Federation University Australia and Deakin University Australia

Case Studies

Delphi Group Round 1

The explanations in these case studies have been reproduced from the original papers to maximize image quality and readability.

Charlotte Young
24 September 2019

Case Study 1

Context for Explanation

An after action review (AAR) board uses an explainable artificial intelligence (XAI) algorithm to create an explanation of why an Unmanned Aerial Vehicle (UAV) deviated from its predefined path.

What is the purpose of explaining?

The purpose of the explanation is to justify the decision of an artificial intelligence algorithm to deviate from a predefined path.

Who is explaining?

An algorithm is explaining the decision using nine sources of information. Seven inputs are used (time, x-coordinate, y-coordinate, heading direction, engage in attack, continue mission, steer UAV); and two outputs are used (weather conditions and distance from enemy).

Who is listening to the explanation?

The explanation's audience is the technicians and mission control team trying to decide why the UAV deviated from its predefined path.

Where is the explanation being presented?

The explanation is designed to be presented to people in a formal, office or military, environment.

When is the explanation being presented?

The explanation will be presented after the unmanned aerial vehicle's mission.

Important (Explanation Specific) Information

Rules

Five input membership functions in input [Weather zones] and three membership functions in input [Distance from enemy] give 15 total fuzzy rules [Output (Decisions/Actions of UAV)]. The fuzzy rules are listed below.

1) If (Weather is Snow) and (Enemy is too close) then (Action is Steer)
2) If (Weather is Snow) and (Enemy is Moderately Close) then (Action is Continue)
3) If (Weather is Snow) and (Enemy is Far) then (Action is Continue)
4) If (Weather is Cloud) and (Enemy is Too Close) then (Action is Attack)
5) If (Weather is Cloud) and (Enemy is Moderately Close) then (Action is Continue)
6) If (Weather is Cloud) and (Enemy is Far) then (Action is Continue)
7) If (Weather is Rain) and (Enemy is too Close) then (Action is Attack)
8) If (Weather is Rain) and (Enemy is Moderately Close) then (Action is Steer)
9) If (Weather is Rain) and (Enemy is Far) then (Action is Continue)
10) If (Weather is Thunderstorm) and (Enemy is too Close) then (Action is Steer)
11) If (Weather is Thunderstorm) and (Enemy is Moderately Close) then (Action is Steer)
12) If (Weather is Thunderstorm) and (Enemy is Far) then (Action is Steer)
13) If (Weather is Wind) and (Enemy is too Close) then (Action is Steer)
14) If (Weather is Wind) and (Enemy is Moderately Close) then (Action is Attack)
15) If (Weather is Wind) and (Enemy is Far) then (Action is Continue)

Event in Mission 1

In the first mission, the UAV is set to navigate in the environment displayed in Figure 9 (a). On the right, Figure 9 (b) shows the final step UAV has taken. As it can be seen from Figure 9, UAV travels taking into consideration five adverse weather conditions displayed in different sized and colored rectangular shapes. Three enemies are also introduced in the system in forms of small dots. The UAV begins its mission to travel in the predefined path displayed below by dashed black lines. Figure 9 (b) shows that when the UAV enters rainy weather condition (shown in blue) and an enemy is moderately close; it decides to steer (went underneath the rainy zone). That episode occurs around (200, 10) x and y coordinates respectively.

Other decisions the UAV makes cannot be displayed in the images below. Rather they are logged along with other important information that was displayed in Figure 12. The simulation set for the first mission logged data of 1990 rows. That data is used in the next section to evolve explanation.

