Heart Failure Prediction

***Abstract*—In our model we have considered making an XAI of cardiac problems since we have observed from daily comments that people tend to complain or ask for help on health advice because the meters seem to have some deficiency or illustrate little information and perhaps cryptic for users, so we have interviewed to determine what factors caused them to require help from others.**

I. Github repository

The implemented prototype and the code to run on can be found at Github repository.

II. Introduction

The adoption of AI-based predictive models in the medical field requires not only high levels of accuracy but also explainability capabilities that allow users to understand and trust the generated outcomes. In this work, we present the development of an XAI prototype aimed at predicting the risk of heart failure. The project followed a user-centered approach, combining qualitative analysis, user story collection, iterative design, and empirical testing cycles involving patients, doctors, and data scientists. Through successive prototyping and validation phases, we identified critical issues and implemented targeted improvements, with the goal of creating a predictive system that is accessible, transparent, and capable of effectively supporting clinical decision-making processes.

III. Related Work

Several existing tools attempt to estimate cardiovascular risk, yet they suffer from limitations in usability, user experience, and explanation quality. Two widely used platforms—HeartScore and Calculadora de Riesgo Cardiovascular—serve as valuable references for understanding these challenges and motivating the development of our user-centered Explainable AI system.

A. HeartScore

HeartScore, developed by the European Society of Cardiology, provides medically validated predictions based on various clinical parameters such as age, sex, smoking status, blood pressure, and cholesterol levels. However, its usability presents several issues. The interface lacks real-time input validation, allowing users to proceed even with missing or implausible values (e.g., extremely high blood pressure or missing cholesterol readings). Moreover, some fields like HDL/LDL cholesterol require medical supervision, making them inaccessible to most users without clinical assistance.

The user experience is further limited by a lack of visual hierarchy and interactive guidance. Critical information such as the absolute risk percentage is embedded within dense text, making it difficult for users to quickly interpret the results. The placement of certain action buttons also adds to the confusion, as options like “Print” and “Modify examination data” are grouped together without clear separation of functionalities.

In terms of explanation quality, while HeartScore provides clinically accurate results, the underlying reasoning remains highly technical and inaccessible to patients. No visual aids are provided to break down how individual risk factors contribute to the overall score. Furthermore, there are no dynamic or personalized recommendations that translate risk scores into actionable next steps for lifestyle modification.

Potential improvements to HeartScore would include inline input validation, constrained input formats (e.g., sliders and numeric-only fields), patient-friendly risk breakdowns using infographics, dynamic recommendations based on patient data, and direct integration of medical guidelines for educational purposes.

B. Calculadora de Riesgo Cardiovascular

The Calculadora de Riesgo Cardiovascular adopts a simpler interface aimed at the general population. Despite its accessibility, it suffers from severe shortcomings in both data validation and explanatory capacity. Users can input incompatible or erroneous values—such as text in the age field—without receiving immediate error messages, potentially compromising the accuracy of risk predictions. Real-time validation and immediate feedback would significantly improve the reliability of user input.

From a user experience perspective, the absence of real-time feedback complicates the interaction for inexperienced users, who may not realize they have entered invalid data until after submission. The lack of contextual guides or tooltips further limits user understanding of the data entry process and the importance of each parameter.

The explanation quality is also limited. While the tool provides a global risk score, it fails to explain how each input variable influences the result. The absence of risk factor breakdowns leaves users with little insight into which aspects of their lifestyle or medical profile contribute most to their overall cardiovascular risk. Moreover, generic advice—such as suggesting to stop smoking even when the user is already a non-smoker—demonstrates a lack of personalized recommendation logic.

Potential improvements include adding visual breakdowns of risk contributions, detailed explanations for each input parameter, and customized recommendations that are dynamically generated based on individual health profiles.

IV. Interview with users

To investigate the issue and understand its scope, we conducted interviews from three different perspectives: patients, doctors, and data scientists. A total of nine volunteers participated—three from each group—each with experience in or exposure to counseling. The interviews explored their past

experiences, typical counseling routines, challenges they faced, what they felt was missing, and what improvements they would like to see. Despite some difficulties, the study provided valuable insights into the relationship between these groups and the tendency to seek counseling.

V. Breakdown of user stories

After the interviews we have translated and determined the goals, frustrations and uses in user stories to determine the characteristics that our XAI model should have, there are some extensively complex or out of place that we cannot currently implement or contemplate, but we believe that we can cover a good part of these expectations that are more frequent and criticized during the interviews, therefore we have:

LEGEND: Numbers in brackets [P,D,DS] represent the number of times the patient, doctor, and data scientist appear in the interviews (respectively); Lines are marked with user persona letters when a single viewpoint is predominant ($\geq$2); they are underlined when they occur more frequently than others ($\geq$4).
Items marked with (*) indicate features approved for implementation.

- **P,D: (*) [2,2,0]** Receive recommendations in addition to advising routines
- **P: (*) [2,0,0]** Make inquiries independently to decrease routine doctor visits
- [1,0,0] Impossibility of completing highly technical medical fields due to dependence on specialized equipment
- **P,D: (*) [2,2,0]** Preference for simple, accessible language over medical terminology
- **(*) [2,2,2]** Support for presenting potential risks even when uncertain
- **P: (*) [2,1,0]** Visual display using traffic light colors for risk assessment
- **DS: (*) [2,0,3]** Concerns about data transmission security and breach scenarios
- [1,1,0] Temporary storage of blood pressure readings
- [1,1,0] Reminders for blood pressure monitoring and medication schedules
- **P: (*) [2,0,0]** Preference for doctor-like explanations of results
- **D: (*) [1,2,1]** Immediate visualization of risk breakdown and intensities
- [1,0,0] Dietary recommendations non-conforming to Mediterranean habits
- **P: (*) [2,0,0]** Difficulty interpreting blood pressure readings (Systolic, Diastolic, Pulse)
- [0,1,1] Integration with diagnostic and medical monitoring systems
- [0,1,0] Export functionality for reports and visual content
- [0,1,0] Minimal user interaction requirements
- **DS: (*) [0,1,2]** Concerns regarding algorithmic bias in predictions
- [0,1,0] Learning capabilities from adverse case studies
- [0,1,1] Model precision, accuracy, and noise resistance requirements
- **DS: (*) [0,0,2]** Implementation of SHAP visualization methods
- **DS: (*) [0,0,1]** Implementation of LIME visualization methods
- [0,0,1] Concern about highlighting non-modifiable factors such as age

