Welcome to today’s ACM Learning Webinar, “Explainable Models for Healthcare AI” The presentation starts at the top of the hour and lasts 60 minutes. Audio and video will automatically play throughout the event. On the bottom panel you’ll find a number of widgets, including Twitter, Sharing, and Wikipedia apps.

If you are experiencing any problems/issues, refresh your console by pressing the F5 key on your keyboard in Windows, Command + R if on a Mac, or refresh your browser if you’re on a mobile device; or close and re-launch the presentation.

To control volume, adjust the master volume on your computer. If the volume is still too low, use headphones.

At the end of the presentation, you’ll see a survey open on your screen. Please take a minute to fill it out to help us improve your next webinar experience. You may also open the survey at any time throughout the presentation from resources window.

This session is being recorded and will be archived for on-demand viewing in the next 1-2 days. You will receive an automatic email notification when it is available, and check http://learning.acm.org/ in a few days for updates. And check out http://webinar.acm.org/ for archived recordings of past webcasts.
Explainable Models for Healthcare AI

Ankur Teredesai  
Muhammad Aurangzeb Ahmad  
Carly Eckert M.D.  
Vikas Kumar  
KenSci Inc. + University of Washington Tacoma

https://www.kensci.com/explainable-machine-learning/
ACM Highlights

• Learning Center tools for professional development: http://learning.acm.org
  • The Safari Learning Platform featuring the entire Safari collection of nearly 50,000 technical books, video courses, O’Reilly conference videos, learning paths, tutorials, case studies
  • 1,800+ Skillsoft courses, 4,800+ online books, and 30,000+ task-based short videos for software professionals covering programming, data management, DevOps, cybersecurity, networking, project management, and more; including training toward top vendor certifications such as AWS, CEH, Cisco, CISSP, CompTIA, Oracle, RedHat, PMI.
  • 1,200+ books from Elsevier on the ScienceDirect platform (including Morgan Kaufmann and Syngress titles)
  • Learning Webinars from thought leaders and top practitioners
  • Podcast interviews with innovators, entrepreneurs, and award winners

• Popular publications:
  • Flagship Communications of the ACM (CACM) magazine: http://cacm.acm.org
  • ACM Queue magazine for practitioners: http://queue.acm.org

• The ACM Code of Ethics, a set of principles and guidelines designed to help computing professionals make ethically responsible decisions in professional practice: https://ethics.acm.org

• ACM Digital Library, the world’s most comprehensive database of computing literature: http://dl.acm.org

• International conferences that draw leading experts on a broad spectrum of computing topics: http://www.acm.org/conferences

• Prestigious awards, including the ACM A.M. Turing Award and ACM Prize in Computing: http://awards.acm.org

• And much more... http://www.acm.org
“Housekeeping”

Twitter: #ACMLearning

• Welcome to today’s ACM Learning Webinar, “Explainable Models for Healthcare AI” The presentation starts at the top of the hour and lasts 60 minutes. Audio and video will automatically play throughout the event. On the bottom panel you’ll find a number of widgets, including Twitter, Sharing, and Wikipedia apps.

• If you are experiencing any problems/issues, refresh your console by pressing the F5 key on your keyboard in Windows, Command + R if on a Mac, or refresh your browser if you’re on a mobile device; or close and re-launch the presentation. You can also view the Webcast Help Guide, by clicking on the “Help” widget in the bottom dock.

• To control volume, adjust the master volume on your computer. If the volume is still too low, use headphones.

• At the end of the presentation, you’ll see a survey open on your screen. Please take a minute to fill it out to help us improve your next webinar experience. You may also open the survey at any time throughout the presentation from resources window.

• This session is being recorded and will be archived for on-demand viewing in the next 1-2 days. You will receive an automatic email notification when it is available, and check [http://learning.acm.org/](http://learning.acm.org/) in a few days for updates. And check out [http://webinar.acm.org](http://webinar.acm.org) for archived recordings of past webcasts.
Talk Back

- Use Twitter widget to Tweet your favorite quotes from today’s presentation with hashtag #ACMLearning

- Submit questions and comments via Twitter to @acmeducation – we’re reading them!

- Use the sharing widget in the bottom panel to share this presentation with friends and colleagues.