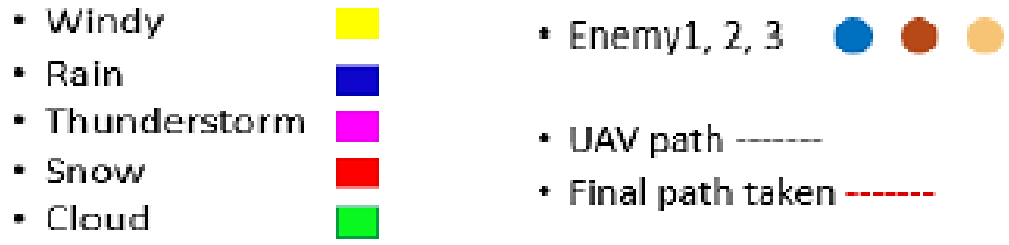


FIGURE 8. Keys for simulation diagrams

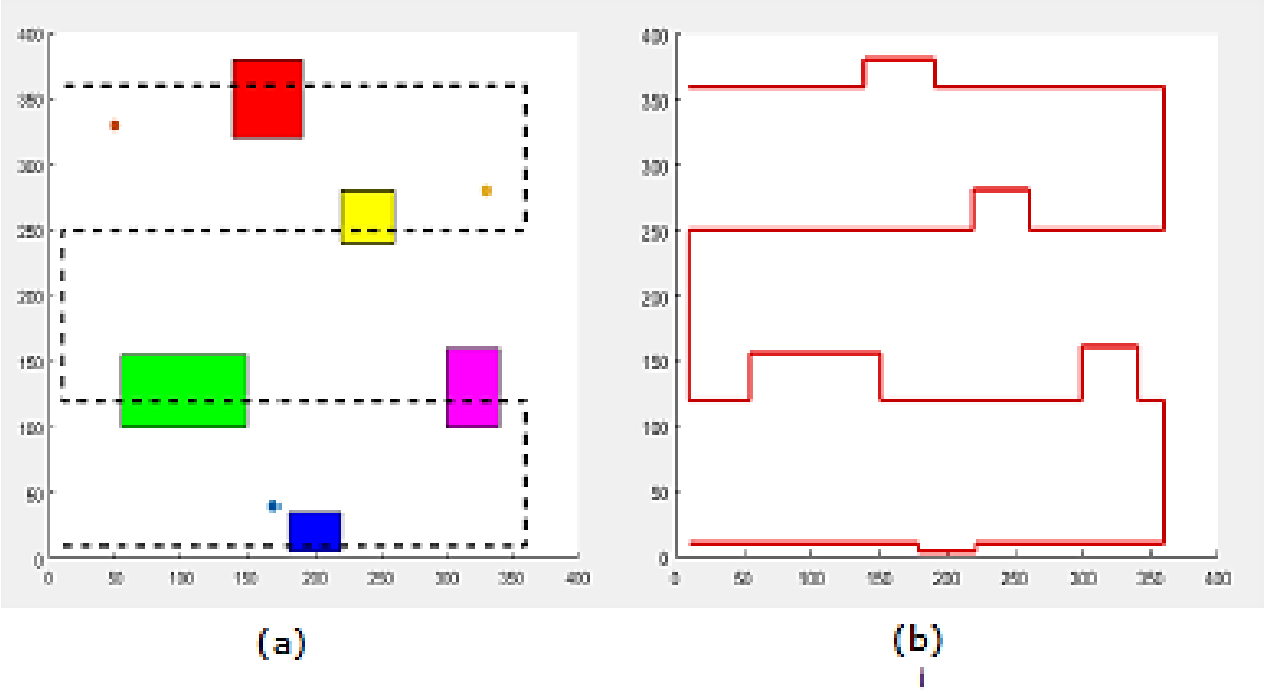


FIGURE 9. Mission 1 (a) mission set up (b) final path taken by UAV

Explanation for an Event in Mission 1

The below is a screenshot from the paper, presenting the proposed explanation

In mission 1, an example to explain what caused the UAV to engage in attack with an enemy is given. The Figure 20 shows a rule view window for Sugeno model that has weather zone as output (ANFIS 1). Whereas in Figure 21 a rule view window for Sugeno model that has UAVs distance from the enemy (ANFIS 2) as output is given. In Figure 20, rule 4 has fired, and in Figure 21, rule 5 has fired.

