Explainable, interpretable, and trustworthy AI for an intelligent digital twin: A case study on remaining useful life

Kazuma Kobayashi, SB Alam

Engineering Applications of Artificial Intelligence. 2024

Georgia State University

Abstract

- ✓Artificial intelligence (AI) and Machine learning (ML) are increasingly used for digital twin development in energy and engineering systems.
- ✓Need to be unbiased, interpretable and explainable.
- ✓Explains XAI and IML.
- ✓Justify the role of XAI/IML for digital twin systems for the prediction of RUL (Remaining useful Life), both locally and globally.

Problem Statement

- ✓Prognostics and health management (PHM) use DTs to monitor and manage system health utilizing statistical algorithms and models.
- ✓DTs help in:
  - ✓Conducting condition based maintenance.
  - ✓Making maintenance decisions.
  - ✓Anticipate future failures in advance.
- ✓XAI can help increase trust the RUL prediction of an Intelligent DT system (physical asset such machinery or infrastructure).

Digital Twin System

- ✓virtual representations of physical systems.
- ✓replicates a physical asset in the virtual environment, including its functionality, features, and behavior.
- ✓DT’s technology can reflect, mimic, and forecast the status of the operating physical system in real-time and is one of the best tools for RUL prediction.

AI/ML in DTs

- ✓Incorporating AI/ML in DTs can facilitate in
  - ✓risk informed decision making related to RUL of a component.
  - ✓Streamlining high performing simulations.
- ✓With ML comes uncertainty and lack of trustworthiness especially in complex non-linear models.
- ✓Need explainability in ML (XAI/IML).
- ✓the use of explainable, interpretable, and trustworthy AI is crucial for accurately predicting the remaining useful life in an intelligent DT system

