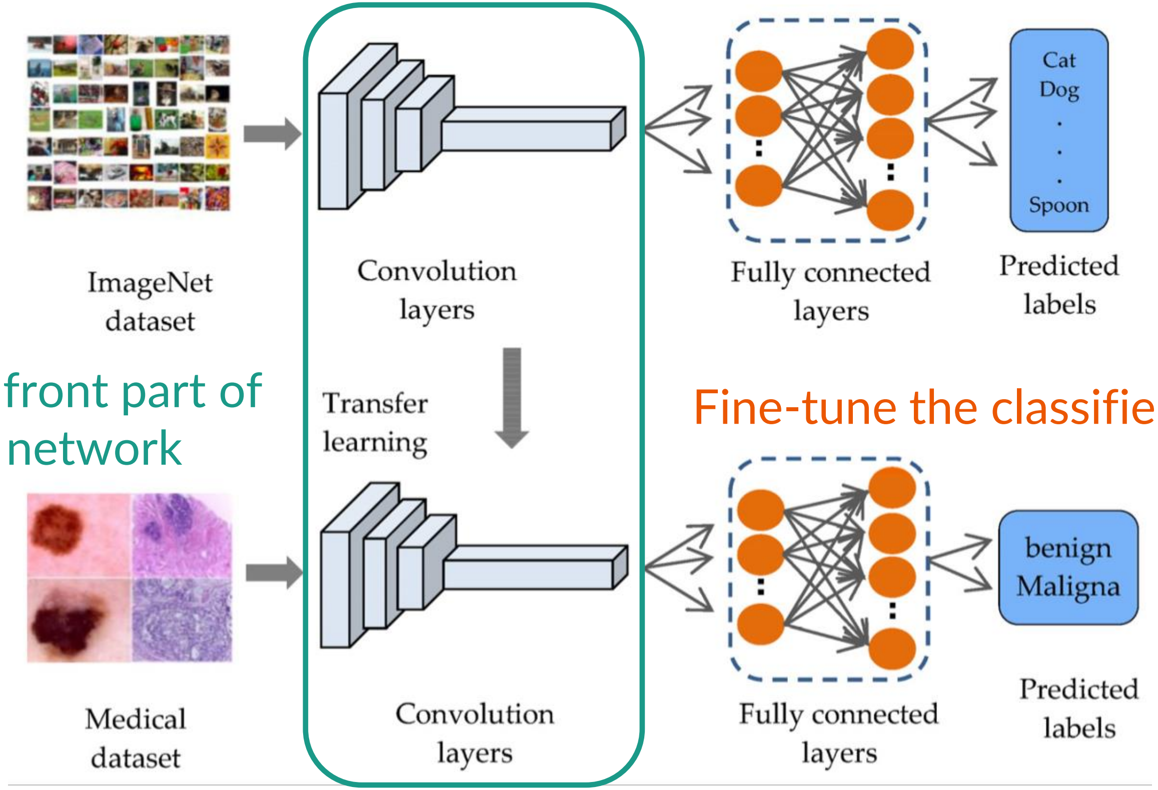


- **Feed-forward, multilayer perceptron:** Any data type
- **Convolutional:** Image data
- **Recurrent:** Time-series data
- **Transformer:** Any data type

A quick rule of thumb on data requirement

Model	Minimal Sample Size	Target Sample Size
Linear	100-200	200
Tree	300	500
Neighbor-based	200-300	300
Neural network	100 to 1,000,000	100 to 1,000,000

