IOT BASED EXPLAINABLE AI METHOD USING IMAGE DATASET

Presented
by
Umesh Kumar
23MT0430

Under Supervision
Of
Prof. Rajendra Pamula

Table of contents

IOT

The **Internet of Things (IoT)** is a network of interconnected physical devices that communicate and exchange data with each other over the internet.

Key Components of IoT

- Devices and Sensors
- Connectivity
- Data Processing
- Storing

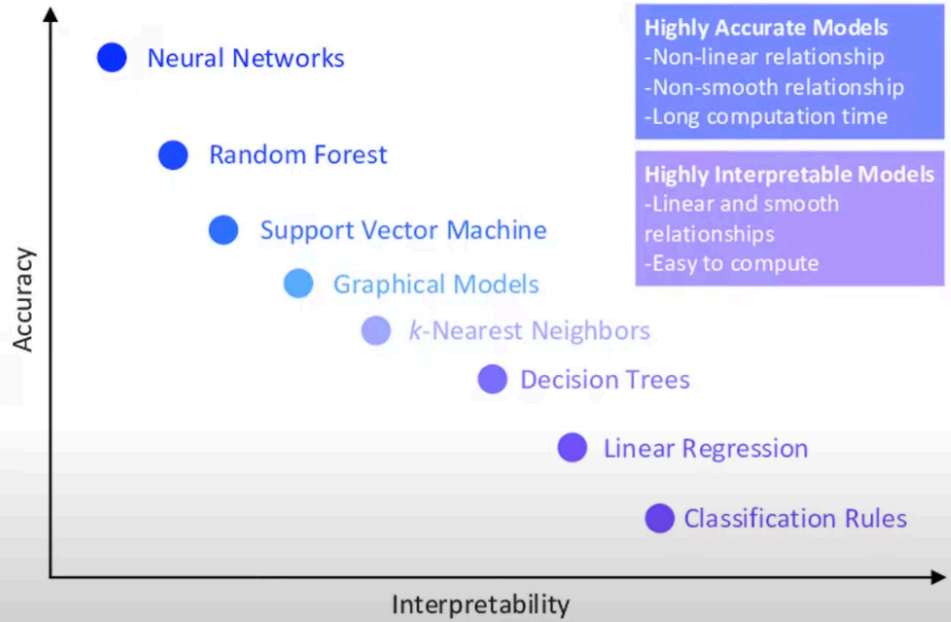
Explainable AI

Explainable AI (XAI) refers to a set of processes and methods that make the outputs and decisions of artificial intelligence (AI) systems understandable to humans. As AI systems, especially those using deep learning and complex algorithms, become more prevalent in critical decision-making areas (like healthcare, finance, and autonomous driving)

Need of Explainable AI

- Explainability
- Interpretability
- Transparency
- Trust
- Debugging and Improvement

What is explainable AI?



