Altruist: Argumentative Explanations through Local Interpretations of Predictive Models

Ioannis Mollas
iamollas@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

Nick Bassiliades
nbassili@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

Grigorios Tsoumakas
greg@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

ABSTRACT

Explainable AI is an emerging field providing solutions for acquiring insights into automated systems' rationale. It has been put on the AI map by suggesting ways to tackle key ethical and societal issues. Existing explanation techniques are often not comprehensible to the end user. Lack of evaluation and selection criteria also makes it difficult for the end user to choose the most suitable technique. In this study, we combine logic-based argumentation with Interpretable Machine Learning, introducing a preliminary meta-explanation methodology that identifies the truthful parts of feature importance oriented interpretations. This approach, in addition to being used as a meta-explanation technique, can be used as an evaluation or selection tool for multiple feature importance techniques. Experimentation strongly indicates that an ensemble of multiple interpretation techniques yields considerably more truthful explanations.

CCS CONCEPTS

• Computing methodologies → Knowledge representation and reasoning; Supervised learning; Machine learning. Philosophical/theoretical foundations of artificial intelligence; • Human-centered computing;

KEYWORDS

Interpretable Machine Learning, Explainable Artificial Intelligence, Local Interpretations, Argumentation, Model-Agnostic, Evaluation, Feature Importance

ACM Reference Format:

Ioannis Mollas, Nick Bassiliades, and Grigorios Tsoumakas. 2022. Altruist: Argumentative Explanations through Local Interpretations of Predictive Models. In 12th Hellenic Conference on Artificial Intelligence (SETN 2022), September 7–9, 2022, Corfu, Greece. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3549737.3549762

1 INTRODUCTION

While we witness a revolutionary adoption of Artificial Intelligence (AI) systems in our everyday activities, it is noticeable that many of them advance through Machine Learning (ML). As a result of the development of AI and ML, several ethical problems affecting our society have arisen, and thus the fields of Explainable AI (XAI) and Interpretable ML (IML) have emerged. Specifically, IML promises approaches for identifying discrimination phenomena in ML models [12, 21], as well as compliance of industry to legal frameworks [19]. Eventually, ML practitioners and researchers, developing stronger and more accurate models through IML, will be able to understand and explain their tasks and even identify issues, for example biases in a model, that would otherwise remain undetected.

Techniques for IML can be classified as global, exposing the entire logic of a model, and local, aiming to explain a single prediction of a model [1]. Moreover, when an interpretation technique can be applied indifferently to any ML model, we speak of a model-agnostic technique, while when it is only applicable to a specific model, we call it model-specific. Feature Importance (FI) interpretation techniques calculate the influence of each feature to the prediction, either at a global or local level. Methods such as LIME [26] (model-agnostic) and GradCam [28] (model-specific for neural networks) are two FI local techniques.

Argumentation concerns the study of how conclusions can be reached through a logical chain of reasoning, that is, claims based, soundly or not, on premises [25]. Based on this concept, different kinds of Argumentation Frameworks (AF) have been designed. IML and argumentation share the same goal of persuading someone to accept the validity of a decision. Many approaches that combine explanation and argumentation towards interactive dialogues, do so in a theoretical way. Whether explanations are arguments or not is a matter of debate in the philosophy of science. An interesting view discriminates between arguments and explanations, provided that arguments are used to justify something in doubt, while explanations are used to express an interpretation of something that is incomprehensible [3].

Several techniques used to acquire interpretations from ML models are approximations of the real interpretation, which is unknown. Therefore, their validity is questionable. This paper introduces Altruist; a preliminary method for transforming FI interpretations of ML models into insightful and valid explanations using argumentation based on classical logic. Altruist extracts the local maximum truthful part of an interpretation, providing reasons for the truthfulness justification. Given multiple interpretations, Altruist can work as a meta-explanation technique, as well as a tool to easily choose between X number of different interpretation techniques. Argumentation works as an explanation to this whole process. Altruist’s power in recognising untruthful parts of interpretations and usefulness as a meta-explanation technique is demonstrated through experimentation on ML models trained on tabular data. Altruist has innate virtues such as truthfulness, transparency and
user-friendliness that characterise it as an apt tool for the XAI community.

2 BACKGROUND

In this section, basic notions of argumentation and IML are introduced. A lot of frameworks have been developed in the area of argumentation, with a similar well-defined mathematical foundation [13, 25, 30]. Abstract [18], Bipolar [9] and Classic Logic-based [2] argumentation are some of the most well-known AFs.