Transfer learning reduces data requirement



Reuse the front part of the neural network

Fine-tune the classifier on target task

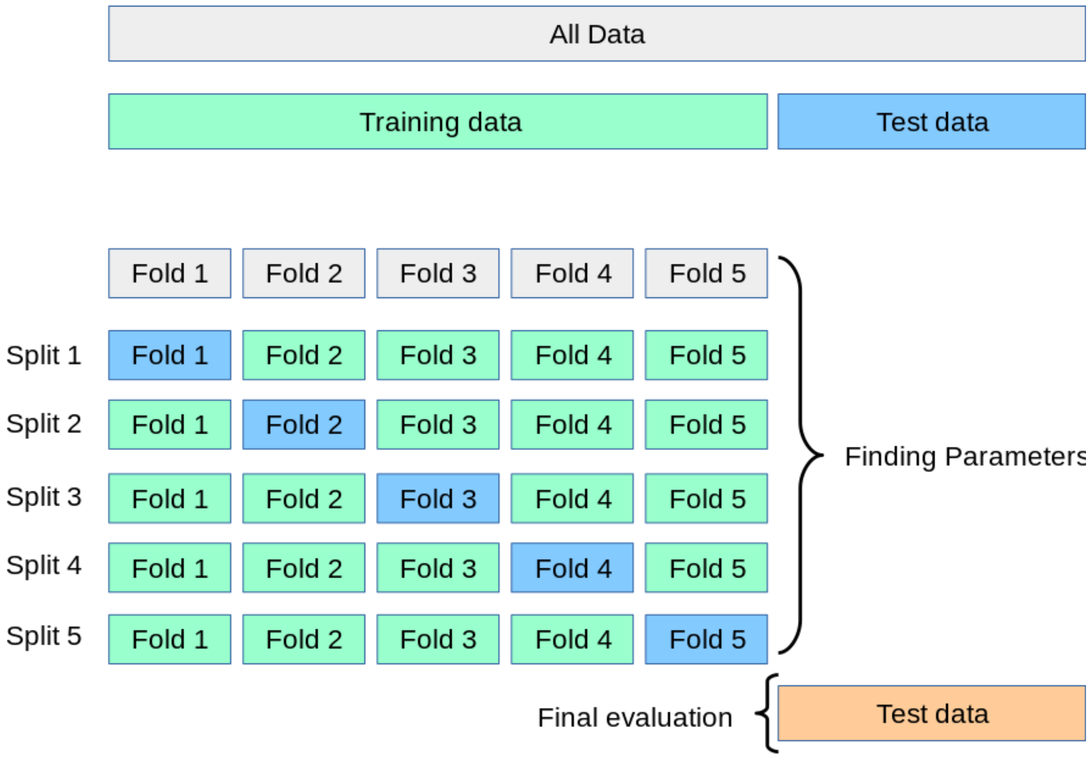
When in doubt, consult an expert

- There are many techniques that can significantly raise or lower the data requirement for machine learning model

- When encountering an unknown approach, check the journal and author’s past works to assess the quality

Good performance numbers do not imply good AI

Internal and external validation schemes



Performance metrics can be misleading

Good

- Accuracy = (25 + 340) / 400 = 91%
- Specificity = 340 / 350 = 97%

Bad

- Precision = 25 / (25 + 10) = 71.4%
- Sensitivity = 25 / 50 = 50%

- Why is accuracy very high while sensitivity and precision are low?

	Predict YES	Predict NO
Known YES	25	25
Known NO	10	340

Metrics must match the question

Would you want to use this model if:

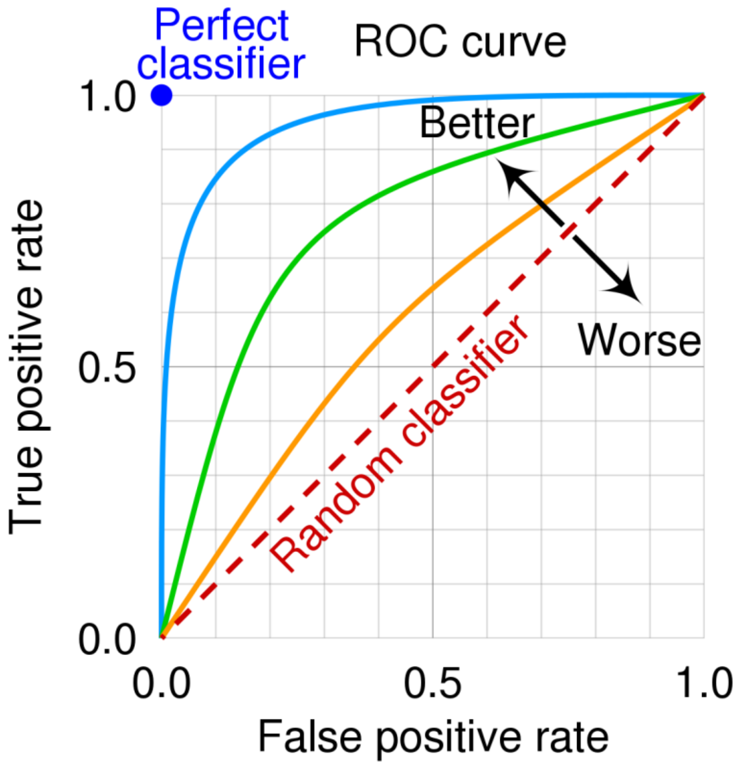
- YES = Patient should undergo a high-risk surgery
- YES = Patient will be allergic to a given drug
- YES = Patient should be called in for a follow-up

