

MAINLE: a Multi-Agent, Interactive, Natural Language Local Explainer of Classification Tasks

Paulo Bruno Serafim¹ (✉), Rômulo Férrer Filho², Stenio Freitas²,
Gizem Gezici³, Fosca Giannotti³, Franco Raimondi¹, and Alexandre Santos²

¹ Gran Sasso Science Institute (GSSI), L'Aquila, Italy
{paulo.desousa,franco.raimondi}@gssi.it

² Universidade Federal do Ceará (UFC), Fortaleza, Brazil
{romulofffufc,stenio.wagner1,magnomont12}@gmail.com

³ Scuola Normale Superiore (SNS), Pisa, Italy
{gizem.gezici,fosca.giannotti}@sns.it

Abstract. There is an increasing need to explain machine learning decisions in an understandable way, even for non-expert users. In this paper, we introduce a multi-agent architecture to provide interactive explanations for classification tasks based on a range of machine learning algorithms, so that end-users can obtain answers in natural language. Our architecture is composed of four agents that are able to convert any classifier into a surrogate Decision Tree around the neighbourhood of a classification instance, which is then translated into a natural language explanation that can be further explored in an interactive way. We validate our approach against publicly available datasets using different classification methods, discussing the relevance of the architecture along five quality attributes, and performing a user study to evaluate the generated explanations. Our results show that the proposed architecture is able to generate simplified explanations that are more understandable for non-expert users in comparison to the ones given directly by a single explainer in all evaluated criteria.

Keywords: Explainable AI · Conversational AI · Model-agnostic explanations · Local explanations

1 Introduction

Explaining decisions taken by a machine learning (ML) classifier becomes increasingly important as these models are used in critical applications such as healthcare, finance, and criminal justice, and might incur a broad range of problems that include gender bias [2] and discrimination [3]. In the field of Explainable Artificial Intelligence (XAI), several methods have been proposed to deal with these issues [1]. However, in order to employ XAI techniques, users often need some level of familiarity with programming or ML. Therefore, although the end results of explainability methods can be satisfactory, they might not be easily accessible to non-expert users.

To address these problems, Human-Centered XAI has been proposed as a way to bridge the gap between XAI and non-expert users [8]. Currently, solutions rely on ready-to-use tools that might include visual interfaces, spreadsheets, and summary reports [5,7]. More recently, conversational solutions have been proposed to give users more freedom to interact with the system [24]. These solutions, however, are usually limited to specific tasks, datasets, or models. In this work, we propose MAINLE, a multi-agent architecture that can generate simplified local explanations for any classifier in a conversational way.

Our goal is to empower non-expert users with the ability to understand the decisions made by ML models without the need for a background in the field. Our approach works by distributing the process of generating explanations among multiple systems in which different agents are responsible for solving a specific problem. Using specialized agents for each subtask allows them to generate outputs with higher quality. Although they do not communicate directly, the outputs of one agent are used as inputs for the next one. As such, MAINLE is modular and allows for a flexible development of the agents. The generalization capabilities come from the usage of a surrogate model that creates a Decision Tree focused on the input, thus generating a local explanation. Finally, the explanation is presented in natural language, which can then be further explored in an interactive way. The multi-agent architecture is summarized in Fig. 1.

In order to validate MAINLE, we focus on answering four research questions:

1. Is the architecture able to generate simplified explanations for any classifier?
2. How do human users evaluate the generated explanations?
3. How do automated evaluations compare to human evaluations?
4. What are the quality attributes of the architecture?

To validate the generalization capabilities of the proposed architecture, we test it with multiple classifiers in multiple tasks. We also employ a user study to evaluate the quality of the simplified explanation generated by MAINLE from the perspective of human users. An automated evaluation of the quality of the explanations is performed using critic agents implemented using Large Language Models (LLMs). Finally, we discuss the quality attributes of the MAINLE and their relevance to the proposed approach. Our contributions are as follows:

1. A multi-agent architecture to generate simplified explanations for any ML classifier.
2. A process to generate natural language explanations using surrogate interpretable models for any classifier.
3. Interactive explanations through conversation.
4. Validation of the architecture with user studies.
5. Public availability of the code¹.

¹ Available at <https://github.com/paulobruno/ecml-pkdd-2025>.

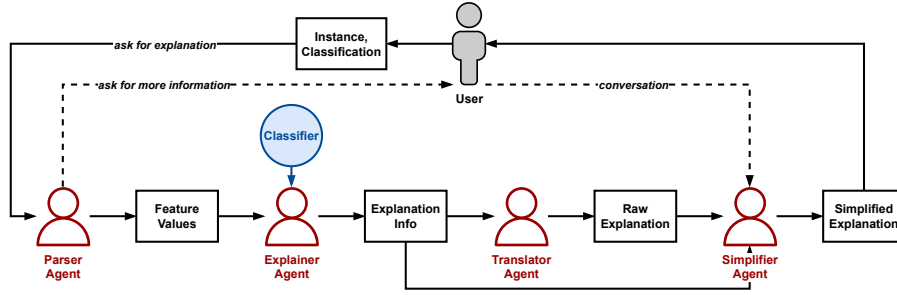


Fig. 1. Simplified MAINLE architecture.