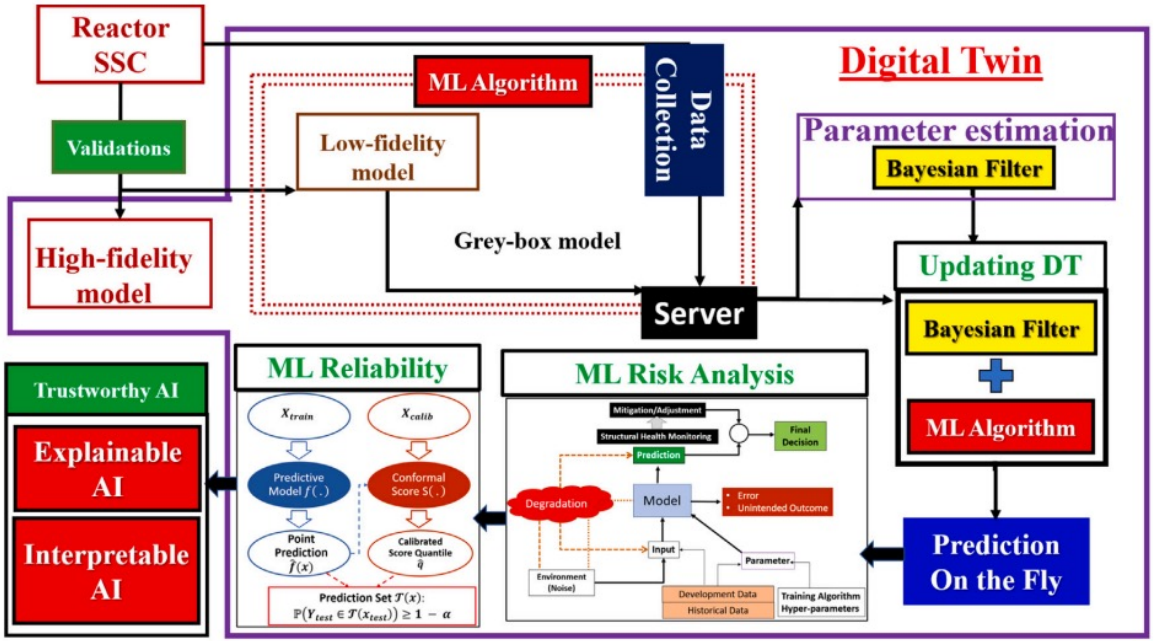
Explainable AI

- ✓Explainable AI is used to describe an AI model, its expected impact and potential biases.
- ✓XAI is important for several reasons:
  - ✓Builds trust by explaining the rational behind a decision.
  - ✓Transparency.
  - ✓International regulations requires to explain decision making process.
  - ✓Improves user's experience.
  - ✓Debug and improve.
  - ✓Increase reliability and accuracy of prediction of a ML model.

- ✓Explainability allows users to understand the factors most important in the model's prediction.
- ✓Interpretability allows the model's predictions to be easily understood by non-technical users.
- ✓Trustworthiness ensures that the model is not making predictions based on irrelevant or misleading features and is unbiased.