- The ACM Discourse Page is available for post-webinar discussion – https://on.acm.org
Explainable Models for Healthcare AI

Ankur Teredesai
Muhammad Aurangzeb Ahmad
Carly Eckert M.D.
Vikas Kumar

KenSci Inc. + University of Washington Tacoma

https://www.kensci.com/explainable-machine-learning/
Conceived in academia. Supported by the federal Health agencies. Trusted by leading health systems globally.

Built by Clinicians, Data Scientists, and Engineers
Explainable Models in AI
Well, the computer assures me that you are fine and it has given me 100 reasons for it.

Doctor, are you sure that I am ok now? My head still hurts.
Learning Objectives of the Tutorial

• Why?
  • The AI & ML needs to provide explanations in Healthcare.

• What?
  • Type of Explanations do we need in Healthcare

• How?
  • To select the right machine learning algorithms when explanations are needed

• Where?
  • Settings and different types of interpretable machine learning models

• When?
  • The past the present and the future of explainable AI in healthcare
Explainable By Any other Name.. is still explainable

• **Explain:** Make (an idea or situation) clear to someone by describing it in more detail or revealing relevant facts.

• **Interpret:** Explain the meaning of (information or actions)

• **Understand:** Perceive the intended meaning of (words, a language, or a speaker)

• **Comprehend:** Grasp mentally; understand

• **Intelligible:** Able to be understood; comprehensible

[Dictionary, Oxford English 2018]
A (Brief) History of Need for Explanations in Statistics

- Foundations of Logic: Clear and explicit reasoning, explainable to almost anyone
- Bayes: Grounding Probability and inference on solid foundations (18th century) [Bayes 1763]
- Goal of (early) AI: Mimic human reasoning mechanically
- Early Expert Systems in healthcare like MYCIN were explainable systems
- Push around explainability when ensemble methods first came about in the 1980s
- Current interest in explainability is because of widespread adoption and deployment of machine learning systems
Explainable AI / Interpretable Machine Learning

• Explanations of AI/machine learning models to humans with domain knowledge [Craik 1967, Doshi-Velez 2014]

• Why is the prediction being made?

• Comprehensible to humans in (i) natural language (ii) easy to understand representations

Standard Model Lagrangian

\[ L_{\text{SM}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - \frac{i}{2} \bar{q} \gamma^\mu (\gamma^\nu D_\nu - m \delta^{\mu\nu}) q - \frac{1}{2} \phi \partial^\mu \phi \partial_\mu \phi - \frac{1}{4} W_{\mu\nu}^a W^{\mu\nu}_a - \frac{1}{2} Z_{\mu\nu}^a Z^{\mu\nu}_a - \frac{1}{2} M^2 \phi^2 - \frac{1}{2} W^a_{\mu\nu} \partial_\nu W^a_{\mu\nu} - W^a_{\mu\nu} W^a_{\mu\nu} - Z_{\mu\nu}^a Z_{\mu\nu}^a - 2 \lambda (\phi^2 - v^2)^2 - \frac{1}{2} M^2 \phi^2 - \frac{1}{2} W^a_{\mu\nu} \partial_\nu W^a_{\mu\nu} \]
Each element constituent of the solution process needs to be explainable for the solution to be truly explainable [Lipton 2016].
What Explanations Are Not?

• **Explanation vs. Justification**
  • Explaining is giving reasons for the prediction [Biran 2014, Biran 2017A, Biran 2017B]
  • Justification is putting the explanation in a context
  • Justification does not have to correspond to how the model actually works

• **Explanation vs. Causality**
  • Explanations are mostly:
    • *Not* Causal
    • *Not* Prescriptive
  • Example: End of life prediction for heart failure patients

© KenSci., Inc. 2018.

Healthcare AI
How Are Decisions Made in Healthcare?

1. Heuristics
Factors: \(7 \pm 2\)

2. Rules based system
Factors: 10s

3. ML based System
Factors: 100s

~80% Care Decisions
~18% Care Decisions
~2% Care Decisions
How **Machine Learning** Can Improve Healthcare Across the **Continuum of Care**?