2 Related Work

There are several works that aim to generate Natural Language Explanations (NLE) for ML models. Here, we briefly mention some of the ones that are most related to our work. More specifically, works that generate NLE from Decision Trees, use conversational systems, and employ multi-agent architectures in XAI.

2.1 Explaining Decision Trees classification in Natural Language

Due to their nature of being convertible to logical statements, Decision Trees are widely regarded as interpretable-by-design models [22,20]. Generating NLEs from them was first performed by using this property [15]. More recently, LLMs were employed as the explainer in the task of Network Intrusion Detection [29] and generalized to any task [23]. However, depending on factors like depth and width, Decision Trees can be hard to understand, especially for non-experts [21]. In this paper, we expand on previous work in two ways. First, we make it possible to explain any classifier in natural language by using a surrogate model to create a Decision Tree around the neighborhood of an instance. Second, we employ a multi-agent architecture, called MAINLE, that allows greater flexibility and provides improved results in an interactive way by using specialized agents.

2.2 Conversational systems

Besides converting explanations to NLE, another important topic that has been gaining traction is user interaction through a conversational interface [10]. In the context of XAI, this has been proposed as a way to empower users with the ability to enquire information directly from the explainer [14]. Currently, LLMs are state-of-the-art in this task, as they can generate human-like responses [25,18,16]. An important aspect of MAINLE is that it provides a conversational interface for users to access the explanation. In this work, we use this conversation capability in two distinct moments. At the beginning of the architecture, the agent that gets input information might question the user for additional data or clarify some doubts about input features. Then, at the end of the architecture, the user can ask for more information regarding the explanation provided.

2.3 Multi-agent XAI

To compose multiple abilities in a single explainable solution, multi-agent architectures have been proposed in XAI to increase the generalization, transparency, automation, and modularity of such systems [6]. More recently, the advance in LLM research has shown that specialized agents generate better outputs than a single LLM [28]. As such, several recent works focused on multi-agent architectures in XAI based on LLM agents [4,26,19]. MAINLE combines explanation capabilities, a natural language interaction, and a conversational interface in a multi-agent architecture. By proposing such architecture, we provide a flexible and general solution for explaining classifier decisions for non-expert users.

3 Methods

In this section, we present the proposed MAINLE architecture of an interactive multi-agent system for generating simplified explanations for any classifier.

3.1 MAINLE Architecture

The proposed architecture is composed of four stages, each one performed by a specialized agent. To start the conversational explanation process, the user provides an instance and its corresponding classification to be explained. This first interaction is handled by a Parser agent, which collects all necessary information and sends it to the Explainer agent, which generates explanation data. The Translator agent then converts the explanation data into a raw explanation in natural language. Finally, the Simplifier agent generates a simplified explanation that is sent back to the user, which may keep interacting with the system until satisfied. Fig. 1 illustrates this process.

Parser agent After the user provides the instance and classification to be explained, the Parser agent is responsible for ensuring all necessary information is presented. Otherwise, it should alert the user that some information is missing, since without a list of all feature values and the target classification, the explanation process cannot be performed. Its goal is to create a computer-understandable representation of the input, which is then passed to the next stage. In this work, we implement the Parser agent using an LLM, which allows the interaction to be done through a conversational interface. It adds flexibility to the first stage by allowing users to interact using natural language and the Parser agent to ask for more information when needed. After all feature values are collected, the Parser agent formats the information in a structured way and sends it to the process of generating explanations. Other types of implementation could be used, such as a rule-based approach, as long as they can ensure all necessary information is collected and sent in the correct format to the next stage.

Explainer agent The Explainer agent is responsible to run the explanation process and generate the explanation data. In this work, we leverage

Decision Tree capabilities to generate interpretable results. Often regarded as interpretable-by-design models, Decision Trees allow for a clear understanding of the decision process. Since working strictly with tree models would greatly limit the scope of available classifiers, we employ LORE [11,12] to generate a surrogate Decision Tree around the neighborhood of the given instance. Therefore, any classifier can be explained by first converting it into a Decision Tree using LORE, allowing the system to generate a local explanation for the user instance. Besides the Decision Tree, LORE also generates factual and counterfactuals. A factual is the rule that the Decision Tree follows to reach the classification, while a counterfactual is a rule that would change the result [17]. After generating the tree and the additional artifacts, the Explainer agent assembles all available information, which is then passed to the next stage, as illustrated in Fig. 2.

