Capstone Project Phase A

Evaluating Explainable AI Methods in Domains of Varying Human Impact

GitHub

Project Number

26-1-R-3

Supervisors:

Dr. Julia Sheidin

Dr. Avital Shulner Tal

Adan Ibrahim adan.ibrahim@e.braude.ac.il

Yarden Nahum yarden.nahum@e.braude.ac.il

Table of Content

1 Introduction

In recent years, AI's role in decision-making has evolved from rule-based systems that automated simple tasks to advanced machine learning models capable of supporting complex and strategic decisions [13].

While AI is widely integrated across various domains and plays an important role in critical decision-making [3], its impact varies significantly by risk level, presenting unique challenges and opportunities [13].

A central challenge is the increasing complexity of AI models, which has raised awareness of the 'black box' problem, where internal decision-making processes are hidden from users, making trust and reliability difficult to assess [22]. This opacity presents a significant challenge for user trust, which varies across domains. While low-risk scenarios often rely on emotional trust based on intuition, high-risk domains demand cognitive trust, which depends on understanding the system's accuracy and reliability [1].

Establishing trust between users and AI-based systems therefore requires explaining both the decision-making process and its outcomes. To address this trust gap, the field of Explainable Artificial Intelligence (XAI) has emerged to improve transparency and interpretability in complex AI systems. XAI provides tools that help users understand how and why a model produces a specific prediction [11]. It offers a range of explanatory approaches that vary in their level of detail and informativeness, including feature importance methods, visualization techniques, and counterfactual explanations. Additionally, explanatory depth is supported by both local explanations, which focus on individual predictions, and global explanations, which provide insights into the overall behavior of the model [16].

Therefore, this research aims to investigate the impact of different types of XAI across domains with varying risk levels, addressing the limited understanding of how user preferences for explanation tools change depending on the severity of the decision.

2 Background and Related Work

2.1 AI in Decision Making

Traditional decision-making often relies on human experience or intuition, which can be subjective and lead to errors. In various sectors, AI can automate processes, identify patterns, and provide actionable insights, thereby improving decision-making and operational efficiency [3]. One of the main advantages of using AI in decision making is its ability to process large datasets quickly and accurately at a scale beyond human capabilities while using data-driven approaches to provide objective insights [13].

AI's role in decision-making has evolved significantly over the past few decades. Initially, applications were restricted to rule-based logic designed for automating tasks. While effective for pre-defined processes, these systems lacked the flexibility and learning capabilities of modern AI. With the advancement of deep learning, AI has moved from performing specific tasks to enabling strategic decision-making through advanced data-driven predictions [13].

Today, AI is widely integrated across various domains and plays an important role in supporting human decision-making. For instance, in the financial domain, AI decision making systems have been integrated into areas such as credit scoring, fraud detection, and investment management by analyzing historical data and assessing risk more efficiently. Similarly, in healthcare, AI assists in disease diagnosis, treatment planning, and resource allocation, while reducing human error and improving patient outcomes through the analysis of large medical datasets and medical images. Another example is in the logistics sector, which relies on AI to optimize supply chains, improve route planning, manage inventory, and enhance customer service through real-time data analysis, leading to improved operational efficiency and reduced costs [3].

However, despite these advantages, the integration of AI into decision-making presents significant challenges. A primary challenge is data quality and availability, since reliable models depend on high-quality input, many organizations struggle with incompatible data, inconsistent formats, and a lack of data standards and supervision. Furthermore, integration with existing systems is often complex and resource-intensive, requiring both technological compatibility and time for employees to learn and adjust [13]. Beyond these technical challenges, the integration of AI presents ethical concerns, particularly regarding bias and fairness. Since AI systems rely heavily on historical data, they can reinforce and even amplify existing inequalities found in the datasets. This bias can arise from various sources, including the training data itself, algorithmic design, or implementation practices. Ultimately, these biases can affect the fairness, accountability, and transparency of the decision-making process [13].

2.2 AI Decision Making in High, Medium, and Low Risk Domains

AI's impact on decision-making varies across domains with different risk levels, each presenting unique challenges and opportunities [13]. High-risk domains are domains where decisions have a significant impact on our lives, such as medical decisions, where the consequences of bias, errors, or lack of transparency can have severe repercussions [2]. Medium-risk domains involve decisions that may not be life-threatening but significantly influence our lives, such as in the Human Resources (HR) field. In this domain, decision-making involves tasks related to recruitment, selection, and training, as well as retention decisions [8]. Finally, low-risk domains typically involve decision-making where the consequences of errors are minimal. For example, in the Text Analysis domain, the objective is to determine whether an input text was produced by a human or by an AI system [14] a task with limited direct impact on individuals' well-being.

To examine these differences, this section focuses on three distinct domains with different risk levels: Healthcare, Human Resources (HR), and AI Text Detection.

2.2.1 Healthcare (High Risk Domain):

Healthcare involves decisions with significant consequences, affecting both individual patients and broader populations [15]. Accordingly, healthcare decision-making can be divided into three main areas: clinical health services, health policy, and health regulation. Clinical health services focus on direct patient care, including diagnosis, treatment, and prevention, while health policy and health regulation address decisions that shape public health outcomes. Health policy defines overarching goals and strategies, whereas health regulation establishes the rules and standards required to ensure safe and effective healthcare [7].

AI is increasingly used in healthcare to support clinical decision-making and improve patient care. A common application is Clinical Decision Support Systems (CDSS), which analyze large volumes of medical data, such as, patient records, clinical guidelines, and scientific literature to assist healthcare professionals in diagnosis, treatment selection, and risk identification. In addition, AI and deep learning models are used to predict disease progression by analyzing historical clinical, laboratory, and imaging data, enabling more proactive and personalized treatment planning. Together, these applications help reduce human error and improve patient outcomes [3].

