XAI in industry

CSEP 590B: Explainable AI

Ian Covert & Su-In Lee
University of Washington

Course announcements

- Today is our last class!
- Homework 3 due on Wednesday

Last lecture

- Recap what we covered in the course
- Discuss examples of how XAI is used in industry
  - We tried reaching out to companies for case studies about their usage
  - For legal reasons, unable to share details 😢
- Instead, we'll discuss:
  - Relevant regulations
  - Landscape of XAI companies/services

Today

- Section 1
  - Course recap 
  - XAI-related regulations
  - XAI services and companies
- Section 2
  - Medical AI examples

- **Finish early:** course evaluation
  - https://uw.iasystem.org/survey/258596

Main topics

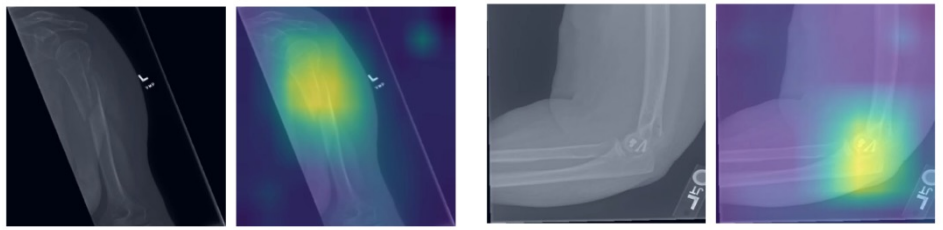
- Feature importance explanations
  - Removal-based methods
  - Gradient-based methods
  - Metrics
- Concept explanations
- Neuron interpretation
- Counterfactual explanations
- Instance explanations
- Human-in-the-loop

Model understanding

- XAI often used to understand existing models
- For model developers:
  - Sanity check dependencies, identify shortcuts
  - Understand underlying data relationships
- For users:
  - Understand decision process
  - Identify changes to alter outcome

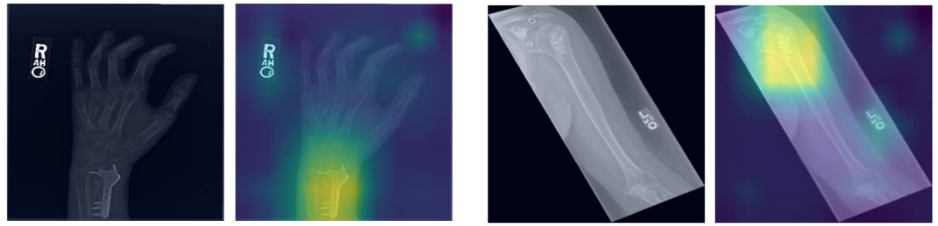
Sanity check dependencies

Model focuses on relevant regions



(a) Frontal radiograph of the left humerus demonstrates a displaced transverse spiral fracture. This area is highlighted by the model CAM.

(b) Lateral radiograph of the left elbow demonstrates transcortical screw fixation of a comminuted fracture in the distal humerus. The model identifies the abnormality as demonstrated by the CAM.



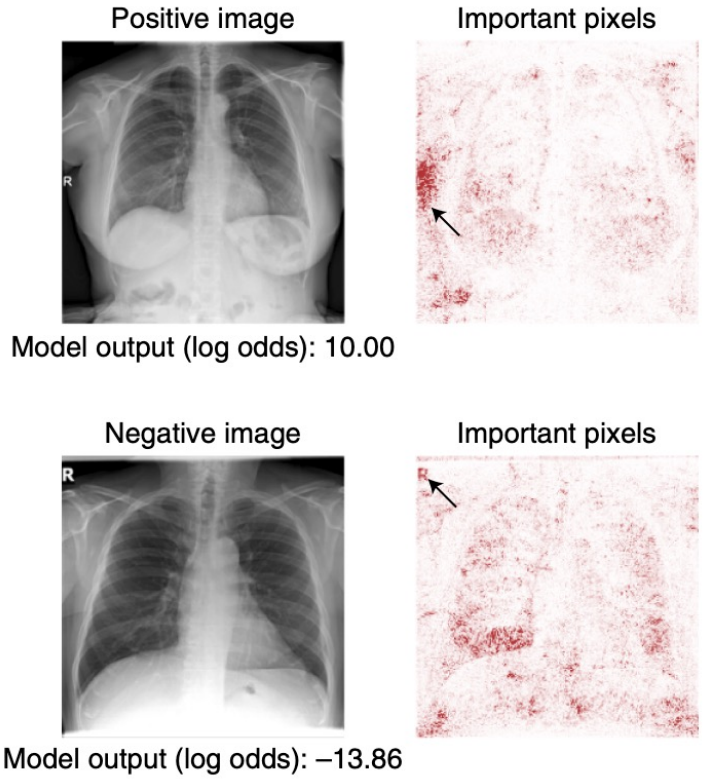
(c) Frontal oblique radiograph of the right hand demonstrates prior screw and plate fixation of a distal radius fracture. This abnormality is localized by the model CAM.