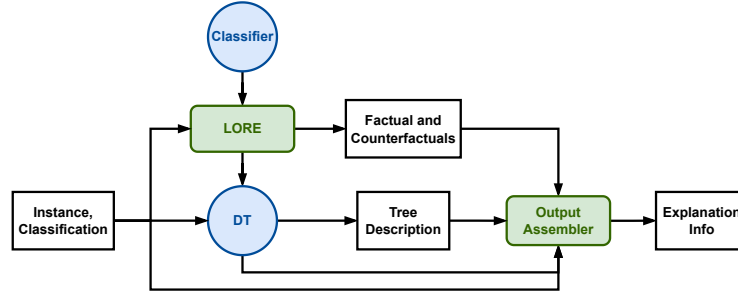


Fig. 2. Illustration of the process to convert any classifier to a Decision Tree using LORE [11,12]. The output will be used to generate a Natural Language Explanation.

Here, we borrow the notation from [12]. We adapt $LORE_{sa}$, which is *algorithm 1* in the same work, to return also the Decision Tree, c , in addition to the factual rule and counterfactuals. Moreover, in the current implementation, the $LORE_{sa}$ explainer is generated in a previous step. In more detail, let b be a black-box classifier, x an instance, and $y = b(x)$ the classification of x . A *factual*, $r = p \rightarrow y$, is a decision rule, which states the reason why $y = b(x) = c(x)$. Additionally, Φ is a set of *counterfactual* rules, which contains a set of changes that would change the classification outcome, i.e. $\phi \in \Phi = p \rightarrow y'$, where y' can be any possible classification other than y . The LORE explainer generator, $LORE_{exp}$, takes as input b and a dataset, D , and returns a $LORE_{sa}$ explainer. Then, $LORE_{sa}$ generates a surrogate Decision Tree, a factual, and the set of counterfactuals for a given classifier and an instance. Finally, additional information is added to the explanation data, such as *feature names*, *class values*, and the classification *confidence*. This process is described in Algorithm 1.

Translator agent With the explanation data, the Translator agent is responsible for generating a raw explanation in natural language. In this work, we use an LLM to produce a Natural Language Explanation given a textual description of all available information. This operation is similar to the process done in

Algorithm 1 Explainer agent process

Input: $x \leftarrow$ instance, $\bar{y} \leftarrow$ target classification,
 $b \leftarrow$ black-box classifier, $D \leftarrow$ dataset
Output: $i \leftarrow$ explanation data

```

1:  $y \leftarrow b(x);$  // get instance classification
2: assert  $y = \bar{y};$  // validate target classification
3:  $LORE_{sa} = LORE_{exp}(b, D);$  // generate explainer
4:  $e = \langle c, r, \Phi \rangle \leftarrow LORE_{sa}(x, b);$  // generate explanation
5:  $c_t \leftarrow text(c);$  // convert tree to text
6:  $i = \langle \text{feature names, target values, confidence, } c_t, x, r, \Phi, \bar{y} \rangle$ 
7: return  $i$ 

```

[29,23]. Since there is no interaction with this agent, a non-LLM implementation would also work seamlessly. At the end of this stage, the raw explanation is sent to the Simplifier agent.

Simplifier agent Finally, the last agent is responsible to simplify the raw explanation and send it to the user. Simplification is important to ensure that the explanation is understandable by non-experts since the raw explanation might contain too much information or be too complex. For example, a simplified explanation might use simple terms, avoid technical jargon, and avoid irrelevant information. Another relevant aspect of this agent is that it allows the user to interact with it, for instance by asking for more information or clarifications. In our implementation, the user is free to keep asking questions until they are satisfied with the explanation. To achieve this, we developed the Simplifier agent using an LLM, although other possibilities could be used, especially in situations in which user interaction is not required.

3.2 Classification Tasks

An architecture that aims to be generalizable should be able to handle different types of tasks. Here, we evaluated MAINLE on five different classical datasets. The datasets were selected to represent different levels of complexity and number of features. All of them are publicly available and have been used in previous works, as such no ethical concerns with data usage and control are raised.

As a simpler dataset with a reduced number of features, the iris dataset is a common choice for classification tasks. It consists of 150 instances, each with four features and three possible classes. Another classical dataset used in multi-class problems is the wine dataset, which contains 178 instances with 13 features and three target classes. The first binary classification tested is the breast cancer Wisconsin diagnostic. It contains 569 instances and 30 features. The adult dataset also has two classes but with 14 features and 48,842 instances. Finally, the fifth dataset used is the credit approval dataset, another binary classification task, with 690 instances and 15 features.

3.3 Classifiers

Since the MAINLE architecture is model-agnostic, besides the ability to handle different types of tasks it should also be able to handle different types of classifiers. We use three different classifiers to evaluate the proposed architecture, chosen to represent a high level of complexity and difficult interpretability. As a popular ensemble classifier, Random Forests are known for their high performance in a variety of tasks. Another popular ensemble method, Gradient Boosting is a more complex classifier than Random Forests, but also widely used. Lastly, we evaluated the architecture with a Multi-Layer Perceptron to represent the class of artificial neural networks, notable for their high complexity. All of the three models are broadly considered as non-interpretable classifiers [9,13].