However, the use of AI in healthcare presents important challenges. A significant challenge is the lack of transparency, often referred to as the black-box nature of AI systems, which makes it difficult for clinicians to understand and interpret how decisions are generated. This can reduce clinicians' confidence in AI recommendations and limit their willingness to rely on such systems in clinical practice. For example, AI-based radiology systems that analyze X-ray and MRI images must provide clear explanations for their diagnoses before making clinical decisions [2].

Another challenge is data privacy, as AI systems require access to large amounts of sensitive medical information, which must be securely stored and protected from data breaches, balancing data accessibility for AI training with privacy protection remains difficult. Furthermore, bias in AI decision-making, if AI models are trained on non-representative or biased datasets, their predictions may be less accurate for certain demographic groups, leading to unfair outcomes.

Therefore, ensuring diverse and high-quality training data is essential to promote fairness and reliability in healthcare decisions [3].

2.2.2 Human Resources (Medium Risk Domain):

In the Human Resources domain, decision-making can be categorized into four primary areas: Strategic Decisions that align HR initiatives with organizational goals, Adaptation Decisions that address organizational change, Operational Decisions related to recruitment, selection, and training, and Retention Decisions aimed at maximizing employee commitment and addressing workforce needs [8]. The impact of these decisions is critical to organizational success, as effective recruitment and retention practices are essential for maintaining a productive workforce and supporting long-term growth. Conversely, ineffective decision-making in these areas may result in high turnover costs, reduced productivity, and the loss of organizational knowledge [4].

Currently, AI-driven tools are increasingly used to support these decision-making areas across Human Resources. In recruitment and selection, AI applications automate resume screening and apply predictive hiring analytics to assess candidate suitability and forecast future performance. In employee training, AI enables adaptive learning platforms that tailor content to individual learning needs, while in performance management, it supports continuous feedback and data-driven evaluations. Additionally, AI applications contribute to retention strategies by monitoring employee engagement and predicting future workforce needs [9].

These technologies offer key benefits, such as the ability to streamline HR operations, saving time and resources by executing tasks with greater speed and precision. Furthermore, AI-driven tools enhance the quality of decision-making by analyzing large datasets to identify patterns and trends that help recruitment and performance strategies. Finally, AI enables a more personalized employee experience through individualized learning and feedback mechanisms, contributing to higher employee engagement and workplace satisfaction [9].

However, the integration of AI also presents several challenges and limitations. A central concern is algorithmic bias, as AI systems rely on historical data that may contain biases related to gender, race, or other factors. As a result, AI-driven tools may produce discriminatory outcomes [9]. Another significant limitation involves data privacy and security. AI-based HR systems depend on large volumes of sensitive employee data, raising concerns about unauthorized access, misuse, and compliance with data protection regulations. In addition, AI deployment in HR raises important ethical and legal considerations. The opaque nature of many AI algorithms can undermine trust and raise concerns regarding bias and discrimination, while organizations must also ensure compliance with labor laws and anti-discrimination regulations [9].

2.2.3 Text Analysis (Low Risk Domain):

In the Text Analysis domain, decision-making focuses on distinguishing between human-written and AI-generated text, a task that has become increasingly important with the widespread use of large language models [14]. Researchers commonly treat this task as a binary classification problem, where the objective is to determine whether an input text was produced by a human or by an AI system using an AI detection tool [14].

The importance of these decisions increases as recent advances have improved the diversity and quality of machine-generated text, enabling the creation of human-like content at scale. These developments increase the difficulty of text detection and raise concerns regarding the misuse of generated text for purposes such as phishing, disinformation, fake reviews, academic dishonesty, and spam. Consequently, decision-making in AI-based text detection systems plays an important role in mitigating potential abuse and supporting the principles of trustworthy AI [5].

However, AI-based text detection faces several challenges. One major limitation is cross-domain performance, as detection models trained on one type of text often perform poorly when applied to different domains. In addition, even minimal human editing of AI-generated text can significantly reduce detection accuracy, making reliable identification difficult. Ethical concerns also arise, including issues related to privacy, bias in detection systems, and unequal access to detection tools across languages and cultural contexts [12].

2.3 Human trust in AI system

The increasing use of AI systems in everyday applications highlights the importance of user trust in AI. Trust is a central component of the interaction between humans and AI systems, as it influences whether users accept and rely on AI-generated decisions. However, trust in AI changes depending on the decision's risk. In low-risk domains, users often rely on emotional trust, which is based on intuition or general feelings. In contrast, in high-risk domains, cognitive trust becomes essential and depends on the user's perception of the system's ability to make accurate and reliable decisions. As a result, trust in AI strongly depends on system-related factors such as accuracy, reliability, transparency, and explainability [1].

However, the need for transparency is directly challenged by the 'black box' nature of modern AI. While observers can witness the inputs and outputs of these systems, the complex and non-linear inner algorithms remain hidden. Consequently, the way AI reaches its conclusions is opaque. This opacity stems from two main causes. In some cases, algorithms are kept hidden simply to protect trade secrets. However, in Deep Learning, the problem is different: it arises from the system's complexity. Unlike simple code, these models use billions of parameters to make decisions. As a result, even if the code were fully visible, the internal logic would remain incomprehensible even to the experts who created it [22].