VI. Features that will cover user stories

From the previous list, we have decided to implement 12 (out of 22) user stories in our XAI to cover the gaps and obstacles that they are currently experiencing. The impact of the user story will also be mentioned in parentheses, with 6 (the maximum) being the most notable impact. We selected the user stories according on two rules: 1º story must have 2 votes from the same group; 2º by similarity or proximity feature:

- List the actual risks and even mention potential risks that might warrant further review if there is concern. (6/6)
- Our XAI will operate without internet dependency, thus eliminating the risk of data breaches and the LOPD (Personal Data Protection Act). However, our XAI will be more limited in features. (5/6)
- Provide recommendations and routines that can be followed based on the risk factors present, emphasizing the risk that can be most reduced. (4/6)
- To provide a more viable medical alternative for widespread use by the public, such as the blood pressure monitor, for more superficial and autonomous monitoring. (4/6)
- After showing the prediction, we will highlight the box with a traffic light color. (3/6)
- Show the risks and their intensities so that they can be appreciated at a glance. (3/6)
- Review the model to ensure that the predictions avoid bias, but we believe this is very difficult to achieve since it is not visible to the naked eye. (3/6)
- Allow patients to self-diagnose, thus alleviating routine recurring visits. (2/6)
- Provide a simplified option for interpreting the results of blood pressure measurements and describe, although very limited, the possible causes of imbalance. (2/6)
- Brief explanation of the guidelines and what you should do, acting as if you were a doctor. (2/6)
- We will implement a visual representation with SHAP format. (2/6)
- We will implement a visual representation with LIME format. (1/6)

VII. Problem statement and motivation

A. Current Challenges in Cardiac Risk Assessment Tools

Our preliminary analysis of existing cardiac risk prediction tools reveals several critical issues. Traditional meters and digital health applications often provide cryptic information that users struggle to interpret, leading to frequent requests for help and reduced confidence in self-monitoring. Through daily observations and user feedback, we identified that people consistently complain about the deficiency of current health

advice systems, which fail to bridge the gap between complex medical data and user understanding.

Existing tools like HeartScore and Calculadora de Riesgo Cardiovascular demonstrate these limitations clearly. HeartScore, while medically sophisticated, requires doctor-exclusive fields that prevent patients from conducting independent self-assessments. Conversely, Calculadora de Riesgo Cardiovascular oversimplifies risk factors, providing age-based predictions (always medium risk at 60, high risk at 80) regardless of the user's actual health status, and offers inappropriate recommendations such as "don't smoke" to non-smokers.

B. The Need for Explainable AI in Healthcare

The complexity of cardiovascular risk assessment creates a unique challenge: medical accuracy must be balanced with user comprehensibility. Traditional black-box AI models, while potentially accurate, fail to provide the transparency necessary for healthcare applications where users need to understand not just their risk level, but also the reasoning behind it.

This transparency gap is particularly problematic in cardiac health monitoring because:

- Patients need to understand which factors contribute to their risk to make informed lifestyle changes
- Healthcare providers require clear explanations to validate AI recommendations and communicate effectively with patients
- Data scientists need interpretable models to ensure bias detection and clinical reliability

C. Motivation for User-Centered XAI Design

The motivation for this work stems from the recognition that effective health monitoring tools must serve diverse user groups with varying levels of medical knowledge. Our approach addresses three key stakeholder needs simultaneously: patients seeking accessible health insights, doctors requiring clinical precision, and data scientists ensuring algorithmic reliability.

By developing an explainable AI system specifically designed through user-centered design principles, we aim to create a tool that not only provides accurate cardiac risk predictions but also delivers clear, actionable explanations tailored to each user's expertise level. This approach promises to reduce the dependency on routine medical consultations for basic risk assessment while maintaining the clinical rigor necessary for healthcare applications.

The ultimate goal is to democratize cardiac health monitoring by making sophisticated risk assessment accessible to general users without sacrificing the interpretability and reliability that medical professionals require.

VIII. Interface Proposal

For the interface we planned to put a menu with the input parameters (at the left), while we keep the most information output, the risks representation and visual at the center page. After tapping the "Predict" button, the model will compute the values and describe the results giving a brief explanation regarding what are each factors, alongside with a semaphore indicator.

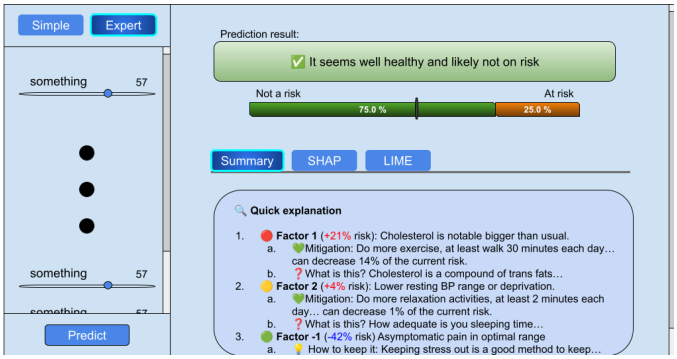

Fig. 1. Interface key concept illustration in general after click "predict" button, default tab

After the explanation summary, there are other tabs in the same page for a deeper review regarding the risk factors, including the negative factors (safety cases) to help the user notice good cases and bad examples. By request, we placed the SHAP with visual representation.