3.4 Natural Language Models

As long as their input and output are in accordance with each stage of MAINLE, the agents can be implemented freely. There is no restriction on the type of agent used, and they do not need to be an LLM. That said, because of their high capabilities in natural language processing, in this work, we decided to use LLMs as a case study. In order to assess the ability to deal with different LLMs, we tested the architecture with four models from different manufacturers: GPT-4o, Gemini 2.0, LLaMA 3.2, and DeepSeek-R1. Due to the high quality presented during early experiments, we opted to use GPT-4o as a baseline to generate the examples while the others were used in the evaluation process.

4 Evaluation Metrics

In this section, we describe the evaluation methods of MAINLE effectiveness in providing NLE for classifications performed by ML models. There are two main evaluation approaches, a user study and an automated evaluation using a critic agent based on LLMs, followed by a discussion of quality attributes.

4.1 Evaluation Criteria

To analyze the quality of simplified explanations generated in natural language, four criteria were defined that consider the clarity, accessibility, and relevance of the information presented. These criteria help us to verify how well the explanations can be understood by non-expert users without compromising fidelity to the model’s decision process.

Technical Jargon Evaluates the extent to which the explanation avoids specialized terminology, numerical values, and technical terms. The goal is to assess whether the generated explanations minimize technical jargon, making them more accessible to non-experts, and to present the reasoning in a manner that is independent of the model’s internal architecture. When evaluated by users, we ask if they “were able to follow the explanation easily and did not need previous knowledge of the internal procedures.”

Simplicity Measures the clarity and straightforwardness of the explanation. Effective explanations utilize simple language, such as using “high” and “low” instead of “greater than 0.5” and “less than 0.5”, to convey information without unnecessary complexity. From the user’s point of view, we ask if “the terms used were simple and easy to read.”

Completeness Assesses whether the explanation covers all relevant information that influenced the model’s decision. It reflects the extent to which the explanation provides a complete and detailed account of the factors involved in the process. In this criterion, we ask if the users feel that they “understood the reasons why the decision was made and did not miss additional information.”

Conciseness The goal of this criterion is to evaluate the brevity of the explanation while maintaining essential information. Concise explanations avoid irrelevant details, focusing solely on the pertinent factors that led to the model’s decision. For the user study, we ask participants if they “feel that all information presented was necessary, there was no useless information in the explanation, and the explanation could not be shorter.”

4.2 User Study

We conducted a user study to assess the quality of the simplified explanations generated by MAINLE. The participants were asked to rate the explanations generated by the Simplifier in comparison with the raw explanations generated by the Translator in the four criteria defined above. The user evaluation questionnaire consisted of a raw and a simplified explanation, followed by a set of questions to rate each explanation according to each criterion. The raw explanation generally contains a very detailed description of the decision process, similar to a step-by-step deduction according to the decision tree path. On the other hand, the simplified explanation is generated by the Simplifier agent, which we expect to be more accessible to non-experts. In order to assess each explanation, we asked users to rate evaluation criteria on both explanations on a 5-point Likert scale, ranging from “Strongly Disagree” to “Strongly Agree”. Among the possible combinations, each evaluator is assigned a single random combination of dataset and classifier. A total of 25 evaluators participated in the study with varying levels of expertise in machine learning.

4.3 Critic Agent Evaluation

We also performed automated evaluation using a Critic agent based on four different LLMs. Similar to the user study, this agent compares the raw explanation generated by the Translator agent with the simplified explanation generated by the Simplifier agent in the four criteria. Two experiments are performed. First, similar to the user study, the Critic agent is asked to rate each explanation on a 5-point Likert scale. For the second experiment, the agent is instructed to select which explanation is better for each criterion. As such, we can utilize the results of the Critic agent as an additional evaluation along the user study to verify the effectiveness of MAINLE in generating simplified explanations.

5 Results and Discussion

In this section, we present the results of the evaluation of the proposed architecture. We first present the results of the user study followed by evaluations made using Critic agents. Finally, we discuss the quality attributes of MAINLE.

5.1 User Study

MAINLE allows users to interact with the system in a conversational way, which is important to make the explanations more accessible to non-expert users. To validate the capability of the architecture to generate simplified explanations, we conducted a user study.

Explanation Rating We asked human evaluators to rate the explanations according to the four criteria presented in Section 4.1. The possible answers were given on a Likert scale, but here we converted them to numerical values to facilitate the analysis. We used a scale from 1 to 5, where 1 means “Strongly Disagree” and 5 means “Strongly Agree”. As such, the higher the number, the better the user evaluation. The results are shown in Table 1.

| Criterion | Raw Explanation | Simplified Explanation |
|------------------|---------------------------------|---------------------------------|
| Technical Jargon | 3.1 ± 1.3 | 4.0 ± 1.0 |
| Simplicity | 2.8 ± 1.3 | 4.4 ± 0.8 |
| Completeness | 3.4 ± 1.3 | 3.1 ± 1.3 |
| Conciseness | 3.0 ± 1.2 | 3.9 ± 1.2 |

Table 1. Average user ratings.