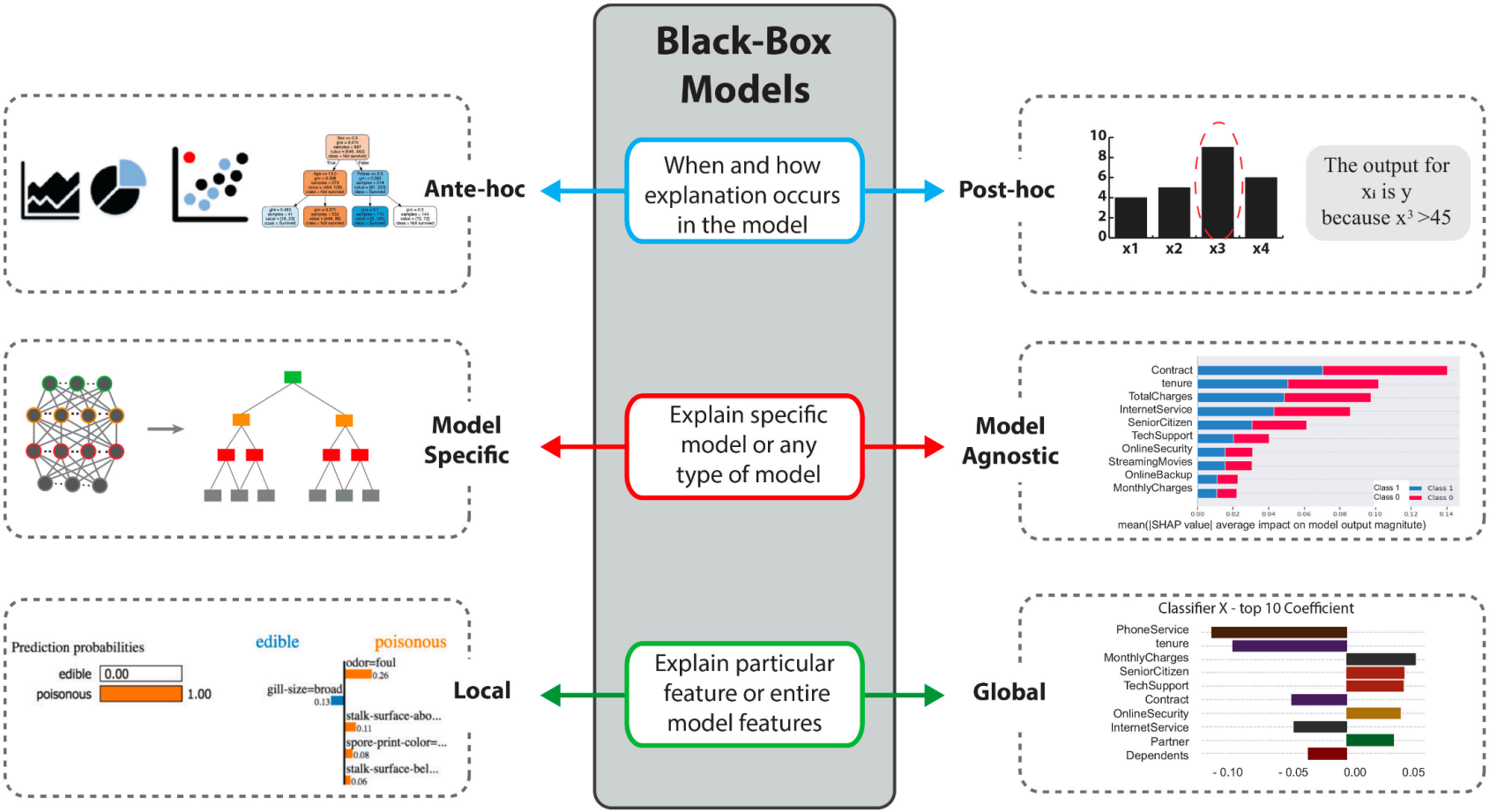


Fig. 2. Taxonomy of XAI in different perspectives.