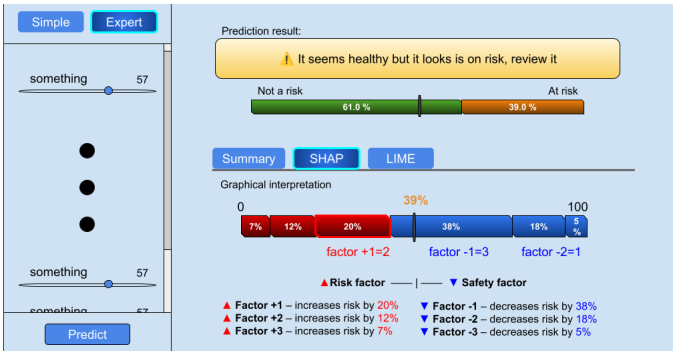

Fig. 2. Interface key concept illustration in general after click "predict" button, SHAP tab

And similarly, we also put LIME representation since it's mostly the same concept.

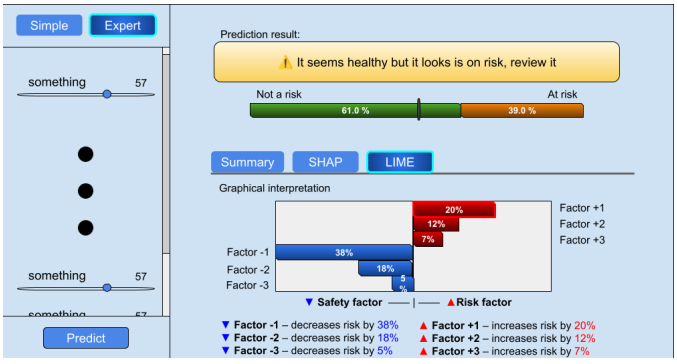

Fig. 3. Interface key concept illustration in general after click "predict" button, LIME tab

To clarify and distinguish patients with doctors, we separated the viewpoints with Simple and Expert, each one has

its own set of parameters. We bet this viewpoint segregation will benefit and cover user needs, with or without medical knowledge.

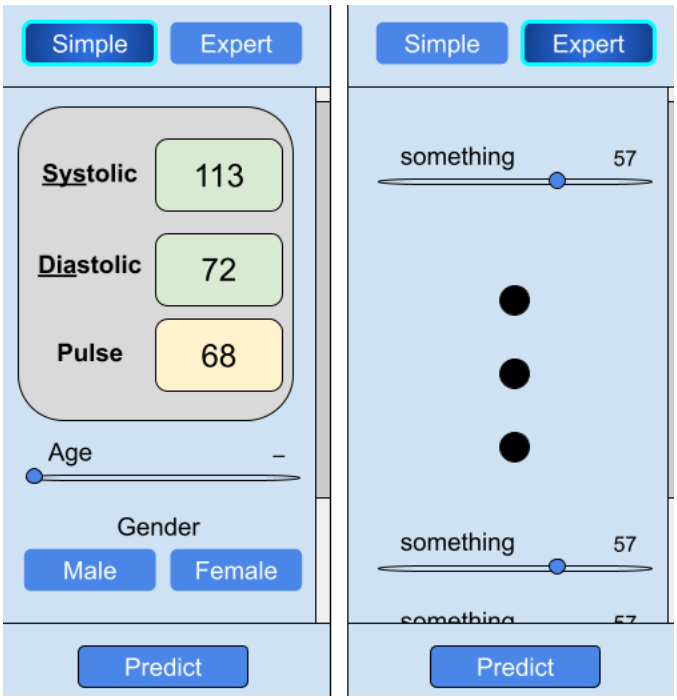

Fig. 4. Interface difference between Simple (Patient viewpoint) and Expert (Doctor viewpoint) inputs

The simple set is just a calculator for blood pressure monitor interpretation, suitable for normal users and allowing to get quick feedback; while it's highly effective the calculation methods are limited and narrow. For experts (initial concept) there are more detailed parameters for more accurate health risk, however this involves medical terms and hence normal users will struggle to use it and may be unable to use it due lack of data or specialized equipment.

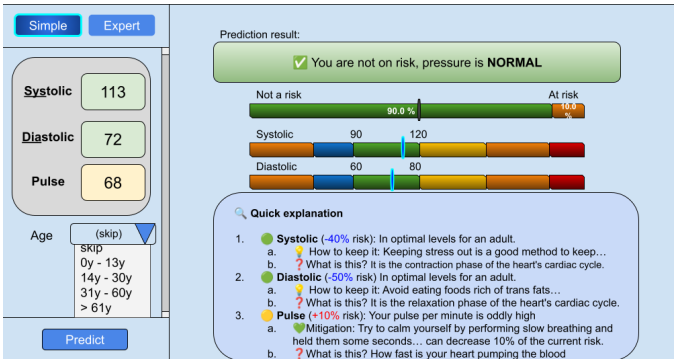

Fig. 5. Interface key concept illustration in simple viewpoint after click "predict" button

IX. Chosen Dataset

We selected the Heart Failure Prediction dataset because according to related work it was shown to be easy to handle, containing no duplicates, no missing values, and consistently yielding high accuracy in experiments.