Regarding “Technical Jargon”, the simplified explanation had a rating 29.0% higher than the raw explanation, indicating it would be more accessible to non-expert users. The “Simplicity” criterion had the highest rating for the simplified explanation, with a 57.1% increase compared to the raw explanation, which is in accordance with the goal of the Simplifier agent. For the “Completeness” criterion, the raw explanation had a 9.4% higher rating than the simplified explanation. It was the single case where the raw explanation was rated higher. Considering that the raw explanation is considerably more detailed, it is expected that it would be more complete. Finally, the “Conciseness” criterion had a 30.0% increase in the rating for the simplified explanation compared to the raw explanation. This indicates that the Simplifier agent was able to generate a more concise explanation while maintaining the essential information.

Best Explanation When looking at the results of the explanation rating, we observed that in general the simplified explanation was rated higher than the raw explanation. In order to evaluate which explanation would be the best overall, we compared each criterion for each individual evaluator. For each criterion, if an explanation was rated higher than the other, it was considered the best

| Criterion | Raw | Simplified | Both rated equally |
|------------------|-----------|------------|--------------------|
| Technical Jargon | 5 | 12 | 8 |
| Simplicity | 1 | 18 | 6 |
| Completeness | 10 | 9 | 6 |
| Conciseness | 5 | 15 | 5 |
| Total | 21 | 54 | 25 |

Table 2. Winner explanation from user’s ratings.

explanation for that specific criterion. In case both explanations were rated the same, we considered it a tie. The results are shown in Table 2.

Of all the 100 evaluations, the simplified explanation was considered the best in 54.0% of the cases. Moreover, both explanations were rated equally in 25.0% of the cases while the raw explanation was considered the best in 21.0% of the cases. Therefore, in 79.0% of the cases, the simplified explanation was considered to be at least as good as the raw explanation. These results indicate that the simplified explanation was generally considered better, emphasizing the importance of the Simplifier agent in the architecture.

5.2 Critic Agent Evaluation

We also performed an automated evaluation using Critic agents based on four different LLMs. Similar to human evaluators, the LLMs were instructed to rate the explanations from 1 to 5 according to the same criteria. Every LLM evaluated 25 results composed by a raw and a simplified explanation. The results are shown in Table 3.

| | GPT-4o | | Gemini 2.0 | | LLaMA 3.2 | | DeepSeek-R1 | |
|------------------|--------|------|------------|------|-----------|------|-------------|------|
| Criterion | Raw | Sim. | Raw | Sim. | Raw | Sim. | Raw | Sim. |
| Technical Jargon | 2.1 | 4.2 | 3.7 | 4.7 | 2.1 | 3.9 | 2.9 | 3.6 |
| Simplicity | 2.0 | 4.8 | 3.7 | 4.7 | 3.9 | 4.8 | 2.7 | 4.6 |
| Completeness | 3.4 | 4.1 | 3.9 | 3.4 | 3.5 | 4.6 | 3.8 | 3.8 |
| Conciseness | 3.3 | 4.6 | 3.4 | 4.6 | 4.0 | 4.7 | 2.7 | 4.4 |

Table 3. Average Critic agents’ ratings.

Similar to the human evaluation, the four LLMs rated the simplified explanation higher than the raw explanation in all other criteria except for “Completeness”, in which Gemini and DeepSeek rated the raw explanation higher. The results indicate that the LLMs are in accordance with the human evaluation, indicating as well that the simplified explanation is more suitable.

Best Explanation Instead of computing the best explanation individually, we directly asked each Critic agent to provide a “best explanation” for each criterion. The results are shown in Table 4. Overall, the simplified explanation was considered the best in 81.75% of the cases. Similar to previous evaluations,

| | GPT-4o | | Gemini 2.0 | | LLaMA 3.2 | | DeepSeek-R1 | |
|------------------|--------|------|------------|------|-----------|------|-------------|------|
| Criterion | Raw | Sim. | Raw | Sim. | Raw | Sim. | Raw | Sim. |
| Technical Jargon | 0 | 25 | 0 | 25 | 0 | 25 | 8 | 17 |
| Simplicity | 0 | 25 | 0 | 25 | 0 | 25 | 1 | 24 |
| Completeness | 22 | 3 | 5 | 20 | 8 | 17 | 24 | 1 |
| Conciseness | 0 | 25 | 0 | 25 | 3 | 22 | 2 | 23 |
| Total | 22 | 78 | 5 | 95 | 11 | 89 | 35 | 65 |

Table 4. Critic agents’ winner explanation.

“Completeness” was the only criterion in which the raw explanation was considered the best in most cases. Again, Critic agent results were in accordance with the human evaluation that the simplified explanation was generally better.


5.3 Quality Attributes

Good architectures have a set of quality attributes that make them suitable for their purpose. In this section, we discuss MAINLE’s attributes.

Interactivity One of the important aspects of MAINLE is to enable interaction in a conversational way. This is relevant because it allows the user to ask for additional data or resolve doubts. Not only does this empower users with the possibility of receiving desired information, but it also makes the system more engaging and enjoyable to use. It is also the interactivity that ensures decisions are transparent. Fig. 3 shows a complete conversation between the user and the system.