(d) Frontal radiograph of the left humerus demonstrates a sclerotic lesion in the humeral head. The model CAM highlights the abnormal region.

Figure 4: Our model localizes abnormalities it identifies using Class Activation Maps (CAMs), which highlight the areas of the radiograph that are most important for making the prediction of abnormality. The captions for each image are provided by one of the board-certified radiologists.

Rajpurkar et al., "MURA: Large dataset for abnormality detection in musculoskeletal radiographs" (2018)

Identify shortcuts



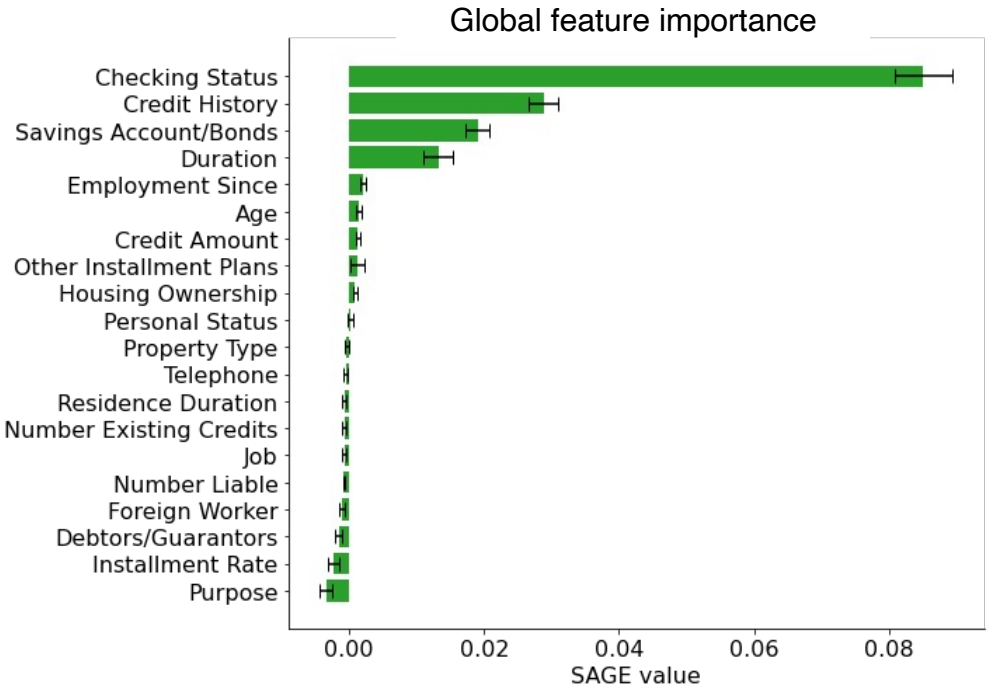
Unexpected dependence on laterality markers

Why? Markers leak label information

DeGrave et al, "AI for radiographic COVID-19 detection selects shortcuts over signal" (2021)

Understand data

What drives credit quality predictions?