Argumentation based on Classical Logic (CL) concerns a framework defined exclusively with logic rules and terms. A sequence of (heatmaps) or bounding boxes. LIME, a model-agnostic local-based interpretation for any ML model there are the global-based variants of feature explanations in the form of FI, when the input data are tabular or numerical. For such image-oriented models, providing saliency maps is an important aspect of the Area Under the Precision-Recall Curve (AUPRC), and the Intersection-Over-Union (IOU), and F1 score. However, we can only assume that humans can provide meaningful rationales if we consider the possible inner bias of each annotator [29]. Based on the concept of rationales, the ERASER benchmark, exclusively for NLP tasks, defines two metrics: comprehensiveness and sufficiency [14]. The former evaluates the interpretation observing the prediction by removing the rationales from the input, while the latter by retaining only the rationales in the input. 

Faithfulness [16], a similar metric to comprehensiveness and sufficiency, does not require annotated rationales to measure the quality of the provided interpretations. Faithfulness is applicable to FI-based techniques tested on models handling textual data, evaluating if the positive importance of a document’s sentences are true. Dealing with the drawbacks of faithfulness, infidelity defines different ways to create perturbations to test the faithfulness of an interpretation [31]. However, infidelity is examined in the context of
image data, and its output ranges in [0, +∞), making comparisons between various algorithms and datasets more challenging.

There are few works focusing on aggregating and ensembling multiple interpretation techniques. A recent work focused on aggregating FI interpretations to eventually produce interpretations with low sensitivity, high faithfulness, and low complexity [4]. Specifically, for a given instance, they identify a set of neighbours. They extract interpretations for all of them, and then they aggregate them. Another study, focusing on image classification tasks and based mainly on the stability metric (also known as robustness), aggregates both FI techniques and rule-based approaches by averaging explanation vectors [6].

Finally, recent work presented a method for providing a human-centric method of justifying a task-related prediction [5]. They examine the feature importance (weights) of features using an ML model and an interpretability technique. They provide each feature with a narrative role descriptor depending on its importance and impact on the prediction. Furthermore, they create core messages based on those narratives, using Natural Language Generation (NLG) to generate simple, brief, qualitative, and intuitive justifications for the predictions. One of the work’s limitations is that it is applied to linear ML models, which are always correct. The problem stems from the assumption that the linear model’s coefficients, or, in the case of a black-box model, the weights generated via techniques such as LIME, are correct. In the second scenario, the weights could be incorrect. In this work, we use Altruist to determine the truthfulness of the weights. Because Altruist acts on a lower level, this work can be used after Altruist.

4 ALTRUIST

Altruist aims to tackle mainly the untruthfulness of FI-based approaches, using logic-based argumentation. Altruist’s ultimate objective is to supply the largest truthful subset of an instance’s interpretation, acting as a curious user who tests various inputs to evaluate the given interpretation. Additionally, Altruist can provide reasons (i.e., arguments) to justify why this maximum subset is truthful, as well as why the features excluded from the set are untruthful. Truthfulness is introduced in Section 4.3. Finally, it can be used as a unified selection or evaluation tool between multiple FI techniques using local perturbations of instances, influenced by faithfulness and infidelity, while it can also be used as a meta-explanation ensemble of FI techniques, similar to aggregation of FIs.

In the present work, Altruist is applied to ML models trained with tabular data. Altruist can indeed be applied to textual, time-series, and even image data in future research. The methodology consists of 5 components (Figure 1). The first component includes the ML model, the second component is the interpretation technique(s), the third component is the truthfulness investigator (TI), the fourth component is the argumentation system, while the fifth component offers the final interpretation.

4.1 Machine Learning Model

The first component of the technique is the ML model to be interpreted. The ML model could be any model trained on tabular data, that is able to provide continuous values as output, e.g. a classification model which can output probabilities or regression model. This component is referred to as M, which is trained on the input dataset \( D = \{x_1, \ldots, x_N\} \), which contains \( N \) instances with \( |F| \) features, where \( F = \{f_1, \ldots, f_{|F|}\} \). Each \( x_i \in D \) instance has a set of values for the \( |F| \) features \( x_i = \{v_{i1}, \ldots, v_{i|F|}\} \). The output of this component is the prediction for an instance \( x_i \), for example \( P_M(x_i) = y_i \) in a supervised learning model.

4.2 Feature Importance Technique(s)

This next component concerns the interpretation technique(s) and is highly correlated with the M component. The interpretation technique(s) must fall within the category of feature importance, and must therefore provide explanations in the form of sets of features accompanied by an indicator of importance.

At this point, we assume that various FI approaches produce weights with a monotonic notion. This intuitive assumption is used by the majority of global and local FI techniques. This means that an end-user can expect monotonic behaviour from the ML model when altering a feature based on the weight provided by the global or local FI technique. However, there are FI techniques, like as

---

**Figure 1: Altruist’s flow chart.** "TI: Truthfulness Investigator"
SHAP, that do not presuppose monotonicity but are perceived as such by the end user.