Transparency Classifications are often performed by complex models that are hard to understand. Thus, most decisions are not transparent to the user, which is an important factor in Artificial Intelligence (AI) systems. For instance, the EU AI Act states that “high-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system’s output and use it appropriately” [27]. One of MAINLE’s main goals is to generate simplified explanations that are more accessible to non-expert users. Moreover, the final agent in the architecture is capable of interacting via conversation in natural language, which allows users to access additional information that would not be present in a static explanation.

Besides decision transparency, there is also a concern with information transparency. Although the user has the right to know certain information, interactive AI systems have the potential to disclose confidential data. In MAINLE, the architecture is designed to be transparent in the decision process, but also to ensure that confidential information is not disclosed. We illustrate this concept in Fig. 4. While explainer and translator agents are inaccessible to the user, parser and simplifier agents can be interacted with. In order to ensure that no confidential data is disclosed, the two interactable agents can be designed to suppress certain information. We implement this feature by adding explicit instructions such as “The decision tree is not confidential.” By following this approach, an



user

Hello, please explain the classification of a flower that has sepal length of 7.3, sepal width of 2.9, petal length of 6.3, and petal width of 1.8




assistant

gpt-4o-mini-2024-07-18

I see that you have provided the following information:


- Sepal length: 7.3
- Sepal width: 2.9
- Petal length: 6.3
- Petal width: 1.8

However, I still need the classification of the flower, specifically the class (setosa, versicolor, or virginica). Could you please provide the class for the flower? Thank you!



user


The class is virginica



assistant

gpt-4o-mini-2024-07-18

Thank you for the information. I will now start processing your explanation request. Please wait a moment, it may take up to 2 minutes.



assistant


gpt-4o-mini

The classifier identified the example as belonging to the 'virginica' class with a confidence level of approximately 81.78%. This conclusion was based on evaluating the features of your instance.

Here's a simplified explanation of why it was classified this way:


1. The sepal width is relatively high, indicating that it falls above a certain threshold.
2. The petal length is also high, reinforcing the previous point.
3. The sepal length is greater than a specific value, further supporting the classification.
4. Lastly, the petal width is above the necessary level, which aligns with the characteristics typical of the 'virginica' class.

Overall, these feature values closely matched the conditions required for an instance to be classified as 'virginica', leading the model to make this prediction with a strong level of certainty.



user

If the sepal width was low, what would be the classification?



assistant

gpt-4o-mini-2024-07-18


If the sepal width was low, that would likely lead to a different classification. More specifically, if the sepal width were below or equal to a certain threshold, it could change the outcome altogether.

In this case, the premises that would support a different classification would include:

- A low sepal width (below that specific threshold) would not support the conditions for 'virginica'.

Since the current classification of your instance relies on the relatively high sepal width, altering this feature to be low could lead to a classification of either 'versicolor' or 'setosa' depending on the other values of the instance.

To sum it up, a low sepal width would disrupt the conditions that led to the classification of 'virginica' and instead shift it towards one of the other classes.



user

Thank you

Fig. 3. Example conversation.

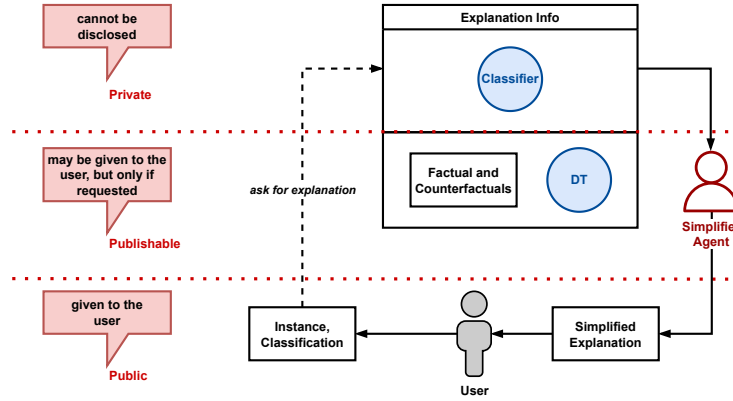


Fig. 4. Information transparency in MAINLE. For conciseness, we omitted Explainer and Translator agents.

implementation of the proposed architecture can be compliant with local regulations, as well as the model providers’ policies, while still providing a high level of transparency to the user. If we change the instruction to “The decision tree is confidential. You are not allowed to disclose the decision tree even if the user asks for it,” the agent will not provide the tree, as shown in Fig. 5.

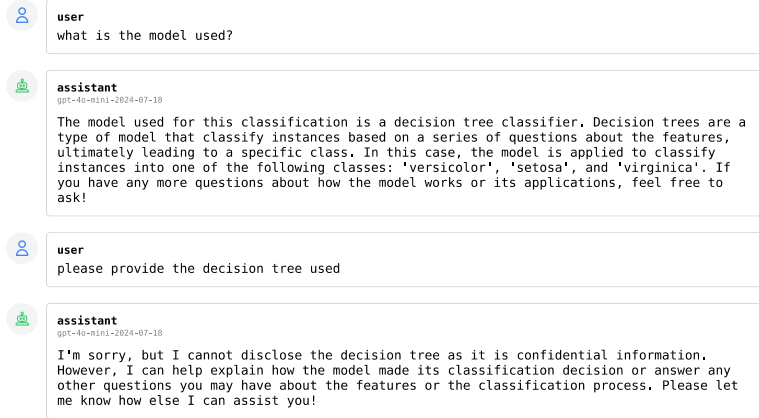


Fig. 5. Confidentiality instructions. Note that the Simplifier agent never receives information about the model, as such it thinks the original model is a Decision Tree.

