Explainable Artificial Intelligence for Stroke Prediction

Pavitra, Jing, Hamza, Hai

Winter Semester 2025/2026

1. Introduction & Research Objectives

In recent years, machine learning has shown remarkable success in medical prediction tasks, such as early diagnosis and risk assessment. However, the increasing use of complex, non-linear models, often referred to as "Black Box models," has raised serious concerns about interpretability and transparency—especially in healthcare, where decisions can directly affect human lives. Clinicians need not only accurate predictions but also clear explanations for why a model classifies a patient as "high risk" or "low risk." This need gave rise to the field of Explainable Artificial Intelligence (XAI), which aims to bridge the gap between model performance and human understanding.

In this project, our research objectives focus on explainability methods applied to a medical prediction problem using the Stroke Prediction Dataset, which contains patient and demographic information (see Section 3). The dataset is used to predict the likelihood of a stroke—an important health outcome where both predictive accuracy and interpretability are essential. We will build an advanced ML model, i.e., XGBoost, along with a simple Decision Tree model. We will then apply several XAI tools and aim to identify which method provides the most trustworthy, interpretable, and clinically meaningful insights into stroke risk prediction.

Recent research in Explainable AI (XAI) has focused on improving the transparency of black-box models, especially in healthcare applications. Ribeiro et al. (2016) introduced LIME, which provides local, interpretable approximations of complex models, while Lundberg and Lee (2017) developed SHAP, a unified framework based on Shapley values ensuring consistent and theoretically grounded feature attributions. More recent studies, such as Karimi et al. (2020), have advanced counterfactual explanations to show how small changes in input features can alter a model's decision altogether. These works form the foundation of modern XAI research and highlight complementary strengths from different methodologies such as LIME, SHAP, PDP & ICE, and Counterfactual Explanations, all of which will be used in this project.

2. Concepts / Methodologies

Pavitra

In this project, I plan to begin with data understanding and preprocessing, where I will explore the dataset, handle missing values, encode categorical features, and ensure balanced classes in training data for effective model training. Next, we will implement and configure two models—a simple, interpretable **Decision Tree** as a baseline and a more complex XGBoost model for comparison. After training, I will apply explainability techniques, focusing on **SHAP (SHapley Additive Explanations)**, to analyze both global and local feature importance and understand how individual attributes contribute to stroke predictions.

Jing

In this project, I will apply **Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE)** to visualize how individual features influence model predictions. These methods provide both global and local insights into model behavior. Using the same trained models and dataset.

Hamza

The main goal of my part of the project is to analyze how **LIME**can generate faithful and local explanations for stroke prediction models and to understand how its parameters affect explanation quality. Here, I specifically aim to study how **LIME** approximates a model's local decision boundary around individual patients, how parameters like kernel width and sampling influence fidelity, and how the resulting explanations align with medical knowledge.

Hai

In this project, I will focus on applying **Counterfactual Explanations** using frameworks such as **DiCE** and **CARLA** for *Random Forest*, *XGBoost*, and *Neural Network* models. The main objective of this part is not accuracy but insight—understanding how small, meaningful changes in patient features (such as blood glucose level, BMI, or hypertension status) could shift a prediction from "high risk" to "low risk." This approach highlights potential intervention points in clinical practice and provides actionable transparency for healthcare professionals. The analysis will compare counterfactual outcomes across different model architectures to assess robustness and clinical plausibility of explanations.

As a team

Finally, we will then evaluate the models using standard classification metrics such as **accuracy, precision, recall, F1-score, and ROC-AUC**, comparing their performance and interpretability.Then our team aims to compare **LIME, SHAP, ICE, PDP** and Counterfactual Explanations to highlight their respective **strengths and weaknesses** in providing transparency, insights.

3. Dataset

For this project, we selected the Stroke Prediction Dataset from the healthcare domain. Initially, we explored the Breast Cancer Wisconsin and Heart Disease datasets; however, we ultimately chose the stroke dataset due to its strong alignment with the goals of explainable AI. The dataset includes a diverse set of demographic, lifestyle, and medical features—such as age, BMI, average glucose level, hypertension, and smoking status—which influence the likelihood of a stroke in different ways. These features are relatively distinct from one another, with limited internal correlation, allowing for a clearer interpretation of each variable's individual contribution to the model's predictions. This makes the dataset particularly suitable for our focus on interpretability, as it enables us to apply explainability methods like SHAP and LIME to uncover and communicate which factors most strongly affect stroke risk in an understandable and transparent manner.

The stroke prediction dataset contains a mix of demographic, lifestyle, and medical features such as age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status, with the target variable being stroke occurrence. These features cover diverse aspects of a person's health profile, making the dataset well-suited for understanding how different risk factors contribute to stroke likelihood. The variety and relatively low correlation among features also allow for a more meaningful application of explainability techniques, as each attribute can be assessed individually for its impact on the model's predictions.

4. Work Structure and Timeline

The project will be carried out over approximately two months during the winter semester. The following schedule outlines the key milestones and deadlines:

Milestone	Date
Kick-off meeting	**Thu, 23 Oct 2025**
Assignment of topics	**Fri, 31 Oct 2025**
Proposal submission	**Fri, 7 Nov 2025**
Literature research	**5–14 Nov 2025**
Introduction to scientific research	**Tue, 11 Nov 2025**
Outline presentation	**Tue, 25 Nov 2025**
Project submission	**Mon, 2 Feb 2026**
Final presentation	**Tue, 10 Feb 2026**

Gantt Chart

Task	Start Date	End Date	Days
Project Planning	2025/10/23	2025/10/31	8
Literature Review	2025/11/1	2025/11/14	13
Model & Data Design	2025/11/15	2025/11/21	6
Outline Presentation	2025/11/22	2025/11/25	3
Implementation	2025/11/26	2025/12/12	16
Analysis & Discussion	2025/12/13	2025/12/20	7
Writing & Submission	2025/12/21	2026/1/5	15
Final Submission & Presentation	2026/1/25	2026/1/29	4