	Predict YES	Predict NO
Known YES	25	25
Known NO	10	340

Some metrics depend on confidence cutoff

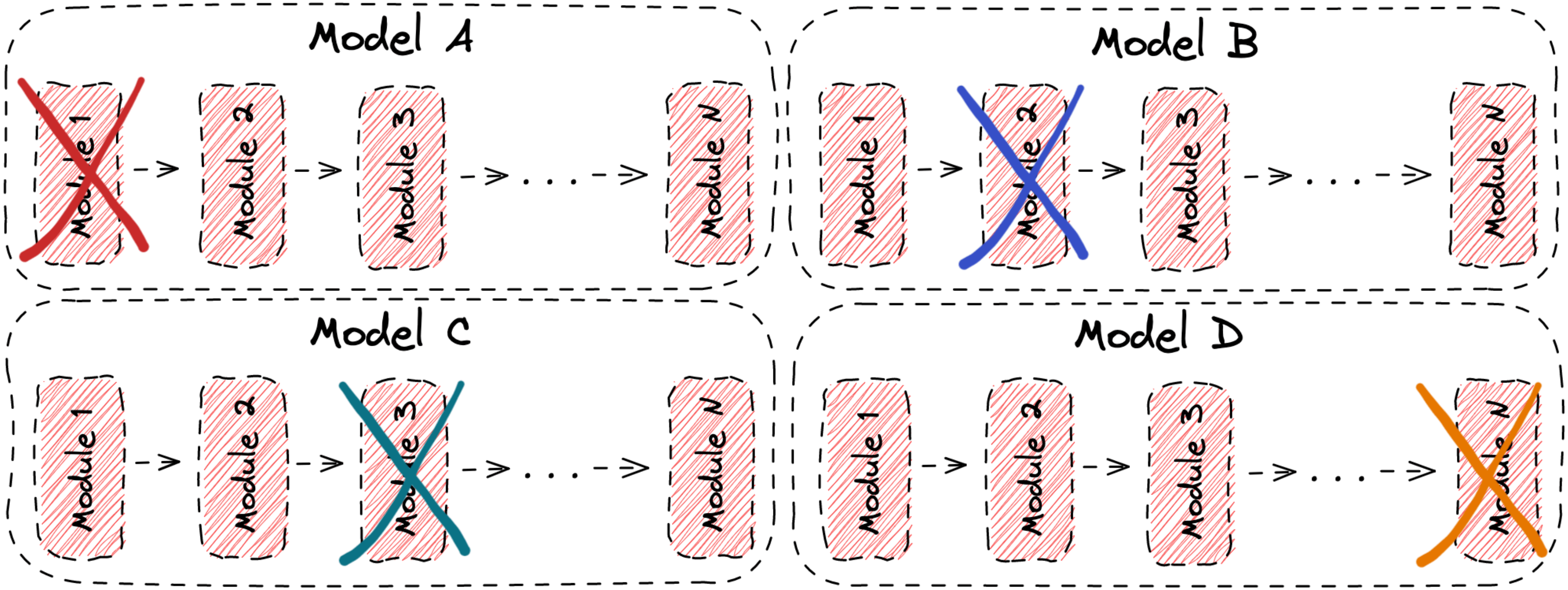
- Receiver Operating Characteristic curve (ROC) show how the model performs at various confidence cutoff
  - **Cutoff = 0** (predict all positive)
  - **Cutoff = 1** (predict all negative)