The English equivalent of this explanation is as follows:

Explanation 1: At time step 1094, x-coordinate 49.95, y-Coordinate 0.109, UAV was headed North, and it decided to engage in attack with enemy because it was in a sunny zone and moderately close to the enemy.

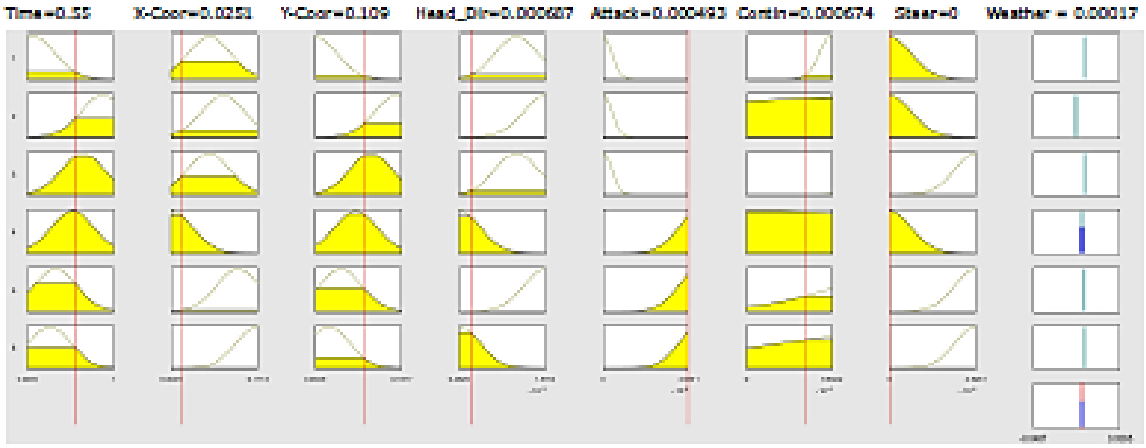


FIGURE 20. Rule viewer for XAI weather output for mission 1

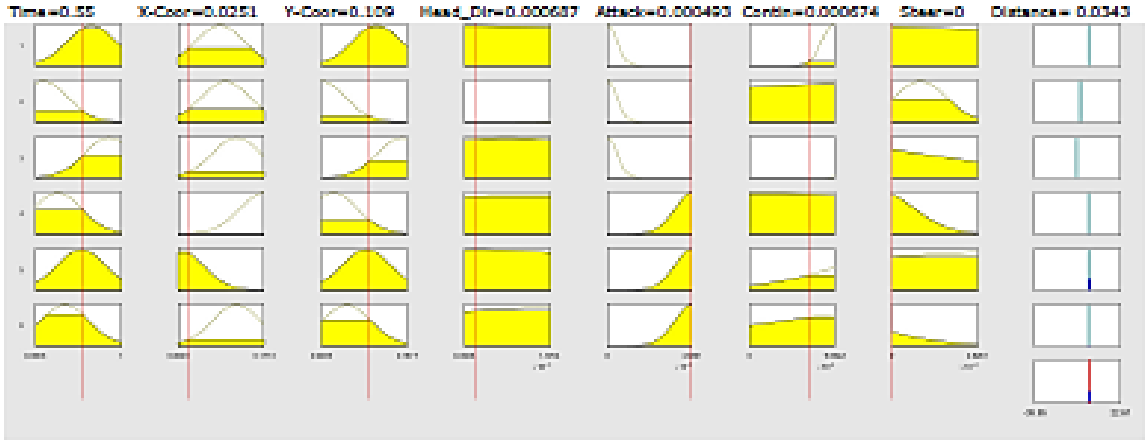


FIGURE 21. Rule viewer for XAI distance output for mission 1 Explanation

Case Study 2

Context for Explanation

A manager wishes to know what the potential issues on a manufacturing plant are, and the likelihood of the issues occurring.

