

Computing Improved Explanations for Random Forests: k -Majoritary Reasons

Louenas Bounia¹ ^a and Insaf Setitra²

¹LIPN-UMR CNRS 7030, Université Sorbonne Paris Nord, Villetaneuse, France

²Heudiasyc-UMR-CNRS 7253 UTC de Compiègne, Compiègne, France

Keywords: Interpretability, Random Forests, Abductive Explanations, Combinatorial Optimization.

Abstract: This work focuses on improving explanations for random forests, which, although efficient and providing reliable predictions through the combination of multiple decision trees, are less interpretable than individual decision trees. To improve their interpretability, we introduce k -majoritary reasons, which are minimal implicants for inclusion supporting the decisions of at least k trees, where k is greater than or equal to the majority of the trees in the forest. These reasons are robust and provide a better explanation of the forest's decision. However, due to their large size and our cognitive limitations, they may be too hard to interpret. To overcome this obstacle, we propose probabilistic majority explanations, which provide a more concise interpretation while maintaining a strict majority of trees. We identify the computational complexity of these explanations and propose algorithms to generate them. Our experiments demonstrate the effectiveness of these algorithms and the improvement in interpretability in terms of size provided by probabilistic majority explanations (δ -probable majority reasons).


1 INTRODUCTION

Context. Understanding the predictions made by machine learning (ML) models is a crucial issue that has prompted significant research in artificial intelligence in recent years (see, for example, (Adadi and Berrada, 2018; Miller, 2019; Guidotti et al., 2019; Molnar, 2019; Marques-Silva, 2023)). This paper focuses on classifications made by random forests, a popular ensemble learning method that builds set of decision trees during the training phase and predicts by taking a majority vote among the base classifiers (Breiman, 2001). The randomization of decision trees is achieved through data subsampling (or bagging), making random forests easy to implement with few parameters to adjust. They often provide accurate and robust predictions, even for small data samples and high-dimensional feature spaces (Biau, 2012). For these reasons, random forests are used in various applications, including computer vision (Criminisi and Shotton, 2013), crime prediction (Bogomolov et al., 2014), and medical diagnostics (Azar et al., 2014).

However, random forests are often considered less interpretable than decision trees. While many XAI

queries (Audemard et al., 2020) are tractable for decision trees, they are not for random forests (Audemard et al., 2021). The prediction on a data instance can be easily interpreted by following the direct reason provided by the classifier (Audemard et al., 2022b). For a decision tree, this corresponds to the unique path from the root to the decision node that covers the instance (or explanation restricted to the path) (Izza et al., 2020). For random forests, the authors of (Audemard et al., 2022c) define the direct reason as the union of the direct reasons from the trees that vote for the predicted class. A key challenge is to formulate abductive explanations, that is, to concisely explain why an instance is classified as positive or negative.

Related Work. Explaining random forest predictions has garnered increasing attention in recent years. Several recent works (Bénard et al., 2021; Audemard et al., 2022c; Audemard et al., 2022a; Choi et al., 2020; Izza and Marques-Silva, 2021) have focused on prime implicant explanations for instances given a random forest, also called sufficient reasons (Darwiche and Hirth, 2020). Simply put, if a random forest classifier is seen as a Boolean function f , a prime implicant explanation for a data instance x classified as positive by f is a minimal implicant for the inclu-

^a  <https://orcid.org/0009-0006-8771-0401>

sion of f that covers \mathbf{x} . For a single decision tree, this explanation can be generated in linear time. However, determining whether a term is a prime implicant explanation for a random forest is a DP-complete problem (Izza and Marques-Silva, 2021). Despite this complexity, algorithms based on **Minimal Unsatisfiable Subset** (Liffiton and Sakallah, 2008) can be efficient in practice. Nevertheless, for high-dimensional instances or large forests, deriving a sufficient reason becomes difficult. To overcome this challenge, (Audemard et al., 2022c) proposed *majoritary reasons*, which are minimal implicants for the inclusion for the majority of the trees and can be derived in linear time. Additionally, there are model-agnostic approaches, such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and Anchors (Ribeiro et al., 2018), although these methods have the drawback of producing explanations that may be consistent with multiple predicted classes, reinforcing the interest in formal methods (Ignatiev et al., 2019; Marques-Silva and Huang, 2023).

A significant limitation of the explanations provided for random forests, including minimum-size majority reasons (Audemard et al., 2022c), is that these explanations can be large, making them difficult for users to interpret. It is essential to remember that explanation is a social process where users, as human beings, have cognitive limitations. As highlighted by the psychologist **G. Miller** in his foundational article on "chunking" (Miller, 1956), human memory is limited to units of 7 ± 2 elements. To make explanations more concise and user-friendly, recent research has turned towards probabilistic explanations. However, computing such explanations remains extremely complex. The problem of deciding whether an instance admits a δ -probable reason of size p under a Boolean function is NP^{PP} -complete (Wäldchen et al., 2021), making this computation inapproximable in practice, even for random forests. Despite recent efforts (Izza et al., 2024) to propose efficient approximations, this problem remains a major challenge when dealing with classifiers that are difficult to explain.

