



KATHOLISCHE UNIVERSITÄT
EICHSTÄTT-INGOLSTADT

CATHOLIC UNIVERSITY
EICHSTÄTT-INGOLSTADT

Explainable AI Techniques on Stroke Risk Prediction

submitted by

Minh Hai Tran

Matriculation number: 11021917

Hamza Muhammad

Matriculation number: 11044170

Pavitra Chandarana

Matriculation number: 202679

Yan Jing

Matriculation number: 11021764

Supervised by

Prof. Dr. Thomas Setzer

WFI - Ingolstadt School of Management

Submission date:

January 28, 2026

Abstract

Machine learning models have been widely used for medical risk prediction; however, there is a lack of transparency due to the black-box nature of these models. In this paper, Explainable Artificial Intelligence (XAI) methods were used to interpret the predictions made by machine learning models on a clinical and demographic dataset for stroke risk prediction. Various machine learning models such as XGBoost, CatBoost, Random Forest were used and evaluated using SHAP, LIME, Partial Dependence Plots, Individual Conditional Expectation plots, and counterfactual explanations. It was found that age and average glucose level were the most important features across all models, while other features were important in a contextual manner. This paper emphasizes the need to consider interpretability while using machine learning models for medical risk prediction.

Key Words: XAI, Machine Learning, Stroke Prediction

Contents

List of Figures	I
List of Tables	II
1 Introduction	1
2 Objectives and Scope	1
3 Team Contribution	2
4 Data Understanding	3
4.1 Data Source	3
4.2 Feature Type	3
4.3 Descriptive Statistics	4
4.4 Missing Values and Outliers	4
4.4.1 Missing Values	4
4.4.2 Outliers	5
4.5 Initial Insights and Limitations	5
4.5.1 Initial Insights	5
4.5.2 Limitations	5
5 Data Cleaning and Preprocessing	6
5.1 Data Cleaning	6
5.2 Feature Engineering and Transformation	6
5.3 Pre-Modeling Preprocessing (SMOTE)	7
6 Shapley Values	7
6.1 Introduction	7
6.2 SHAP: Methodology and Mathematical Background	8
6.2.1 Shapley Values from Game Theory	8
6.2.2 Mapping Shapley Values to Machine Learning	8
6.2.3 SHAP Additive Explanation Model	8
6.2.4 Desirable Properties of SHAP	9
6.2.5 Efficient Computation for Tree-Based Models	9
6.2.6 Interpretation of SHAP Values	9
6.3 SHAP Implementation in This Project	9
6.3.1 Model and Explainer Selection	9
6.3.2 Global Explanation: Feature Importance	9
6.3.3 Local Explanation: Individual Predictions	10
6.4 Feature Effects and Interaction Analysis	10
6.5 Summary	11
7 Local Interpretable Model-Agnostic Explanations	11
7.1 Introduction	11
7.2 Motivation behind LIME	11
7.3 Math behind LIME	12
7.4 Algorithm of LIME	13
7.4.1 Algorithm	13
7.5 LIME Metrics: Weights and Fidelity	14

7.5.1	LIME Weights	14
7.5.2	Fidelity	14
7.6	Submodular Pick LIME (SP-LIME)	15
7.6.1	SP-LIME Algorithm	15
7.7	Practical Implementation	16
7.7.1	Insights	16
7.7.2	LIME Weights and Fidelity results	16
7.7.3	Model Comparisons and Final Statistics	17
8	Counterfactual Explanations	19
8.1	Understanding Counterfactual Explanations (CE)	19
8.2	Diverse Counterfactual Explanations (DiCE)	19
8.3	Strength and Weakness	24
9	Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE)	25
9.1	Introduction to PDP and ICE	25
9.2	Methodology and Implementation	26
9.3	Global Feature Effects using PDP	26
9.3.1	Effect of Age on Stroke Risk	26
9.3.2	Effect of Average Glucose Level	27
9.3.3	Effect of BMI	27
9.3.4	Clinical Threshold	27
9.4	Relationship with SHAP Analysis	27
9.5	Individual Heterogeneity Revealed by ICE	28
9.6	Model Comparison: Random Forest vs XGBoost	30
9.7	Clinical Implications	30
9.7.1	Three-Tier Risk Stratification Framework	30
9.8	Conclusion	31
10	Analysis and Comparison of Results	32
11	Conclusion	35
	References	36

List of Figures

1	Approximating the tangent to the curve. To understand the shape of the ML function LIME generates points around the red cross (we have an idea of the boundary $f(x)$ thanks to the colors of the generated points). From LIME: explain Machine Learning predictions	13
2	PDP and ICE for Key Risk Factors based on XGBoost (for comparison)	31

List of Tables

1	LIME Feature Contributions for Stroke Prediction (Representative Instance)	17
2	Model Performance Metrics	18
3	Original Patient Profile	20
4	Generated Counterfactuals (Noisy / Counterintuitive Paths) . . .	20
5	Original Patient Profile (High Risk)	21
6	Counterfactuals (Extreme BMI and Glucose Shifts)	21
7	Original Patient Profile (Low Risk / Youth)	21
8	Counterfactuals	21
9	Original Patient Profile	22
10	Counterfactuals	22
11	Original Patient Profile	22
12	Counterfactuals Generated	22
13	Female Subgroup Patterns by Age Range	29
14	Male Subgroup Patterns by Age Range	29
15	Personalized Fasting Glucose Targets by Risk Profile	32
16	Comparative Analysis of XAI Interpretability Frameworks	34

1 Introduction

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. In sectors such as healthcare, where decisions can directly affect human lives, professionals 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 Chapter 4). The dataset is used to predict the likelihood of a stroke, an important health outcome where both predictive accuracy and Interpretability is essential. We will build advanced ML models, i.e., XGBoost, Catboost, and Random Forest, 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, Singh, & Guestrin, 2016) introduced LIME, which provides local, interpretable approximations of complex models, while (Lundberg & Lee, 2017) developed SHAP, a unified framework based on Shapley values, ensuring consistent and theoretically grounded feature attributions. More recent studies, such as (Karimi, Schölkopf, & Valera, 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 Objectives and Scope

The primary objective of this project is to analyze how black-box machine learning models make predictions and determine what factors impact these predictions the most. Although it is important for models to make good predictions, this project is especially concerned with interpretability because it is important that the reasoning behind predictions is understood, especially in areas such as medicine, where decisions have serious consequences.

For this, various Explainable Artificial Intelligence (XAI) methods are used on stroke risk prediction models developed on the Stroke dataset, which is publicly available on Kaggle. The aim here is not only to produce explanations but also to critically examine the validity, reliability, and relevance of those

explanations.

In particular, the following are the objectives of the project:

1. Model Development and Comparison

To create and compare different models of different complexities, such as Random Forest, XGBoost, and CatBoost, for their prediction capabilities on the stroke prediction problem.

2. Local Interpretability Analysis

In order to interpret the predictions made by the models at the level of the individual predictions, techniques such as LIME and SHAP can be used, which would allow us to understand the rationale behind the predictions of certain patients as being at high or low risk.

3. Global Interpretability Analysis

To investigate the overall model behavior employing global explanation techniques, such as SHAP summary plots, SP-LIME, Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) plots, in an attempt to comprehend the effect of the features on the predictions made on the data set as a whole.

4. Counterfactual Explanation and Validation

To create and interpret counterfactual explanations that show how small modifications to patient attributes can lead to different predictions, and to judge the reality and believability of these counterfactual explanations.

5. Comparative Evaluation of xAI Methods

For comparing the pros and cons of various interpretability methodologies (LIME, SHAP, PDP/ICE, and Counterfactual Explanations) in terms of interpretability, stability, computational requirements, and applicability for clinical decision support systems.

The scope of the project is restricted to the explanation methods applied after the training process. This implies that the explanation task is performed without altering the models. This research focuses on structured medical data. This is because unstructured medical data, which includes medical imaging or text, is not considered.