What is the purpose of explaining?

The purpose of explaining is to provide managers with enough information to resolve potential issues in a timely manner

Who is explaining?

The explanation is not provided by a person or an AI; rather it is displaying in real time a graphical representation of what is happening. In essence, a computer program, not an AI, is providing an explanation to the viewer.

Who is looking at the explanation?

Line and product managers are looking at the explanation to give them adequate time to resolve potential issues.

Where is the explanation being presented?

The explanation is being presented as part of an office control panel

When is the explanation being presented?

It is presented in real time

Important (Explanation Specific) Information

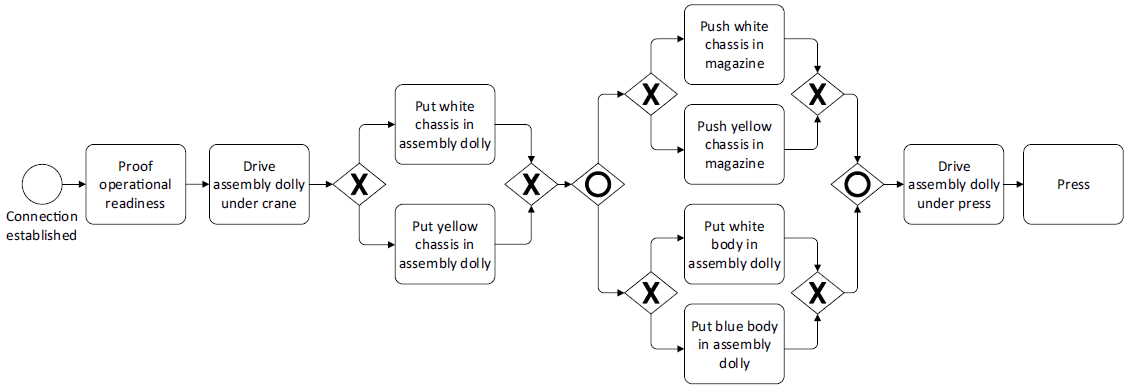
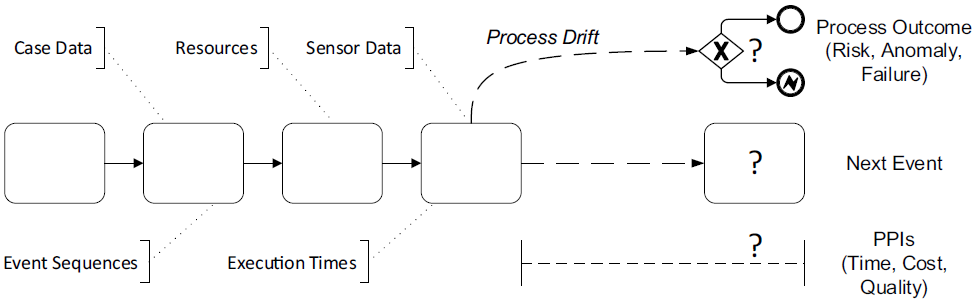
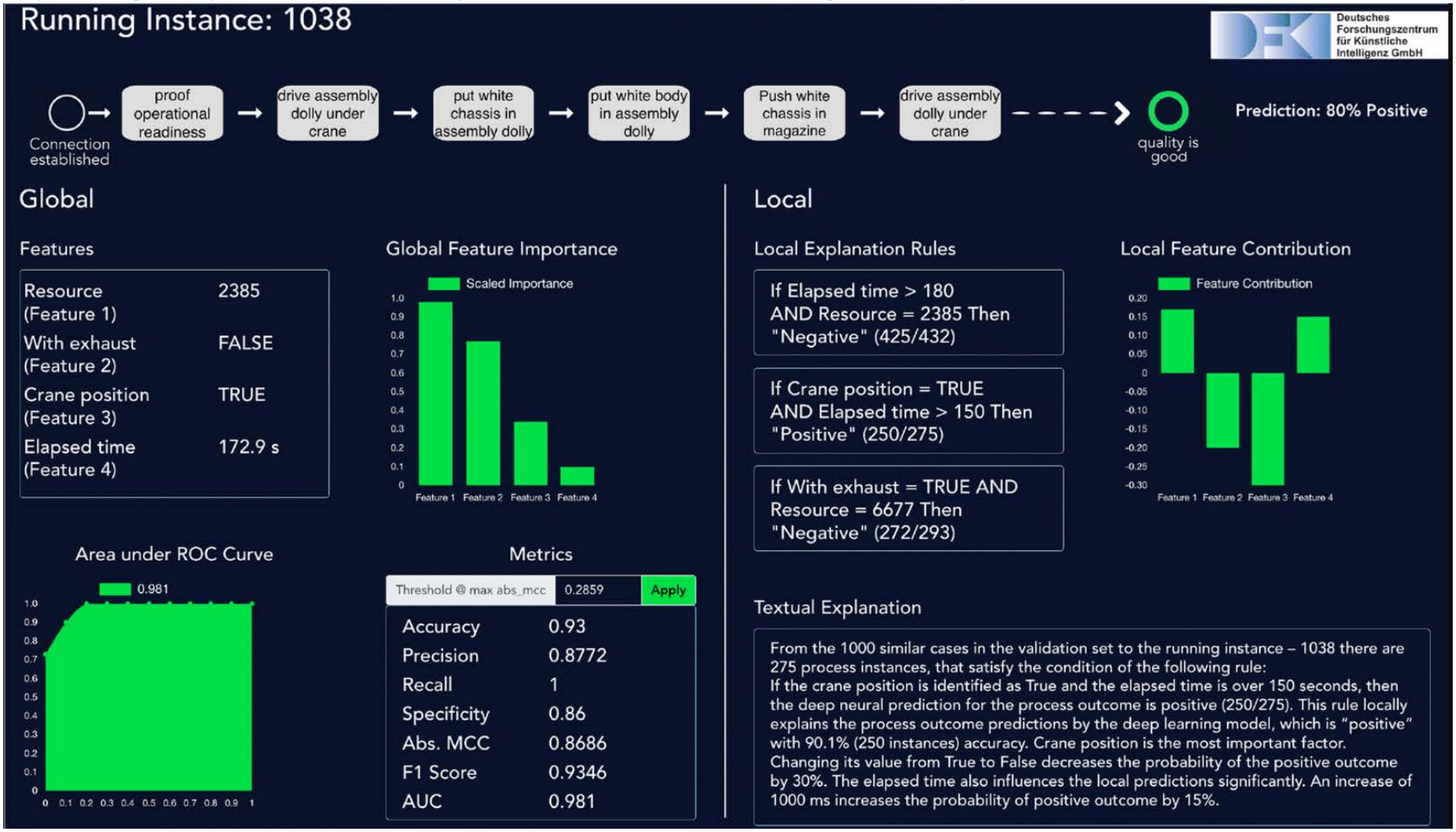