X. Interface Implementation

A. Advanced Mode disabled

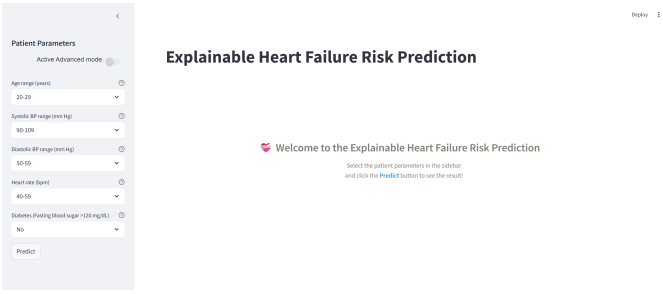

Fig. 6. Dashboard with a sidebar for selecting parameters using ranges

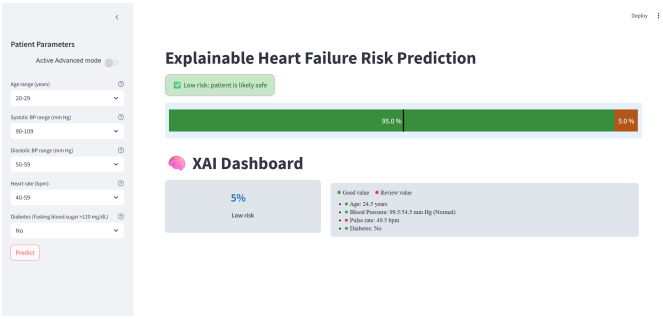

Fig. 7. Prediction with "low risk" of heart failure

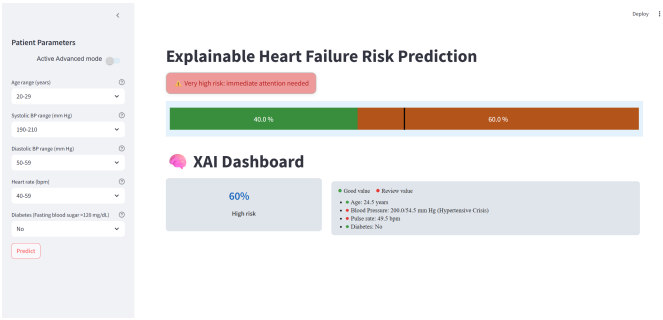

Fig. 8. Prediction with "very high risk" of heart failure

B. Advanced Mode activated

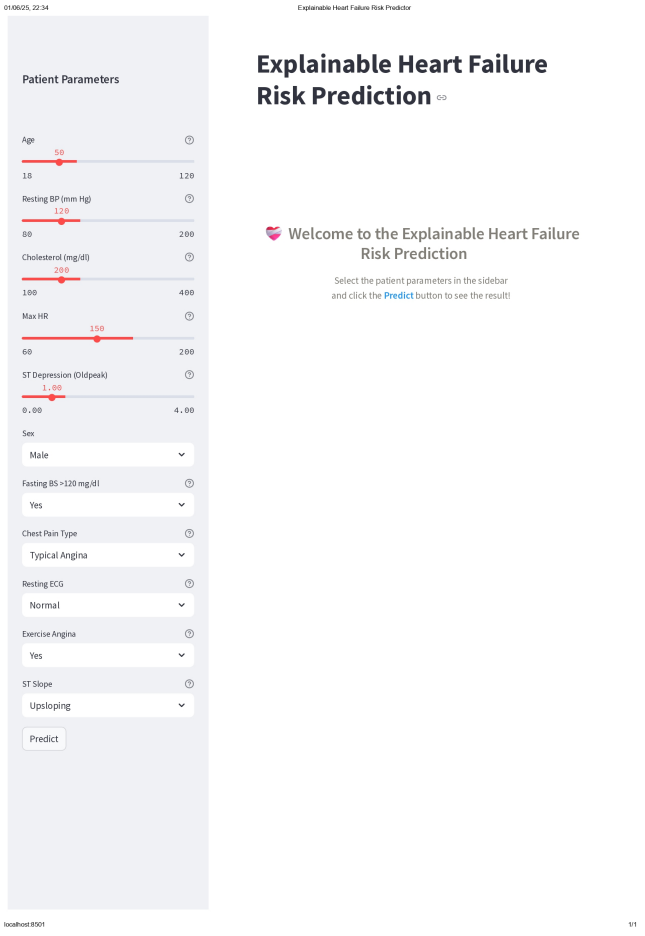

Fig. 9. Dashboard with a sidebar for selecting parameters in a specific way

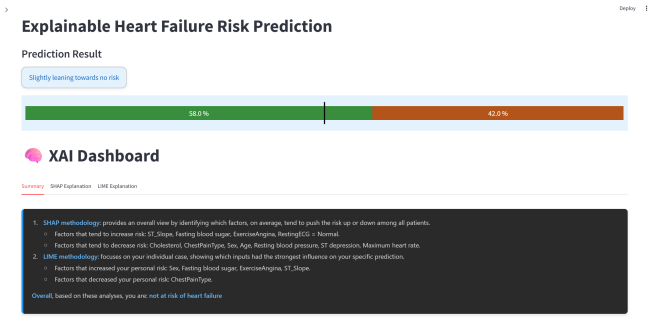

Fig. 10. Prediction with "slightly leaning towards no risk" of heart failure and summary Tab