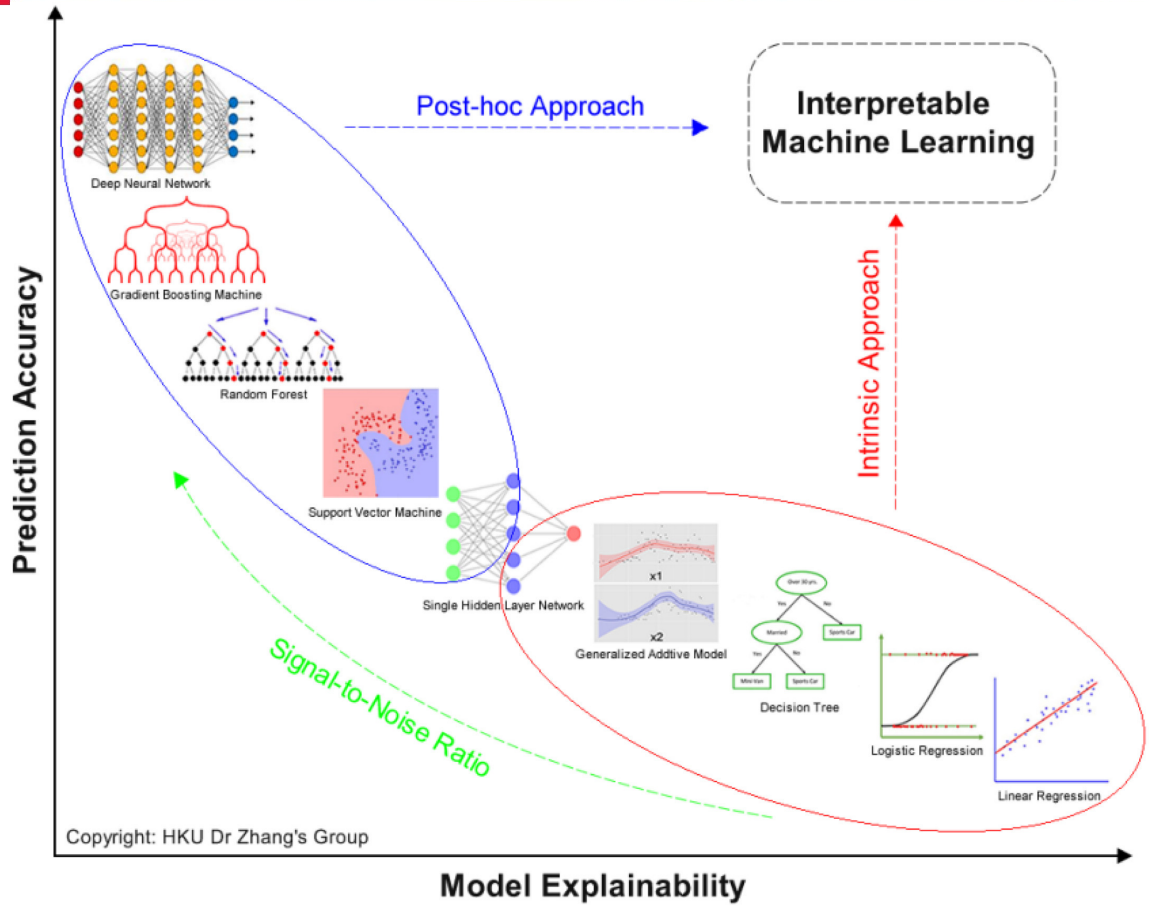
Intelligent Digital Twin framework

- ✓An Intelligent DT framework has five essential components.
  1. Prediction module
  2. System Update Module.
  3. Data processing module
  4. Visualization module
  5. Decision Making module.
- ✓System update and prediction module require sophisticated ML algorithms.


Reactor SSC
Validations
High-fidelity model
ML Algorithm
Low-fidelity model
Data Collection
Grey-box model
Server
Digital Twin
Parameter estimation
Bayesian Filter
Updating DT
Bayesian Filter
ML Algorithm
Prediction On the Fly
ML Reliability
ML Risk Analysis
Trustworthy AI
Explainable AI
Interpretable AI
Predictive Model $f(.)$
Conformal Score $S(.)$
Point Prediction $\hat{f}(x)$
Calibrated Score Quantile $\hat{q}$
Prediction Set $\mathcal{T}(x)$:
$\mathbb{P}(Y_{test} \in \mathcal{T}(x_{test})) \geq 1 - \alpha$
Mitigation/Adjustment
Structural Health Monitoring
Final Decision
Prediction
Degradation
Model
Error
Unintended Outcome
Input
Parameter
Environment (Noise)
Development Data
Historical Data
Training Algorithm Hyper-parameters

Interpretable AI

- Interpretable AI refers to the AI system's capacity to justify its judgement and predictions.
- Enables to comprehend the decision-making process for the user.
- Two strategies to develop such systems
  - Use transparent models (decision trees, linear models).
  - Post hoc explanations (feature importance).


Post-hoc Approach
Interpretable Machine Learning
Intrinsic Approach
Deep Neural Network
Gradient Boosting Machine
Random Forest
Support Vector Machine
Single Hidden Layer Network
x1
x2
Generalized Addtive Model
Decision Tree
Logistic Regression
Linear Regression
Signal-to-Noise Ratio
Prediction Accuracy
Model Explainability
Copyright: HKU Dr Zhang's Group

XAI and Interpretable AI

- XAI means creating AI systems that can explain their actions (transparent models or post hoc explanations).
- Interpretable AI refers to interpreting and comprehending a model's outcomes.
- XAI is a broader concept of developing explainable and transparent AI system whereas Interpretable explains the relationship between input and output variables related to decision making.

Interpretable methods

- ReLu-DNN
- Explainable boosting machine (EBM).
- Fast interpretable greedy-tree sums (FIGS).
- Partial dependence plot (PDP).
- Explainable neural networks(XNN).

- This study uses inherently interpretable models.
- Dataset used : PHM08 (Generated by C-MAPSS simulation tool developed by NASA to mimic turbofan engine degradation under different operating condition)

Feature selection

- Reduces a large set of features to a manageable subset.
- It can improve model's precision and interpretability.
- Removes noise and reduce overfitting.
- Methods:
  - Pearson correlation
  - Distance correlation
  - Feature importance

Pearson correlation

- Statistical technique to determine the linear relationship between two continuous variables.
- Computed for each feature wrt target variable.
- Simple and straightforward method to implement.
- Most frequently used metric.
- Not good with non-linear interactions in the data.