Fig. 2 The Smart-Lego-Factory production process (excerpt)



Explanation

Explaining the process outcome prediction in the Smart-Lego-Factory



"**Textual Explanation**"

"From the 1000 similar cases in the validation set to the running instance – 1038 there are 275 process instances, that satisfy the condition of the following rule:

If the crane position is identified as True and the elapsed time is over 150 seconds, then the deep neural prediction for the process outcome is positive (250/275). This rule locally explains the process outcome predictions by the deep learning model, which is "positive" with 90.1% (250 instances) accuracy. Crane position is the most important factor.

Changing its value from True to False decreases the probability of the positive outcome by 30%. The elapsed time also influences the local predictions significantly. An increase of 100ms increases the probability of positive outcome by 15%."

From the Explanation Figure

Case Study 3

Broader Context for Case Study 3

An explainable clinical decision support visualization was created to alleviate cognitive biases and to help clinicians make reliable decisions after inputting the necessary data into the program.

What is the purpose of explaining?

The purpose of explaining is to provide doctors and nurses with enough information to resolve potential issues in a timely manner

Who is explaining?

The explanation is not provided by a person or an AI; rather it is displaying in real time a graphical representation of what is happening. In essence, a computer program, not an AI, is providing an explanation to the viewer.

Who is looking at the explanation?

Case managers are looking at the explanation to give them adequate information to alleviate cognitive biases and make reliable decisions.

Where is the explanation being presented?

The explanation is being presented as part of a design making graphical user interface panel

When is the explanation being presented?

It is presented in real time

Explanation

AI-driven medical diagnosis tool

Screenshot of the AI-driven medical diagnosis tool with explanation sketches showing a patient with high-predicted risk of acute myocardial infarction (AMI), heart disease, diabetes with complications, shock, etc. Explanations include (top left) feature value time series, (top right) class attribution of predicted disease risk, (middle right) feature attribution by vitals, and (bottom) counterfactual rules indicating key rules for each prediction. Interpretation: e.g., explanations suggest that the AI thinks that the patient has shock because of low oxygen saturation and blood pressure.

Case Study 4

Broader Context for Case Study 4

A clinician requests the reasoning behind the suggested treatment and diagnosis of a patient.

What is the purpose of explaining?

The purpose of explaining is to provide doctors and nurses with enough information to resolve potential issues in a timely manner

Who is explaining?

The explanation is not provided by a person or an AI; rather it is displaying in real time a graphical representation of what is happening. In essence, a computer program, not an AI, is providing an explanation to the viewer.

Who is looking at the explanation?

Case managers are looking at the explanation to give them adequate information to alleviate cognitive biases and make reliable decisions.

Where is the explanation being presented?

The explanation is being presented as part of a design making graphical user interface panel

When is the explanation being presented?

It is presented in real time

Explanation

AI-driven medical diagnosis tool

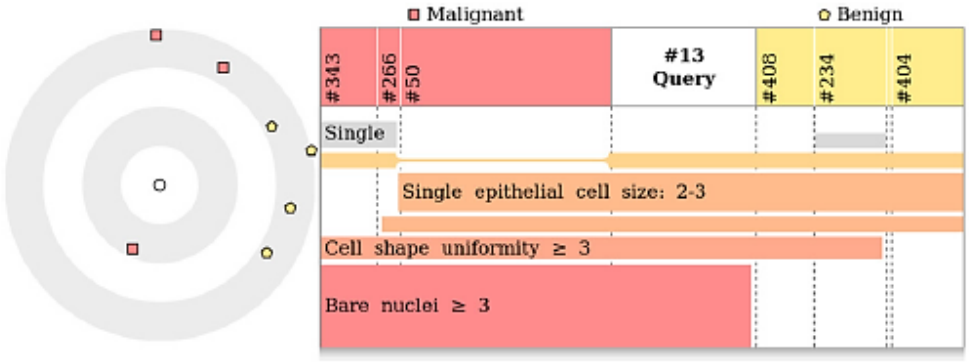


Fig. 6. Visualization of a case in the BCW dataset, with $n = 8$ cases and 7 boxes. Numbers (*e.g.* "#13") are case identifiers.

Fig. 6 shows the application of our visual interface to a case of the BCW dataset (the query case was extracted from the dataset, but its class was ignored). The "benign" class was associated with yellow, and the "malignant" one with red. We extracted 7 similar cases (n=8, counting q) and we selected at most m=11 boxes (however, fewer boxes are displayed due to the low number of attributes). In Fig. 6, in both scatter plot and rainbow boxes, we can see that 4 similar cases were benign, while 3 were malignant.