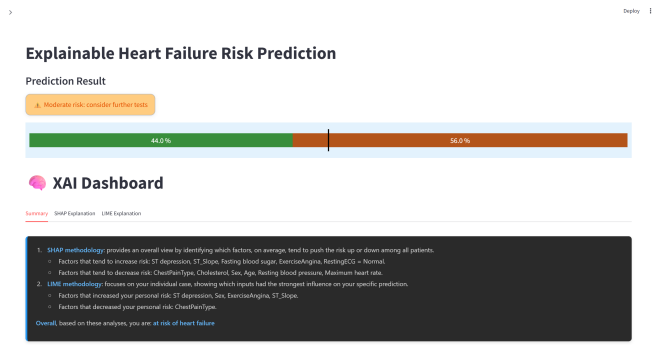

Fig. 11. Prediction with "moderate risk" of heart failure

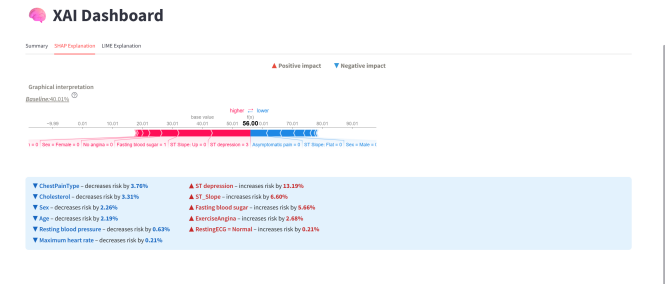

Fig. 12. SHAP explanation tab

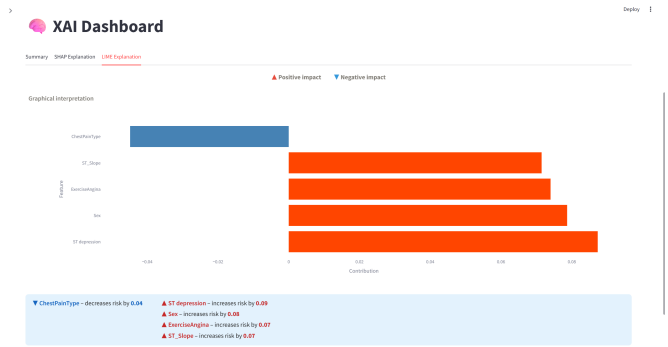

Fig. 13. LIME explanation tab

XI. XAI PROCESS AND METHODOLOGY

This section describes how we implemented explainable AI techniques in our heart failure prediction system and how these techniques specifically help different types of users understand the AI model's decisions. Our XAI approach combines two complementary explanation methods: SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), each serving distinct user needs identified through our interview process.

A. Overview of XAI Integration

Our XAI methodology follows a user-centered approach where explanation techniques are not simply added as an afterthought, but are integrated into the core system design based on specific user requirements. The process involves three main stages:

1) **Prediction Generation**: The heart failure prediction model processes input parameters and generates a risk assessment
2) **Explanation Creation**: Both SHAP and LIME algorithms analyze the prediction to create complementary explanations
3) **User-Tailored Presentation**: Explanations are formatted and presented according to the user's expertise level (Simple vs. Advanced mode)

B. SHAP Implementation and User Benefits

SHAP provides a global view of feature importance across all patients in our dataset, helping users understand general patterns in cardiac risk assessment. In our implementation, SHAP analysis serves multiple user needs:

For Patients: SHAP visualizations show which health factors generally increase or decrease heart failure risk. For example, a patient can see that "high cholesterol typically increases risk by 15%" or "regular exercise typically decreases risk by 20%." This helps patients understand which lifestyle changes might have the most impact on their health.

For Doctors: Medical professionals use SHAP results to validate the model's clinical reasoning. The global feature importance rankings help doctors identify whether the AI model aligns with established medical knowledge. If SHAP shows that age has the highest importance followed by blood pressure, doctors can confirm this matches clinical understanding.

For Data Scientists: SHAP values enable bias detection and model validation. Data scientists can examine whether certain demographic factors are inappropriately weighted or if the model shows unexpected feature interactions that might indicate training data issues.

Technical Implementation: Our SHAP implementation calculates Shapley values for each feature across the entire dataset, then presents aggregated importance scores with confidence intervals. The visualization uses a horizontal bar chart where positive values (shown in red) indicate risk-increasing factors and negative values (shown in blue) indicate risk-decreasing factors.

C. LIME Integration and Interpretation

While SHAP provides global insights, LIME focuses on explaining individual predictions, answering the question "Why did the model predict this specific risk level for this particular patient?" This local explanation approach addresses different user needs:

For Patients: LIME explanations help patients understand their personal risk factors. Instead of general statistics, LIME shows "Your high blood pressure contributes +25% to your risk" or "Your regular exercise reduces your risk by -15%." This personalized feedback makes the explanation directly relevant to the individual's situation.

For Doctors: LIME enables case-by-case validation of AI predictions. When consulting with a patient, doctors can review the LIME explanation to understand which specific factors drove the prediction, allowing them to provide targeted medical advice and validate whether the AI's reasoning aligns with their clinical assessment.

For Data Scientists: LIME helps identify edge cases and model behavior on individual instances. Data scientists can examine whether the model makes consistent predictions for similar patients or if there are unexplained variations that might indicate model instability.

Technical Implementation: Our LIME implementation creates local linear approximations around each prediction instance. It perturbs input features and observes prediction changes, then fits a linear model to explain the local behavior. The result is presented as a contribution chart showing how each feature pushes the prediction toward or away from heart failure risk.

D. Explanation Generation Pipeline

The explanation generation process follows a structured pipeline designed for real-time user interaction:

1) **Input Processing**: User parameters are validated and preprocessed to match model requirements
2) **Prediction Calculation**: The trained heart failure model generates a risk probability score
3) **SHAP Analysis**: Global feature importance is calculated and compared against the current prediction
4) **LIME Analysis**: Local explanations are generated specific to the current input values
5) **Explanation Synthesis**: Both explanation types are combined into a coherent narrative
6) **User-Appropriate Formatting**: Explanations are formatted based on the selected user mode (Simple/Advanced)

This pipeline ensures that explanations are generated consistently and efficiently while maintaining the real-time responsiveness required for interactive use.