Such techniques may be global or local, as well as model-agnostic or model-specific. The output of this component, given a specific \( x_i \), and the \( M \) component, is denoted as \( Z = \{ z_1, \ldots, z_|F| \} \), where \( z_j \in \mathbb{R} \). It is possible to have multiple interpretation techniques to let Altruist choose the best (most truthful) interpretation. Then, for \( T \) different techniques, one would have \( Z_t \), where \( t \in \{1, T\} \).

### 4.3 Truthfulness Investigator

The third component of this methodology is the Truthfulness Investigator (TI). For a specific \( x_i \), this component takes as input the FI interpretation(s) of \( x_i \) from the previous component. Based on the faithfulness and infidelity metrics mentioned in Section 3, TI investigates locally (e.g., around an instance) if the FIs are truthful or not. Altruist mimics human behaviour with this component. When an end-user receives an explanation, either local or global, they can alter values to see if the prediction changes. TI exhibits the same behaviour, but only locally, i.e., the alterations it performs are relatively small.

**Definition 3.** The expected behaviour of an \( M \) component can be \( \text{EXP} \in \{1 = \text{Increasing} (P_M(x'_i) > P_M(x_i)), -1 = \text{Decreasing} (P_M(x'_i) < P_M(x_i)), 0 = \text{Remaining Stable} (|P_M(x'_i) - P_M(x_i|) \leq \delta) \} \), where \( x'_i \) the instance with the altered value, while tolerance \( \delta \) is defined either manually by the user or is set to a default value (0.01).

| \( \text{ALT} \) | \( \text{EXP} \) |
|---|---|
| 1 | -1 |
| 1 | 0 | -1 |
| 0 | u | t | u | u |
| -1 | u | u | t | u |

Table 1: Truthfulness matrix [(t)truthful and (u)untruthful states]

**Definition 4 (Truthfulness).** The importance assigned to a feature can be defined as **truthful** when the expected changes to the output of the \( M \) model \( P_M(x'_i) \) are correctly observed with respect to the alterations that occur in the value of this feature. Thus, for both values of \( \text{ALT} \) and a given \( \text{IMP} \), the \( \text{IMP} \times \text{ALT} = \text{EXP} \) must be in accordance with the truthfulness matrix (Table 1).

Therefore, for a positive \( \text{IMP} \), these alterations should increase the prediction, for the increased value modification, and decrease the prediction, for the decreased value. For features with negative IMP, the inverse behaviour is expected. If the IMP was neutral, \( z_i = 0 \), we would expect the prediction to remain stable for both the increased and decreased values, \( v_{i,inc}^\text{inc} \) and \( v_{i,dec}^\text{inc} \), respectively, or to be altered within a very tight range \( \delta \), e.g., 0.749 to 0.750. Tolerance \( \delta \) is defined either manually by the user or is set to a default value 0.01 was selected after experimentation as default value because it represents an insignificant alteration of a prediction.

It is worth demonstrating this with an example. For a random instance \( x_i \) assigned to class \( Y \) with probability \( P_M(x_i) = 0.7 \), the feature \( f_j \), with a value of \( v_{i,j} = 1 \), has acquired an IMP \( z_j = 0.5 \) (Positive). Altruist seeks to increase and decrease the value of the feature by using Gaussian noise based on its distribution, \( v_{i,inc}^\text{inc} = 1.21 \) and \( v_{i,dec}^\text{dec} = 0.85 \). By querying the M ML model, it observes the alteration of the model’s output. In this example, for the \( v_{i,inc}^\text{inc} \) the output was raised to 0.85, and for the \( v_{i,dec}^\text{dec} \) remained stable.

This component does not judge the truthfulness of a feature, but it generates predicates in the output. The following is a specific predicate example that could be generated by this component: “The model’s behaviour by increasing \( f_2 \)’s value is not according to its Positive importance”. Such predicates are generated and used as input to the argumentation system and are fully described in the following section.

The reason why we need the following component is that it is simpler to turn this problem into an argumentation problem and solve it by using logic instead of manually coding it. In fact, using this approach, the results for the selection of features are also justifiable, so we have an all-inclusive transparent method. Finally, using logic and argumentation, the output of this component is a set of natural language arguments that can be easily used by the user, or can be even utilised in a user-chatbot dialogue.
4.4 Argumentation System

The fourth component, having as input the predicates generated by the TI component, formulates atoms of the following types:

- $a$: The explanation is untrusted
- $b$: The explanation is trusted
- $c_j$: The importance $z_j$ is untruthful
- $d_j$: The importance $z_j$ is truthful
- $e_{j,ALT}$: The model’s behaviour by altering $f_j$’s value is not according to its importance
- $f_{j,ALT}$: The evaluation of the alteration of $f_j$’s value was performed and the model’s behaviour was as expected, according to its importance.