One notable example of a 'black box' is the 'Deep Patient' system, which was designed to analyze health records from approximately 700,000 patients. The system proved to be highly effective, predicting severe diseases like schizophrenia and cancer with greater accuracy than doctors. Yet, because of the system's opacity, researchers could not determine how or why it reached these diagnoses, leaving them unable to explain the medical reasoning to patients. This example highlights the distinction between reliability and trustworthiness. While systems like Deep Patient are reliable, reliability alone is not enough for trust. Trust requires understanding the 'why' behind a decision to ensure it aligns with human interests. Since these 'black box' systems cannot explain their reasoning, they fail to meet the necessary criteria for trustworthiness [22].

2.4 Explainable Artificial Intelligence (XAI)

Establishing trust between users and AI based systems requires explaining the decision-making process and its result. Explainable Artificial Intelligence (XAI) refers to a set of methods and techniques designed to make the decisions and outputs of AI systems transparent, interpretable, and understandable to users. As AI models become increasingly complex and are used in critical

decision-making domains, many of these systems operate as “black boxes”, where the internal reasoning behind predictions remains opaque. This lack of transparency limits users’ ability to understand, trust, and effectively evaluate AI-driven decisions. XAI addresses the black-box problem by providing explanations that reveal how models arrive at specific conclusions, thereby enabling users to interpret, verify, and assess AI behavior. By improving transparency and interpretability, XAI supports trust, accountability, and ethical AI use, while also helping to identify potential biases and unintended consequences in high-stakes domains [16]. As a result, XAI is now recognized as a fundamental requirement for justifying the trustworthiness of high-performance models rather than merely a technical enhancement. To achieve this, XAI must satisfy several essential goals beyond simple transparency, including fairness, privacy awareness, and regulatory compliance, thereby ensuring that users can confidently rely on AI-driven decisions in high-risk domains [17].

XAI development typically progresses through three key stages. The first stage uses visualization tools to help non-technical users understand how the system works, which builds initial trust. The second stage focuses on generating clear explanations for specific decisions using XAI methods. This is critical for verifying results in tasks such as fraud detection. The final stage integrates these explainable AI insights directly into business operations, allowing organizations to combine AI predictions with business rules for more advanced, personalized applications [17].

XAI does not provide a single type of explanation, rather, it offers a range of explanatory approaches that vary in their level of detail and informativeness. In this context, explanatory depth refers to the level of detail provided by AI systems to justify their decisions. This concept reflects how thoroughly a system can explain its reasoning, as not all AI explanations are equally informative. While some systems offer explanations with limited value, others deliver detailed and insightful information that enhances user trust and understanding [16].

Achieving explanatory depth involves various XAI techniques designed to make AI more understandable. These include feature importance methods that identify the most influential factors in a decision, visualization techniques that illustrate decision pathways and feature interactions, and counterfactual explanations that demonstrate how changes in input features could lead to different outcomes. Additionally, explanatory depth is supported by both local explanations, which focus on individual predictions, and global explanations, which provide insights into the overall behavior of the model. Together, these approaches contribute to an understanding of AI systems [16].

2.4.1 XAI Benefits And Challenges

XAI addresses the challenge of opaque models by making the system's reasoning transparent, enabling users to confirm that decisions are safe and correct. Primarily, it increases trust, as users are more likely to accept an AI decision, such as a loan rejection, if they understand the specific reasons behind it. Furthermore, XAI helps organizations prove that their systems are following strict laws and ethical standards. From a technical perspective, it enables easier debugging, by revealing the internal logic behind errors, developers can identify exactly why a model failed and fix bugs faster than in "black box" systems. Additionally, XAI helps reduce bias by revealing if a decision was based on prejudiced data, allowing operators to spot and fix unfair outcomes. Finally,

it enables better actions by identifying the specific factors causing a result, allowing users to address the root cause of a problem rather than relying on generic solutions [19].

Despite the advantages of XAI, its integration faces several challenges. A primary issue is the lack of expertise, as many people do not have the technical skills to understand AI explanations or judge if a decision is fair. Additionally, explaining AI is not a neutral process. When experts try to simplify complex logic for the users, they must make specific choices about what to include. These choices can be biased, which might make the system look fair even when it still uses biased data. The dynamic nature of algorithms also presents a challenge because these systems learn and change over time, making older explanations obsolete. This issue is further complicated by the interference of algorithms, where a decision is influenced by a long chain of different data sources and many different models working together. Furthermore, AI outcomes are highly context-dependent, meaning it is difficult to explain the logic in a general way when it affects each person differently. This is especially problematic when dealing with complex decisions that are naturally unclear and do not have one single right answer. Finally, explaining the link between inputs and outputs does not guarantee that the AI used that logic to reach its decision, as the inner workings of these "black box" systems often remain hidden [6].

2.4.2 XAI Tools

The following section reviews three main XAI tools: LIME for local explanations, SHAP for global explanations, and DICE for counterfactual explanations.

2.4.2.1 LIME (Local Interpretable Model-Agnostic Explanations)

LIME is a widely used and flexible tool designed to interpret the predictions of complex "black box" systems. As a model-agnostic technique, it can be applied to any AI model because it does not need to see the model's internal parts, such as its layers or weights. Instead, LIME explains a specific decision by focusing on a local surrogate model [11].

To generate an explanation, LIME creates small variations of a specific input and observes how the target model reacts to them. It then fits a simpler, linear model to these variations to estimate the complex model's behavior in that specific area. Consequently, LIME does not explain the entire system at once, rather it provides a local explanation that highlights specific features that led to a particular decision. While effective, the quality of the explanation depends on how well the simple model fits the complex one, and it can sometimes produce slightly different results for the same input because of the way it uses random sampling [11].