3 Team Contribution

We began with the development of an initial project plan through collaborative efforts among the members and was recorded in the file “Project Initial Outline.” This provided an outline of the tasks to be done during the entire project process on a week-by-week basis. During this process, responsibilities were well distributed among all the members of the group.

To have a clear understanding of the dataset, every team member carried out their own process of data preprocessing. Through this process, different members of the team were able to bring different insights to handling the dataset and preparing the features from this dataset. Subsequently, a consensus of the most efficient methods of preprocessing this dataset was achieved by combining all of these methods in a singular process. All of these individual processes were then combined to give a singular notebook that can be accessed on a shared GitHub repository here. Additionally, the repository contains the LaTeX template used in the creation of this report, thus guaranteeing the availability of all files relating to the project in a single location.

In what followed, version control as well as collaboration was done via GitHub, each of us following the best practices in software development. Each team member worked on their own branch while contributing their parts, with the contributions being reviewed and merged into the main branch. This approach made the integration process easy with minimal conflict. Overall, the level of collaboration in the group was both effective and successful. The members were able to respect each other's responsibilities and communicate effectively throughout the duration of the project. The tasks were shared equally among all members, and everyone was able to contribute equally to the completion of the project with mutual trust and respect.

4 Data Understanding

4.1 Data Source

We chose the healthcare-dataset-stroke-data.csv available freely on Kaggle.

4.2 Feature Type

After preprocessing, the dataset contains 5110 observations and 11 features. The identifier variable `id` was removed as it does not carry predictive information.

The features can be categorized as follows:

- **Categorical features:**

`gender, ever_married, work_type, Residence_type, smoking_status`

These variables describe demographic and lifestyle characteristics. They were converted to the categorical data type for efficient storage and later encoded.

- **Numerical features:**

`age, avg_glucose_level, bmi`

These variables represent continuous medical measurements and are suitable for statistical analysis and modeling.

- **Binary features:**

hypertension, heart_disease, stroke

These variables indicate the presence or absence of medical conditions.
The variable `stroke` is the target variable.

Categorical Mappings Key:

- `gender`: {0: Female, 1: Male}
- `ever_married`: {0: No, 1: Yes}
- `work_type`: {0: Govt_job, 1: Never_worked, 2: Private, 3: Self-employed, 4: children}
- `residence_type`: {0: Rural, 1: Urban}
- `smoking_status`: {0: Unknown, 1: formerly smoked, 2: never smoked, 3: smokes}

4.3 Descriptive Statistics

Descriptive statistics were computed to understand the distributions of numerical variables.

- **Age** ranges from **0.08** to **82** years, with a mean of approximately **43.2** years. Very small values represent infants and are therefore valid observations.
- **Average glucose level** has a mean of **106.1 mg/dL** and a maximum of **271.7 mg/dL**, indicating a right-skewed distribution with high glucose values for some individuals.
- **BMI** has a mean of **28.9**, with values ranging from **10.3** to **97.6**, showing a skewed distribution and potential extreme values.
- The **stroke** variable has a mean of **0.0487**, indicating that stroke cases are relatively rare in the dataset.

Overall, the statistics reveal heterogeneous health conditions and non-normal feature distributions.

4.4 Missing Values and Outliers

4.4.1 Missing Values

Among all variables, only **BMI** contains missing values, with **201** missing entries. Given the skewed distribution of BMI, missing values were imputed using the **median**, which is more robust than the mean.

After imputation, the dataset contains no missing values.

4.4.2 Outliers

Histogram and box plot analyses show that:

- **Age** does not exhibit problematic outliers; extreme values are clinically plausible.
- **Average glucose level** shows a long right tail with several high-end outliers.
- **BMI** also exhibits high-value outliers.

These outliers were **not removed**, as they likely represent real medical conditions rather than data errors.

4.5 Initial Insights and Limitations

4.5.1 Initial Insights

- The dataset is **highly imbalanced**, with stroke cases accounting for less than **5%** of observations.
- Medical factors such as age, glucose level, BMI, hypertension, and heart disease are likely important predictors of stroke.
- Several features capture lifestyle and socioeconomic information, which may contribute additional predictive value.

4.5.2 Limitations

- The strong class imbalance may bias predictive models toward the majority class.
- Some categorical values (e.g., `smoking_status = Unknown`) may reduce interpretability.
- BMI values were imputed, which may introduce uncertainty.
- The dataset is cross-sectional, so causal conclusions cannot be drawn.

5 Data Cleaning and Preprocessing

5.1 Data Cleaning

To ensure data quality and reliability, the following cleaning steps were executed:

- **Handling Unvalidated Values (Age):** An initial investigation revealed that the minimum value for the `age` feature is **0.08**, which appears suspicious. Further analysis confirmed that these values represent children under one year old and are therefore valid observations.
- **Missing Value Imputation (BMI):** The `bmi` feature contained **201** missing values. Analysis of the data distribution showed that `bmi` is right-skewed. Consequently, the **median** value was chosen for imputation instead of the mean to prevent bias from outliers.
- **Outlier Removal:** Exploratory data analysis, including distribution plots and box plots, revealed the presence of outliers in the `bmi` and `avg_glucose_level` features. Since tree-based models were primarily used in this project and such models are generally robust to outliers, these extreme values were initially retained as they may contain useful information for stroke prediction.

However, counterfactual explanations revealed a small number of clinically implausible cases. Two specific types of outliers were identified and removed to improve clinical plausibility and interpretability:

1. **Extremely high BMI values (≥ 50):** A small number of instances exhibited very high BMI values while belonging to the non-stroke class. Given the rarity of such cases and their potential to introduce bias due to class imbalance, these instances were removed.
2. **Extremely low average glucose levels in non-stroke cases:** Some counterfactual explanations suggested that reducing glucose levels would increase stroke risk. This contradicts established medical knowledge and was observed only in a very small subset of patients with unusually low glucose values. These cases were therefore considered spurious outliers and removed.

Overall, outlier removal was applied selectively and guided by both domain knowledge and explainability analysis, rather than purely statistical criteria.

5.2 Feature Engineering and Transformation

Transforming the data into a format suitable for machine learning involved the following operations:

- **Type Casting:** All category type columns (e.g., `gender`, `ever_married`, `work_type`, `Residence_type`, and `smoking_status`) were converted to categorical data types to optimize memory usage and prepare for encoding.
- **Feature Encoding:** Categorical variables were converted into numerical categories for modeling via One Hot Encoding. Special handling was noted for the `smoking_status` category `Unknown`, which remains a distinct value in the dataset.

5.3 Pre-Modeling Preprocessing (SMOTE)

Due to the high class imbalance—where only approximately **5%** of the patients in the dataset had experienced a stroke—synthetic data generation was required.

- **Class Imbalance Resolution:** SMOTE (Synthetic Minority Over-sampling Technique) was employed during the preprocessing phase to oversample the minority class (`stroke = 1`). This ensures that the model learns the characteristics of stroke cases effectively rather than simply predicting the majority class.

6 Shapley Values

6.1 Introduction

Shapley Additive exPlanations (SHAP) was introduced in the paper “A Unified Approach to Interpreting Model Predictions” by Scott M. Lundberg and Su-In Lee in 2017. In some applications of machine learning, interpretability becomes more vital than accuracy. In terms of accuracy, complex machine learning models often perform exceptionally well. However, even experts struggle to interpret complex models such as ensemble or deep learning models. For this reason, SHAP was introduced.

SHAP assigns each feature an importance value for a particular prediction (Lundberg & Lee, 2017).

One major advantage of SHAP is that its additive explanation model follows the property that the sum of all feature contributions equals the prediction difference. This property ensures that SHAP explanations possess local accuracy and consistency, so that feature attributions correctly represent changes in model behavior (Lundberg & Lee, 2017). It is important to note that SHAP explains model behavior rather than establishing causality.

