Explainable AI for Cyber Security Applications

ABSTRACT

The goal of Explainable AI(**XAI**), a branch of Artificial Intelligence, is to make AI systems more **transparent** and **interpretable**. AI models are made to perform human tasks with very high computational speed and accuracy. Over the years, the quest to **improve** the **accuracy** of these models **increased** the **complexity** of the system and created "**black boxes**". The "black box" model (inability to explain the working of the model) **raised concerns** about the system regarding **accountability**, **trust**, **transparency,** and **fairness**. Regulatory bodies are now attempting to address these concerns by **implementing laws** guiding the use of these "black box" models. The European Union, General Data Protection Regulation **(GDPR)** has included that all AI models must be interpretable to be used in decision-making, the **European Commission** and United states are also working on legislation to address these issues and **regulate the use of the "black box" models** in industries.

This research will solve the problem of the "black box" in AI by developing an Explainable AI model for cybersecurity applications. This will **improve trust in AI** models while **meeting regulatory requirements**. AI is frequently used in cybersecurity to assist corporations in safeguarding their systems and data. The methodology of this research will utilize a chosen public dataset (CIC-Bell-DNS 2021[2] Dataset from the Canadian Institute for Cybersecurity) illustrating one of such applications to build and demonstrate our explainable model.

USE CASE

XAI-based model for Detection and Classification of Malicious domains.

- The dataset contains DNS (Domain Name System) features of a **benign** (non-threat) domain.
- DNS features are also recorded for types of **malicious** (threats) domains.
    - Malware
    - Phishing
    - Spam

Ante-hoc: Model is interpretable
Post- hoc stage: Explanation of a complex model (see Fig. 1).
Model Agnostic: The algorithm can be applied to all models
Global scope: Explaining the whole model.
Local scope: Explaining individual predictions
Mixed output format: Combines Numbers, Visuals, Texts, etc

METHODOLOGY

- **Machine Learning Model** - **Random Forest**
- **XAI libraries** (Explanators) – **SHAP** (Shapley Additive exPlanations) and **LIME** (Local Interpretable Model-Agnostic Explanations)

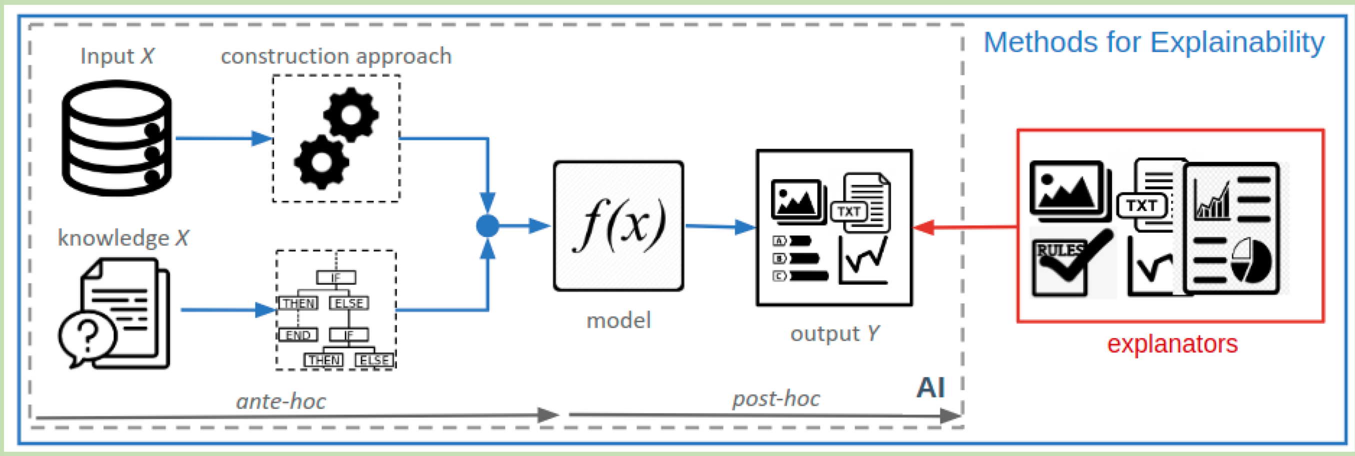


Figure 1. Diagrammatic view of how an explainable artificial intelligence (XAI) solution is typically constructed.[3]