2.4.2.2 SHAP (Shapley Values)

SHAP is a tool that explains AI predictions using rules from Game Theory. It treats the prediction process like a game where all input features work together as a "team" to produce a result. SHAP calculates the credit each feature deserves, ensuring that if two features have the exact same effect, they receive equal credit. However, if one feature causes the other, SHAP cannot tell the difference and will simply split the credit equally instead of giving it all to the source feature [11].

To generate an explanation, SHAP breaks down the model's prediction into the individual contributions of each feature. It starts with a baseline value, which represents the average prediction of the model. Then, it calculates how much each specific feature pushes the result away from this average, either increasing the score or decreasing it. The final explanation is simply the sum of these individual contributions, showing exactly how the combination of features led to the specific decision [11].

2.4.2.3 DiCE (Diverse Counterfactual Explanation)

DiCE is a XAI method used to provide insights into machine learning decisions by generating counterfactual explanations. The primary goal of this method is to identify the minimum changes needed in the input data to modify a model's specific prediction. By doing so, DiCE offers actionable insights, showing users exactly what would need to be different for the model to reach a desired outcome [21].

The process works by using a loss function that measures the distance between the model's prediction and the desired result, and the distance between the original instance and the counterfactual example. While DiCE helps address the problem of opaque decision-making, it faces certain limitations. Specifically, it may struggle with high-dimensional data or very complex models, which can limit the diversity and variety of the explanations it generates [21].

2.5 Summery

In conclusion, the literature demonstrates that AI systems are increasingly central to decision-making processes across diverse domains, each associated with different levels of risk and potential impact. Despite their growing adoption, these systems face significant challenges, most notably the opacity of black-box models, which limits user trust, accountability, and ethical oversight. To address these concerns, Explainable Artificial Intelligence (XAI) has emerged as a framework for enhancing transparency and interpretability, enabling users to better understand and evaluate AI-driven decisions. Furthermore, XAI techniques such as LIME, SHAP, and DiCE provide approaches to explaining model behavior through local explanations, feature attribution, and counterfactual reasoning. Collectively, the reviewed literature highlights explainability as an essential requirement for the trustworthy and responsible deployment of AI systems, particularly in domains where decision-making risk is high.

3 Expected Achievements

The first expected achievement is to develop a decision-making system that operates across three distinct domains: Healthcare (High Risk), Human Resources (Medium Risk), and Text Analysis (Low Risk). This system will serve as a foundation for our research, allowing us to generate predictions and compare how different risk levels influence the decision-making process.

Secondly, we aim to successfully integrate three different explanation tools SHAP, LIME, and DiCE into the system. By doing so, the system will produce three unique types of explanations for every decision. This will demonstrate the ability to combine different transparency methods to give users a complete picture of how the AI reached its conclusion.

Finally, the project will provide clear insights into user preferences through a user study. We expect to identify which explanation method is most effective and trustworthy for each specific risk level.

We anticipate that in high-risk domains, such as Healthcare, users will prefer in-depth and comprehensive explanations. Given the critical nature of medical decisions, users are likely to require detailed evidence to verify the system's reasoning before they can trust the result.

In contrast, for low-risk domains like Text Analysis, we expect users to prefer summarized and concise explanations. Since the consequences of an error are minor, we think that users will prioritize efficiency and speed, favoring short, simple explanations over complex data.

4 Research Process

The research process began with the selection of datasets representing three distinct risk levels: Healthcare (High Risk), Human Resources (Medium Risk), and Text Analysis (Low Risk), analyzing their applications to ensure they align with the classification tasks required for our study. Following this, we conducted an extensive literature review to examine the role of AI in decision-making across these domains, focusing on the challenges of "black box" opacity and the varying requirements for user trust and transparency. Next, we analyzed specific Explainable AI (XAI) tools LIME, SHAP, and DiCE to understand their mechanisms for providing local explanations, feature attribution, and counterfactuals.

The practical phase involved developing predictive models for each domain, creating a system that processes a selected record and generates a prediction accompanied by all three explanation types. Finally, utilizing existing tools for explanation evaluation, we designed a user study to assess user preferences. In this study, participants are presented with specific records to determine which explanation method best fits the context of each domain and to provide the reasoning behind their choices.

4.1 Chosen Datasets

We selected the following datasets to represent the three risk levels

4.1.1 Healthcare (High Risk Domain)

In the Healthcare domain, we chose the dataset: **Diabetes Health Indicators Dataset.**

The dataset contains 100,000 patient profiles and 35+ features that include health and lifestyle indicators such as BMI, physical activity, smoking, and general health measures that are well-established indicators of diabetes risk.

The primary use of this dataset is predictive modeling and analysis, including classification tasks that aim to identify individuals at higher risk of diabetes based on health indicators. It can also be used to analyze the relationship between health behaviors and diabetes outcomes.

4.1.2 HR (Medium Risk Domain)

In the Human Resources (HR) domain, we chose the dataset: **IBM HR Analytics Attrition Dataset.**

This dataset contains 1,470 employee records with 35 features, including demographics details, job roles, satisfaction levels, and monthly income.

The primary use of this dataset is to predict which employees are likely to quit. Companies use it to understand the reasons for leaving—such as low salary or overtime—and to create plans to keep their workers.

4.1.3 Text Analysis (Low Risk Domain)

In the Text Analysis domain, we chose the **LLM - Detect AI Generated Text Dataset**.

This dataset contains more than 28,000 essays, including both human-written and AI-generated texts.