„Model based“

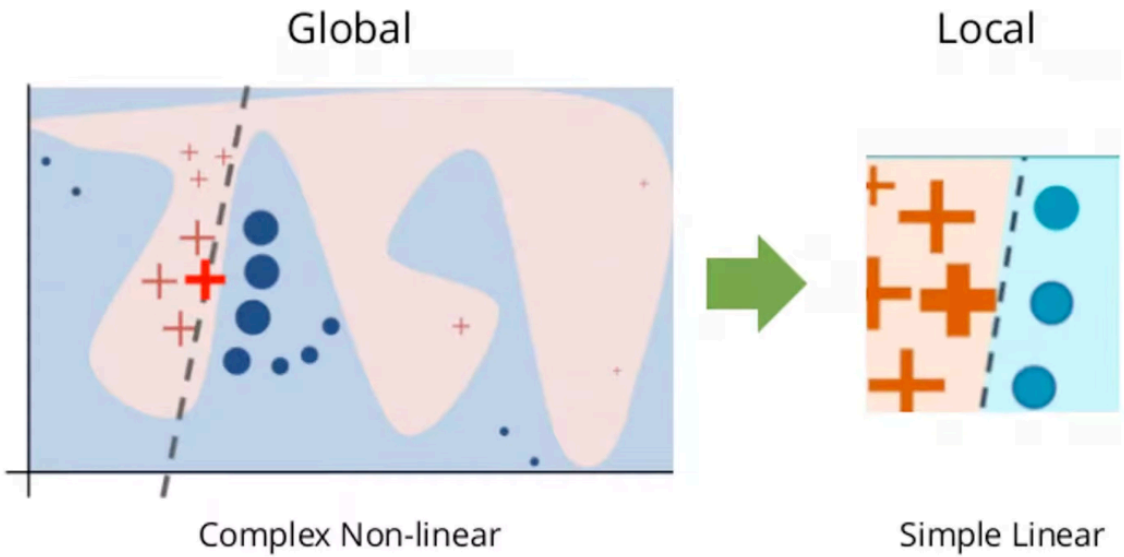
Build **interpretable** ML models

„Post-hoc“

Derive explanations for **complex** ML models

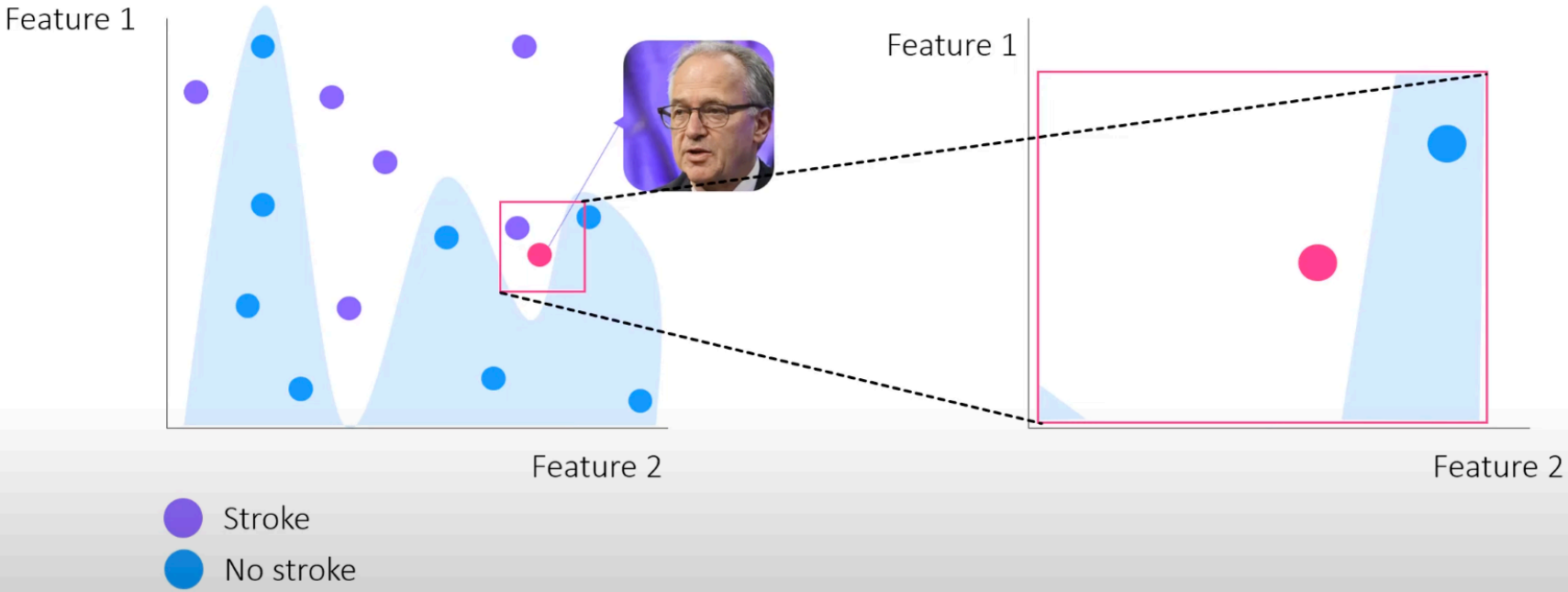
