Master's Thesis : Research on Explainable Artificial Intelligence — Informational Leaflet

Author: Martin Katona
Supervisor: prof. Ing. Peter Sinčák, CSc.
Institution: Technical University in Košice,
Faculty of Electrical Engineering and Informatics,
Department of Cybernetics and Artificial Intelligence
Programme: Intelligent Systems (Computer Science)

Motivation

Artificial Intelligence (AI) has achieved remarkable predictive capabilities, but its adoption in healthcare and other critical domains is often hindered by a lack of transparency. Many AI models act as “black boxes,” producing predictions without clear reasoning, which reduces trust and usability. Explainable Artificial Intelligence (XAI) addresses this challenge by revealing decision-making factors and providing human-understandable explanations.

Methods & Framework

This work introduces an explainable diagnostic framework for diabetes prediction that integrates two core techniques. Shapley Additive Explanations (SHAP) are applied to assess feature importance, illustrating how individual factors such as age, BMI, or blood pressure influence the model's predictions. Retrieval-Augmented Generation (RAG) combined with Large Language Models (LLMs) is then used to produce natural language explanations grounded in verified medical sources, reducing the risk of hallucinations. An interactive graphical interface allows users to explore predictions, visualize feature contributions, receive explanations with direct references, and ask additional questions for deeper clarification.

Key Results

The research successfully delivered a framework capable of producing transparent and user-oriented outputs. Feature engineering enhanced model efficiency by focusing on the most relevant inputs, improving usability without compromising performance.

The project also validated the role of natural language outputs and visual explanations in building user trust, supported by evidence retrieval from reliable sources. Evaluation of multiple language models ensured that the generated insights remain consistent, grounded, and suitable for real-world applications.

Applications in Real Life

While diabetes prediction served as the case study, the proposed framework is applicable across various domains where transparency is essential. Potential use cases include any field requiring interpretable AI outputs with verifiable sources.