E. Validation of XAI Effectiveness

To ensure our XAI techniques effectively address the identified user stories, we implemented targeted validation measures during our two-round testing process:

- **Simple Language Validation** [P,D: 2,2,0]: We verified that both SHAP and LIME explanations use accessible terminology instead of medical jargon, testing whether patients could understand terms like "heart stress indicators" rather than "ST Depression values"
- **Uncertainty Communication** [2,2,2]: We validated that explanations clearly present potential risks even when model confidence is low, ensuring users understand both the prediction and its reliability level
- **Risk Breakdown Clarity** [D: 1,2,1]: Medical professionals tested whether SHAP and LIME visualizations provide immediate understanding of risk factors and their relative contributions
- **Bias Detection Capability** [DS: 0,1,2]: Data scientists evaluated whether SHAP analysis effectively reveals inappropriate weighting of demographic factors versus modifiable health indicators

- **Doctor-like Explanation Quality** [P: 2,0,0]: We assessed whether patients found the AI explanations as comprehensible and trustworthy as doctor consultations

F. Validation of XAI Effectiveness

To ensure our XAI techniques effectively support user understanding, we implemented several validation measures:

- **Consistency Checking**: SHAP and LIME explanations are cross-validated to ensure they provide complementary rather than contradictory information
- **Medical Validation**: A cardiologist reviewed explanation accuracy to confirm that highlighted risk factors align with clinical knowledge
- **User Comprehension Testing**: During our two-round testing process, we specifically evaluated whether users could correctly interpret the provided explanations
- **Explanation Stability**: We tested explanation consistency across similar input cases to ensure reliable interpretation

This comprehensive approach ensures that our XAI implementation not only provides technically sound explanations but also delivers practical value to real users in healthcare contexts.

XII. First round

Prior to starting, we recruited a total of five participants representative of our user personas (some of them are the same volunteers from our interview), so we had: three patients, two of whom have no specific medical knowledge, the third is on a high-risk profile, one cardiologist doctor with their medical portable tools, and one data scientist.

A. Testing Protocol

At the start, we talked with the participants at the same time and let them know the protocol will be performed in two steps. The first step is taking a blood pressure monitor for each one, and after they get the results from the monitor, try to use our XAI in "simplified mode" to get the feedback and evaluate if the information results were effective or not. The second step is the common roles, like patients requesting health issues while the doctor has to use our XAI in "advanced mode" to determine if it is effective. Our scientist will also evaluate and guide our doctor for common usage mistakes or for not knowing the input fields.

After performing the testing and taking notes about the key findings, we requested to meet them again in the next week (round two) to test the XAI with the proposed key findings if it goes better.

B. Key Findings and Insights

In the first step and to shorten the timing, the doctor placed the blood pressure monitor cuff on the patient and took one time lecture; then the patient tried to input the values on the interface, but they experienced some issues because they didn't really know the Sys (blood pressure monitor) was the equivalent of "Systolic" in our XAI and other values. So our doctor had to guide them, and finally they had been able to receive the feedback with some effort, but it still needed some help anyway. Naturally, our doctor and scientist already knew the input fields and filled them correctly.

For the second step, the doctor (using medical tools) attempted to diagnose the patients using our XAI with the help of our scientist. The results were disappointing; moreover, the patients admitted that they did not know how to use it.

At the end of the first round, we noticed heavy issues such as the contrast in light theme mode as the texts were practically illegible, the scroll panel was making it difficult to distinguish the patient inputs vs. doctor inputs, struggle for the doctor and scientist to determine a verdict due to the lack of confidence indicator and long series of percentages, lastly there has been some parameters where the doctor and scientist were unable to figure out which one was in reality.

C. Specific changes implemented based on feedback

While the first step went notable well, the second did need notable changes, we considered to improve our XAI based on the feedback with these changes in order to fix them:

- **Basic/Advanced mode**
  A switch has been added to the top right-hand corner to switch from a simplified to an advanced compilation with a single click.
  - **Basic**: each parameter (Age, Blood pressure, Cholesterol, Max HR, ST Depression) is selected via dropdown menus with preset ranges (e.g. 20-30, 30-40 years; 120-140 mm Hg; etc.). This eliminates typing in exact numbers, reducing anxiety and the possibility of error.
  - **Advanced**: the same entries become free numeric fields to ensure maximum clinical accuracy for practitioners.
- **Prototype introductory block**
  Just below the header we have inserted a dedicated information block, designed to contextualise the app: it concisely explains the purpose of the tool and its main functionalities.
- **Two-level explanation of results**
  In order to combine clarity and technical depth, the results section has been reorganised into two sections: a 'Quick Explanation' offers an immediate and easy-to-understand summary of the result, while an 'Advanced Interpretation' provides a more in-depth analysis of the inner workings of the model.
- **Adjustment of the light theme**
  All interface components, text and icons, have been revised for the light theme.
- **Input information tooltips**
  An information icon has been added next to the title of each field: on mouseover or tap, a short descriptive tooltip appears to explain in simple terms the meaning of each parameter.
- **Model confidence badge**
  Below the model result, the probability that the patient is at risk of cardiac arrest is plotted. In this way, the user

immediately perceives not only the estimated risk level, but also how certain the model is of its assessment.

