CSCI 4800/5800 Explainable AI

Explainable AI (XAI): Core Ideas, Techniques, and Solutions

Denver

Explainable AI (XAI): Core Ideas, Techniques, and Solutions

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Comput. Surv. 55, 9, Article 194 (January 2023), 33 pages.

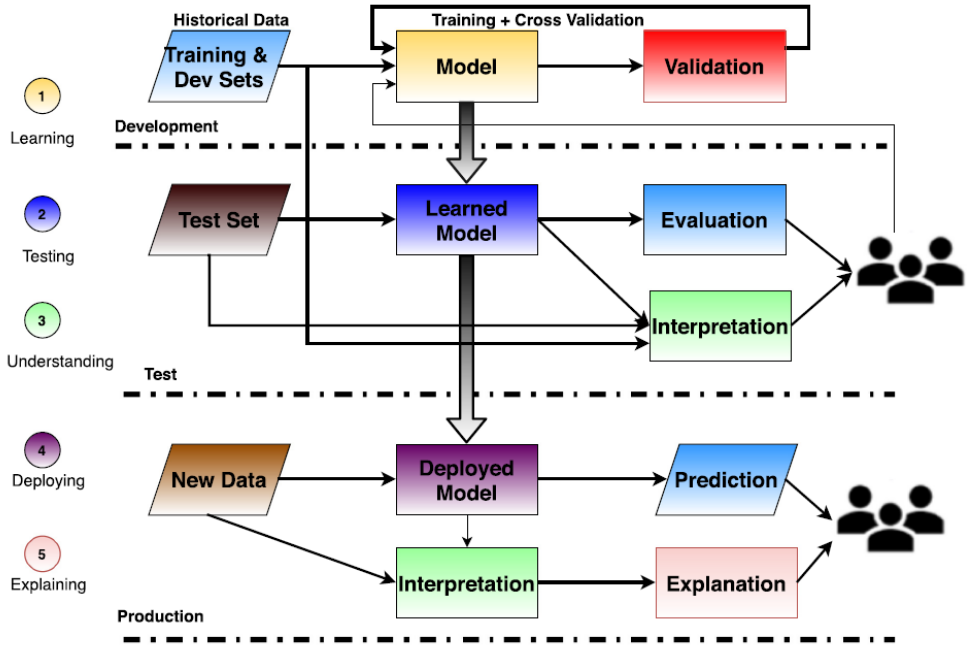
https://doi.org/10.1145/3561048

Introduction

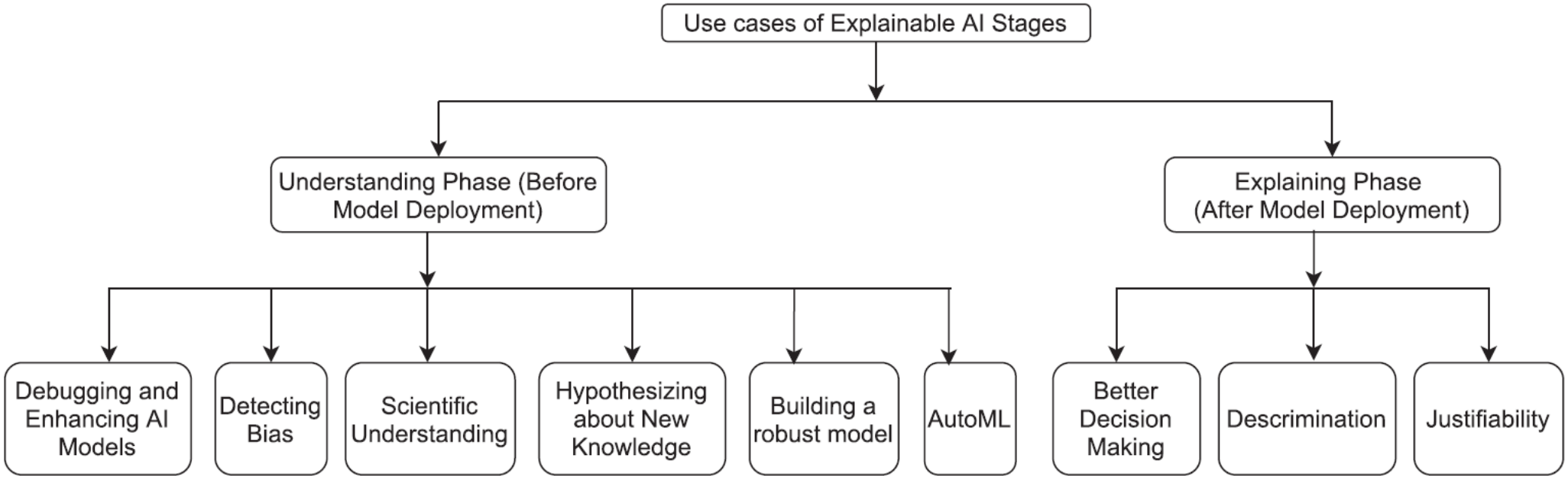
- The prime reason for rapid growth in XAI is the increased robustness of artificial intelligence (AI) systems in business, enterprise computing, and critical industries.
- For tech companies, a false prediction can lead to the application user being shown a wrong recommendation / decision / result.
- In critical sectors such as healthcare, finance, and the military, inaccurate predictions can have serious consequences on human life.
- Therefore, it is cruci.al to understand how these systems make their decisions

A pipeline for building ML models with explanation.