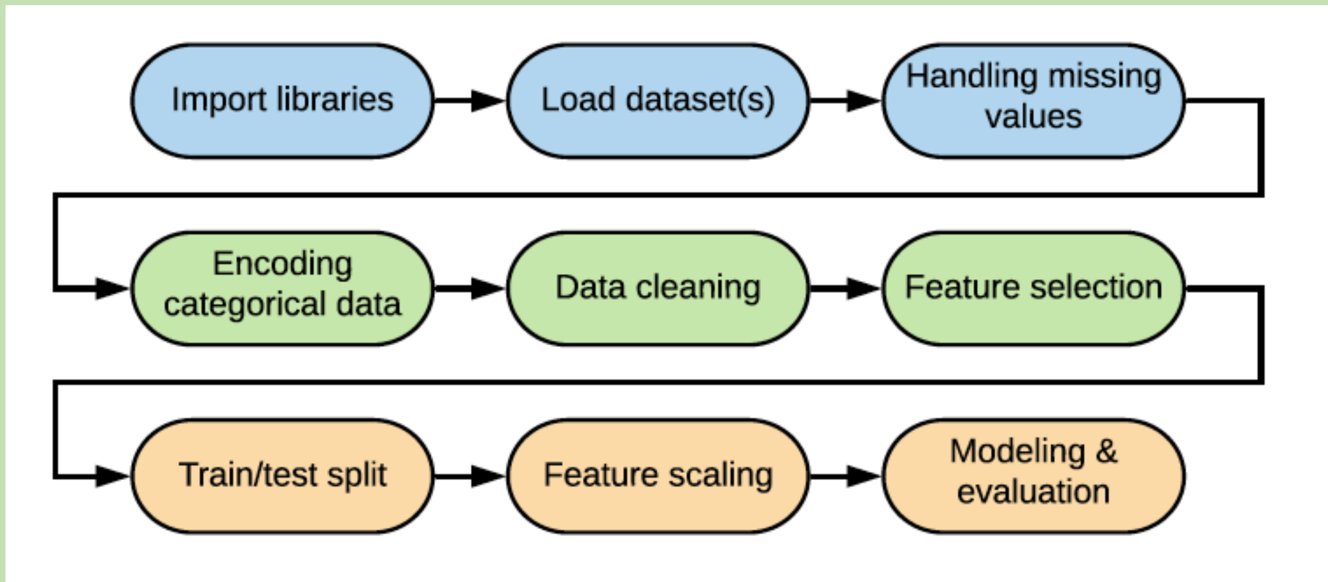


Figure 2: Machine Learning Lifecycle[4]