6.2 SHAP: Methodology and Mathematical Background

6.2.1 Shapley Values from Game Theory

In cooperative game theory, a Shapley value represents the fair contribution of a player to the overall outcome of a game. Given a set of players $N = \{1, 2, \dots, n\}$ and a value function $v(S)$ that assigns a payoff to each subset $S \subseteq N$, the Shapley value for player i is defined as:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

This formulation computes the marginal contribution of player i averaged over all possible coalitions S in which i can participate.

6.2.2 Mapping Shapley Values to Machine Learning

In the context of machine learning, each feature is considered a player, and the model prediction corresponds to the game outcome. Let $f(x)$ be the prediction of a trained model for input instance x . The value function $v(S)$ is defined as the expected model output when only the subset of features S is known:

$$v(S) = \mathbb{E}[f(X) \mid X_S = x_S]$$

where X_S represents the values of features in subset S , and the remaining features are marginalized over their empirical distribution.

6.2.3 SHAP Additive Explanation Model

SHAP approximates the model prediction using an additive explanation model:

$$f(x) = \phi_0 + \sum_{i=1}^n \phi_i$$

where:

- $\phi_0 = \mathbb{E}[f(X)]$ is the baseline prediction (expected model output),
- ϕ_i is the SHAP value representing the contribution of feature i .

This ensures that the sum of all feature attributions exactly equals the difference between the instance-specific prediction and the baseline.

6.2.4 Desirable Properties of SHAP

SHAP satisfies several important theoretical properties:

- **Local Accuracy:** The sum of feature contributions equals the model prediction.
- **Consistency:** If a feature's contribution increases in the model, its SHAP value does not decrease.
- **Missingness:** Features that are not present receive zero attribution.

These properties guarantee fair and consistent feature attributions across different models.

6.2.5 Efficient Computation for Tree-Based Models

While computing Shapley values directly is computationally expensive due to the exponential number of feature subsets, SHAP introduces efficient algorithms for specific model classes. For tree-based models such as Random Forests, XGBoost, and CatBoost, the TreeExplainer algorithm computes exact SHAP values in polynomial time by leveraging the structure of decision trees.

6.2.6 Interpretation of SHAP Values

A positive SHAP value indicates that a feature increases the model's prediction relative to the baseline, while a negative value indicates a decreasing effect. Importantly, SHAP values explain model behavior rather than causal relationships, and should be interpreted as contributions to the model's decision process.

6.3 SHAP Implementation in This Project

6.3.1 Model and Explainer Selection

In this project, SHAP was applied to models such as XGBoost and CatBoost, which are classification models. A tree-based explainer was used to explain features. The explainer was initialized using trained models, and SHAP values were then computed on the test dataset. This approach enabled interpretation at both patient and dataset levels without requiring any changes to the models.

6.3.2 Global Explanation: Feature Importance

To inspect model behavior globally, SHAP summary plots were generated. These plots rank features based on their mean absolute SHAP values and also

illustrate the distribution of feature effects.

The results indicated that **age** was the most important factor influencing stroke risk prediction, followed by **average glucose level** and **body mass index (BMI)**. Increased values of age and average glucose level were associated with higher stroke risk, while lower age values corresponded to reduced stroke risk.

Other features, such as smoking status, nature of employment, and gender, exhibited relatively lower and context-dependent effects. Some features appeared to behave as proxy variables for age-related factors, highlighting the importance of understanding feature attribution.

6.3.3 Local Explanation: Individual Predictions

Moreover, SHAP was utilized to explain individual predictions using patient-level visualizations. Waterfall plots were created to illustrate the contribution of individual features to specific predictions (patients numbered 50 and 100). These plots demonstrate the additive nature of positive and negative feature contributions leading to the final model output. The same patient instances will be used in all other xAI methods in this report for ease of comparison.

For high-risk individuals, the major contributors were primarily increased age and elevated glucose levels. In contrast, protective factors contributed to reducing risk for low-risk individuals. This enhanced transparency by clearly illustrating the reasons behind each prediction.

6.4 Feature Effects and Interaction Analysis

To further investigate feature behavior, SHAP dependence plots were employed to examine interactions between feature values and corresponding SHAP contributions. These plots revealed linear effects for both age and average glucose level.

Interaction effects were further explored by visualizing SHAP interaction values. It was observed that the combined effect of age and glucose levels on stroke risk increased in the presence of hypertension, indicating that the model effectively captured interaction relationships.

Additionally, gender-based SHAP scatter plots showed separation between encoded gender categories. The female category exhibited higher SHAP scores than the male category, implying a degree of gender-related contribution to the model output. However, the overall contribution remained relatively small compared to dominant clinical factors such as age and glucose level.

This observation aligns with clinical evidence suggesting that while men tend to experience higher stroke rates at younger and middle ages, women face higher lifetime stroke risk and mortality rates, with approximately one in five women

experiencing a stroke during their lifetime (Yoon & Bushnell, 2023). When combined with other features, gender demonstrated a contextual rather than dominant influence on stroke prediction.

6.5 Summary

This chapter demonstrated the application of SHAP as the primary interpretability technique for stroke prediction models. SHAP provided valuable insights into how complex models generate predictions and confirmed that clinically relevant features such as age and glucose levels played dominant roles.

7 Local Interpretable Model-Agnostic Explanations

7.1 Introduction

Local Interpretable Model-Agnostic Explanations (LIME) was first introduced in the University of Washington paper “Why Should I Trust You: Explaining the Predictions of Any Classifier”.

LIME aims to explain any black box model by creating local approximations of a complex decision boundary function, relying on the idea that even if a model is complex globally (most Deep Neural Networks), it is often simple locally. The internals of the model predictions are hidden from LIME, only being able to interpret the input and outputs of a model. LIME works with many data types i.e., text, tabular data, images, or even graph curves. Thus LIME:

- is Post Hoc (Explains decisions after a model has made its predictions)
- is Locally Interpretable
- is Model Agnostic (works with any kind of ML model)

7.2 Motivation behind LIME

As mentioned earlier in Chapter 1, Machine Learning models have been excellent in aiding medical decision making. However, research supports that humans do not easily rely on a model’s decision without some prior reasonable trust in it. (J. L. Herlocker & Riedl, 2000)

Thus, as is the case with other xAI methods, the recent research boom in the field of xAI methods stems from humility on our end; from humans understanding but also appreciating the limitations of machine learning models. Instead of just discarding an entire model because of a few wrong predictions

(which of course may be fatal), LIME, for instance, provides us a lens to zoom in into decision boundaries, tweak variables, and understand exactly what variables influences the most/least towards a prediction

7.3 Math behind LIME

The core of LIME is defined by the the following equation:

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

The equation defines the optimal explanation, denoted as $\xi(x)$, as the result of a minimization problem. It seeks a model g (from a family of interpretable models G) that minimizes two competing objectives: the local unfaithfulness \mathcal{L} and the model complexity Ω .

$\xi(x)$ represents the final explanation for the specific instance x . It is the specific interpretable model (e.g., a specific linear regression equation with defined weights) that best satisfies the criteria of fidelity and simplicity.

x is the specific data instance being explained (e.g., a single patient's medical record). It serves as the center of the local neighborhood where the explanation is valid.

The $\operatorname{argmin}_{g \in G}$ operator indicates that the goal is to search through all possible interpretable models G (such as linear models or decision trees) to find the single model g that results in the lowest combined value of loss and complexity.

Now as for the two minimizers in the equation:

$\mathcal{L}(f, g, \pi_x)$ is defined the fidelity loss function. It measures the error between the complex black-box model f and the simple surrogate model g . Crucially, this error is weighted by π_x , a proximity measure that assigns high importance to data points close to x and low importance to those far away, ensuring the explanation is locally accurate.

The simple surrogate model g that is generated is a sparse linear model. An advantage of this being that g being sparse automatically takes care of minimizing the second loss term i.e. $\Omega(g)$. Note that the aim of a sparse linear model is to produce as many zero weights as possible. In practice, this is accomplished via regularization techniques such as done for Lasso Regularization. This way we ensure to get a simple explanation, with only a few relevant variable. This is the super power behind LIME.