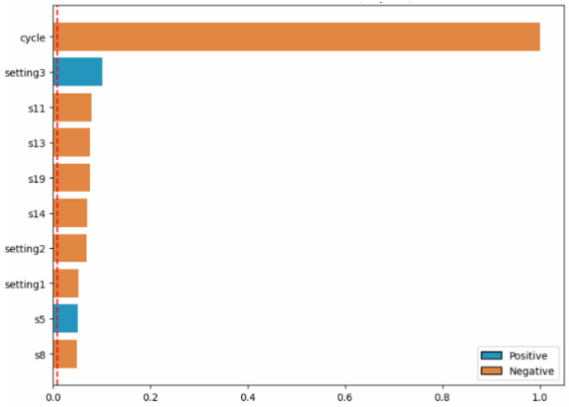
Distance correlation

- Quantify relationship between two variables.
- Determine the most significant parameter for predicting the target variable.
- Can reveal non-linear relationships between variables.
- Computationally demanding.

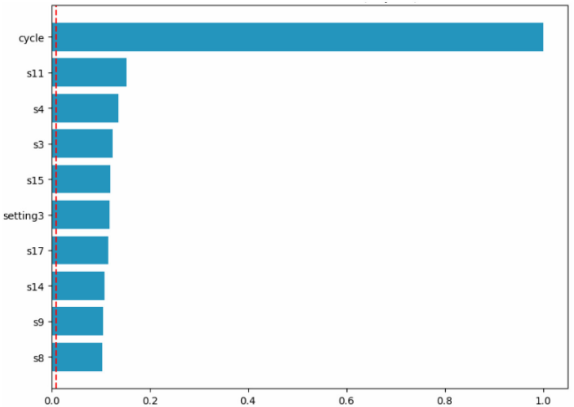
LGBM based feature importance

- Light gradient boosting machine.
- LGBM utilizes decision trees as its base model and train multiple trees through an iterative process.
- The algorithm assigns higher weights to observations poor predicted by previous trees, which helps improve accuracy.
- Capable of handling large datasets, missing values and categorical features.

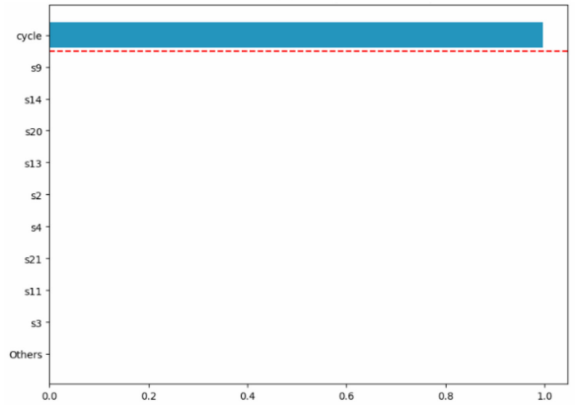
(a)


cycle
setting3
s11
s13
s19
s14
setting2
setting1
s5
s8
Positive
Negative
0.0
0.2
0.4
0.6
0.8
1.0

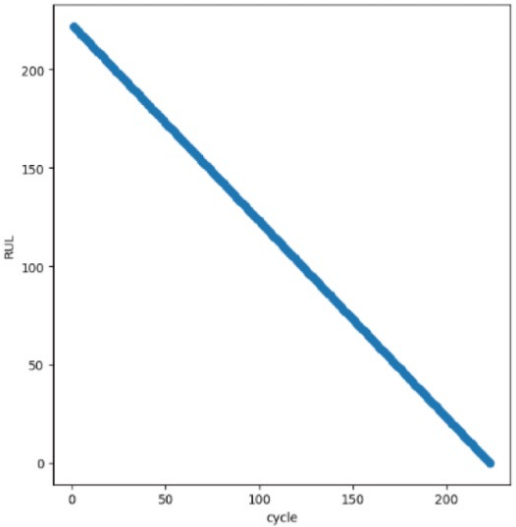
(b)