Black-box approach

White-box approach

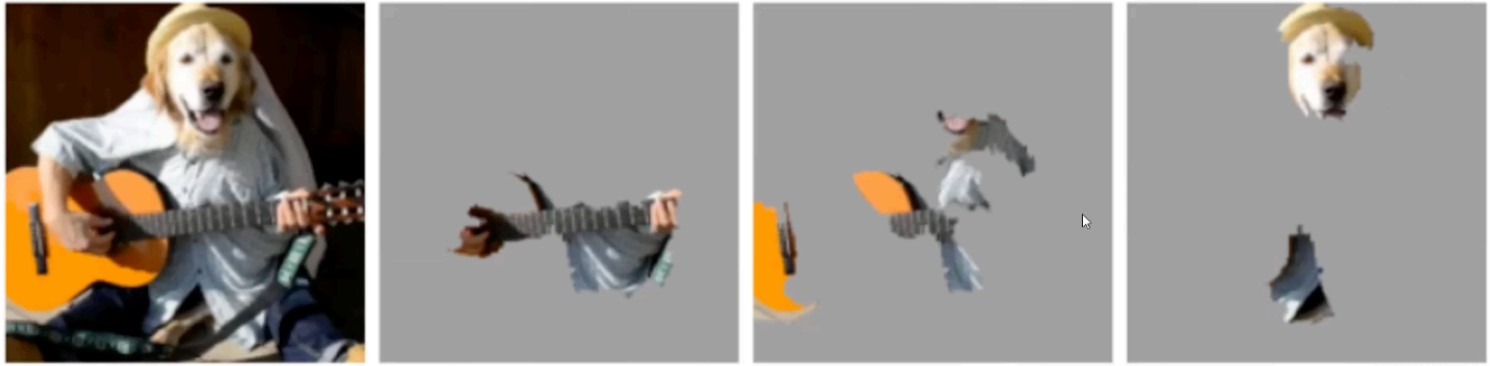


Source: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Ribeiro et al.