- **Optimize hospital efficiency**
  - Prevent sequelae of disease
  - Improve operations and diagnose illness sooner
  - Prevent post operative complications
  - Chronic disease management (DM)
  - Chronic disease management (CHF)
  - Help patient plan for end of life needs

- **Variation analysis**
  - No show prediction
  - Rx noncompliance prediction
  - LOS prediction
  - ROR prediction
  - Care management RecSys
  - Prevent post operative complications
  - Predict disease complications

- **Optimize clinic operations and benefit to patient**
  - Prevent sequelae of disease
  - Improve operations and diagnose illness sooner
  - Prevent post operative complications
  - Chronic disease management (DM)
  - Chronic disease management (CHF)
  - Help patient plan for end of life needs

- **Optimize inpatient flow**
  - Prevent sequelae of disease
  - Improve operations and diagnose illness sooner
  - Prevent post operative complications
  - Chronic disease management (DM)
  - Chronic disease management (CHF)
  - Help patient plan for end of life needs

- **Preventive care**
  - Prevent sequelae of disease
  - Improve operations and diagnose illness sooner
  - Prevent post operative complications
  - Chronic disease management (DM)
  - Chronic disease management (CHF)
  - Help patient plan for end of life needs

- **Help patient plan for end of life needs**
  - Prevent sequelae of disease
  - Improve operations and diagnose illness sooner
  - Prevent post operative complications
  - Chronic disease management (DM)
  - Chronic disease management (CHF)
  - Help patient plan for end of life needs

- **Variation analysis**
  - No show prediction
  - Rx noncompliance prediction
  - LOS prediction
  - ROR prediction
  - Care management RecSys
  - Prevent post operative complications
  - Predict disease complications

- **ED flow suite**
  - SSI prediction
  - Other predictions
  - EOL prediction

© KenSci., Inc. 2018.
Operationalizing AI in Healthcare

Only a small fraction of real-world machine learning systems actually constitutes machine learning code [Sculley 2015].
Data Complexity in **Operationalizing AI in Healthcare**

- **Syntactic Correctness**
  - Is the data in the correct format e.g., if the AI data pipeline requires sex for males as ‘m’ and the input data encodes it as ‘1’

- **Morphological Correctness**
  - Is the data within the range of possible values e.g., a blood pressure of 500 does not make sense

- **Semantic Correctness**
  - Do the variables actually correspond to what the semantics that are being ascribed to them e.g., a variable which encodes blood pressure as high vs. low will have different semantics for children as compared to adults
Why Do We **Need** Explanations in Healthcare AI Now?

More Data

More AI

More Implications (known/unknown)

More Data

© KenSci., Inc. 2018.
Do We **ALWAYS** Need Explanations in Healthcare AI?

**When fairness is critical:**
- Any context where humans are required to provide explanations so that people cannot hide behind machine learning models [Al-Shedivat 2017B, Doshi-Velez 2014]

**When consequences are far-reaching:**
- Predictions can have far reaching consequences e.g., recommend an operation, recommend sending a patient to hospice etc.

**When the cost of a mistake is high:**
- Ex: misclassification of a malignant tumor can be costly and dangerous

**When a new/unknown hypothesis is drawn:**
- “It's not a human move. I've never seen a human play this move.” [Fan Hui]
- Pneumonia patients with asthma had lower risk of dying [Caruana 2015]

**Compliance is key:**
- GDPR
- Right to Explanation
- Predictive performance is not enough [Doshi-Velez 2017]
Why Do We Need Explanations Now?

Risk Prediction with Blackbox Models

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Has Asthma</th>
<th>Risk of Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>Yes</td>
<td>5%</td>
</tr>
<tr>
<td>85</td>
<td>Yes</td>
<td>6%</td>
</tr>
<tr>
<td>86</td>
<td>No</td>
<td>12%</td>
</tr>
<tr>
<td>87</td>
<td>No</td>
<td>15%</td>
</tr>
</tbody>
</table>

- Feature Importance (Higher risk of death): Low — High
- Feature Importance (Lower risk of death): Low — High

With Context: Patients with asthma have a lower risk of death from pneumonia because they receive more intensive care.

Misdiagnosis via Adversarial Attacks

- Fundoscopy: Absent/mild DR — Moderate/Severe DR
- Chest X-Ray: Normal — Pneumothorax
- Dermoscopy: Nevis — Melanoma

[Caruana 2015, Caruana 2017]

[Finlayson 2018]
Pillars of Explainable AI
7 Pillars of Explainable AI in Healthcare

- Transparency
- Domain Sense
- Consistency
- Parsimony
- Generalizability
- Trust/Performance
- Fidelity

© KenSci., Inc. 2018.
Transparency
Ability of the machine learning algorithm, model, and the features to be understandable by the user of the system.