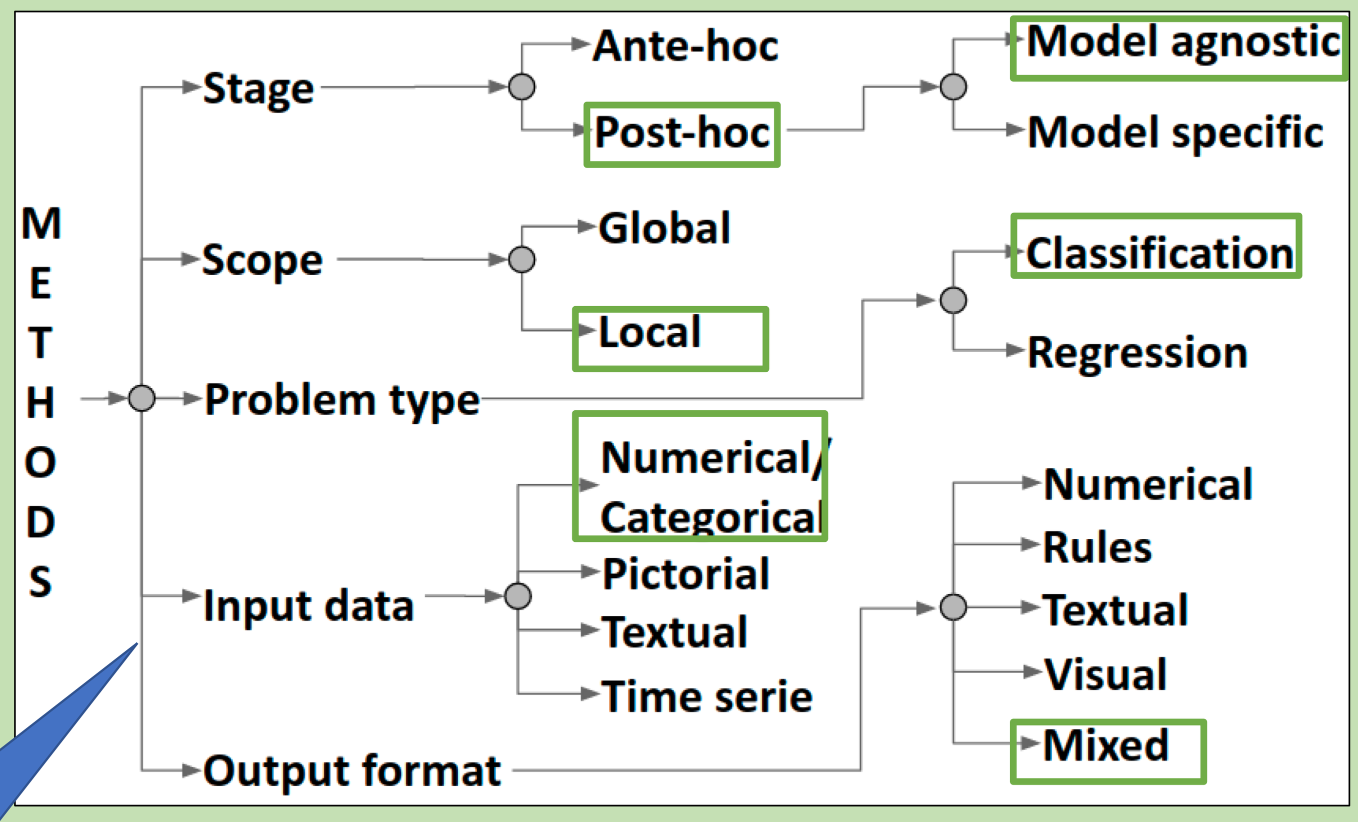
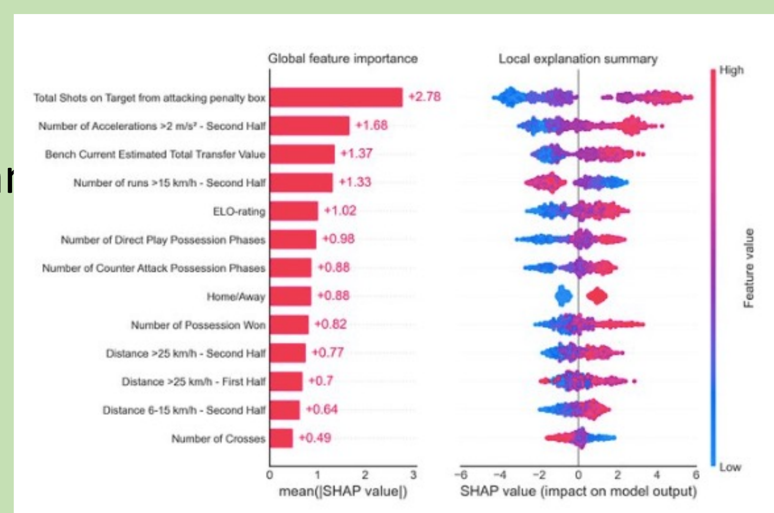


Figure 3. Classification of XAI methods into a hierarchical system.[3]

RESULTS

- Model performance evaluation: Accuracy, F1-Score, Precision, Recall
- Explaining the decisions using XAI (SHAP & LIME)[1]
    - SHAP summary plot- The SHAP value for each feature indicates the impact of the feature on the predicted label of the model.
    - Sample Figure 4 [5]
    - Compare XAI algorithm performance



References

[1] https://doi.org/10.1109/TIFS.2022.3183390

[2] https://www.unb.ca/cic/datasets/dns-2021.html

[3] https://doi.org/10.3390/make3030032

[4]https://dspace5.zcu.cz/bitstream/11025/41766/1/Explainable_Artificial_Intelligence.pdf

[5] https://doi.org/10.3390/ app11052378

By Ukamaka Nkechi Oragwu
M.Sc. Artificial Intelligence, Brunel University, 2023