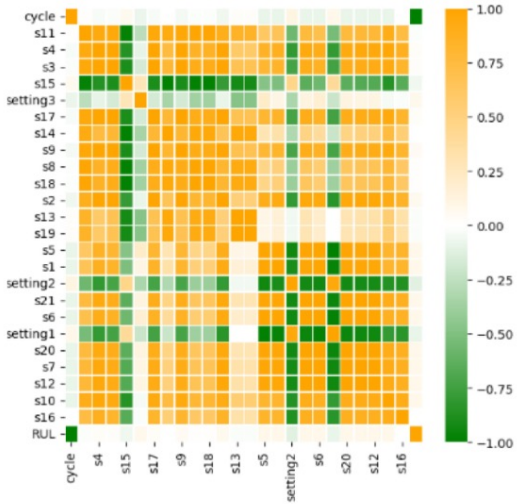

cycle
s11
s4
s3
s15
setting3
s17
s14
s9
s8
0.0
0.2
0.4
0.6
0.8
1.0

(c)


cycle
s9
s14
s20
s13
s2
s4
s21
s11
s3
Others
0.0
0.2
0.4
0.6
0.8
1.0

Multivariate correlation Heatmap



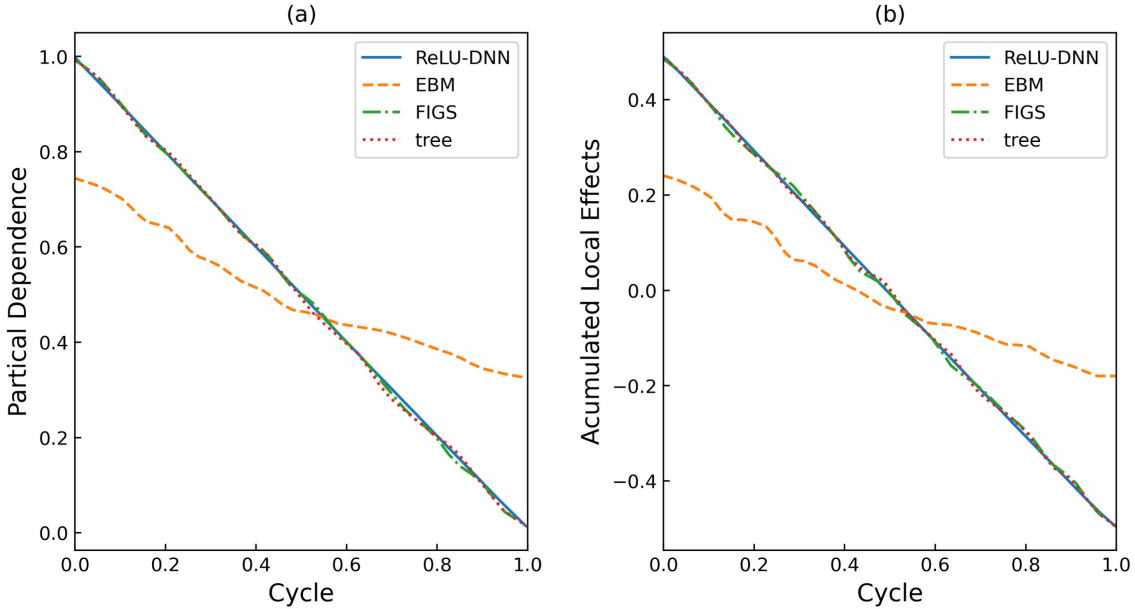


Global explainability

Partial Dependence plot

- PDPs are an interpretability ML method to understand the relationship between a single feature and the target variable.
- Capable of finding linear or more complex relationships in the data.
- Identify cofounding factors that might obscure the connection between an intriguing quality and the desired result.
- Easy to understand (works well for non-technical audiences).
- Provide only partial picture if a feature is linked with other features.

- Accumulated Local effects
- Plot the model's average prediction as a function of the feature of interest while maintaining the observed values of all other features.
- Capable to reveal linear and more complex relationship between features and target variable.
- Rely on observed values rather than mean and median to fix the values of other features, hence less sensitive to this decision.