The primary use of this dataset is to identify whether a text was written by a human or generated by an AI system. It can be used for classification tasks that detect AI-generated text and analyze differences between human and machine-written content.

4.2 User Evaluation of XAI Methods

To assess the explanations, we will use the System Causability Scale (SCS). Unlike general usability metrics, the SCS specifically measures whether an interface helps users understand the underlying causality of an AI decision [10].

We have selected the items that specifically measure the balance between explanatory depth and ease of use.

Procedure: Participants will view a decision record accompanied by an explanation (SHAP, LIME, and DiCE) and rate their agreement with the following statements on a 5-point Likert scale.

- **Causal Factors**: I found that the data included all relevant known causal factors with sufficient precision.
- **Causality**: I found the explanations helped me to understand the cause-and-effect relationship in the prediction.
- **Context**: I understood the explanations within the context of the domain.
- **Support**: I did not need support or external help to understand the explanations.
- **Learnability**: I think that most people would learn to understand these explanations very quickly.
- **Efficiency**: I received and understood the explanations in a timely and efficient manner.

In addition to the SCS, we will also use the FATE framework (Fairness, Accountability, Transparency, and Explainability). Explainability and causability play a key role in shaping how users perceive and trust AI systems, to reflect these principles, we include FATE that evaluate how users perceive the explanations [21]:

- **Causability**: evaluates how effectively the explanations help users understand the reasoning behind the algorithm's outputs.

- **Trust**: assesses the level of user trust in the system, focuses on confidence in recommendations, trustworthiness of outputs, and perceived reliability of results.
- **Transparency**: examines users' perceptions of how understandable, explainable, and observable the algorithmic processes are.
- **Satisfaction**: measures users' overall satisfaction with the algorithmic services and their outputs.

4.3 System Workflow

The following activity diagram, presented in figure 1, shows how the system will work when presenting explanations to the user. The same flow is applied separately to each of the three domains: Healthcare, HR, and Text Analysis.

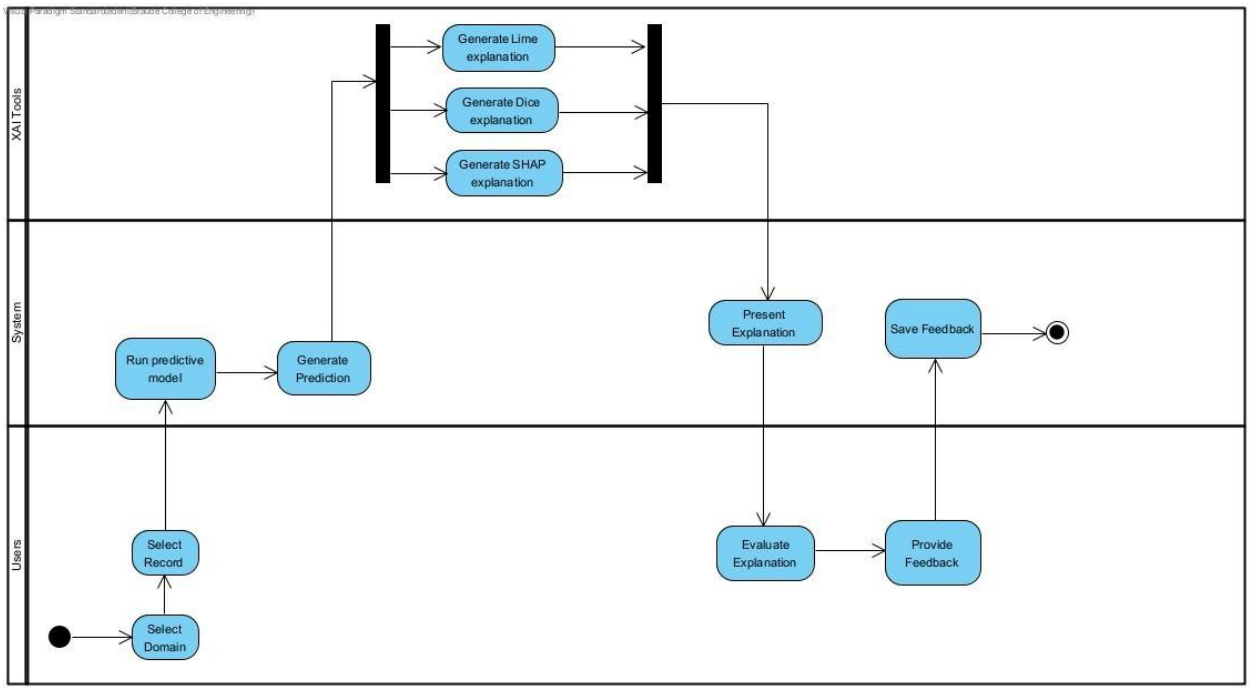

Figure 1 - Activity Diagram

After the users provide their feedback, the results will be analyzed both within each domain and across domains. This will allow us to compare users' explanation preferences and evaluation outcomes inside each domain (differences between different XAI methods) as well as between domains with different risk levels (differences between domains for the same XAI method).

4.4 Architecture and tools

The following tools and technologies will be used in our project:

4.4.1 Backend (AI and Data Processing)

Programming Language:
Python: Widely used for machine learning, data processing, and API integration.

Data Processing Libraries:
Pandas: Used for cleaning and preprocessing of the raw datasets.
NumPy: For high numerical computations and handling multi-dimensional arrays required by the AI models.

AI Models:
ML/DL Frameworks: TensorFlow or PyTorch: For custom machine learning models.