- Although a traditional ML pipeline (1, 2, and 4) can provide accurate predictions, it lacks two important phases: understanding (3) and explaining (5).
- The understanding phase involves the training and quality assurance of an AI model.
- The explaining phase is important when an ML model is deployed and used in real-world applications.
- The figure on the right illustrates a revised ML life cycle with the additional steps.

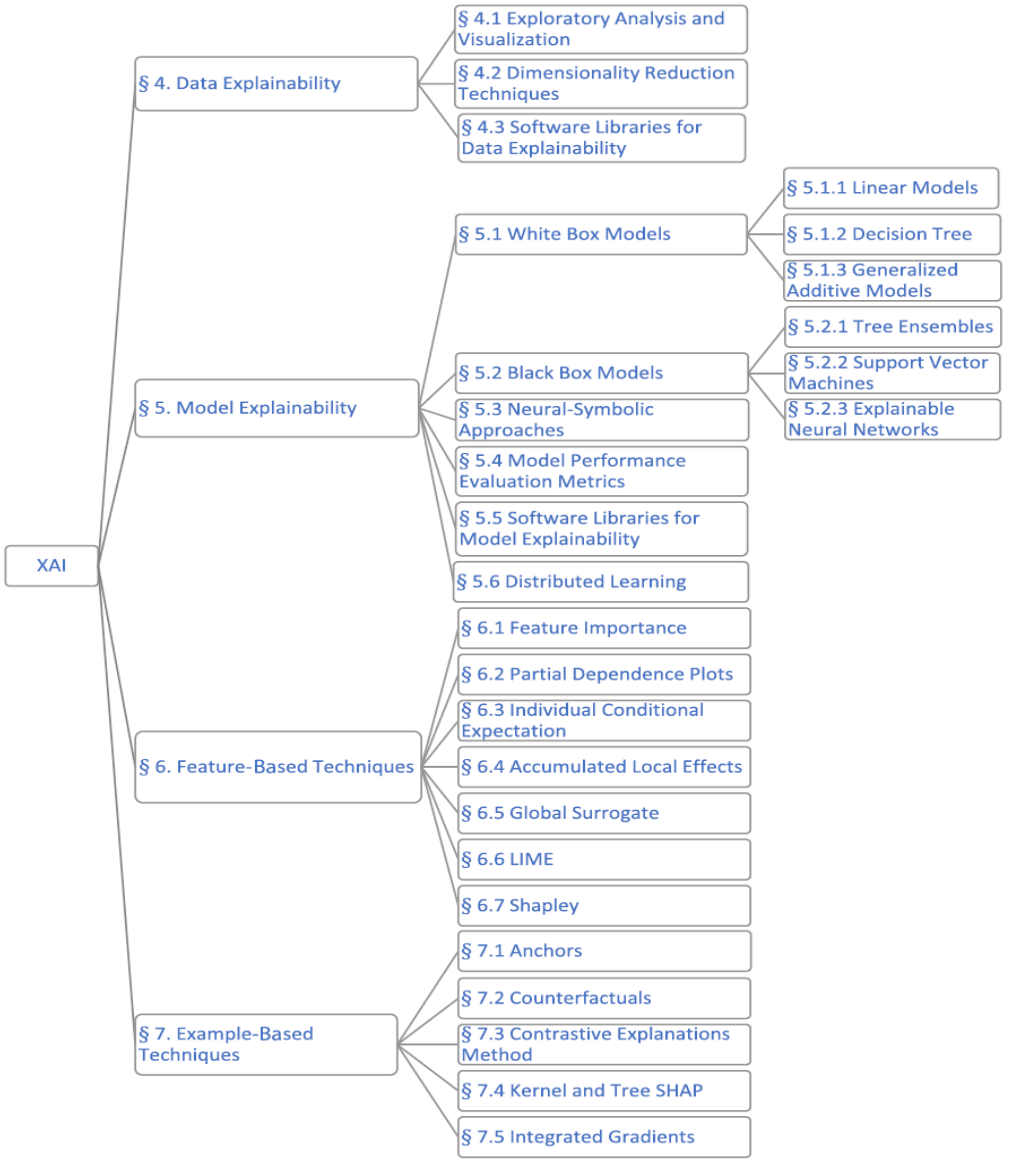


Use cases of XAI at different phases.