- **Sensitivity-Specificity** tradeoff
  - High sensitivity for screening task
  - High specificity for recommending high-risk procedure



The best way to understand a model is to compare to others

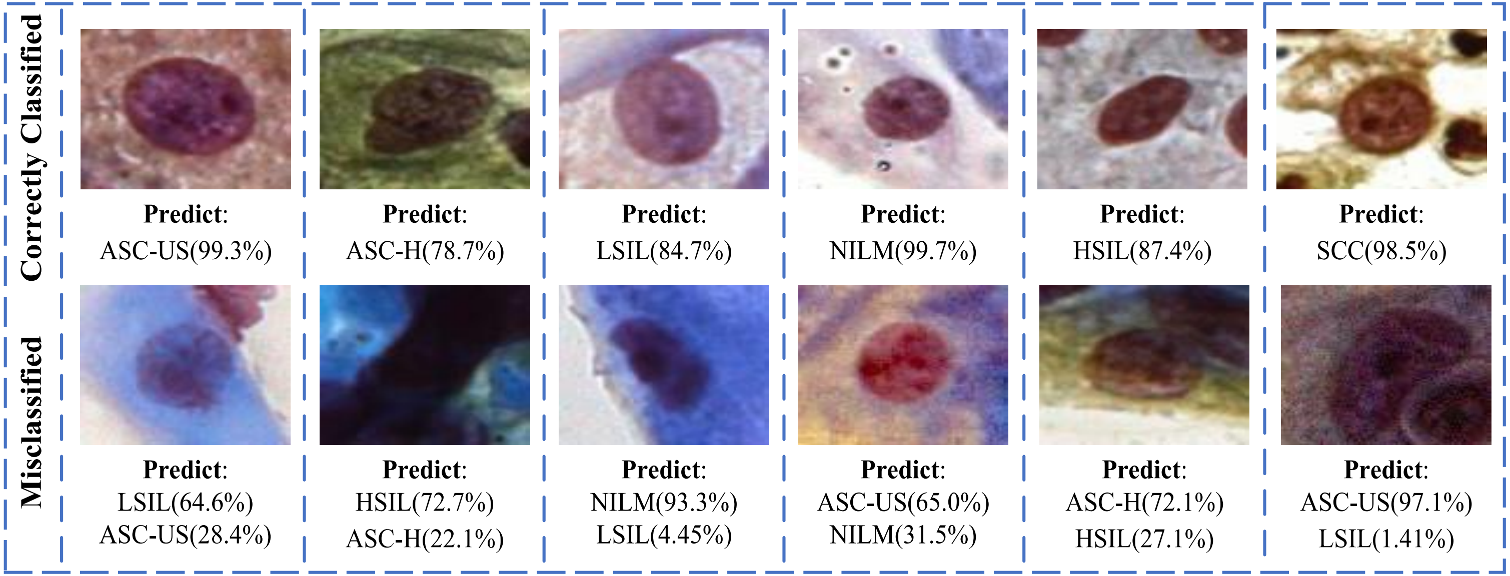
Ablation analysis



Contribution of each part on the performance

model	#points	relation	BN	DP	scale	voting	acc.
A	1k				1		87.2
B	1k	✓			1		89.9
C	1k	✓	✓		1		91.9
D	1k	✓	✓	✓	1		92.2
E	1k	✓	✓	✓	2		92.5
F	1k	✓	✓	✓	3		92.9
G	1k	✓	✓	✓	3	✓	**93.6**
H	2k	✓	✓	✓	3	✓	**93.6**
I	1k		✓	✓	3	✓	90.1

Error analysis



- Understanding the model through mistakes
- Compare to human errors and knowledge