XAI Tools:
SHAP (SHapley Additive exPlanations): Provides insights into individual predictions by calculating the contribution of each feature using principles from cooperative game theory.
LIME (Local Interpretable Model-Agnostic Explanations): Explains the predictions of any classifier by approximating it locally with an interpretable model, helping users understand model behavior in specific instances.
DiCE: dice_ml, a python library for counterfactual explanation and recourse models.

Backend Service / API:
FastAPI / Node-JS : Serves to retrieve the study scenarios and store user responses.

Database:
MongoDB (NoSQL): Selected for its flexibility in handling JSON documents. Can be used to store: (1) Study Records: The pre-computed records, model predictions, and XAI visualization data. And (2) User Data: participant responses, timestamps, and survey results.

4.4.2 Frontend (User Interaction and Visualization)

Programming Language:
JavaScript: Standard for interactive web applications.

Frameworks/Libraries:
React.js: For building responsive, component-based web interfaces.

Survey Implementation:

React Hook Form: Utilized to manage the SCS and FATE questionnaires.

Visualization Tools:
Plotly: For interactive dashboards and graphs.
Chart.js: For lightweight charts (e.g., bar charts, pie charts).
Recharts: For rendering interactive data visualizations.

Styling:
Tailwind CSS: For responsive and visually appealing designs.

4.4.3 System Architecture:

The architecture of the system we intend to build is presented in figure 2.

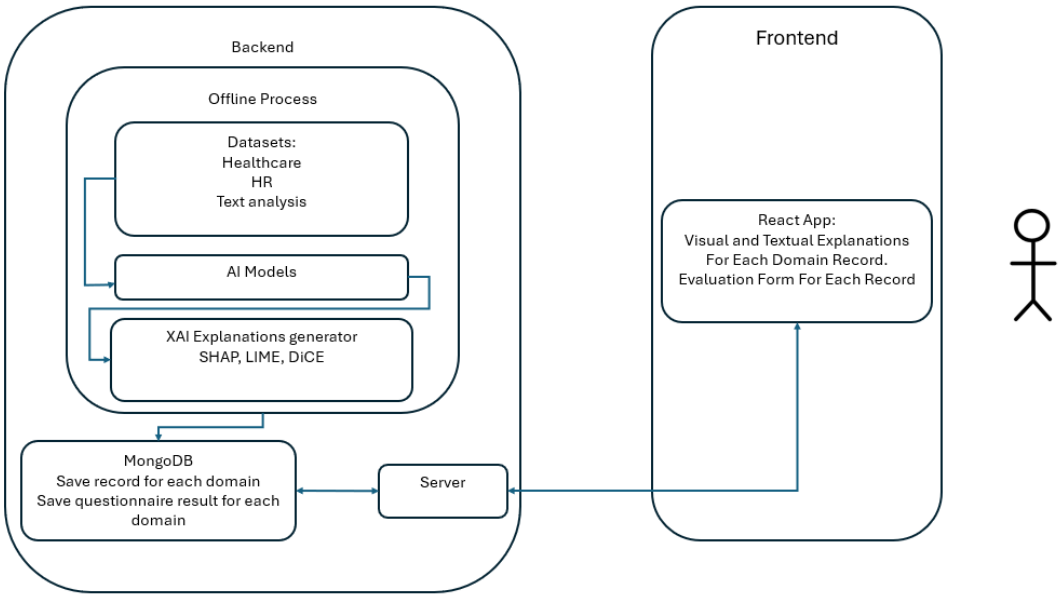

Figure 2 - System Architecture

For the research purpose, the system will display a pre-determined record to each user. This ensures that all participants evaluate the same scenario, allowing for consistent comparison across the study.

The diagram illustrates the system's architecture:

1. **Offline Phase (Left):** Datasets from three domains (Healthcare, HR, Text) are processed by the AI models. The system then generates explanations using SHAP, LIME, and DiCE, storing the results in MongoDB.

2. **Online Phase (Right):** The React Frontend fetches these scenarios via the Server. Users review the visual explanations and submit their evaluations via forms. These responses are sent back through the server and stored in the database for analysis.