Based on the aforementioned atoms, we can present the arguments of our AF in the form of $(\phi, \alpha)$, where $\alpha$ is the claim of the argument and $\phi$ is the support:

- $\alpha_1$: $(\{a\}, a)$
- $\alpha_2$: $(\{b, b \rightarrow \neg a\}, \neg a)$
- $\alpha_3$: $(\{c, \{c \vee \cdots \vee c_j\}, (c \vee \cdots \vee c_j) \rightarrow \neg b\}, \neg b)$
- $\alpha_4$: $(\{d_j, d_j \rightarrow \neg c_j\}, \neg c_j)$
- $\alpha_5$: $(\{e_{j,inc} \vee e_{j,dec}),(e_{j,inc} \vee e_{j,dec}) \rightarrow \neg d_j\}, \neg d_j)$
- $\alpha_6$: $(\{f_{j,ALT}, f_{j,ALT} \rightarrow \neg e_{j,ALT}\}, \neg e_{j,ALT})$

Most arguments are self-explanatory. We only briefly explain $\alpha_3$, which states that if the importance of one feature is untruthful, then the explanation is not trusted, and $\alpha_5$, which states that if one of the applicable alterations of $c_j$’s value is not according to its importance, then the importance $z_j$ is not truthful. These arguments are explained in Section 4.6. It is important to notice that the first argument, $\alpha_1$, is trivial, provided only to ease conversion of arguments into discussions.

| Types of Arguments | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ |
|--------------------|------|------|------|------|------|------|
| Rebuttals          |      |      |      |      |      |      |
| Undercuts          |      |      |      |      |      |      |
| $a_2$              |      |      |      |      |      |      |
| $a_3$              |      |      |      |      |      |      |
| $a_4$              |      |      |      |      |      |      |
| $a_5$              |      |      |      |      |      |      |

Table 2: Attack relations between arguments

We also define the attack relations between such arguments. We use the special cases of attacks, undercut and rebuttal as discussed in Section 2, presenting them in Table 2.

We can proceed to the definition of the argumentation tree $Tr$. A $Tr$ begins when an initial argument is presented as a claim and is called root argument. In the form of a counterargument, an objection (or objections) is raised. This is articulated in turn, eventually leading to a counterargument if it is feasible. In Altruist, the root argument is always $\alpha_1$. Thus, the $Tr$ is similar to the structure of the tree in Figure 2. The goal is to decide whether the root argument of this tree is defeated. We use the following judge function:

$$J(Tr) = \begin{cases} \text{Unwarranted} & \text{if Mark}(\alpha_1) = (D)efeated \\ \text{Warranted} & \text{if Mark}(\alpha_1) = (U)ndefeated \end{cases}$$

The root argument ($\alpha_1$) is U if the attacking argument ($\alpha_2$) is D. Thus, every argument is marked:

$$\text{Mark}(\alpha_j) = \begin{cases} U & \text{if Mark}(\alpha_j) = (D), \forall \alpha_j \in \text{opp}(\alpha_i) \\ D & \text{otherwise} \end{cases}$$

where the acceptable arguments (arguments without opponents/ conflicts) are marked as U, while $\text{opp}(\alpha_i)$ are the attacking arguments of $\alpha_i$.

To judge the $Tr$, we utilise a Prolog script, which outputs the arguments in a natural language form. In case the Prolog program judges the root argument $\alpha_1$ as $U$, and therefore the $Tr$ as Warranted, this means that one or more features are untruthful. Then, these features are discarded, and by re-examining the $Tr$, we expect to be Unwarranted. The output to the following component is the new reduced interpretation, or interpretations in case of many techniques, which will be $Z' = [z_1', z_2', z_3', \ldots, z_{|F|}']$, where $z_i'$ the truthful feature importances, and $z_i''$ the untruthful. An explanatory example is presented in Section 4.6.

4.5 Maximum Truthful Calculator

The previous component provides information about which features’ importances are truthful and untruthful in the interpretation, for each FI technique if more than one is used. Then, it reforms the interpretation excluding all the untruthful features $z_i''$. If there are multiple FI techniques, it chooses as the final interpretation the one with the minimum number of untruthful features, $\text{argmin}_i \{|z_i''| : z_i'' \in Z_i', i \in [0,|Z_i'|]\}$, hence the maximum number of truthful features. This interpretation provides richer details, as more features appear, and more accurate results, that can be tested by the end user. Moreover, due to the transparent nature of argumentation, Altruist can explain why a feature is excluded.
or included. A detailed qualitative experiment takes place in the following Section 5. The results of this component could be used by a system designer in a textual format, by converting the arguments into phrases in natural language. The output of Altruist is not intended to be presented directly to end users. A higher-level application, such as a chatbot, is expected to use Altruist.