Appraising AI in a workplace

Huge gap between AI development and deployment

Healthcare, Law, Regulation, and Policy, Machine Learning

“Flying in the Dark”: Hospital AI Tools Aren’t Well Documented

MODEL REPORTING GUIDELINES	EPIC MODEL BRIEFS											
	Deteriorati on Index	Early Detec tion of Sepsi s	Risk of Unpl anne d Read missi on	Risk of Patie nt No-Show	Pediatri c Risk of Hospital Admissi on or ED Visit	Risk of Hospit al Admiss ion or ED Visit	Inpatie nt Risk of Falls	Proje cted Block Utiliz ation	Remai ning Lengt h of Stay	Risk of Admiss ion of Heart Failure	Risk of Hospital Admissi on or ED Visit for Asthma	Risk of Hyper tensio n
TRIPOD	63%	63%	61%	48%	42%	61%	47%	36%	55%	48%	44%	51%
CONSORT-AI	63%	43%	63%	60%	33%	67%	53%	47%	47%	49%	42%	51%
SPIRIT-AI	61%	55%	54%	54%	38%	61%	44%	49%	51%	41%	39%	46%
Trust and Value	46%	33%	39%	50%	29%	42%	38%	46%	46%	25%	33%	46%
ML Test Score	27%	15%	33%	24%	9%	33%	15%	6%	18%	12%	9%	15%

Evaluation of sepsis diagnosis AI

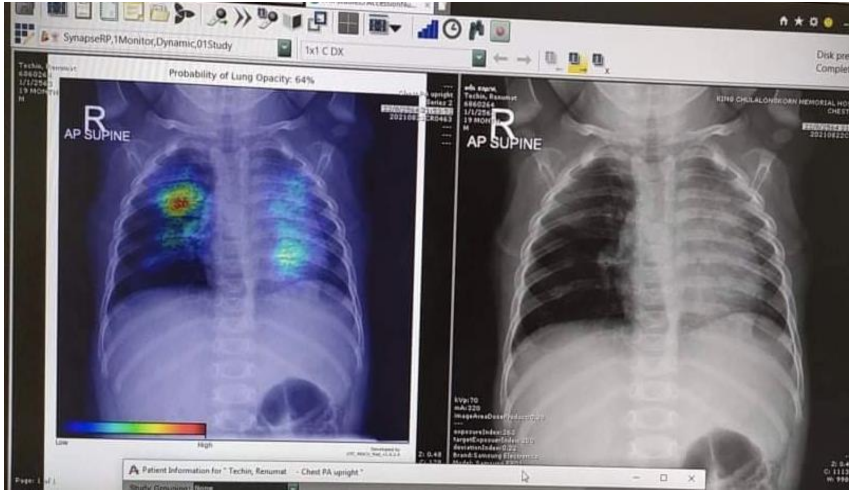
Results We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclusion criteria, of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level area under the receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM identified 183 of 2552 patients with sepsis (7%) who did not receive timely administration of antibiotics, highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating a large burden of alert fatigue.

- AUC of 0.63 in practice
- Missed 67% of sepsis

Points to look out for when evaluating an AI

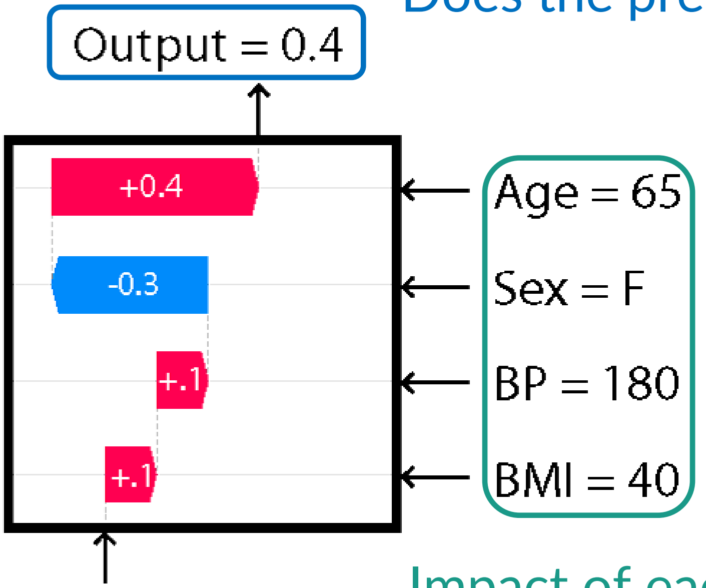
- Was the model developed using data that match your task?
  - Patient population
  - Data collection protocol and devices
  - Definition of output