4.5 GUI

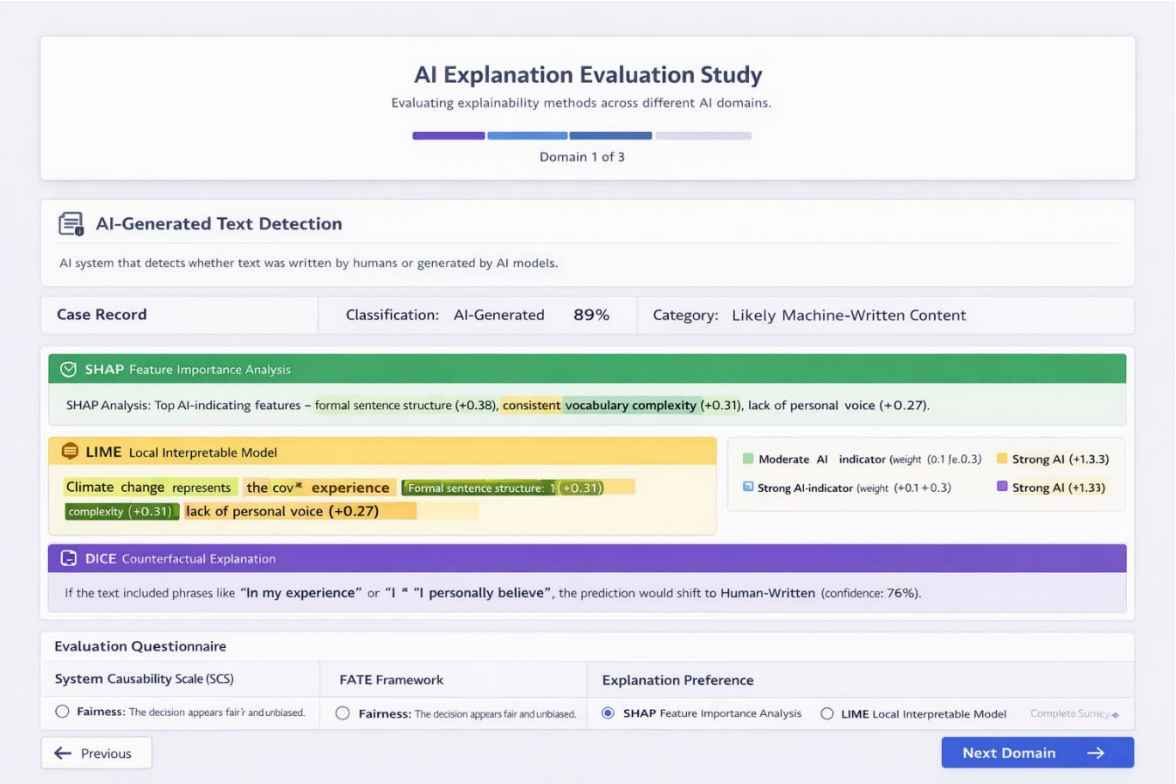

Figure 3 - Mock Of The GUI

The user GUI guides participants through a user study across three distinct risk domains. For each domain, the system presents a domain-specific record alongside the AI's prediction and three corresponding explanations generated by SHAP, LIME, and DiCE. Users evaluate the clarity and utility of these insights via a questionnaire (SCS and FATE), after which their responses are stored in the MongoDB database, and the system automatically transitions to the next domain scenario.

4.6 Challenges

4.6.1 Finding Relevant Academic Literature

One of the main challenges was finding strong and relevant academic sources that fit our exact topic. Many papers discuss AI decision-making in general, and others focus only on XAI methods, but fewer sources connect all parts of our work together: decision-making, trust, explainability,

and comparing domains with different human impacts. In addition, some available sources were not peer-reviewed or did not provide enough detail for academic writing, so we had to carefully filter and choose references that were both reliable and closely related to our project.

4.6.2 Selecting Appropriate Datasets

Choosing datasets that accurately represented each domain and aligned with the intended decision-making tasks was challenging. The datasets needed to be suitable for classification and interpretable by explanation tools, while also reflecting realistic decision contexts. Another significant challenge was ensuring the specific task within each domain matched the required human impact level. Since risk varies even within a single field, we had to carefully select datasets where the consequences of an error aligned with our High, Medium, or Low risk categories.

4.6.3 Writing a Coherent and Well-Organized Report

Finally, a practical challenge was organizing the report so that all parts connect smoothly. Making the report structured, clear, and consistent in writing style required several iterations and careful editing.

5 Evaluation Plan

To ensure the system effectively evaluates user preferences for Explainable AI across different risk domains, we will evaluate it through a structured testing process. The primary goal of this evaluation is to collect user's feedback regarding trust, clarity, and the perceived suitability of different explanation methods.

5.1 Testing Procedure

Test ID	Description	Expected Result	Precondition	Comments
1	Running Python explanation scripts	Python scripts generate SHAP, LIME, and DiCE explanations for each dataset.	Datasets and trained models are available.	One instance per domain.
2	Saving explanations to database	All generated explanations are saved in MongoDB.	Explanations are successfully generated.	Same explanations used for all users.
3	Presenting all domains	The system presents all three domains: Healthcare, HR, and Text Analysis.	Explanations are stored in the database.	Same domains shown to all users.
4	Loading predefined instances	The system loads one predefined decision instance for each domain.	Domains are presented.	Identical instances shown to all users for comparison.
5	Displaying SHAP explanations	The system displays SHAP explanations for each domain.	SHAP explanations are pre-generated.	Ensure clarity and consistent format.
6	Displaying LIME explanations	The system displays LIME explanations for each domain.	LIME explanations are pre-generated.	Same explanation structure across domains.

7	Displaying DiCE explanations	The system displays DiCE-based explanations for each domain.	DiCE explanations are pre-generated.	Explanations refer to the same decision instance.
8	User review of explanations	The user can review all explanations for each domain.	All explanations are displayed.	Explanations are presented without guidance.
9	Filling the evaluation form	The user fills in the evaluation form based on SCS and FATE criteria for the domain.	Explanations have been reviewed.	Measures trust, clarity, and explainability.
10	Saving user responses	The system saves all user responses successfully.	Form is submitted.	Data stored for later analysis.

5.2 User-Based Evaluation

The evaluation will involve real users who will review explanations generated by SHAP, LIME, and DiCE across domains with different risk levels. The goal of this evaluation is to assess user understanding, trust, and perceived suitability of the explanations, rather than technical model performance. User feedback will be collected through structured questionnaires based on the System Causability Scale (SCS) and the FATE framework. We will collect the responses and analyze them to identify user preferences and compare explanation effectiveness across domains. The questions are rated on a 5-point Likert scale (from strongly disagree to strongly agree) and focus on trust, clarity, and suitability.

