{% include breadcrumbs.html %}

{% include icon.html icon="fa-solid fa-heart-pulse" %} Explainable Artificial Intelligence for Trustworthy and Transparent Decision-Making in Medical Applications

Project Description

The project seeks to address the growing need for **transparency, accountability, and interpretability** in artificial intelligence (AI) systems used in healthcare. As deep learning and other machine learning techniques become integral to medical diagnostics, prognosis, and treatment planning, the *black-box* nature of many AI models poses significant challenges for clinical adoption, regulatory approval, and patient trust.

This research focuses on developing and evaluating **Explainable Artificial Intelligence (XAI)** methodologies tailored to medical contexts. The goal is to ensure that AI-driven decisions are not only accurate but also **interpretable by clinicians**, **understandable to stakeholders**, and **compliant with ethical and legal standards**.

Core Framework: Multi-Dimensional Explainability

The project adopts a **multi-dimensional framework** for explainability, structured around four core axes:

1. Data Explainability

- Enhancing the transparency of input features and their individual contributions to model decisions.
- Identifying data biases and improving feature relevance interpretation.

PROF

2. Model Explainability

- Designing or adapting models that are inherently interpretable (e.g., decision trees, rule-based models).
- Exploring **hybrid approaches** that balance predictive accuracy with interpretability.

3. Post-hoc Explainability

- Applying **state-of-the-art interpretability techniques** such as:
    - SHAP (SHapley Additive exPlanations)
    - Grad-CAM (Gradient-weighted Class Activation Mapping)
    - LIME (Local Interpretable Model-Agnostic Explanations)
    - LRP (Layer-wise Relevance Propagation)
- Making complex model predictions more understandable to non-technical users.

4. Assessment of Explanations

- Developing **robust evaluation methodologies** including human-centered studies.
- Measuring explanation quality, usability, trustworthiness, and clinical relevance.

Project Objectives

- Bridge the gap between **algorithmic complexity** and **clinical insight**.
- Empower healthcare professionals with **transparent and reliable** AI systems.
- Develop XAI methods that align with **ethical standards** and **regulatory requirements**.
- Explore explainability from **technical, clinical, and regulatory** perspectives.

Research Impact

This research contributes to the **foundation of trustworthy AI in medicine**, promoting the safer and more responsible use of intelligent systems in healthcare environments. By integrating explainability at all stages—data, model, interpretation, and evaluation—this project supports the development of AI solutions that clinicians can understand, trust, and effectively use in real-world medical decision-making.