- When you try using it:
  - Prediction confidence
  - Feature importance
  - Saliency map
  - Error analysis



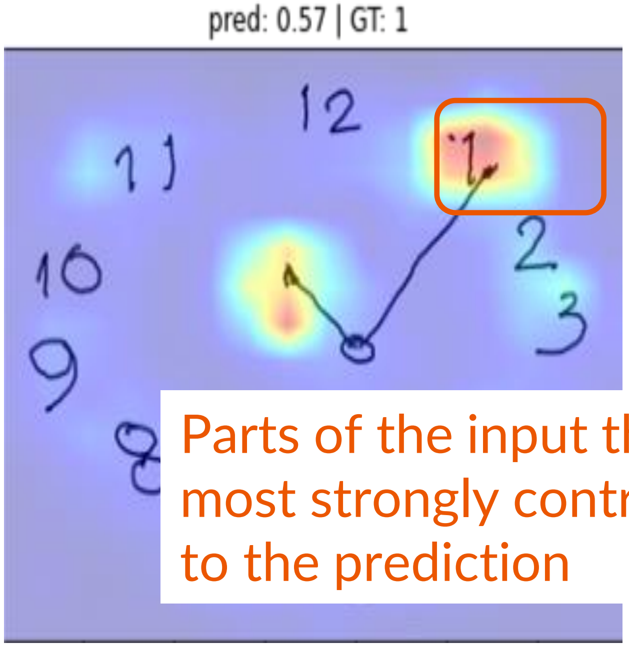
Explainability

Does the predicted confidence match your expectation?