$\Omega(g)$ is the complexity penalty. It quantifies how complicated the explanation model g is (e.g., the number of non-zero coefficients in a linear model). Minimizing this term ensures the explanation remains simple enough for human interpretation.

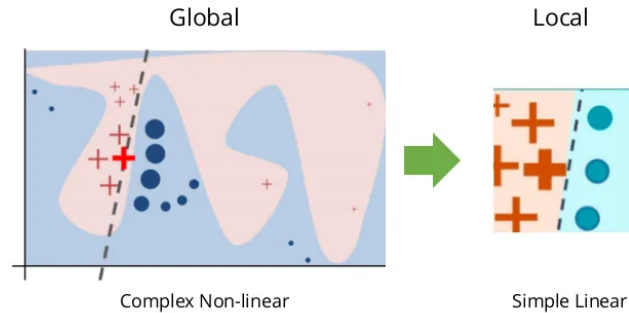


Figure 1: Approximating the tangent to the curve. To understand the shape of the ML function LIME generates points around the red cross (we have an idea of the boundary $f(x)$ thanks to the colors of the generated points). From LIME: explain Machine Learning predictions

7.4 Algorithm of LIME

LIME does not solve anything analytically, rather it works by approximating the solution (as seen in the last section) via sampling. The core idea here is the perturbation (small changes) LIME adds to the instance's features (see Figure 1). It then uses this perturbed instance and queries the black box model. The algorithm is roughly described below with an example.

7.4.1 Algorithm

Input: Patient X with features: Age 72, Smoker Yes, BMI 32, Glucose 210.

1. **Segmentation:** Treat each feature (Age, Smoker status, BMI, Glucose) as an interpretable component.
2. **Perturbation:** Create 5,000 "fake" patient profiles by randomly changing feature values (e.g., changing Smoker to "No" or Glucose to 180).
3. **Black Box Query:** Ask the Model: "What is the probability of stroke for this patient?" for all 5,000 variations.
4. **Weighting:** Give higher importance (weight) to fake profiles that are numerically closest to the original Patient X (e.g., a profile with Age 71 is weighted higher than Age 40).
5. **Linear Fit:** Fit a weighted linear regression to predict the Model's stroke probability based on the varied feature values.

Output: The features with the highest coefficients (e.g., Glucose Level) are highlighted as the primary reasons for the high-risk prediction.

7.5 LIME Metrics: Weights and Fidelity

7.5.1 LIME Weights

The average absolute LIME weight indicates how strongly a feature influences the model’s predictions across multiple local explanations. The absolute value is taken since the magnitude of influence is of interest regardless of direction. Averaging across instances allows identification of features that are typically most impactful at the local level.

In our analysis, both Random Forest and XGBoost models consistently identified age and average glucose level as the two most influential features. This aligns with findings in existing medical literature and supports the clinical relevance of the model predictions (J. L. Herlocker & Riedl, 2000). These results contribute to building trust in the model.

However, prior hypertension and heart disease exhibited minimal influence (less than 0.1) on the predictions. It should be noted that LIME weights are local in nature and depend on feature binning strategies. Therefore, these values should be interpreted in conjunction with fidelity measures (see below).

Mathematically, let $w(i, f)$ denote the LIME weight of feature f for the instance i . The average absolute weight for feature f is defined as:

$$\text{AvgAbs}(f) = \frac{1}{N_f} \sum_{i \in I_f} |w(i, f)|$$

where I_f is the set of instances in which feature f appears in the top- k LIME explanations and $N_f = |I_f|$.

7.5.2 Fidelity

Fidelity is a trust measure for local explanations. It tells us how well the LIME surrogate model represents the original model’s outputs on the samples. A high fidelity (i.e., close to 1) indicates that the surrogate matches the original model fairly well, and vice versa.

With a high fidelity (i.e., > 0.6), we can trust both the direction and the relative magnitude of the LIME weights. A moderate fidelity (i.e., 0.3–0.6) suggests that LIME should be treated mainly as providing directional insight, indicating which features push the predictions up or down. As for low fidelity (i.e., < 0.3), we should realize that the local linear surrogate does not do a very good job of fitting the original model.

Mathematically, Fidelity is formulated as weighted R^2 for probabilities, defined as following:

Let y_i be the original model's probability on sample i , \hat{y}_i the surrogate's prediction, w_i the kernel weight (where closer samples receive higher weight), and \bar{y} the weighted mean of y_i . Then:

$$R^2 = 1 - \frac{\sum_i w_i (y_i - \hat{y}_i)^2}{\sum_i w_i (y_i - \bar{y})^2} \quad (1)$$

7.6 Submodular Pick LIME (SP-LIME)

In the same first paper on LIME, an extension framework of LIME was introduced as well. It is explained as *"a method that selects a set of representative instances with explanations to address the "trusting the model" problem, via submodular optimization."* (Ribeiro et al., 2016)

SP-LIME tries to generalize the behavior of the model, helping us on answering questions like what features does this model generally rely on. To trust the model globally, we can not just look at one explanation as LIME does. SP-LIME instead, adapts by using the Greedy Algorithm to find the perfect set of instances whose explanations cover the most important features. Note that mathematically this is an NP-hard problem, however, since the problem is submodular (adding more of one input has a decreasing additional benefit), a simple Greedy Algorithm is mathematically guaranteed to be near optimal. Further note, however, that this makes SP-LIME nevertheless easy computationally expensive (Decrypting your Machine Learning model using LIME).

7.6.1 SP-LIME Algorithm

The algorithm operates in three main phases.

1. It first runs the standard LIME "explain" function on every instance in the dataset (X) to create a matrix of feature weights (\mathcal{W}).
2. It then computes a global importance score (I_j) for every feature by aggregating how strongly that feature appears across all the generated explanations.
3. Greedy Selection: It iteratively picks instances one by one to add to the set V . In each step, it chooses the specific instance that maximizes the coverage function c —meaning it picks the instance that best explains important features that have not yet been covered by the instances already picked. (Ribeiro et al., 2016)

Algorithm 1 Submodular pick (SP) algorithm

Require: Instances X , Budget B

```
1: for all  $x_i \in X$  do
2:    $\mathcal{W}_i \leftarrow \text{explain}(x_i, x'_i)$ 
3: end for
4: for all  $j \in \{1 \dots d'\}$  do
5:    $I_j \leftarrow \sqrt{\sum_{i=1}^n |\mathcal{W}_{ij}|}$  ▷ Compute feature importances
6: end for
7:  $V \leftarrow \{\}$ 
8: while  $|V| < B$  do ▷ Greedy optimization
9:    $V \leftarrow V \cup \text{argmax}_i c(V \cup \{i\}, \mathcal{W}, I)$ 
10: end while
11: return  $V$ 
```

7.7 Practical Implementation

7.7.1 Insights

The LIME analysis revealed several critical insights about the model's decision-making process:

- **Age as a Dominant Factor:** Consistently, Age appears as a top contributing factor. Higher age strongly pushes the prediction toward "Stroke" risk, while younger age pushes toward "No Stroke."
- **Glucose Levels:** High average glucose levels were identified as a significant risk factor, positively correlating with stroke prediction.
- **BMI Ambiguity:** Interestingly, the analysis showed that while BMI is a factor, its impact was sometimes less straightforward or lower in magnitude compared to age and glucose, highlighting potentially complex non-linear relationships or data noise handled by the Random Forest.
- **Confirmation of Medical Knowledge:** The model's reliance on Age, Glucose, and Heart Disease aligns well with established medical literature, validating that the model is learning relevant biological signals rather than spurious correlations.
- **Local vs. Global** While the global metrics (accuracy/AUC) were high, LIME showed that for *specific* borderline patients, the model could be swayed by factors like "Residence Type" or "Work Type," which might be less medically causal but statistically correlated in this specific dataset.