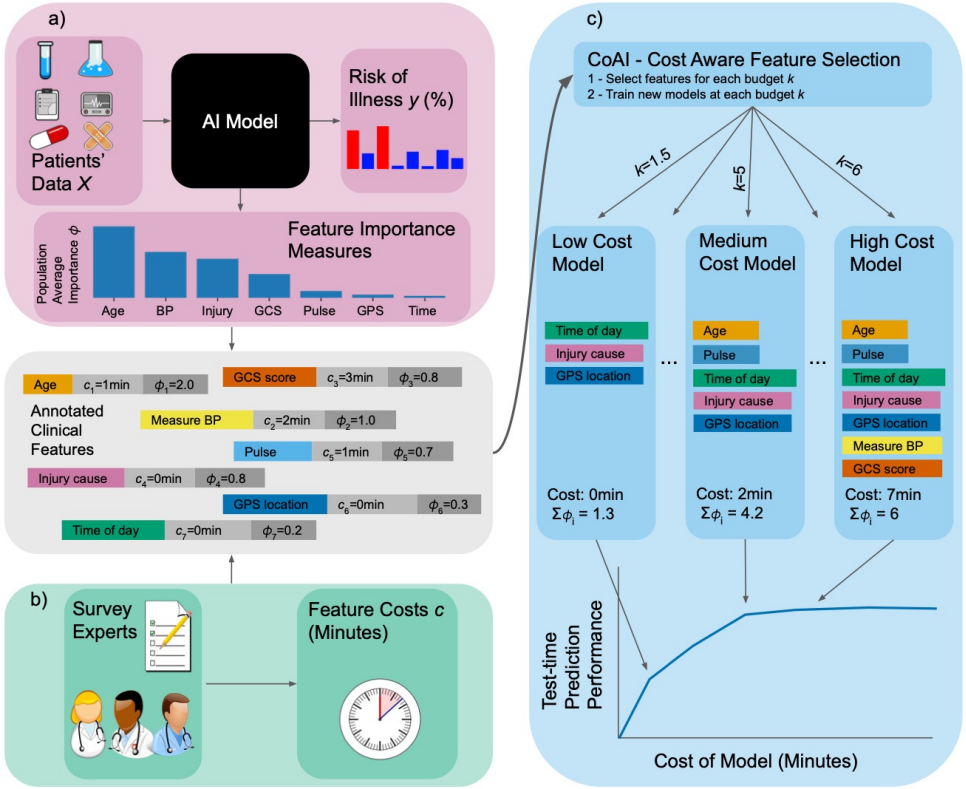
Covert et al., "Understanding global feature contributions with additive importance measures" (2020)

Model improvement

- Feature selection
  - Reduce data collection cost while preserving accuracy (SAGE)
- Dataset refinement
  - Find training data to remove (Data Shapley, TracIn)
- Other use cases:
  - Explanation regularization
    - Erion et al., "Improving performance of deep learning models with axiomatic attribution priors and expected gradients" (2021)
  - Model distillation
    - Pruthi et al., "Evaluating explanations: How much do explanations from the teacher aid students?" (2021)

Cost-aware feature selection

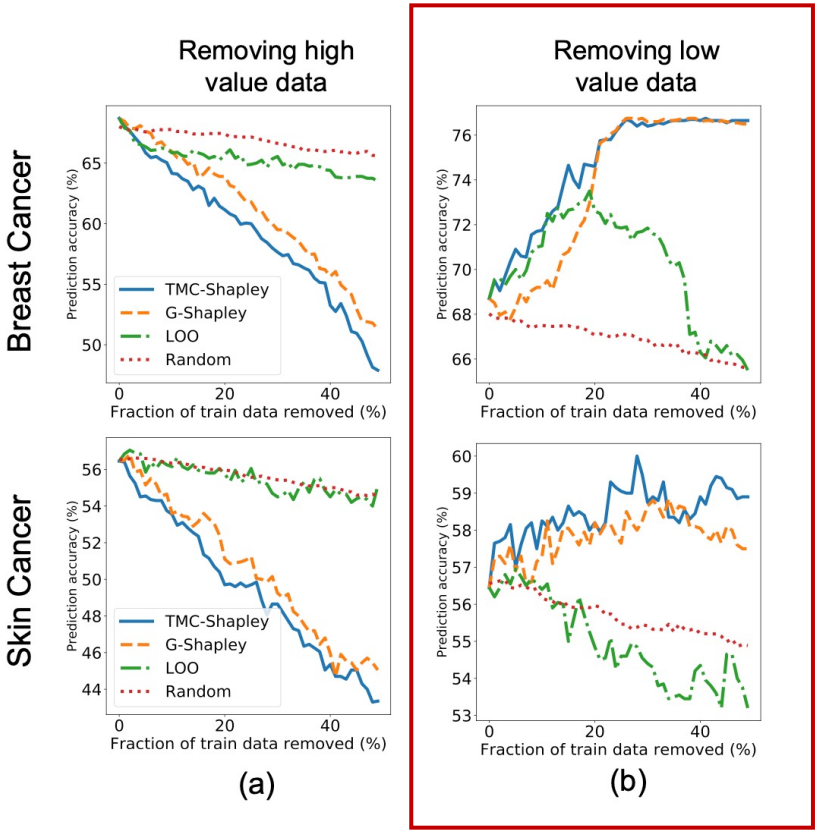
Associate cost (time to measure) and SAGE value with each feature