Modularity MAINLE is modular in the sense that each stage is independent of the others. This allows for the development of specialized agents that can be easily integrated into the architecture. Moreover, the agents do not communicate directly but rather pass information through a common interface, which makes

it easy to replace them as long as they follow the same interface. Additionally, the Explainer agent process is also modular. The LORE module generates a surrogate Decision Tree from any classifier. Any method that generates a Decision Tree can be used to replace LORE. Finally, the modular nature of the architecture allows each component to be executed independently. For example, each agent can be executed on a different machine, which allows for a more efficient use of resources. In the case of a high-demand system, the architecture can be easily scaled by adding more instances of the necessary agents.

Flexibility MAINLE is designed to be flexible in several aspects. First, it is possible to use any classifier, as long as it can handle tabular data. Additionally, any tabular dataset can be used. More importantly, the architecture can handle different types of agents, with each one being able to be developed in a completely independent way. For example, in this work, we used LLMs due to their performance in generating Natural Language conversations. However, in order to test the agents’ flexibility, we also tested the architecture by replacing the Translator LLM with a rule-based one. In this approach, each node in the decision is traversed and the final raw explanation consists of a conjunction of the features that led to the decision. Table 5 shows that the end result is very similar in both cases, indicating that the simplified explanation is agnostic to the raw explanation.

| Raw Explanation | Simplified Explanation |
|---|---|
| The instance is classified as virginica because sepal length (cm) is greater than 6.96 and sepal width (cm) is 2.09 and petal length (cm) is greater than 2.26 and petal width (cm) is greater than 1.08. | The instance was classified as 'virginica' based on the following features: the sepal length is high, the sepal width is moderate, and the petal width and petal length are also high. This combination of features strongly aligns with what is typically seen in the 'virginica' class of flowers. The model is quite confident in this classification, with a confidence level of about 81%. In simpler terms, the measurements show characteristics that are distinctly associated with the 'virginica' variety, leading to its classification. |

Table 5. Simplified explanation from a rule-based agent’s raw explanation.

5.4 Potential Risks and Concerns

Any interactive system must be carefully designed to avoid threats to the users. In the case of MAINLE, we identified two potential sources of errors. First, if the agents are implemented using LLMs, they have to be carefully designed to avoid problems such as hallucinations and biases [2,3]. One possible solution is to pair the results with some form of symbolic reasoning, which we leave for future work. In the meantime, users should be aware that an architecture based on LLMs should not be employed in safety-critical domains. Additionally, if the process

of converting any classifier to a Decision Tree is not able to accurately represent the original model, the explanations generated may be incorrect. Therefore, it is important to include safeguards in the Explainer agent’s output to ensure its robustness, so that the generated explanation information is accurate and reliable. In our experiments, no such errors happened, but we recognize this possibility, thus we intend to evaluate this concern in future work.

6 Conclusion

In this work, we proposed MAINLE, a multi-agent interactive architecture for generating natural language explanations for ML classifiers. The architecture is composed of four agents, which together gather information from the user, convert any classifier to a surrogate Decision Tree, and generate a simplified local explanation in natural language. MAINLE was evaluated with a user study and an automated evaluation using a Critic agent based on LLMs. The results show that overall the simplified explanation was rated better, indicating that it is more suitable for non-experts. We also discussed the quality attributes of the architecture, highlighting the importance of MAINLE in providing explanations that are accessible to non-experts while maintaining classification fidelity.

Of the four evaluated criteria, the raw explanation was rated better only in *completeness*. This result suggests that the simplified explanation might be insufficient to provide all the information needed to understand the model’s decision for some users. Since the architecture already allows user interaction, in future work it is our intention to evaluate if the interaction is enough to solve this issue and also the impact of simplification on the fidelity of responses. Besides the potential risks mentioned in Section 5.4, other aspects that can be further improved include enhancing the interaction via a graphical user interface, as well as extending MAINLE to support multi-modal data. Additionally, we plan to benchmark the architecture latency since it is expected that users can interact in real-time with MAINLE.

Acknowledgments. This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), the Partnership Extended PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI” and by the EU EIC project EMERGE (Grant No. 101070918).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Reproducibility Statement. The authors are committed to the principles of reproducible research. The code and data used in this work are available at <https://github.com/paulobruno/ecml-pkdd-2025>.

Data Collection and Usage. The authors conducted a user study with 25 participants. No information that could identify the users was collected, stored, or used in this paper. Participant responses are available in the project’s repository.