7.7.2 LIME Weights and Fidelity results

LIME Weights

The analysis of average absolute LIME weights revealed that **age** and **average glucose level** emerge as the two most influential features in predicting stroke risk across both Random Forest and XGBoost models (see Table 1). This alignment with scientific literature significantly enhances the trustworthiness of both models and validates their clinical relevance. However, a notable concern is the surprisingly low influence of prior hypertension and heart disease history (weights < 0.1), which are traditionally considered important stroke risk factors. This discrepancy underscores the importance of contextualizing LIME weights within their local nature—these weights are binning-dependent and should be interpreted in conjunction with fidelity scores rather than in isolation.

Table 1: LIME Feature Contributions for Stroke Prediction (Representative Instance)

Feature	Condition / Value	Weight (Contribution)
Age	> 61.00	+0.28
Avg. Glucose Level	> 180.50	+0.15
BMI	$30.00 < \text{BMI} \leq 35.00$	+0.08
Heart Disease	1 (True)	+0.05
Work Type	Private	-0.02
Residence Type	Urban	-0.01

LIME Fidelity:

LIME fidelity, as mentioned in Equation 1, measures the local surrogate R^2 score, revealing critical differences in the interpretability of the two models. From the two models used as black boxes, the Random Forest model exhibited a clustering pattern with higher fidelity scores (R^2 of 0.5-0.7) for many predictions, particularly in the low-risk probability range, indicating that LIME’s linear surrogate effectively captures RF’s decision-making in these regions.

Conversely, XGBoost showed more dispersed and generally lower fidelity scores across the probability spectrum, suggesting that XGBoost’s decision boundaries are more nonlinear and complex, making them harder to approximate with LIME’s linear local surrogate model. This finding is crucial since LIME provides reliable feature weight magnitudes and directions only in high-fidelity regions ($R^2 > 0.6$), while in moderate-fidelity zones (0.3-0.6), LIME should be treated as a directional guide for understanding feature influence. In low-fidelity regions (< 0.3), LIME weights should be treated with caution and require cross-validation with clinical expertise.

7.7.3 Model Comparisons and Final Statistics

In this section, we list further performance metrics, namely, predicted stroke probabilities and LIME surrogate fidelity scores for our Random Forest and XGBoost models for each test instance.

Model	Proba-Mean	Proba-Median	Proba-Std	Lime Fidelity Mean	Lime Fidelity Median
Random Forest	0.375	0.246	0.293	0.343	0.229
XGBoost	0.429	0.285	0.418	0.289	0.161

Table 2: Model Performance Metrics

Interpreting the results

- Stroke probability (original model outputs)
 - XGBoost shows higher mean/median stroke probability than Random Forest. This essentially tells us XGBoost is more risk-sensitive and will flag more patients as likely stroke. This is good! Recall over Precision is best in medical fields
 - XGBoost’s probabilities are also more spread out (higher variability), while Random Forest looks more conservative/stable. Expect XGBoost to boost recall but potentially lower precision.
- LIME fidelity (how well the local surrogate mimics the model)
 - Average fidelity is modest (roughly in the 0.27–0.33 range, while the median is lower). Normal for fast, lightweight settings and reminds us LIME is an approximation.
 - Use LIME primarily for the direction and main drivers (which features push toward Stroke vs No Stroke). Avoid over-interpreting tiny weight differences.
 - We are aware that fidelity can be improved by adjusting kernel width, or allowing more features in the local surrogate, but as far as analysis has to go, this suffices + kept it small for speed

The mean/median of Random Forest and XGBoost probability shows how conservative or aggressive each model is. Higher averages imply the model is assigning more cases near the decision threshold.

- LIME fidelity:
 - Higher average fidelity implies explanations more faithfully capture local behavior. Low-fidelity instances deserve caution; the local rules may be unstable or sensitive to perturbations.
 - In our case, RF has a higher fidelity score implying that its local explanations are more reliable for decision support.
- How to use these results?

- We should prefer the model that achieves acceptable probability calibration and higher average LIME fidelity in cohorts of interest. If RF is more faithful locally for older patients, doctors should use RF-driven explanations for that subgroup.
- Having low fidelity does not mean we should discard the LIME approximation altogether; rather we should flag low fidelity instances for additional review. For these cases, we should avoid relying solely on local rules; consider global checks or alternative explanation methods (SHAP etc.)

8 Counterfactual Explanations

In Chapter 8, we will explore **Counterfactual Explanations**, a powerful model-agnostic technique in explainable AI (xAI) that provides human-centric insights into machine learning predictions. This chapter covers the theoretical foundations, the specific implementation using the **DiCE (Diverse Counterfactual Explanations)** framework, and the application of these methods to our stroke prediction analysis.

8.1 Understanding Counterfactual Explanations (CE)

Counterfactual explanations answer “what-if” questions. Unlike SHAP or LIME, which explain why a model produced a given prediction, counterfactuals describe the minimal changes required to input features to achieve a desired outcome.

For example, if a model predicts a high stroke risk for a patient, a counterfactual might state:

If the patient’s average glucose level were reduced from 220 mg/dL to 100 mg/dL, the predicted stroke risk would drop to low.

Such explanations provide actionable insights for stakeholders such as clinicians and patients.

8.2 Diverse Counterfactual Explanations (DiCE)

DiCE is a library developed by Microsoft that specializes in generating a set of diverse counterfactual examples. The “diversity” aspect is crucial because there is rarely just one way to change an outcome. DiCE attempts to find multiple paths to the desired prediction (e.g., changing BMI versus changing smoking status) while ensuring that the suggested changes are:

1. Proximity

- DiCE tries to make counterfactuals close to the original input, so the suggested changes are minimal and realistic.
- Example: If your BMI is 30, suggesting a BMI of 18 might technically change the outcome but is too drastic. DiCE prefers smaller, feasible changes.

2. Feasibility

- DiCE ensures counterfactuals make sense according to the data.
- For example, it will not suggest “age = -5” or “married = yes AND single = yes.”
- This is achieved by respecting constraints and distributions in the training data.

3. Flexibility

- DiCE can be used with any predictive model and thus is model agnostic (tree-based, neural networks, etc.) since it relies on model outputs rather than internal structures.
- It supports continuous, categorical, and ordinal features.

Counterfactual Analysis Results

CEs 1

Table 3: Original Patient Profile

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Female	43	No	No	Yes	Private	Urban	86.67	33.3	never smoked	0

Table 4: Generated Counterfactuals (Noisy / Counterintuitive Paths)

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Female	43	0	0	Yes	Private	Urban	110.87	33.3	formerly smoked	1
Female	43	0	0	Yes	Private	Urban	69.48	33.3	never smoked	1
Female	43	0	0	Yes	Private	Urban	60.55	33.3	never smoked	1

Actionable Insight: The model produces counterfactuals where **lower glucose values increase stroke risk**, which is clinically implausible. This indicates the presence of **spurious correlations** learned from rare or noisy data points. These instances with low glucose in the non-stroke cases were removed.

CEs 2

Table 5: Original Patient Profile (High Risk)

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Male	61.0	1	1	Yes	Govt_job	Rural	86.06	34.8	never smoked	1

Table 6: Counterfactuals (Extreme BMI and Glucose Shifts)

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Male	61.0	1	1	Yes	children	Rural	86.06	96.1	never smoked	0
Male	32.9	1	1	Yes	Govt_job	Rural	86.06	34.8	never smoked	0
Male	61.0	1	1	Yes	Govt_job	Rural	67.00	34.8	never smoked	0

Actionable Insight: The model produces counterfactuals in which extreme changes in BMI disproportionately increase stroke risk, despite these scenarios being clinically unrealistic. This suggests the presence of spurious patterns learned from rare or noisy observations. Consequently, implausible BMI instances were removed to ensure that counterfactual explanations reflect realistic patient conditions, thereby improving model reliability and supporting meaningful clinical interpretation.