Find feature combinations by solving knapsack problem, use cost-aware models

Erion et al., "A cost-aware framework for the development of AI models for healthcare applications" (2021)

Dataset refinement



Removing low-value data can improve performance

Ghorbani et al., "Data Shapley: Equitable valuation of data for machine learning" (2019)

Today

- Section 1
  - Course recap
  - XAI-related regulations 
  - XAI services and companies
- Section 2
  - Medical AI examples

Right to explanation

- In the regulatory context, the right to receive explanation for an algorithm's output
- Implemented to some extent by several recent laws
  - Equal Credit Opportunity Act (ECOA, USA)
  - General Data Protection Regulation (GDPR, EU)
  - Digital Republic Act (France)

GDPR

- Original text suggests a right to explanation:
  - "…ensure fair and transparent processing"
  - "…meaningful information about the logic involved"
  - "…obtain an explanation of the decision reached"

- Subsequent debate over precise interpretation
  - Final decision belongs to regulators (data protection authorities) who enforce the law

Casey et al., "Rethinking explainable machines: The GDPR's 'right to explanation' debate and the rise of algorithmic audits in enterprise" (2019)

Adverse action notices

- ECOA mandates notification and explanation for adverse actions
  - E.g., denial of loan request, account termination
  - "The bank must also either provide the applicant with the specific principal reason for the action taken or disclose [...] the right to request the reason(s) for denial within sixty days" (ECOA Regulation B)
- Previously viewed as an obstacle for adopting advanced AI/ML solutions

BLDS, Discover Financial Services & H2O.AI, "Considerations for fairly and transparently expanding access to credit" (2020)

Adverse action notices (cont.)

- CFPB published a reminder about the use of AI/ML last week (May 26, 2022)
  - "ECOA and Regulation B do not permit creditors to use complex algorithms when doing so means they cannot provide the specific and accurate reasons for adverse actions."
  - "Whether a creditor is using a sophisticated machine learning algorithm or more conventional methods to evaluate an application, the legal requirement is the same: Creditors must be able to provide applicants against whom adverse action is taken with an accurate statement of reasons."

XAI for adverse action notices?

- Recent whitepaper from FinRegLab (nonprofit innovation center) investigates the use of model diagnostic tools (XAI)
  - Focuses on two consumer protection goals: adverse action notices and disparate impact
  - Among other things, finds encouraging results when using SHAP for adverse action notices

Blattner et al., "Machine learning explainability & fairness: insights from consumer lending" (2022)

More regulatory interest

- Federal Reserve (USA)
  - Along with other regulatory agencies (FDIC, CFPB, etc.), requested information on usage of AI/ML and XAI at banks for any operational purposes

- Veritas Initiative (Singapore)
  - Consortium to determine proper evaluation for AI systems according to principles of fairness, ethics, accountability and transparency (FEAT)

Surkov et al., "Unleashing the power of machine learning models in banking through explainable artificial intelligence" (2022)

Problems

- Can regulators enforce how explanations are provided, or which algorithms are used?

- Can companies provide explanations that are vague, or that mask negative aspects of their model (e.g., dependence on racial proxies)?

Today

- Section 1
  - Course recap
  - XAI-related regulations
  - XAI services and companies 
- Section 2
  - Medical AI examples

XAI companies

XAI companies

- For these companies, XAI falls under the umbrella of **MLOps** or **ML observability**
- They often provide other services, including:
  - Accuracy monitoring
  - Data drift detection
  - Bias mitigation

Platforms supporting XAI

Google Cloud Platform

Azure Machine Learning

Amazon SageMaker

XAI services

- Customers prefer widely known and accepted techniques
- Mostly implement famous methods
  - Local feature importance: SHAP, LIME
  - Global feature importance: permutation tests
  - Other simple approaches: partial dependence plots (PDPs), etc.
- Several companies also contribute new research

Related initiatives

- Responsible AI, Trustworthy AI
  - Encompass a variety of positive, loosely related objectives
  - Advocate for:
    - Fairness
    - Ethics
    - Accountability
    - Explainability/transparency
    - Robustness
    - Sustainability

Today

- Section 1
  - Course recap
  - XAI-related regulations
  - XAI services and companies
  - **10 min break**
- Section 2
  - Medical AI examples