Admission Prediction
What is the likelihood of the patient being admitted to the hospital from the emergency department.
The ML model for predicting Katherine’s likelihood of admission gives her a low likelihood (0.32)

Katherine’s physician has noted her age, health history and vital signs and is reassured by her relatively low risk score

The physician knows that the risk model is a deep learning model so he cannot understand how it is working

But, he can examines the top factors associated with prediction

- Heart rate
- Creatinine
- Family history
- No PCP
- Age
- Albumin WNL
- Absence of leukocytosis
• Transparency may mean different different things to different people

• Understanding Model Outputs:

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \]

• Understanding Algorithms:

\[ a_j^l = \sigma \left( \sum_k w_{jk}^l a_{k}^{l-1} + b_j^l \right) \]
**Transparency | Examples**

**Transparent**
- Falling Rule Lists
- GAM (Generalized Additive Models)
- GA2M (Generalized Additive Models with interactions)
- LIME (Locally Interpretable Model Agnostic Explanations)
- Naïve Bayes
- Regression Models
- Shapley Values

**Semi-Transparent**
- Shallow Ensembles

**Non-Transparent**
- Deep Learning
- SVM (Support Vector Machines)
- Gradient Boosting Models

© KenSci., Inc. 2018.
Domain Sense
The explanation should make sense in the domain of application and to the user of the system

ED Census Prediction
Predict the number of patients in the emergency department (ED) at a given time

Several years later, Katherine revisits the emergency department due to abdominal pain. She has an elevated temperature and is dehydrated. She is at the ED on a Friday after work. The ED is very crowded and she must wait several hours to be seen.
Making Sense of Model Output

• Interpreting output from machine learning models may also has an element of subjectivity

• Example: If a patient has a readmission risk score 0.62, what does that mean?

Actionability

• Actionability does not presume causality but can be used to devise interventions

• Actionability is temporal e.g., flu vaccine administration is a 1 minute task vs. weight loss programs which can take months

<table>
<thead>
<tr>
<th>Mutability</th>
<th>Interveniability</th>
<th>Actionability</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immutable</td>
<td></td>
<td></td>
<td>Age, Sex, Ethnicity</td>
</tr>
<tr>
<td>Mutable</td>
<td>Non-Interveniable</td>
<td></td>
<td>Intrinsic Heart rate variability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Marital Status</td>
</tr>
<tr>
<td></td>
<td>Interveniable</td>
<td>Signal</td>
<td>Temperature (in Appendectomy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>Appendicitis</td>
</tr>
<tr>
<td></td>
<td>Post-Interveniabe</td>
<td></td>
<td>Immunization</td>
</tr>
</tbody>
</table>

Figure 1: Factors for Predicting Risk of Readmission
**Consistency**

The explanation should be consistent across different models and across different runs of the model.

**LWBS**

Left without being seen refers to a patient leaving the facility without being seen by a physician.

**Factors**

- Prior LWBS
- Insurance status
- Past history
- Temperature
- Chief complaint
- Age
Consistency | Model Multiplicity

• Given the same dataset, multiple machine learning algorithms can be constructed with similar performance
• The explanations from multiple explainable algorithms should be very similar
• Divergent Explanations symptomatic of problem with explanations and/or algorithm(s)
• Evaluation:
  • Human Evaluation: Expert Agreement
  • Machine Evaluation: Ranked List Comparison

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model A</th>
<th>Model B</th>
<th>Model C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Gender</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Diabetic</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Race</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Smoker</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Alcoholic</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Variables for Length of Stay Prediction Ranked

\[ W = \frac{12S}{m^2(n^3 - n)} \]

Kendall's W: [Kendall 1939]
Pillar 4: Parsimony

The explanation should be as simple as possible

Applies to both the complexity of the explanation and the number of features provided to explain

Admission Disposition

Where in the hospital the patient should go once they are admitted
Parsimony

• MDL (Minimum Description Length) and Occam’s Razor

• **Occam’s Razor**: To derive a unifying diagnosis that can explain all of a patient’s symptoms