CEs 3: Reliable Interventions After Outlier Removal

Table 7: Original Patient Profile (Low Risk / Youth)

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Male	17	No	No	No	Govt_job	Urban	68.91	23.0	Unknown	0

Table 8: Counterfactuals

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Male	76.6	0	0	No	Govt_job	Urban	222.16	23.0	Unknown	1
Male	57.1	0	0	No	Govt_job	Urban	179.49	23.0	Unknown	1
Male	65.8	0	0	No	Govt_job	Urban	227.21	23.0	Unknown	1

Actionable Insight: By removing outliers from the dataset, the counterfactuals now reflect plausible and medically meaningful scenarios. This suggests that careful preprocessing, such as outlier removal, is essential for trustworthy counterfactual explanations in healthcare models. Clinically, this allows doctors to provide specific guidance to patients. For example: older patients with poorly controlled glucose levels are at significantly higher risk of stroke. By targeting these modifiable risk factors—through weight management, diet, exercise, or glucose control—clinicians can help reduce an individual patient’s likelihood of experiencing a stroke.

CEs 4

Table 9: Original Patient Profile

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Female	5	No	No	No	children	Rural	102.04	18.5	Unknown	0

Table 10: Counterfactuals

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Female	49.9	0	0	No	Private	Rural	102.04	18.5	Unknown	1
Female	74.4	0	0	No	Never_worked	Rural	102.04	18.5	Unknown	1
Female	76.3	0	0	No	Govt_job	Rural	102.04	18.5	Unknown	1

Actionable Insight: Counterfactual analysis highlights that non-traditional risk factors such as age-related stress and occupational exposure can strongly influence predicted stroke risk. Proper handling of stress and lifestyle factors is crucial, especially for individuals in high-risk age groups.

CEs 5

Table 11: Original Patient Profile

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Female	69.0	No	Yes	No	Govt_job	Urban	202.38	34.6	Unknown	1

Table 12: Counterfactuals Generated

Gender	Age	Hypertension	Heart Disease	Married	Work Type	Residence	Avg Glucose	BMI	Smoking Status	Target
Female	50.0	0	1	Yes	Private	Urban	193.80	26.4	never smoked	0
Female	52.0	0	0	Yes	Private	Rural	200.46	25.0	Unknown	0
Female	53.0	1	1	Yes	Private	Urban	196.25	24.9	smokes	0

Counterfactual analysis suggests that stroke risk is primarily influenced by BMI and lifestyle-related factors, rather than glucose levels alone.

Clinically actionable recommendations include:

- Weight management to reduce BMI from obese to healthy ranges.
- Stress-aware lifestyle interventions, particularly addressing occupational pressures and work-related stress.

Model caution: Some counterfactuals assign lower stroke risk to patients with active heart disease. This likely reflects a data skew, where the majority of non-stroke cases do not have heart disease. Consequently, the model may have captured spurious correlations influenced by data imbalance or rare patterns, leading to counterintuitive risk associations.”

Interestingly, the analysis also suggests that being married may correlate with lower stroke risk, suggesting a potential association between marital status and lower predicted risk, possibly reflecting social support or unobserved socioeconomic factors.

Summary of Counterfactual Explanations (CEs)

From analyzing multiple patient profiles and their counterfactuals, several clear patterns emerge regarding how the model predicts stroke risk:

1. Age dominates the model predictions

- Across nearly all CEs, **age is the most influential feature**.
- While this drives target changes effectively, it is **not clinically actionable**, as we cannot modify a patient’s age.

2. Metabolic factors (BMI and average glucose) are predictive when age is fixed

- When age is held constant in counterfactual generation, the model often adjusts **BMI** or **average glucose** to flip predictions.
- Keeping these features within **normal or lower ranges** decreases the predicted stroke risk.
- These are **clinically actionable interventions**, such as weight management and glucose control.

3. Stress-related or lifestyle factors

- Some CEs suggest that changing **work type** (e.g., Self-employed → Private) or **residence type** can reduce stroke risk.
- This aligns with real-world evidence: **stress and lifestyle influence cardiovascular health**.

4. Marriage appears protective

- Interestingly, many CEs show that **being married** slightly lowers predicted stroke risk.
- A “fun insight”: **the power of love may be real**, at least according to the model’s learned patterns.

5. Challenges with actionable feature constraints

- If **age and gender are fixed** (to focus on modifiable features), DiCE must rely on other variables like **BMI, glucose, or smoking** to flip the prediction.
- If these features have **low feature importance**, the Dice may **fail to generate a feasible counterfactual**, resulting in an **under-constrained problem**.

6. Detecting outliers

- CEs are powerful for identifying **rare outliers or extreme feature combinations** that may not be obvious in standard analysis.
- This helps uncover **data issues or rare risk patterns** that could be clinically relevant.

Key Takeaways

- **The main actionable insights from CEs are:**
 1. Maintain healthy BMI
 2. Control glucose levels
 3. Address stress-related lifestyle factors (work, residence, daily habits)
- **Immutable features** like age and gender should generally be fixed to focus on clinically relevant interventions.
- In small datasets where **age is extremely predictive**, the DiCE may struggle to find feasible counterfactuals if age is fixed.
- Overall, CEs provide a **unique window into the model’s learned patterns**, offering both **actionable clinical guidance** and the ability to **detect outliers or unusual cases**.

Conclusion: Counterfactual explanations transform complex model behavior into **practical insights**, highlighting which features can be realistically modified to reduce stroke risk, while also revealing hidden patterns like stress, marriage, or outlier profiles.

8.3 Strength and Weakness

Strengths of Counterfactual Explanations (CEs)

1. **Actionable Insights:** CEs show how small changes in input features can alter model predictions, making them highly useful in domains such as healthcare for risk mitigation strategies.
2. **Model-Agnostic:** Most CE methods, including DiCE, can be applied to any predictive model, whether tree-based, ensemble, or neural network models.

3. **Local Interpretability:** They provide explanations at the individual prediction level, supporting human decision-making.
4. **Bias and Plausibility Diagnostics:** Counterfactuals can reveal unrealistic feature changes or spurious correlations, helping diagnose dataset bias or modeling artifacts.

Weaknesses of Counterfactual Explanations

1. **Plausibility Issues:** CEs may suggest implausible or unrealistic scenarios.
2. **Sensitivity to Data Bias:** Skewed datasets can lead to spurious correlations, causing misleading counterfactuals (e.g., suggesting that heart disease or hypertension lowers stroke risk).
3. **Computational Complexity:** Generating multiple counterfactuals can be computationally expensive, especially for large datasets or complex models.
4. **The “Rashomon Effect” (Multiple Truths):** There is rarely a single correct counterfactual, which can create user confusion and cognitive overload if not carefully constrained. Choosing the “best” option is often subjective. However, in our context, this diversity is an advantage, as it allows doctors and patients to select interventions that best fit individual lifestyles and preferences, supporting personalized and actionable care.
5. Only provides **local**, not **global**, explanations.
6. Explains **how** to change the outcome, not **why** the model predicts it.
7. Limited library support compared to SHAP and LIME.

9 Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE)

9.1 Introduction to PDP and ICE

In Chapter 6, SHAP identified age, average glucose level, and body mass index (BMI) as the dominant contributors to stroke risk predictions. While SHAP explains feature attributions for individual outcomes, it does not directly characterize how variations in these features influence model predictions across their ranges.

To complement SHAP, this chapter applies Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots to examine marginal feature effects and individual-level heterogeneity.

PDP visualizes the average effect of selected features on the predicted response, revealing linear, monotonic, and nonlinear relationships (Molnar, 2019). ICE plots further disaggregate these averages by displaying individual prediction trajectories, with PDP representing their mean (Molnar, 2019).

A hierarchical analysis was conducted, progressing from global PDP trends to ICE-based heterogeneity detection, followed by subgroup and gender-age interaction analyses within the obese population ($\text{BMI} \geq 30$).