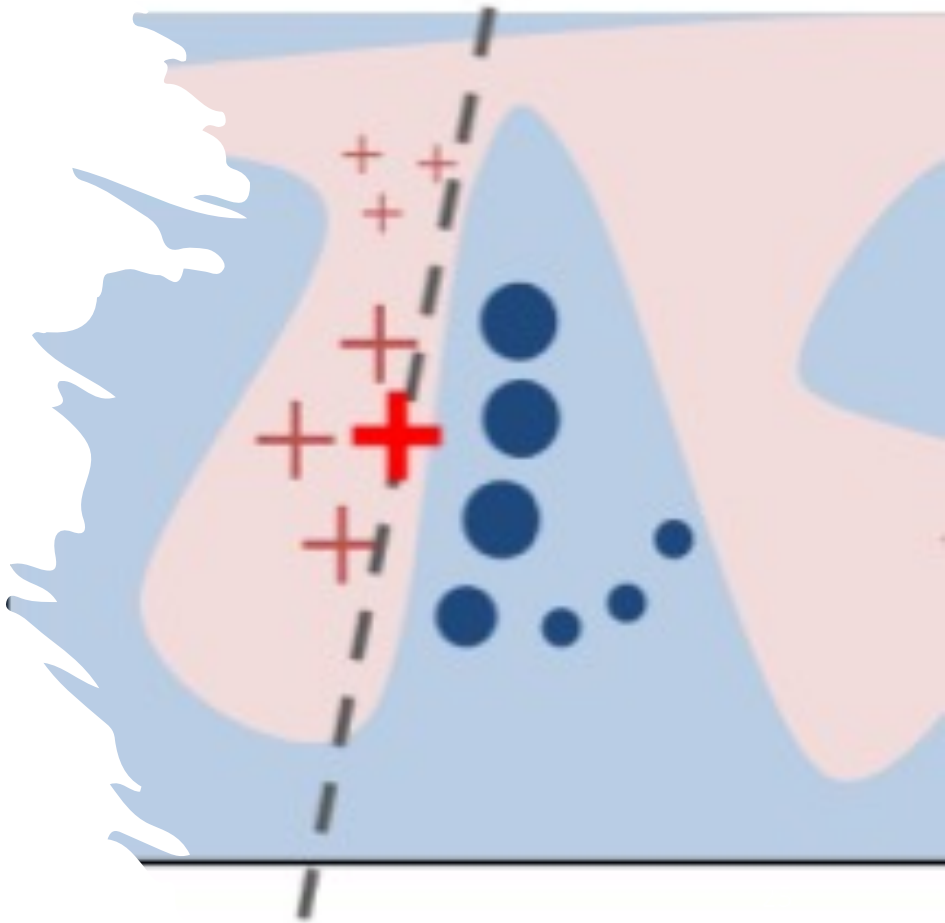

(a)
ReLU-DNN
EBM
FIGS
tree
Partical Dependence
Cycle
0.0
0.2
0.4
0.6
0.8
1.0
(b)
ReLU-DNN
EBM
FIGS
tree
Acumulated Local Effects
Cycle
−0.4
−0.2
0.0
0.2
0.4