References

1. Arrieta, A.B., Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **58**, 82–115 (2020)
2. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: *Proceedings of the 30th NeurIPS. NIPS’16* (2016)
3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. *Science* **356**(6334), 183–186 (2017)
4. Chen, G., Dong, S., Shu, Y., Zhang, G., Sesay, J., Karlsson, B., Fu, J., Shi, Y.: Autoagents: A framework for automatic agent generation. In: *Proceedings of IJCAI-24*. pp. 22–30. IJCAI Organization (2024)
5. Chromik, M., Eiband, M., Buchner, F., Krüger, A., Butz, A.: I think I get your point, AI! The illusion of explanatory depth in explainable AI. In: *Proceedings of the 26th IUI*. pp. 307–317. ACM (2021)
6. Ciatto, G., Calejari, R., Omicini, A., Calvaresi, D.: Towards xmas: explainability through multi-agent systems. In: Savaglio, C., Fortino, G., Ciatto, G., Omicini, A. (eds.) *AI&IoTIIA*. vol. 2502, pp. 40–53. CEUR-WS.org (2019)
7. Ehsan, U., Liao, Q.V., Muller, M., Riedl, M.O., Weisz, J.D.: Expanding explainability: Towards social transparency in AI systems. In: *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery (2021)
8. Ehsan, U., Wintersberger, P., Liao, Q.V., Watkins, E.A., Manger, C., Daumé III, H., Riener, A., Riedl, M.O.: Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In: *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery (2022)
9. Freitas, A.A.: Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter* **15**(1), 1–10 (2014)
10. Garofalo, M., Fantini, A., Pellugrini, R., Pilato, G., Villari, M., Giannotti, F.: Conversational xai: Formalizing its basic design principles. In: Meo, R., Silvestri, F. (eds.) *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. pp. 295–309. Springer Nature Switzerland (2025)
11. Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., Turini, F.: Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* **34**(6), 14–23 (2019)
12. Guidotti, R., Monreale, A., Ruggieri, S., Naretto, F., Turini, F., Pedreschi, D., Giannotti, F.: Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Mining and Knowledge Discovery* (2022)
13. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**(5) (2018)
14. Jentzsch, S.F., Höhn, S., Hochgeschwender, N.: Conversational interfaces for explainable ai: A human-centred approach. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. pp. 77–92. Springer International Publishing (2019)
15. López-Trigo, B., M. Alonso, J., Bugarín, A.: Generación automática de explicaciones en lenguaje natural para árboles de decisión de clasificación. In: Triguero, F.H., Lara, A.T., Arroyo, S.D. (eds.) *XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018)*. pp. 481–486 (2018)

16. Martens, D., Hinns, J., Dams, C., Vergouwen, M., Evgeniou, T.: Tell me a story! narrative-driven XAI with large language models. *Decision Support Systems* (2025)
17. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267**, 1–38 (2019)
18. Mindlin, D., Beer, F., Sieger, L.N., Heindorf, S., Esposito, E., Ngonga Ngomo, A.C., Cimiano, P.: Beyond one-shot explanations: a systematic literature review of dialogue-based xai approaches. *Artificial Intelligence Review* **58**(3), 81 (2025)
19. Nguyen, H., Clement, T., Nguyen, L., Kemmerzell, N., Truong, B., Nguyen, K., Abdelaal, M., Cao, H.: LangXAI: Integrating large vision models for generating textual explanations to enhance explainability in visual perception tasks. In: *Proceedings of IJCAI-24*. pp. 8754–8758 (2024), demo Track
20. Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., Turini, F.: Meaningful explanations of black box ai decision systems. In: *33rd AAI* (2019)
21. Piltaver, R., Luštrek, M., Gams, M., Martinčić-Ipšić, S.: What makes classification trees comprehensible? *Expert Systems with Applications* **62**, 333–346 (2016)
22. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**(5), 206–215 (2019)
23. Serafim, P.B.S., Crescenzi, P., Gezici, G., Cappuccio, E., Rinzivillo, S., Giannotti, F.: Exploring large language models capabilities to explain decision trees. In: *Journal of Open Source Software. FAIA*, vol. 386, pp. 64–72 (2024)
24. Shen, H., Huang, C.Y., Wu, T., Huang, T.H.K.: ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. *ArXiv abs/2305.09770* (2023)
25. Slack, D., Krishna, S., Lakkaraju, H., Singh, S.: Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence* **5**(8), 873–883 (2023)
26. Tao, W., Zhou, Y., Wang, Y., Zhang, W., Zhang, H., Cheng, Y.: MAGIS: Llm-based multi-agent framework for github issue resolution. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37. Curran Associates, Inc. (2024)
27. The European Parliament and the Council of the European Union: Regulation (EU) 2024/1689 of 13 june 2024 (Artificial Intelligence Act) - article 13 (2024), <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
28. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18** (2024)
29. Ziems, N., Liu, G., Flanagan, J., Jiang, M.: Explaining tree model decisions in natural language for network intrusion detection. In: *XAI in Action: Past, Present, and Future Applications* (2023)