4.6 Illustrative Example

To demonstrate the functionality of our methodology, we present a simple but complete example. The example depicts a user interacting with a system via dialogue (e.g., a chatbot) to understand why a prediction is made and to investigate the truthfulness of the provided explanation. We will demonstrate how employing Altruist on top of another interpretation technique, in this case, LIME, can help to prevent presenting misleading information to the end-user by detecting truthful and untruthful feature importance scores. Figure 3 illustrates this example.

Suppose there is a system solving a classification problem with only three features; Age (‘A’), Weight (‘W’) and Height (‘H’), which predicts the probability of ‘Author’s Paper Approval’. A probability of [0, 0.5) means that the paper will be rejected, while a probability of [0.5, 1] means that the paper will be accepted. John is a PhD student who is 25 years old, his weight is 62 kg, and his height is 170 cm and asks the system for a prediction. The M component of the system predicted a probability of 0.75 of his paper to be accepted. John is also given an explanation through LIME, corrected by Altruist, suggesting that his Age is positively influencing \( z_1 = 0.5 \) the probability of his paper to be accepted, while his Height has neutral influence \( z_1 = 0 \). The actual LIME explanation also included an importance value for \( W (z_2 = 0.1) \), but Altruist considered it untruthful and chose not to present it.

John is presented with arguments generated by Altruist. The first argument \( a_1 \) claimed: \( a = “The explanation is not truthful” \). John can generally raise this argument to derogate the truthfulness of the explanation. Subsequently, the second argument claims that \( b = “The explanation is truthful” \). This argument, \( a_2 \), is provided by the system, and it is a rebuttal to \( a_1 \). At this point, we can observe that, as indicated in Section 4.4, the argument \( a_1 \) is trivial, but appears crucial for the discussion’s flow.

For each one of the \( |F| \) features appearing in the explanation a claim \( \alpha_j, j \in [0, |F|] \) can be raised by John, stating that “The \( z_j \) is untruthful”. Specifically, two claims are raised \( \alpha_1, \alpha_3 \), composing argument \( a_3 \) which is an undercut attack to the argument \( a_2 \). These claims are \( \alpha_1 = “The importance of A is untruthful” \) and \( \alpha_3 = “The importance of H is untruthful” \). Note here that \( \alpha_2 = “The importance of W is untruthful” \) is not stated by John, as Altruist opted for this feature importance, and never reached John. In the end of this example, the omission of this feature is explained as well. Altruist creates another 2 claims to answer, \( d_1 = “The importance of A is truthful” \) and \( d_3 = “The importance of H is truthful” \). Two arguments \( \alpha_4, \alpha_5 \) are compounded undercutting \( \alpha_3 \).

To prove these claims, John raises four new claims \( e_1,Inc, e_1,Dec, e_3,Inc, e_3,Dec \). For example, we present two of them: \( e_1,Inc = “The model’s behaviour by Increasing A’s value is not according to its Positive importance” \) and \( e_1,Dec = “The model’s behaviour by Decreasing A’s value is not according to its Positive importance” \). Each pair of these claims (e.g., \( e_1,Inc, e_1,Dec \)) form an argument \( \alpha_{5,1} \), which is an undercut attack to \( \alpha_4 \). Finally, the last four claims are generated \( f_1,Inc, f_1,Dec, f_3,Inc, f_3,Dec \), each pair is forming an argument \( \alpha_{6,2,ALT} \) which undercuts the argument \( \alpha_{5,1} \). These claims are for example: \( f_1,Inc = “The evaluation of the alteration (Increased) of A’s value is performed and the model’s behaviour is as expected (Increased), according to its importance” \) and \( f_1,Dec = “The evaluation of the alteration (Decreased) of A’s value is performed and the model’s behaviour is as expected (Decreased), according to its importance”. However, if any of these last arguments are missing, this means that the root of the Tr, which is the argument \( a_1 \) is judged as Warrented, and therefore the explanation is untrusted. Otherwise, if all

\[
\begin{align*}
\alpha_1 & \rightarrow \alpha_2 \\
\alpha_3 & \rightarrow \alpha_4 \\
\alpha_5 & \rightarrow \alpha_6 \\
\end{align*}
\]

**Figure 3:** Argumentation tree for the simple example
the final four arguments are generated, the Tr will be Unwarranted, and the argument α₁ will be defeated.

Let’s get back to the point where Altruist chose to hide W. If the user was given the whole interpretation of LIME, including W, neither argument α₆, inc nor α₆, dec would be available to support that the importance of W was truthful. This happens because Altruist evaluated this positive importance and found it untruthful for both alterations. Then, if W was given to the end user, the claim (c₂) could have been raised, and the α₅,₂ argument would have no opponent and would attack α₄,₂, allowing the claim c₂ to support α₁ via α₃, resulting in an untruthful interpretation. Figure 4 depicts this alternative scenario, considering the elements inside the red boxes as apparent, regarding the third untruthful feature importance score.