9.2 Methodology and Implementation

The analysis was conducted on the test dataset ($n = 1,001$). The main features examined were the following, quite similar to our results from LIME and SHAP methodologies:

- Age
- Average glucose level
- BMI

Two models were interpreted:

- Random Forest (primary interpretation model)
- XGBoost (comparison)

Key implementation parameters included a grid resolution of 50 and subsampling for ICE visualization. Subgroup analyses were conducted only when sample size exceeded 20 to ensure reliability.

The interpretability workflow followed four levels:

1. Global PDP analysis
2. ICE-based heterogeneity detection
3. Subgroup PDP/ICE (high BMI population)
4. Gender \times age stratification

9.3 Global Feature Effects using PDP

9.3.1 Effect of Age on Stroke Risk

The PDP for age exhibited a nonlinear S-shaped curve:

- Low and stable risk before age 30
- Gradual increase from 30 to 50
- Rapid acceleration between 50 and 70
- Plateau beyond 70

Interpretation: Age 50 emerged as a critical inflection point where stroke risk begins to rise sharply, consistent with epidemiological findings.

9.3.2 Effect of Average Glucose Level

The glucose PDP showed a monotonic increase:

- Steep rise from normal to prediabetic range
- Continued increase into diabetic range

Even moderate glucose elevations significantly increased predicted risk, indicating no clear “safe threshold.”

9.3.3 Effect of BMI

Risk increased gradually in normal and overweight ranges and rose sharply once BMI exceeded 30.

9.3.4 Clinical Threshold

Obesity ($\text{BMI} \geq 30$) represented a major risk inflection point.

9.4 Relationship with SHAP Analysis

SHAP analysis in Chapter 6 identified:

1. Age
2. Average glucose level
3. BMI

as the most important predictors. PDP effect ranges confirmed the same ranking, validating global consistency.

SHAP	PDP/ICE
Explains feature contribution	Shows functional effect
Individual attribution	Marginal response curves
Captures interactions	Reveals thresholds and nonlinearities

Methodological Complementarity SHAP answers *why* predictions are high, while PDP/ICE explains *how* changes in features alter risk and *for whom* effects differ.

PDP further refined insights by identifying:

- Age 50 as a nonlinear turning point
- Continuous glucose sensitivity
- Obesity as a structural risk threshold

These patterns are not directly visible in SHAP values.

9.5 Individual Heterogeneity Revealed by ICE

To explore individual-level variability beyond average effects captured by PDP, ICE plots were analyzed within clinically relevant subgroups, with particular focus on obese patients ($\text{BMI} \geq 30$), who constituted approximately 40% of the test set.

Gender Differences in the High BMI Group

Among obese patients, clear gender-specific patterns emerged.

Female patients exhibited:

- High baseline predicted stroke risk
- Relatively weak sensitivity to glucose variation
- ICE curves closely clustered around the PDP

This indicates a relatively homogeneous risk response, where glucose plays a limited but consistent role. As for the general higher baseline, this is supported by scientific literature in this field. This enhances interpretability. (Yoon & Bushnell, 2023)

Male patients showed:

- Similar average predicted risk levels

- Extremely dispersed ICE curves
- Substantial variation in glucose impact across individuals

For some males, glucose strongly increased risk, while for others the effect was minimal or nonlinear.

Table 13: Female Subgroup Patterns by Age Range

Age Range	Sample Size	PDP Pattern	ICE Heterogeneity	Clinical Significance
< 40 years	63	Non-monotonic (peak at 80-85)	High (0.10-0.30)	May reflect subgroup-specific variability or unobserved confounding factors
40-60 years	104	High plateau (0.50)	Extreme (0.30-0.85)	Elevated risk patterns observed in this age group may be influenced by multiple physiological and lifestyle factors
≥ 60 years	70	Steady rise (0.70→0.80)	Low, curves concentrated	Elderly obese females show predictable response

Table 14: Male Subgroup Patterns by Age Range

Age Range	Sample Size	PDP Pattern	ICE Heterogeneity	Clinical Significance
< 40 years	46	Low-risk U-shape (0.15-0.18)	Moderate	Young obese males relatively protected
40-60 years	69	Flat PDP (0.50-0.55)	Massive (0.20-0.80)	Glucose non-predictive, need personalized assessment
≥ 60 years	49	Rising (0.70→0.82)	Low, consistent upward	Converges with females, age dominates risk

Interpretation: Obese male patients demonstrated strong heterogeneity, suggesting that glucose is not a uniform risk driver within this subgroup. Individual metabolic responses varied widely, highlighting the limitation of relying solely on average trends.

Age-Stratified Heterogeneity Patterns

Further stratification by age revealed dynamic interactions between age, gender, and glucose sensitivity.

Under 40 years:

- Generally low baseline stroke risk
- Female ICE curves showed irregular glucose sensitivity

Age 40–60:

- Females exhibited consistently elevated baseline risk
- Males displayed maximal heterogeneity in glucose response

Above 60 years:

- Risk levels converged across genders
- Age became the dominant determinant, overshadowing glucose and BMI effects

Key Insight: Gender modifies risk most strongly in middle age, while in older populations risk becomes increasingly age-driven and homogeneous.

9.6 Model Comparison: Random Forest vs XGBoost

Interpretability stability was assessed by comparing PDP and ICE outputs from Random Forest (RF) and XGBoost models.

Random Forest

- Smooth, continuous PDP curves
- Coherent and structured ICE patterns
- Clinically plausible relationships between features and risk

XGBoost

- Jagged and irregular PDP behavior
- Highly unstable ICE trajectories
- Implausible peaks and sudden reversals

Conclusion:

Random Forest produced more reliable and clinically interpretable explanations. XGBoost appeared sensitive to noise and prone to overfitting in local explanations, reducing trust in its interpretability outputs despite strong predictive performance.

9.7 Clinical Implications

The combined PDP and ICE analysis enabled a structured risk stratification framework.

9.7.1 Three-Tier Risk Stratification Framework

Tier 1: Age-Dominant Risk (≥ 60 years + BMI ≥ 30)

- **Characteristics:** High baseline risk (0.70–0.80) regardless of gender
- **Glucose target:** < 100 mg/dL

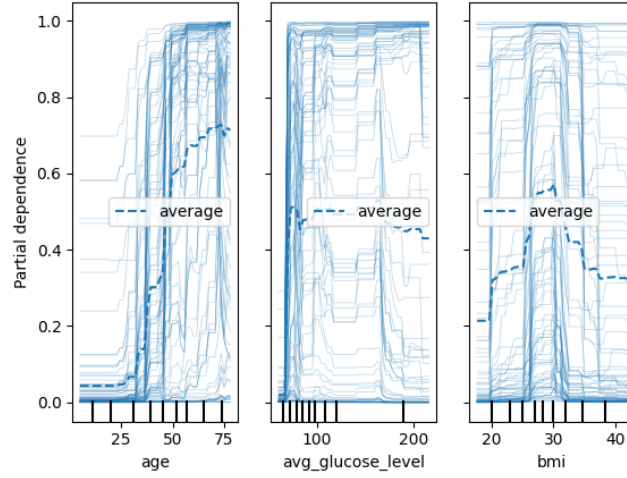


Figure 2: PDP and ICE for Key Risk Factors based on XGBoost (for comparison)

- **Potential focus:** Enhanced glucose monitoring and risk communication strategies. These findings are intended to support interpretability rather than direct clinical treatment decisions.

Tier 2: Gender-Modified Risk (40–60 years + BMI \geq 30)

- **Female subtype:** Moderate-high baseline (0.50), modest glucose effect
 - **Potential focus:** Closer metabolic monitoring and further clinical assessment.
- **Male subtype:** Highly variable risk (0.20–0.80), glucose non-predictive
 - **Potential focus:** Individualized risk assessment considering multiple cardiovascular indicators.