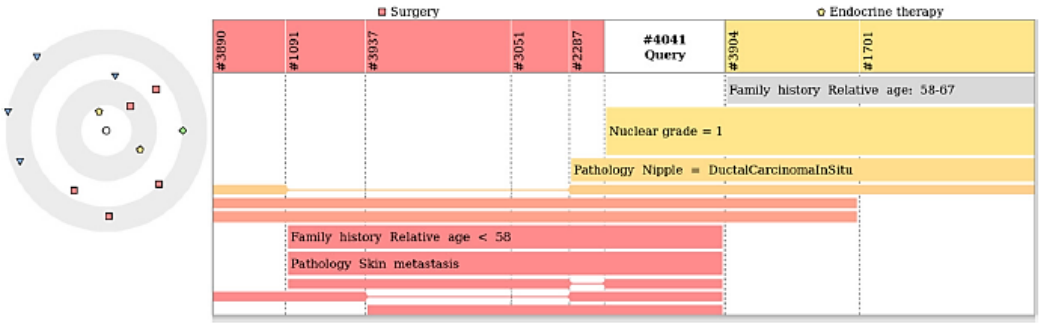


Fig. 7. Screenshot of the visual interface for CBR on the simulated dataset for the therapeutic decision.

Fig. 7 shows an example, with n=13 and m=11. The scatter plot shows that there are 5 similar cases treated by surgery, 4 by radiotherapy, 2 by endocrine therapy and 1 by chemotherapy. The closest case was treated by endocrine therapy. In rainbow boxes, only the two main classes are retained: surgery and endocrine therapy. The red color is dominant, hence, the visual interface advocates for prescribing surgery. However, the "nuclear grade=Grade1" criteria may be considered by clinicians, orienting toward endocrine therapy.

Case Study 5

Broader Context for Case Study 5

A bank customer requests the reasoning behind her bank loan rejection in order to understand why she has been rejected.

What is the purpose of explaining?

The purpose of explaining is to provide a reason for the bank loan rejection.

Who is explaining?

The explanation is provided by an algorithm that provides a graphical and textual explanation of the rejection.

Who is looking at the explanation?

A bank customer, someone who may have no special training in loans, algorithms, or A.I.

Where is the explanation being presented?

Unknown. Could be a letter or a webpage

When is the explanation being presented?

The explanation is presented after the algorithm has made its decision.

Explanation

AI Explainability 360 - Demo

Data — Consumer — Explanation

A Bank Customer wants to understand:

Why was my application rejected?
What can I improve to increase the likelihood my application is accepted?

Providing Contrastive Explanations for Insight into Loan Application Outcomes

The Bank Customer wants to know how and why the decision was made to accept or reject their loan application. The explanation given will help them understand if they've been treated fairly, and also provide insight into what – if their application was rejected – they can improve in order to increase the likelihood it will be accepted in the future. To help provide that insight and suggest avenues for improvement, we will use the Contrastive Explanations Method (CEM) algorithm available in AI Explainability 360. This algorithm sits on top of an existing predictive model and helps detect both the features that a bank customer could improve (e.g., amount of time since last credit inquiry, average age of accounts), and also further detects the features that will increase the likelihood of approval and those that are within reach for the customer. See examples below.

Select a customer asking for explanations

Jason	Ann	Julia
Denied	Denied	Denied

Several features in Jason's application fall outside the acceptable range. All would need to improve before acceptance was recommended.

Factors contributing to Jason's application denial

1. The value of **Consolidated risk markers** is **65**. It needs to be around **72** for the application to be approved.
2. The value of **Average age of accounts in months** is **52**. It needs to be around **68** for the application to be approved.
3. The value of **Months since most recent credit inquiry not within the last 7 days** is **2**. It needs to be around **3** for the application to be approved.

Relative importance of factors contributing to denial

While all three factors need to improve as indicated above, the most important to improve first is the Consolidated risk markers. Jason now has insight into what he can do to improve his likelihood of being accepted.