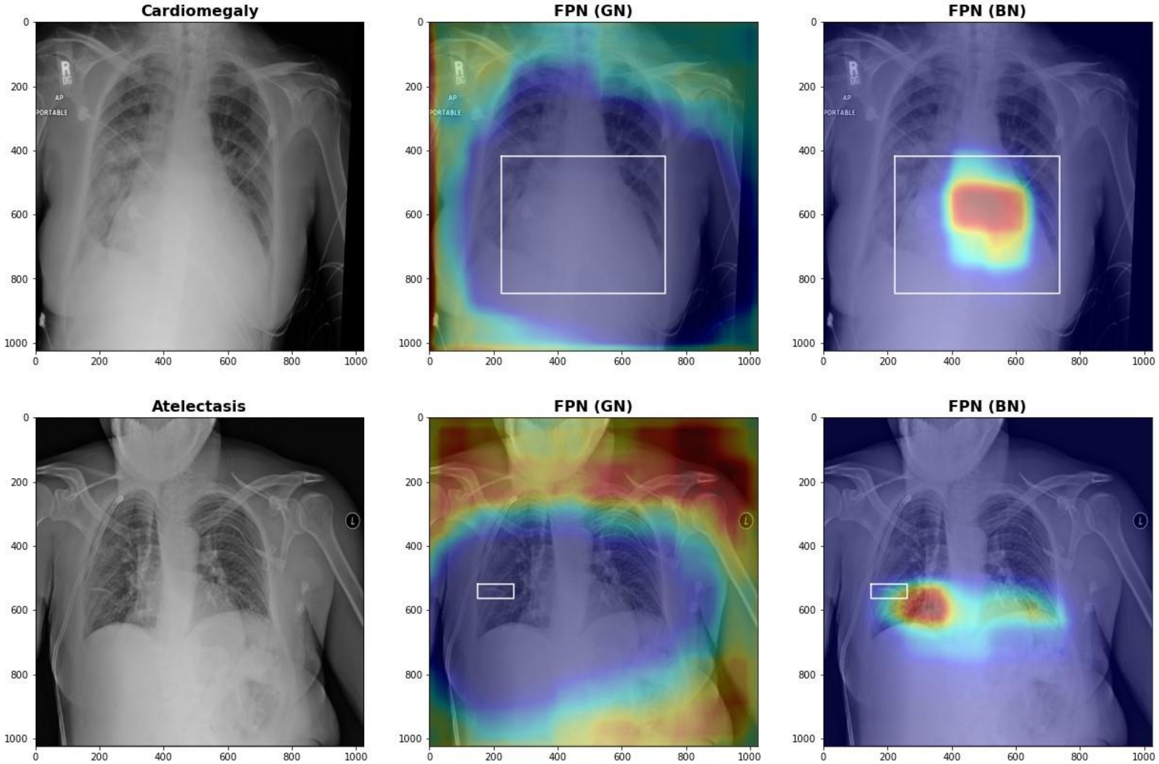
Base rate = 0.1

Impact of each input on the prediction

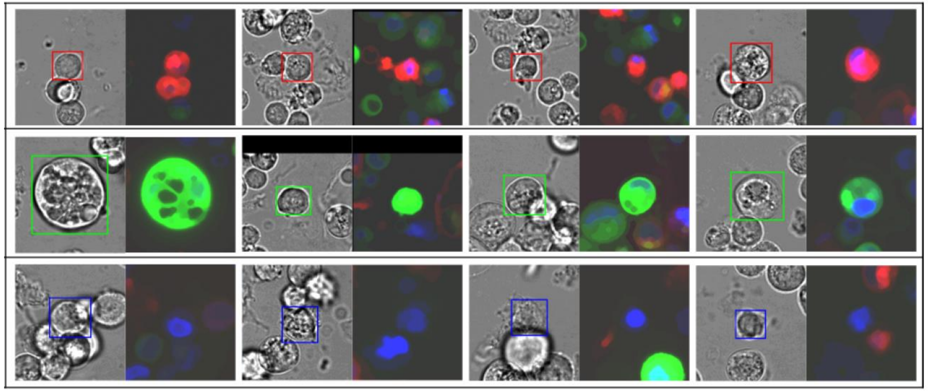


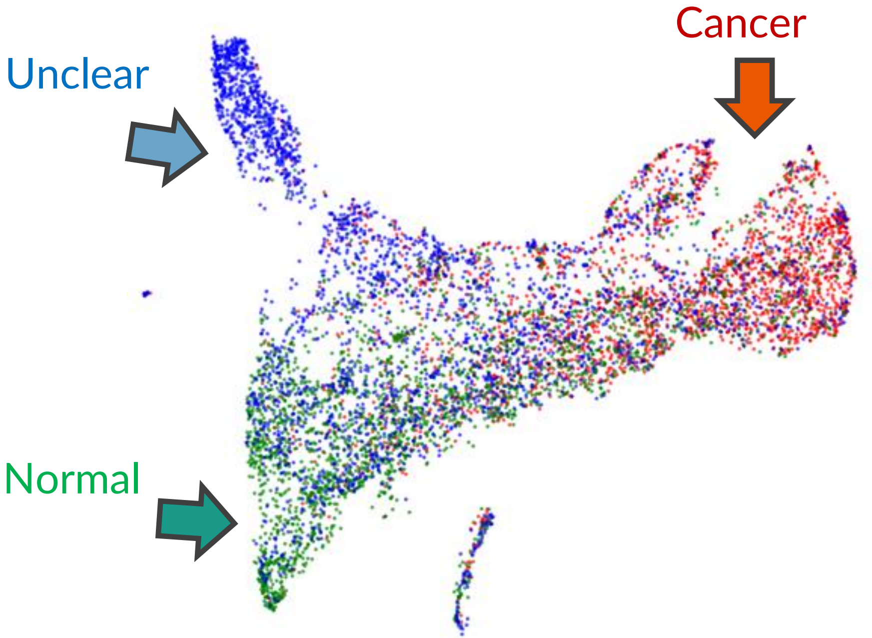
Parts of the input that most strongly contribute to the prediction

Correct prediction doesn't imply correct reasoning



Visualization with dimensionality reduction





- Identify whether the model stumbles on hard cases and whether the errors are systematic or random

Things to try during evaluation

- Test the model on **easy / medium / difficult samples**

- **Intentionally alter input values** to observe how the prediction changes

- Test the model on **edge cases**

Summary

- Appraising AI is like appraising all other things
- Don’t give AI too much benefits of the doubt
- Consult experts if needed
- Always look for explanation