D. Rationale for Design Decisions

Design choices are based on the distinct needs of our personas and the usability principles that emerged during the research:

- **Cognitive Load Reduction**
  The "Basic/Advanced" toggle lowers entry anxiety for less experienced patients by offering selections at predefined intervals, without sacrificing the clinical precision needed by clinicians.
- **Orientation and context**
  The introduction at the top of the prototype addresses the lack of a workflow "map." A short descriptive block immediately contextualizes the tool's purpose and use cases, reducing initial disorientation and increasing user confidence before each interaction.
- **Balancing simplicity and transparency**
  The two-level explanation (Quick Explanation vs. Advanced Interpretation) reflects the principle of progressive clarity: first an immediate and understandable message is offered, then on demand the technical details. In this way, patients and professionals each find the most appropriate level of insight, without overburdening anyone.
- **Visual consistency**
  Clear theme correction and the introduction of contextual tooltips ensure readability and inline support exactly where you need it. Tooltips reduce the need for external documentation, while appropriate contrast ensures a consistent experience across both themes.

All of these decisions were guided by the goal of maximizing the effectiveness of interaction for each person, while maintaining a modular and easily extensible architecture.

E. Comparison of New Findings with Previous User Research Insights.

Early qualitative studies had already highlighted the importance of simple language for patients, transparent explanations for physicians, and XAI visualizations for data scientists. The first round of testing confirmed and expanded these insights:

- To avoid confusion with patients and doctors, we changed the buttons "simple" and "advanced" as a toggleable to select them.
- We added a welcome page in our XAI.
- After the prediction, it will show a Quick explanation and Advance interpretation.
- Revamped the light theme and contrast.
- Parameters now have a label to describe it better.
- Model now outputs the summary confidence of cardiac arrest risk.

XIII. Second round

After the week, we met them again with the five participants with potential changes we gathered from their feedback; we refreshed the protocol we did in the last week to try it again and see if our improvements worked for them.

A. Testing Protocol

We repeated the same protocol and steps in the same way we did it in round one with our XAI improvements and see if the changes were better or not.

The first step is taking a blood pressure monitor each one, get the results, use our XAI in "Simple" and see if they can use it without help. The second step is the common roles, patients attend with health issues while the doctor has to use our XAI in "advance" to resolve them while being supervised by our scientist.

After performing the testing and taking notes about the key findings, we have thanked them for their patience and time so that these improvements can make our XAI more usable and helpful in the future

B. Key Findings and Insights

Prior starting, our volunteers did already figure out there's a welcome page in our XAI, the parameters panel were changed with better visual contrast, including the toggleable button that made it more palatable.

During the first step, again with doctors taking the pressure of the patients and inserting the input values on the interface (since they already remember what is each parameter), the surprise came from the predict results as they found it changed with the quick explanation giving a confusing content, the contrast was better and did liked it, only one complain was regarding the text being somewhat smaller.

On the second step, the doctor invented another diagnose and better prepared to input the values (has previous experience and researched the meaning during the week), the toggleable did surprise because the other parameters (simple ones) was hide it and avoided more confusion, however the doctor had confusion with the predict button clicks because there's no visual info if the XAI is processing or not as it took longer than before; we figure up by the feedback from our scientist since we didn't notice the doctor struggle.

Regarding the first round, the improvements with the new findings was mostly positive:

- The toggleable button and hide other panels totally separated patient vs doctor.
- Finding a welcome page did helped to notice faster what has to do.
- The quick explanation was useful, however it didn't work well as it lacks details and tends to confuse the readers, so we had to retry again.
- The contrast change was far enough to see it clearly, however we haven't tested with color-blind user persona since none of the volunteers had it.
- Parameter tooltips information can be useful, but it needs more refinement.
- The cardiac arrest summary is currently suffice descriptive.

In addition, we have also spot more potential improvements, such as:

- **Unclear "Quick Explanation" and "baseline":** some participants hesitated when confronted with the word

"baseline", wondering what exactly the displayed percentage referred to, while some had difficulty interpreting the "Quick Explanation" box.
- **Need for a process indicator:** in the absence of visual feedback during processing, some users expressed after a few seconds, 'Is it still running?', revealing that even a short delay without any indication can weaken the perception of reliability of the tool.

C. Specific changes implemented

With the changes many being positive and barely finding anything new, our XAI it seems to be closer our user histories goal, at some point it would be convenient to address accessibility but finding a volunteers is harder that it looks; therefore our next improvements and fixes goes:

- **Redo the quick explanation**
  We've restructured the Quick Explanation box into clear bullet points under a concise heading, using everyday language to walk users step by step through the main takeaways. We hope it fixes the improvement gap from round one.
- **Loading spinner**
  We've added a visual loading indicator around the model's prediction and explanation processes, so users instantly see that the app is working and don't click repeatedly in uncertainty.
- **Baseline tooltips**
  We've introduced hover-over definitions for the term "baseline" that permit the users to have a simple definition about this parameter.

D. Rationale for Design Decisions

After working with the improvements and polishing the unsolved gaps from round one, we made the following changes:

- **Quick Explanation as a bullet list**
  Transforming the Quick Explanation into a clear heading leverages progressive disclosure: users receive the most critical insights immediately, in plain-spoken steps, without being overwhelmed.
- **Loading spinner**
  Adding a visible loading indicator around prediction and explanation routines addresses the need for clear system-status feedback. By showing an immediate "predicting" cue, we prevent repeated clicks.
- **Baseline tooltips**
  We've added a hover-over definition for the term "baseline," so users can instantly see that it represents the feature means without hunting through documentation.

E. Comparison with Previous User Research Insights

In the second round, all participants immediately identified the welcome page, appreciated the improved contrast and Basic/Advanced toggle, and navigated the interface without help. However, the revised Quick Explanation was still unclear in layout and content, users froze in the absence of visual feedback during processing, and the term "baseline" was not sufficiently explicit. Based on these findings, we implemented three key improvements:

- **Quick Explanation redesign**
  Introduced a concise header and plain-language bullet points to walk users step-by-step through the most important insights.
- **Loading spinner**
  Added a visible spinner around the "Predict" button and results section to signal immediately that the app is processing.
- **Baseline tooltips**
  Implemented hover-over definitions for "baseline," providing a contextual explanation of the feature's mean value.