• **Hickam’s Dictum**: A man can have as many diseases as he damn well pleases

• Occam’s Razor in Machine Learning [Domingos 1999]
  • Occam’s First Razor
  • Occam’s Second Razor

• The simplest explanation is not always the best explanation
Pillar 5: Generalizability | Length of Stay Prediction

Generalizability
Models and explanations should be generalizable across problem whenever possible

Length of Stay
The time that a patient will spend at a particular healthcare facility

Katherine eventually develops diabetes. It is well controlled and she takes her medications as directed. One afternoon, she is admitted from clinic due to highly elevated glucose levels and a urinary tract infection. Her nurse tells her that based on her illness and other factors, her predicted length of stay is 3 days.
Model Generalizability

- **Local Models**: Models that give explanations at the level of an instance e.g., LIME, Shapley Values etc.
- **Cohort Level Models**: A type of global models where the explanations are generated at the level of cohort
- **Global Models**: Models that give explanations e.g., Decision Trees, Rule Based Models etc.

Algorithm Generalizability

- **Model Agnostic Explanations**:  
  - Examples: LIME, Shapley Values etc.
- **Model Class Specific Explanations**:  
  - Examples: Tree Explainers, Gradient Boost Explainers
- **Model Specific Explanations**:  
  - Examples: CENs, Decision Trees, Random Forest Explainer etc.
Trust / Performance

• The expectation that the corresponding predictive algorithm for explanations should have a certain performance

ICU Transfer Prediction

• Predict if a patient on the hospital ward will require transfer to the intensive care unit due to increasing acuity of care needs
• Expectation that the predictive system has a sufficiently high performance e.g., precision, recall, AUC etc. [Lipton 2016, Hill 2018]

• Explanations accompanied with sub-par predictions can foster distrust

• The model should perform sufficiently well on the prediction task in its intended use

• The model has at least parity with the performance of human practitioners

• Trauma patients: vital signs and lab criteria fulfill criteria to trigger alarm, leads to increasing numbers of false positives [Nguyen 2014]
Pillar 7: Fidelity | Risk of Readmission

Fidelity
• The expectation that the explanation and the predictive model align well with one another

Risk of Readmission
• Predict if the patient will be readmitted within a particular span in time, i.e. 30 days from time of discharge
Fidelity | Explanations

- An explanation is **Sound** if it adheres to how the model actually works.
- An Explanation is **Complete** if it encompasses the complete extent of the model.
- **Ante-Hoc**: Models where the predictive model and the explanation model is the same.
- **Post-Hoc Models**: Models where the predictive model and the explanation model are different.
- Special Case: **Mimic Models**

*Example*: Readmission model which uses lunar cycles as a feature.
Fidelity | Mimic Models

- Also known as Shadow, Surrogate or Student Models
- Use the output (instead of the true labels) from the complex model and the training data to train a model which is explainable [Bucilua 2006, Tan 2017]
- The performance of the student model is usually quite good
- Example: Given a highly accurate SVM, train a decision tree on the predicted label of the SVM and the original data
Conclusion: Explainable ML in Healthcare AI

Artificial Intelligence + Explanations → Assistive Intelligence

Clinical Expertise
HEALTHCARE NEEDS HELP.
AND HOPE.

DEATH VS. DATA SCIENCE
HELP US IN THIS MISSION

@KenSci
Ankur@KenSci.com
References

Ahmad 2018  Muhammad Aurangzeb Ahmad, Carly Eckert, Greg McKelvey, Kiyanu Zolfaghari, Anam Zahid, Ankur Teredesai. Death vs. Data Science: Predicting End of Life IAAI February 2-6, 2018


References


Gunning 2016 David Gunning Explainable Artificial Intelligence (XAI) DARPA/I2O 2016


References


Ross 2018, Ross, Casey. IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show Stat News July 25, 2018


© KenSci., Inc. 2018.
ACM: The Learning Continues...

• Questions/comments about this webcast? learning@acm.org

• ACM Code of Ethics: https://ethics.acm.org

• ACM’s Discourse Page: http://on.acm.org

• ACM Learning Webinars (on-demand archive): http://webinar.acm.org

• ACM Learning Center: http://learning.acm.org

• ACM SIGKDD: http://www.kdd.org/