Example questions include:

- The explanation clearly shows which factors influenced the decision. (Rate from 1-5)
- I trust the AI decision based on the explanation provided. (Rate from 1-5)
- The explanation provides sufficient information to justify the decision. (Rate from 1-5)

References

[1] Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11, Article 1568. https://www.nature.com/articles/s41599-024-04044-8

[2] Ahadian, P., Xu, W., Liu, D., & Guan, Q. (2025). Ethics of trustworthy AI in healthcare: Challenges, principles, and practical pathways. *Neurocomputing*, 661, 131942. https://www.sciencedirect.com/science/article/pii/S0925231225026141

[3] Alabi, M. (2024). AI and deep learning for decision-making in healthcare, finance, and logistics. ResearchGate. December 1, 2024. https://www.researchgate.net/publication/386329740_AI_and_deep_learning_for_decision-making_in_healthcare_finance_and_logistics

[4] Căvescu, A.M.; Popescu, N. Predictive Analytics in Human Resources Management: Evaluating AIHR's Role in Talent Retention. *AppliedMath* 2025, 5, 99. https://doi.org/10.3390/appliedmath5030099

[5] Crothers, E. N., Japkowicz, N. and Viktor, H. L. (2023). Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods," in *IEEE Access*, vol. 11, pp. 70977-71002, 2023, doi: 10.1109/ACCESS.2023.3294090. https://ieeexplore.ieee.org/document/10177704

[6] de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2), 101666. https://doi.org/10.1016/j.giq.2021.101666

[7] Doreswamy, N., & Horstmanshof, L. (2023). Attributes That Influence Human Decision-Making in Complex Health Services: Scoping Review. *JMIR Human Factors*, 10, e46490. https://humanfactors.jmir.org/2023/1/e46490

[8] Friedman, B. A. (2007). Globalization implications for human resource management roles. *Employee Responsibilities and Rights Journal*, 19(3), 157–171. https://link.springer.com/article/10.1007/s10672-007-9043-1

[9] Gupta, R. (2024). Impact of Artificial Intelligence (AI) on Human Resource Management (HRM). *International Journal for Multidisciplinary Research*, 6(3). https://pdfs.semanticscholar.org/777c/20d96221f47f83130be9c5cf6a9986759917.pdf

[10] Holzinger, A., Carrington, A. & Müller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS). *Künstl Intell*, 34, 193–198. https://doi.org/10.1007/s13218-020-00636-z

[11] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W. (2022). Explainable AI Methods - A Brief Overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) *xxAI* -

Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science, vol 13200. Springer, Cham. https://doi.org/10.1007/978-3-031-04083-2_2

[12] Kehkashan, T., Riaz, R. A., Al-Shamayleh, A. S., Akhunzada, A., Ali, N., Hamza, M., & Akbar, F. (2025). AI-generated text detection: A comprehensive review of methods, datasets, and applications. *Computer Science Review*, 58, 100793. https://doi.org/10.1016/j.cosrev.2025.100793

[13] Kumar, B. R., Madhuri, A., & Shireesha, B. (2024). The Role of Artificial Intelligence in Decision-Making Processes. *African Journal of Biological Sciences*, 6(6), 6344–6362. https://www.researchgate.net/publication/381767448_The_Role_of_Artificial_Intelligence_in_Decision-Making_Processes

[14] Liu, X., Li, Y., & Li, K. (2025). Enhancing the robustness of AI-generated text detectors: A survey. *Mathematics*, 13(13), 2145. https://www.mdpi.com/2227-7390/13/13/2145

[15] Masic, I. (2022). Medical Decision Making – an Overview. *Acta Informatica Medica*, 30(3), 230–235. https://pmc.ncbi.nlm.nih.gov/articles/PMC9560052/

[16] Mathew, D. E., Ebem, D. U., Ikegwu, A. C., Ukeoma, P. E., & Dibiaezue, N. F. (2025). Recent emerging techniques in explainable artificial intelligence to enhance the interpretability and understanding of AI models for humans. *Neural Processing Letters*, 57, 16. https://link.springer.com/article/10.1007/s11063-025-11732-2

[17] Mohamed, A., & El-Ghamry, H. (2025). Explainable Artificial Intelligence: A systematic review of systematic reviews. *Artificial Intelligence Review*. https://doi.org/10.1016/j.iswa.2025.200595

[18] Ottosen, M. J., Sedlock, E. W., Aigbe, A. O., Bell, S. K., Gallagher, T. H., & Thomas, E. J. (2021). Long-Term Impacts Faced by Patients and Families After Harmful Healthcare Events. *Journal of Patient Safety*, 17(8), e1145–e1151. https://pmc.ncbi.nlm.nih.gov/articles/PMC6050155/

[19] Praveen, S., Joshi, K. (2023). Explainable Artificial Intelligence in Health Care: How XAI Improves User Trust in High-Risk Decisions. In: Hassanien, A.E., Gupta, D., Singh, A.K., Garg, A. (eds) *Explainable Edge AI: A Futuristic Computing Perspective*. Studies in Computational Intelligence, vol 1072. Springer, Cham. https://doi.org/10.1007/978-3-031-18292-1_6

[20] Raufi, B., Finnegan, C., & Longo, L. (2024). A comparative analysis of SHAP, LIME, ANCHORS, and DICE for interpreting a dense neural network in credit card fraud detection. In L. Longo (Ed.), *Explainable Artificial Intelligence* (pp. 365–383). Springer. https://doi.org/10.1007/978-3-031-63803-9_20

[21] Shin, D. (2021). Effects of explainability and causability on perception, trust, and acceptance. *International Journal of Human-Computer Studies*, 146, 102551. https://www.sciencedirect.com/science/article/pii/S1071581920301531

[22] von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34(4), 1607–1622. https://link.springer.com/article/10.1007/s13347-021-00477-0