Taxonomy of XAI Techniques

- Data Explainability
- Model Explainability
- Feature-Based Techniques
- Example-Based Techniques



White Box Versus Black Box Model Techniques

- Black box models are non-transparent in nature, whereas white box models are transparent and comparatively easy to understand.
- The white box model is also termed post hoc, as it is applied on the model after training.
  - Some AI models are simple and self-explanatory.
  - For example, the predicted outcome y can be mathematically expressed as a weighted sum of all of its features $\bar{x}$.
- The black box model is also termed intrinsic, as it is achieved by limiting the complexity of an AI model.
  - Black box models such as neural networks or complex ensembles of much lower complexity.
  - The architecture of these models is hard to decipher, as it is not clear how important a role any given feature plays in the prediction model or how it interacts with other features.
- *Note that different authors may use different definitions of white box and black box based on visibility into the models.*

Model-Specific Techniques Versus Model-Agnostic Techniques

- Model-specific techniques deal with inner working of a model to interpret its results.
  - Model-specific interpretation tools are designed purely to interpret models with specific features and capabilities.
  - They can be used only for a single algorithm class.
- Model-agnostic techniques deal with analyzing features, their relationship with outputs, and the data distribution.
  - The interpretation techniques classified as model agnostic can be used on any ML model.
  - The widely used Local Interpretable Model Explanations (LIME) technique is model agnostic and can be used to analyze and interpret any set of ML inputs and corresponding predictions (outputs).
  - *Note that while model-agnostic techniques can be used on any ML model, their implementations generally cannot – for example, there are important differences in implementing image-based techniques versus text-based techniques.*

Global Interpretation Versus Local Interpretation

- The global interpretation analyzes the decision-making process at a broader level and is goal oriented.
  - Global interpretation methods involve an overall analysis of a model and its general behavior.
  - The process of defining variables, their dependency, and their interactions goes alongside with the process of assigning importance to these components.
- The local interpretation gives detailed explanations for every decision made.
  - Local interpretation involves an analysis of individual predictions and decisions made by the model, to clarify why the model suggested a particular course of action.
  - When a data point prediction/decision is analyzed, the focus is on the subregion around that data point.
  - It enables us to understand the contextual importance of the data point output in that space.

XAI Techniques Versus Taxonomy of XAI Techniques

Classification	XAI Techniques	Global	Local	Model Specific	Model Agnostic	White Box	Black Box
Data explainability	Commonly used data visualization plots	✓	✗	✗	✓	N.A.	N.A.
	Dimensionality reduction techniques	✓	✗	✗	✓	N.A.	N.A.
White box models	Linear model (Section 5.1)	✓	✗	✗	✓	✓	✗
	Decision tree (Section 5.1)	✓	✗	✗	✓	✓	✗
	Generalized additive models (GAMs) (Section 5.1)	✓	✗	✗	✓	✓	✗
	Tree ensembles (Section 5.1)	✓	✗	✗	✓	✓	✗
Artificial neural networks	Neural networks (Section 5.2)	✓	✗	✗	✓	✗	✓
	Neural-symbolic (Section 5.3)	✓	✓	✓	✓	✓	✗
Evaluation metrics	Model evaluation metrics (Section 5.4)	✓	✗	✗	✓	✓	✗
Feature-based XAI techniques	Feature importance (Section 6.1)	✓	✗	✗	✓	✗	✓
	Partial dependence plots (Section 6.2)	✓	✗	✗	✓	✗	✓
	Individual conditional expectation (Section 6.3)	✓	✗	✗	✓	✗	✓
	Accumulated local effects (ALE) (Section 6.4)	✓	✗	✗	✓	✗	✓
	Global surrogate (Section 6.5)	✓	✗	✗	✓	✗	✓
	Local interpretable model-agnostic explanations (LIME) (Section 6.6)	✗	✓	✗	✓	✗	✓
	Shapley value (Section 6.7)	✓	✓	✗	✓	✗	✓
Example-based XAI techniques	Counterfactuals (Section 7.2)	✗	✓	✗	✓	✗	✓
	Anchors (Section 7.1)	✗	✓	✗	✓	✗	✓
	Contrastive explanations method (Section 7.3)	✗	✓	✗	✓	✗	✓
	Prototype counterfactuals (Section 7.2)	✗	✓	✗	✓	✗	✓
	Integrated gradients (Section 7.5)	✗	✓	✗	✓	✓	✗
	Kernel SHAP (Section 7.4)	✓	✓	✗	✓	✗	✓
	Tree SHAP (Section 7.4)	✓	✓	✓	✗	✓	✗

N.A., not applicable.