Tier 3: Low-Risk Surveillance (< 40 years)

- **Characteristics:** Low baseline (0.10–0.20)
- **Caution:** Young females with glucose 70–90 + BMI \geq 30 (possible PCOS)
- **Focus:** Lifestyle modification, monitor metabolic syndrome progression

9.8 Conclusion

PDP and ICE effectively transformed black-box model predictions into clinically meaningful insights.

Table 15: Personalized Fasting Glucose Targets by Risk Profile

Patient Profile	Target Fasting Glucose	Rationale
Elderly obese (≥ 60 , BMI ≥ 30)	< 100 mg/dL	Steep PDP slope above 100
Middle-aged obese female	< 110 mg/dL	Moderate effect, balance with QoL
Middle-aged obese male	< 120 mg/dL	Weak glucose effect, avoid overtreatment
Young adults	< 126 mg/dL	Standard prediabetes threshold

Key contributions include:

- Confirmation of age, glucose level, and BMI as primary stroke risk drivers
- Identification of nonlinear thresholds and inflection behaviors
- Revelation of strong heterogeneity among obese middle-aged males
- Demonstration of Random Forest’s superior interpretability stability

When combined with SHAP:

- SHAP identifies feature contributions to individual predictions
- PDP and ICE illustrate effect directions, nonlinear trends, and variability

Together, they form a comprehensive explainable AI framework for health-care applications.

Overall Insight Interpretability is essential for trustworthy medical machine learning. PDP and ICE bridge predictive performance and clinical decision-making by uncovering global patterns, nonlinear effects, and patient-specific variability.

10 Analysis and Comparison of Results

This chapter discusses the comparison of the prediction models and the explainability methods used during the project. The ensemble models XGBoost and CatBoost were more accurate than the Random Forest model, and the complexity of the models was addressed through the explainability methods.

Among the explainability methods used during the project, SHAP was the most comprehensive method for obtaining explanations. This is due to the fact that it provides both global and local explanations with theoretical guarantees. SHAP and LIME were able to consistently identify the most important features for the prediction models. This is evident in the fact that they consistently identified age and average glucose level as the most important features. This is consistent with the knowledge that these two features are the most important

risk factors for the disease. (Feigin, Norrving, & Mensah, 2017) (Benjamin et al., 2019)

Counterfactual explanations provided actionable insights into the model’s behavior by identifying how predictions could be changed. Unlike other methods that only highlight important features, CEs show specific alterations in input variables that would lead to different outcomes, allowing clinicians to understand what factors could increase or decrease risk. For example, CEs can reveal patient characteristics that, if modified, could potentially prevent a high-risk prediction. While many of these explanations captured realistic scenarios, some were clinically implausible, highlighting the need to filter out extreme or unlikely cases. Despite this limitation, CEs are valuable because they offer a direct path from model predictions to potential interventions, making them particularly useful for decision support in healthcare settings.

The following table summarizes the key characteristics, strengths, and limitations of the interpretability methods evaluated in this project.

Table 16: Comparative Analysis of XAI Interpretability Frameworks

Method	Local	Global	Key Strengths	Primary Limitations
SHAP	✓	✓	Theoretically grounded in game theory; provides consistent and additive feature attributions.	High computational cost for complex models; assumes feature independence.
LIME	✓	–	Model-agnostic and computationally efficient for individual instance explanations.	Explanation instability due to local sampling; lacks global consistency.
SP-LIME	✓	✓	Selects a diverse, representative set of instances to provide a global overview of model behavior.	Computationally expensive (requires running LIME on many instances); inherits LIME’s stability issues.
CEs	✓	–	Highly actionable; identifies minimal changes needed to flip a prediction and detects data outliers.	May suggest clinically unrealistic scenarios; explains ‘how’ to change rather than ‘why’.
PDP	–	✓	Visualizes average marginal effects and general model trends for sanity checks.	Cannot explain individual cases; heavily biased if features are highly correlated.
ICE	✓	–	Visualizes individual-level prediction trajectories; reveals heterogeneity and interactions hidden by PDP averages.	Can become cluttered with many instances; sensitive to correlated features; harder to summarize globally.

Summary Recommendation: For clinical stroke risk prediction, no single explainability method is sufficient on its own. **SHAP** is the most reliable method for understanding the model’s overall logic, as it provides consistent and theoretically grounded insights at both the global and local levels. **Counterfactual Explanations (CEs)** complement SHAP by offering highly actionable, patient-specific guidance, making them particularly suitable for patient-facing consultations where concrete risk-reduction strategies or lifestyle modifications are required.

LIME serves as a lightweight tool for rapid, local approximations of model behavior but should be used cautiously due to its instability and lack of theoret-

ical guarantees. LIME Weights and Fidelity Scores should always be looked at when understanding the prediction. SP-LIME serves to present a global view of the instances by generating several instances that maximize covering the most important features. **PDP/ICE** methods are best suited for validating global feature trends and performing sanity checks, though they are not appropriate for individual-level clinical decisions.

Overall, a **combined approach**—using SHAP for model transparency, CEs for actionable insights, and LIME and PDP/ICE for complementary validation—provides the most robust and clinically meaningful interpretability framework for stroke prediction models.

11 Conclusion

This project aimed to investigate the application of Explainable Artificial Intelligence approaches in stroke risk prediction using machine learning models. The project used various machine learning models such as XGBoost, CatBoost, and Random Forest to predict stroke risk. The results showed that these models performed well in predicting stroke risk. However, due to the black-box nature of these models, it was necessary to use explainability tools to ensure the transparency and trustworthiness of the results.

The project used various explainability approaches such as SHAP, LIME, SP-LIME, PDP/ICE, and Counterfactual Explanations to show how different approaches provide different insights into the results of the machine learning models.

Comparative analysis revealed that while LIME effectively highlighted local decision boundaries for individual patients, SHAP offered the most consistent global interpretation of feature importance. The analysis confirmed that **age** and **average glucose level** are the dominant determinants of stroke risk, with **heart disease** playing a significant contributing role.

Importantly, these findings align with established medical literature, validating that the models are leveraging causal biological signals rather than spurious correlations. Ultimately, this project demonstrates that integrating xAI is not merely a technical enhancement but a fundamental requirement for deploying machine learning in healthcare, transforming opaque algorithms into transparent decision-support tools."

References

- Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., ... others (2019). Heart disease and stroke statistics—2019 update: A report from the american heart association. *Circulation*, 139(10), e56–e528. doi: 10.1161/CIR.0000000000000659
- Feigin, V. L., Norrving, B., & Mensah, G. A. (2017). Global burden of stroke. *The Lancet*, 390(10113), 1423–1459. doi: 10.1016/S0140-6736(17)30819-5
- J. L. Herlocker, J. A. K., & Riedl, J. (2000). Explaining collaborative filtering recommendations. in conference on computer supported cooperative work.
- Karimi, A.-H., Schölkopf, B., & Valera, I. (2020). *Algorithmic recourse: from counterfactual explanations to interventions*. Retrieved from <https://arxiv.org/abs/2002.06278>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. , 30, 4765–4774.
- Molnar, C. (2019). Interpretable machine learning. (Section 5.1)
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"why should i trust you?": Explaining the predictions of any classifier*. Retrieved from <https://arxiv.org/abs/1602.04938>
- Yoon, C. W., & Bushnell, C. D. (2023). Stroke in women: A review focused on epidemiology, risk factors, and outcomes. *Journal of Stroke*, 25(1), 2–15.

Affidavit

We hereby declare that this group project was completed independently by the undersigned authors and without the use of assistance from third parties, except where explicitly stated in the document and communicated to the supervisor. No sources or aids other than those properly cited and referenced were used in the preparation of this project. All content taken directly or indirectly from external sources has been clearly identified and appropriately acknowledged.

We further confirm that the work presented reflects the collaborative effort of all group members, with responsibilities and contributions shared equally and transparently throughout the project.

28.01.2026,

Hamza Muhammad, Minh Hai Tran, Pavitra Chandarana, Yan Jing