Legend: numbers show how many times the issue was mentioned by patients, doctors, and data scientists, respectively.

2,2,0 **Restructure Quick Explanation**
Transformed the Quick Explanation into a concise heading with clear bullet points in everyday language, walking users step-by-step through the main takeaways.

1,1,0 **Add Loading Spinner**
Introduced a visual spinner around the Predict button and result area so users immediately see that the app is processing and avoid repeated clicks.

2,0,0 **Introduce Baseline Tooltips**
Implemented hover-over tooltips for the term "baseline," providing a simple in-context definition of feature means.

XIV. Persisting gaps and future work

While we planned and covered many user histories in our XAI, there were only two of them that we were not able to implement, in total we fulfilled 10 of the 12 histories. The ones we have not been able to cover were:

- **3º - Provide recommendations and routines that can be followed based on the risk factors present, emphasizing the risk that can be most reduced. (4/6).**
- **10º - Brief explanation of the guidelines and what you should do, acting as if you were a doctor. (2/6).**

Personalized Recommendation Engine

- Add a "What to Do" section right after the risk result, suggesting concrete actions (e.g., "Walk 30 minutes daily" or "Reduce salt intake") tailored to the user's highest risk factor.
- Use an algorithm to prioritize these recommendations by impact, showing first the actions that will lower risk the most.

Doctor-Style Explanations

- Introduce a "Doctor's Advice" box in everyday language, summarizing in 2–3 sentences what steps to take (for example, "To lower your blood pressure, try these changes…").
- Provide step-by-step guidelines (e.g., how often to measure, recommended medications, when to see a cardiologist) as if spoken by a medical professional.

XV. Conclusion

Throughout the development of our XAI-based heart failure prediction system, we have followed a user-centered and iterative approach, which allowed us to refine both the interface and the functionality based on real-world feedback. The initial phase of interviews with patients, doctors, and data scientists provided us with valuable and diverse insights into the specific needs, frustrations, and expectations of each user group. These user stories guided the selection of features to be implemented, ensuring that our model would not only offer accurate predictions, but also present them in an understandable and actionable manner.

The two rounds of usability testing played a crucial role in validating our design choices and revealing unforeseen usability challenges. In the first round, we discovered significant obstacles related to terminology, interface clarity, confidence indicators, and visual consistency. These findings motivated targeted improvements, such as introducing the Basic/Advanced mode toggle, contextual tooltips, clearer contrasts, and dual-level explanations (Quick Explanation and Advanced Interpretation). Each modification was carefully reasoned based on usability principles like cognitive load reduction, progressive disclosure, and visual feedback consistency.

The second round of testing confirmed the effectiveness of most of these improvements, but also highlighted new issues, such as the ambiguity of the ”Quick Explanation” section and the absence of a system feedback indicator during processing. Once again, we iterated on the design by restructuring explanations into bullet-point lists, adding a visible loading spinner, and providing clear tooltips for technical terms like ”baseline.” These refinements substantially improved the system’s clarity and user experience, enabling participants to interact with the tool more autonomously and confidently, regardless of their medical expertise.

Despite the progress achieved, some limitations remain. Two user stories—personalized recommendations and doctor-style explanations—could not yet be fully integrated, primarily due to the complexity of dynamically generating medically sound and personalized advice. Nevertheless, these gaps have been clearly identified as priority areas for future development. In the upcoming iterations, we plan to enrich the system with tailored recommendation engines and simplified expert guidance that simulate real-world counseling, thus further narrowing the gap between automated prediction and human-centered support.

In conclusion, this project demonstrates the importance and feasibility of embedding explainability and usability into medical AI systems. By involving stakeholders from multiple domains and repeatedly validating our design choices through iterative testing, we have built a solution that balances technical sophistication with user accessibility. The final prototype stands as a solid foundation for future improvements, with the potential to assist both patients and healthcare professionals in managing heart failure risks more effectively, while maintaining transparency and trust—two key pillars of any clinical decision support system.

References

[1] Dataset https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

[2] HeartScore https://heartscore.escardio.org/Calculate/quickcalculator.aspx?model=low

[3] Calculadora de Riesgo Cardiovascular https://fundaciondelcorazon.com/prevencion/calculadoras-nutricion/riesgo-cardiovascular.html

XVI. Biography Section

Gioele Modica My name is Gioele Modica, I am originally from Sicily (Italy) and I am a 23-year-old on an Erasmus exchange. Currently in my second year of a master’s degree in Artificial Intelligence at the University of Pisa. I hold a bachelor’s degree in Computer Science and work as a backend developer for a startup in Italy.

Giovanni Criscione I am Giovanni Criscione, a 23-year-old from Sicily in Italy. I completed my Bachelor’s degree in Computer Science at the University of Pisa, where I then continued my studies with a Master’s in Artificial Intelligence. During my second year of graduate school, I joined the Erasmus program, which brought me here to Palma de Mallorca, where I am now finishing the last courses required for my degree. Meanwhile, I have been working as a web developer in a remote capacity, a position I continue to hold while attending classes in Palma.

Christian Ortega My name is Christian Ortega, I am from the Balearic Islands (Spain) and I am 30 years old. Currently, I am studying for a Master’s Degree in Intelligent Systems at the University of the Balearic Islands (UIB). I hold a Degree in Informatics Engineering.

XVII. Acknowledgments

This article used artificial intelligence tools to support the writing process and the search for sources. In particular, ChatGPT-4 (OpenAI, https://openai.com/chatgpt) was used to help refine the writing style and identify relevant sources of information.