5 EXPERIMENTS

In this section, we qualitatively showcase the ability of Altruist to identify untruthful interpretations, followed by an explanation. Moreover, we quantitatively test Altruist on a range of FI techniques. Altruist is evaluated in 3 separate datasets, 3 uninterpretable ML models and 1 interpretable (Table 3), as well as 4 FI techniques. Nevertheless, the following experiments are not intended to identify the best model or the best FI technique, but to highlight the fact that untruthful features can be identified using Altruist. The experiments’ source code are available in Altruist’s public repository (https://github.com/iamollas/Altruist).

We are utilising the datasets: Banknote (identify real/fake banknotes) and Heart Statlog (predict absence/presence of a heart’s disease) [17], and Adult Census [20] (predict if income exceeds 50K/yr or not), and the ML models: Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF) and Neural Networks (NN). To provide the unbiased performance of each algorithm, 10-fold cross-validation grid searches are performed¹. The results are presented in Table 3. The interpretation techniques selected for this set of experiments are PI, LIME, SHAP, and when available the models’ intrinsic interpretation. Only LR and RF can provide intrinsic and pseudo-intrinsic² interpretations.

5.1 Qualitative

For the qualitative experiments, we select the Banknote dataset, due to the small number of features, which makes the example easier to follow. The ML model selected is the SVM, which achieves the perfect F₁ score (100%). The Banknote dataset contains instances with 4 features \(F = [f₁, f₂, f₃, f₄]\), where \(f₁\) is the variance, \(f₂\) the skew, \(f₃ \) the curtosis and \(f₄ \) the entropy. We did various experiments, two of which are presented here.

We select the following random instance: \(xᵢ = [0.38, 0.78, 0.76, -0.45]\). The SVM classified the \(iᵢ\) banknote as fake with a 81.45% probability. In Figure 5, original interpretations provided by LIME and PI for the SVM’s prediction are shown. Altruist discovered untruthful importance given to variance and skew in LIME’s result, as well as entropy’s importance provided by PI. To evaluate the judgement of Altruist, we conducted the following manual experiments.

¹The grid search parameters can be found in GitHub: https://github.com/iamollas/Altruist. In addition, the optimal set of parameters for each model per dataset, the selection and engineering of features and the undersampling strategies (used in the Adult Census, 1000 instances, 80 features) can also be found in the repo.

²A procedure similar to PI, as proposed in the original paper.
| LIME | SHAP | PI | IN | AL | LIME | SHAP | PI | IN | AL | LIME | SHAP | PI | IN | AL |
|------|------|----|----|----|------|------|----|----|----|------|------|----|----|----|
| LR   | 49.74% | 37.30% | 0.00% | 0.00% | 0.00% | 61.71% | 51.79% | 38.46% | 0.00% | 36.87% | 18.54% | 23.96% | 17.18% | 0.00% | 13.36% |
| SVM  | 41.09% | 56.45% | 29.19% | -    | 25.91% | 59.20% | 48.80% | 46.21% | -    | 40.91% | 25.23% | 12.05% | 10.79% | -    | 10.06% |
| RF   | 91.56% | 77.90% | 87.54% | 91.27% | 74.00% | 76.61% | 75.98% | 84.87% | 73.87% | 66.10% | 70.01% | 13.08% | 24.83% | 62.61% | 13.08% |
| NN   | 39.89% | 47.83% | 17.26% | -    | 14.43% | 58.86% | 47.58% | 69.23% | -    | 43.73% | 21.51% | 26.64% | 17.89% | -    | 14.66% |

Table 4: Percentage of untruthful feature importances per interpretation technique, for the 3 dataset, among the 4 different classifiers. The most truthful technique for each of the models per dataset is denoted with bold. Altruist is the ensemble of the LIME, SHAP and PI techniques. (IN = Intrinsic, AL = Altruist)

Figure 5: Interpretation of SVM’s classification

LIME assigned a negative importance to the variance feature. However, Altruist found it untruthful. We have therefore proceeded to two tests to ensure that Altruist was correct. With respect to the distribution of the variance feature, with range [−7.04, 6.83], mean = 0.43, and std = 2.84, initially, we increased the value of variance from 0.38 to 0.74 and the probability increased from 81.45% to 93.49% instead of decreasing. We also altered the variance’s value from 0.38 to 0.02, and the probability decreased from 81.45% to 56.46% instead of increasing. Therefore, Altruist correctly concluded that the importance of this feature was untruthful.

With respect to the distribution of the entropy feature, with range [−8.55, 2.45], mean = −1.19, and std = 2.1, initially, we increased the value of entropy from −0.45 to −0.08 and the probability decreased from 81.45% to 71.19% instead of increasing. We also altered the entropy’s value from −0.45 to −2.23, and the probability increased from 81.45% to 97.57% instead of decreasing. Hence, Altruist correctly concluded that the importance of entropy given from the PI, was untruthful.