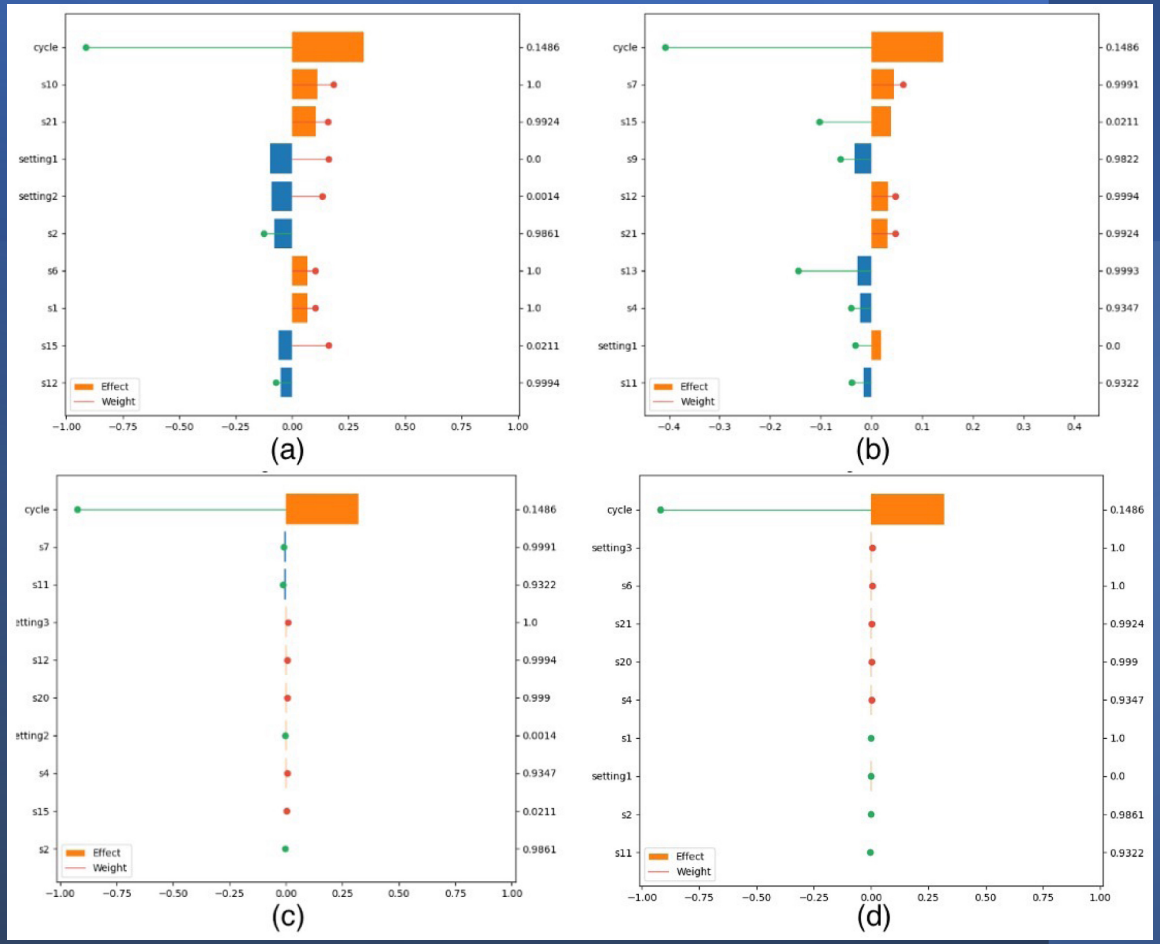
Local Explainability

Local explainability model agnostic explanations (LIME).

- Model agnostic.
- Provides local explanations for a single data point or a small set of data points.

- Toy example to present intuition for LIME.
- The black box model's complex decision function f is represented by the blue/pink background.
- The bold red cross is the instance being explained.
- LIME samples instances, gets predictions using f and weighs them by the proximity to the instance being explained (represented by size).
- The dashed lined is the learned explanation.




cycle
s10
s21
setting1
setting2
s2
s6
s1
s15
s12
0.1486
1.0
0.9924
0.0
0.0014
0.9861
1.0
1.0
0.0211
0.9994
Effect
Weight
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
(a)
cycle
s7
s15
s9
s12
s21
s13
s4
setting1
s11
0.1486
0.9991
0.0211
0.9822
0.9994
0.9924
0.9993
0.9347
0.0
0.9322
Effect
Weight
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
(b)
cycle
s7
s11
etting3
s12
s20
etting2
s4
s15
s2
0.1486
0.9991
0.9322
1.0
0.9994
0.999
0.0014
0.9347
0.0211
0.9861
Effect
Weight
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
(c)
cycle
setting3
s6
s21
s20
s4
s1
setting1
s2
s11
0.1486
1.0
1.0
0.9924
0.999
0.9347
1.0
0.0
0.9861
0.9322
Effect
Weight
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
(d)


Thank
you!!