Local interpretable model-agnostic explanations



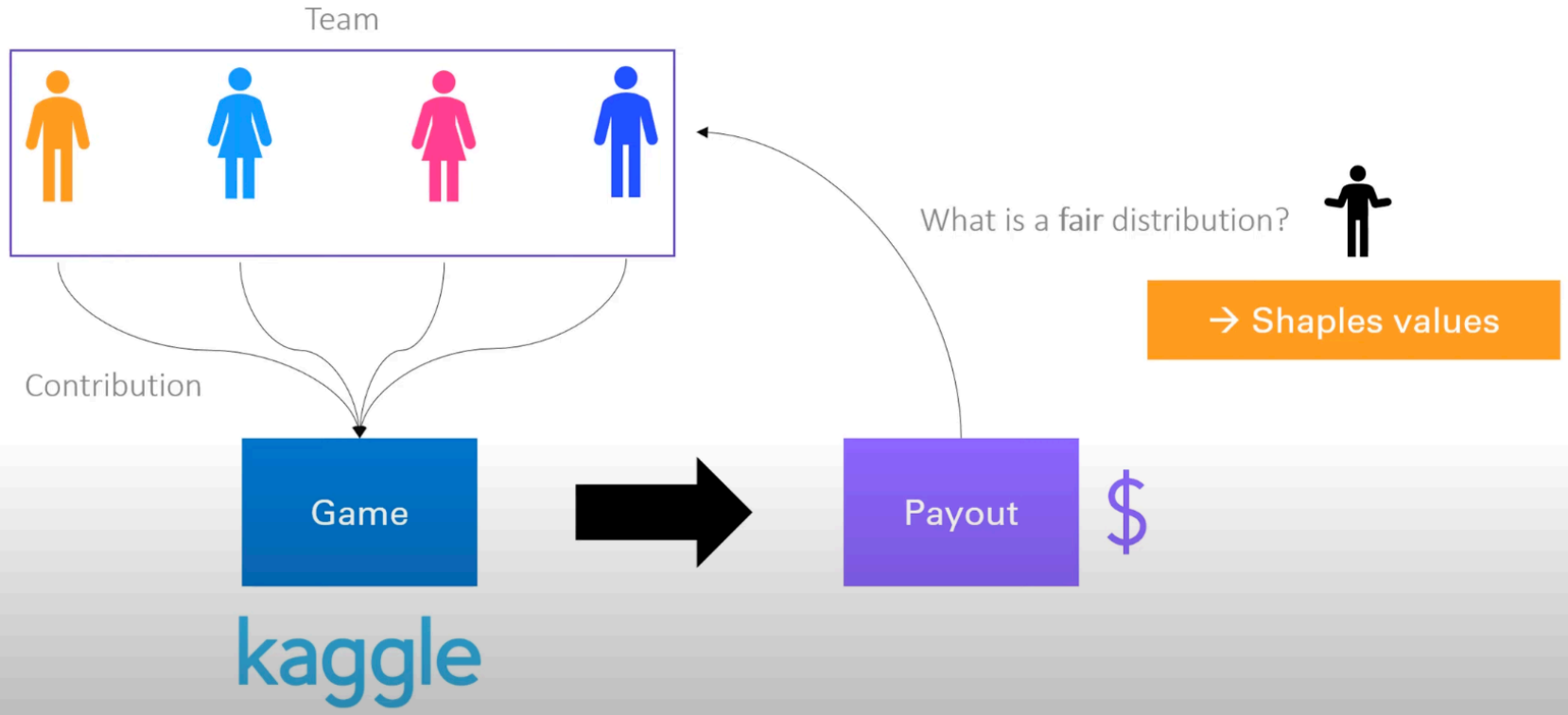
Other examples for LIME



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$)**, "Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

SHapley Additive exPlanations



XAI IN IOT APPLICATION DOMAINS

- Autonomous Systems and Robotics
- Smart Agriculture
- Environmental Monitoring
- Financial System
- Healthcare

CHALLENGES, OPEN ISSUES, AND FUTURE RESEARCH DIRECTIONS

- XAI methods does not measure expected error rates
- XAI models are vulnerable to adversarial attacks
- Lack of proper structure to combine multiple XAI methods
- Non-standardised terminology
- Absence of a quantitative measure of the completeness and accuracy of systems
- It explains what are important features, but does not explain what is the exact relationships between features and why they are important

CONCLUSION

XAI methods provide a spectrum of tools and techniques to interpret and understand AI models. Choosing the right XAI method depends on the model type, data type, and the specific needs of the stakeholders (e.g., doctors, patients, regulators). For IoT-based systems, especially in sensitive fields like healthcare, combining several XAI techniques (e.g., SHAP and uncertainty quantification) can provide a more holistic understanding of model behavior and reliability.

References

D. S. Watson, “Conceptual challenges for interpretable machine learn- ing,” *Synthese*, vol. 200, no. 2, p. 65, 2022.

A. Das and P. Rad, “Opportunities and challenges in explainable artificial intelligence (XAI): A survey,” 2020, *arXiv:2006.11371*.

G. Fidel, R. Bitton, and A. Shabtai, “When explainability meets adversarial learning: Detecting adversarial examples using SHAP sig- natures,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2020, pp. 1–8.

A.B.Arrietaetal.,“Explainableartificialintelligence(XAI):Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

-

Thank you!