We are presenting one more example using the LR model for the same dataset. We select one more random instance: \( x_j = [-3.38, -13.77, 17.92, -2.03] \). The LR classified the \( j \)th banknote as fake with a 47% probability. In Figure 6, original interpretations provided by LIME and the actual weights of the LR (intrinsic) are shown. Altruist discovered untruthful importance given to curtosis and entropy in LIME’s result, and no untruthful importances to the intrinsic interpretation, as it was expected considering the interpretable nature of LR. To evaluate the judgement of Altruist, we conducted the following manual experiments.

We will test the importance given to curtosis by LIME. The distribution of the curtosis feature has a range of [−5.29, 17.92], mean = 1.4, and std = 4.31. We cannot further increase the value of curtosis because it already has the maximum value (17.92). Thus, we will only test it by decreasing slightly the value from 17.92 to 16.77. We see that the probability dropped from 47% to 8.01% rather than increasing, as implied by the positive importance value. Furthermore, we already know that curtosis is a significantly influential feature of LR based on its real weights. We may indeed say that Altruist correctly detected LIME’s erroneous importance assigned to curtosis.

These experiments demonstrate how Altruist attempts to mimic human behaviour to evaluate an explanation. If given such an explanation, a user would, naturally, attempt two alterations to see how the model behaves. It is worth noting that the alterations made by Altruist are indeed small, in terms of the feature’s distribution. As seen in the preceding examples, we are slightly altering the feature’s value each time.

5.2 Quantitative

To quantitatively evaluate Altruist’s ability to detect untruthful features, as well as to select the best interpretation technique among many, we will test it in 3 datasets, 3 uninterpretable ML models and 1 interpretable (Table 3). The results are visible in Table 4, describing the mean ratio of untruthful features’ importance per interpretation.

Banknote: In this case, the SVM achieves the higher \( F_1 \) (100%). Among the 4 models, LR provides the most truthful interpretations, second comes the NN, and third the SVM. All interpretation techniques struggle to provide truthful explanations for the RF.

Heart (Statlog): Here, the SVM achieves the higher \( F_1 \) (82%). Among the 4 models, LR provides the most truthful interpretations, second comes the SVM with PI, and third the NN
with SHAP. At the same time, every interpretation technique struggles to provide truthful explanations for the RF.

**Adult Census:** For this dataset, the SVM achieves the higher $F_1$ score (96%). Among the 4 models, LR provides the most truthful interpretations, second comes the SVM with PI technique, and third the RF with SHAP. In contrast to the other 2 test cases, SHAP provides reasonable interpretations for the RF.

The effect of Altruist is not mentioned in the aforementioned. Based on the experiments referred to above, we can infer that Altruist is a critical tool for evaluating interpretations given by non-intrinsic techniques such as LIME, SHAP, and PI, detecting on average 43.79% of untruthful features in 35 out of 36 tests. Altruist can be used effectively as a selection tool (an ensemble), achieving the lowest percent in every case. Provided that no interpretation approach has prevailed over the others, it appears to be an ideal tool for selecting the best technique automatically in a setup where several techniques are used in parallel to interpret an instance’s prediction.

In terms of Altruist’s scalability, we can state that its performance is proportional to the model’s inference time and the response time of the interpretation techniques. However, we measured Altruist’s response time, which, on average, evaluates an FI in 0.05 seconds (measured across the 4 models in the 3 datasets). The experiments run in a personal PC with an Intel i7-9700 CPU@3.00GHz, with 16GB of RAM.

### 6 CONCLUSION

IML has emerged as an important research area to interpret ML algorithms. A lot of IML approaches are presenting their interpretations in a feature importance manner. Argumentation is the study of how conclusions can be reached through logical reasoning; that is, claims based, soundly or not, on premises. This paper presents Altruist, a preliminary technique which combines FI interpretation techniques and logic-based argumentation, to provide truthful interpretations on the decision-making of ML models to the end users. It provides the local maximum truthful interpretation, as well as the justification for the truthfulness. Altruist is also presented as an evaluation metric, strongly influenced by the metrics of faithfulness and infidelity. Moreover, it can be used as a tool for automatic selection of the most truthful interpretation among a variety of multiple different interpretations. In future work, the meta-explanation aspect of Altruist is going to be explored. Altruist will also be evaluated in other ML tasks (e.g., regression, multilabel classification) and data types (text, time-series and possibly image). Another aspect that could be investigated is the alteration of categorical feature values, and the effect of the localness of features. Finally, a human-oriented evaluation will be conducted to validate the usefulness of the meta-explanations and the available arguments.

### ACKNOWLEDGMENTS

This paper is supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825619 (AIEU Project).

### REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6 (2018), 52138–52146. https://doi.org/10.1109/ACCESS.2018.2870952

[2] Philippe Besnard and Anthony Hunter. 2009. Argumentation Based on Classical Logic. Springer US, Boston, MA, 133–152. https://doi.org/10.1007/978-0-387-98197-0_7

[3] Floris Ber and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. Argument & Computation 7, 1 (2016), 55–68.

[4] Umang Bhatt, Adrian Weller, and José M. F. Moura. 2020. Evaluating and Aggregating Feature-based Model Explanations. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI 2020. 3016–3022. https://doi.org/10.24963/ijcai.2020/417

[5] Or Biran and Kathleen R. McKeown. 2017. Human-Centric Justification of Machine Learning Predictions. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, Carlos Sierra (Ed.). ijcai.org, 1461–1467. https://doi.org/10.24963/ijcai.2017/202

[6] Szymon Bobek, Pawel Balaga, and Grzegorz J Nalepa. 2021. Towards Model-Agnostic Ensemble Explanations. In International Conference on Computational Science. Springer, 39–51.

[7] Leo Breiman. 2001. Random Forests. Machine Learning 45, 1 (01 Oct 2001), 5–32. https://doi.org/10.1023/A:1010933408432

[8] Lucas Carpentis and Francesca Toni. 2017. Using argumentation to improve classification in natural language problems. ACM Transactions on Artificial Intelligence Technology (TOMT), 17 (3) (2017), 1–23.

[9] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Springer, 378–389.

[10] Kristijonas Cyra, Oana Cocarascu, and Francesca Toni. 2018. Explanatory Predictions with Artificial Neural Networks and Argumentation. In IJCAI/ECAI Workshop on Explainable Artificial Intelligence (XAI 2018).

[11] Kristijonas Cyra, Ken Satoh, and Francesca Toni. 2016. Abstract Argumentation for Case-Based Reasoning. In Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016. Cape Town, South Africa, April 25-29, 2016, Chitara Baral, James P. Delgrande, and Frank Wolter (Eds.). AAAI Press, 549–552. http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12879

[12] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. San Francisco, CA: Reuters. Retrieved on September 9 (2018), 2018.

[13] Martijn Demolin, Qurat-ul-ain Shaheen, Katarzyna Budzynska, and Carles Sierra. 2020. Argumentation Theoretical Frameworks for Explainable Artificial Intelligence. In NLA@I/BGL, Association for Computational Linguistics, 44–49.

[14] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-9, 2020. Dan Jurafsky, Joyce Chai, Natalie Schlieter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 4443–4458. https://doi.org/10.18653/v1/2020/acl-main.408

[15] Pedro Domingos. 1998. Knowledge discovery via multiple models. Intelligent Data Analysis 2, 1-4 (1998), 187–202.

[16] Mengnan Du, Ninghao Liu, Fan Yang, Shuiwang Ji, and Xia Hu. 2019. On attribution of recurrent neural network predictions via additive decomposition. In The World Wide Web Conference. 383–393.

[17] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[18] Phan Minh Dung. 1995. An argumentation-theoretic foundation for logic programming. The Journal of Logic Programming 22, 2 (1995), 151–177.

[19] GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union (OJ) 59, 1-88 (2016), 294.

[20] Ron Kohavi. 1996. Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision-Tree Hybrid. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, to appear.

[21] Jeff Larson, Surya Matta, Lauren Kirchner, and Julia Angwin. 2020. How We Analyzed the COMPAS Recidivism Algorithm. shorturl.at/gwEU

[22] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 4765–4774. shorturl.at/bdCT4

[23] Alexander Moore, Vanessa Murdoch, Xaoyong Cai, and Kristine Jones. 2018. Transparent tree ensembles. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 1241–1244.

[24] Martin Mozina, Jure Zakhar, and Ivan Bratko. 2007. Argument based machine learning. Artificial Intelligence 171, 10-15 (2007), 922–937.
[25] Iyad Rahwan and Guillermo R Simari. 2009. *Argumentation in artificial intelligence.* Vol. 47. Springer.
[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2 2016).
[27] Tulio Ribeiro, Marco, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *32nd AAAI Conference on Artificial Intelligence.* 1527–1535.
[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7
[29] Chenchao Tan. 2021. On the Diversity and Limits of Human Explanations. arXiv:2106.11988 [cs.CL]
[30] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36 (2021).
[31] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. 2019. On the (In)Fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, Vancouver, BC, Canada.* 10965–10976. https://proceedings.neurips.cc/paper/2019/hash/a7471fde77b3435276507cc8f2dc2569-Abstract.html
[32] Omar Zaidan, Jason Eisner, and Christine D. Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA,* Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai (Eds.). The Association for Computational Linguistics, 260–267. https://aclanthology.org/N07-1033/