Contributions. In this paper, we introduce the notion of k -majoritary reasons, which are majority reasons involving at least k trees, where $k \geq \lfloor \frac{m}{2} \rfloor + 1$ (with m representing the number of trees in the forest f). A k -majoritary reason for an instance \mathbf{x} given a random forest f is a term t that covers \mathbf{x} and constitutes a minimal implicant for inclusion for at least k trees in forest. We also define the notion of a δ -probable majority reason, which is a δ -probable reason (Louenas, 2024) for a strict majority of trees in the forest, where these trees classify the instance in

the same way as the forest.

The k -majoritary reasons are particularly interesting because they more robustly support the decision made by the forest, making them more useful than a simple arbitrary majority reason (denoted MAJ). Although they can be derived in polynomial time, identifying minimum-size k -majoritary reasons is an NP-complete problem. To achieve this, an approach based on a PARTIAL MAXSAT solver can be used. While the δ -probable majority reasons offer gains in intelligibility and size, as they are, by construction, smaller than the majority or k -majoritary reasons, making them more interpretable while improving the intelligibility of random forests. We subsequently propose algorithms to derive k -majoritary reasons and δ -probable majority reasons to enable empirical comparison. Our experiments on standard benchmarks show that the PARTIAL MAXSAT solver generally allows for the derivation of minimum-size k -majoritary reasons, comparable in size to minimum-size majority reasons (denoted minMAJ) while involving a greater number of trees. The δ -probable majority reasons, on the other hand, provide a significant reduction in size. Moreover, the computational effort required to derive minimum-size k -majoritary reasons is similar to that of minimum-size majority reasons, while obtaining a δ -majoritary reason is less costly than obtaining a probabilistic explanation (the NP^{PP} -complete problem (Wäldchen et al., 2021)). A greedy algorithm will be employed to derive the δ -probable majority reasons.

2 PRELIMINARIES

Preliminaries. Let $[n]$ be the set $\{1, \dots, n\}$. We denote by \mathcal{F}_n the class of all Boolean functions from $\{0, 1\}^n$ to $\{0, 1\}$, and use $X_n = \{x_1, \dots, x_n\}$ to represent the set of Boolean variables. Any assignment $\mathbf{x} \in \{0, 1\}^n$ is called an *instance*. A *literal* ℓ is either a variable x_i or its negation $\neg x_i$, also denoted \bar{x}_i . x_i and \bar{x}_i are complementary literals. A *term* t is a conjunction of literals, and a *clause* c is a disjunction of literals. $\text{Lit}(f)$ denotes the set of all literals in f . A DNF formula is a disjunction of terms, and a CNF formula is a conjunction of clauses. The set of variables appearing in a formula f is denoted by $\text{Var}(f)$. A formula f is *consistent* if and only if it has at least one model. A formula f_1 *implies* a formula f_2 , denoted $f_1 \models f_2$, if and only if every model of f_1 is also a model of f_2 . Two formulas f_1 and f_2 are *equivalent*, denoted $f_1 \equiv f_2$, if and only if they have the same models.

Given an assignment $z \in \{0, 1\}^n$, the corresponding term is defined as $t_z = \bigwedge_{i=1}^n x_i^{z_i}$ where $x_i^0 =$

\bar{x}_i and $x_i^1 = x_i$. A term t covers an assignment z if $t \subseteq t_z$. An *implicant* of a Boolean function f is a term that implies f . A *prime implicant* of f is an implicant t of f such that no subset of t is an implicant of f .

A partial instance is a tuple $\mathbf{y} \in \{0, 1, \perp\}^n$. Intuitively, if $y[i] = \perp$, the value of the i -th feature is undefined. We say that \mathbf{y} is subsumed by \mathbf{x} if it is possible to obtain \mathbf{y} from \mathbf{x} by replacing some undefined values of \mathbf{y} with values from \mathbf{x} , denoted $\mathbf{y} \subseteq \mathbf{x}$. We define $|\mathbf{y}|_{\perp} = |\{i \in \{1, \dots, n\} : y[i] = \perp\}|$, where \perp represents a missing value. The restriction of \mathbf{x} to S , denoted \mathbf{x}_S , is the partial instance in $\{0, 1, \perp\}^n$ such that, for every $i \in [n]$, $(\mathbf{x}_S)_i = \mathbf{x}_i$ if $i \in S$, and $(\mathbf{x}_S)_i = \perp$ otherwise. Any instance $\mathbf{y} \in \{0, 1\}^n$ is covered by \mathbf{x}_S if and only if $\mathbf{y}_S = \mathbf{x}_S$. We define $\mathbb{P}_{[f(z)=f(\mathbf{x})|t_S \subseteq t_z]}$:

$$\begin{aligned} \mathbb{P}_{[f(z)=f(\mathbf{x})|t_S \subseteq t_z]} &= \frac{|\{z \in \{0, 1\}^n : f(z) = f(\mathbf{x}), z_S = x_S\}|}{2^{n-|S|}} \\ &= \frac{h_{f,\mathbf{x}}(S)}{2^{n-|S|}} \end{aligned} \quad (1)$$

$\mathbb{P}_{[f(z)=f(\mathbf{x})|t_S \subseteq t_z]}$ can be seen as the probability of classifying an instance \mathbf{x}' , which shares a subset of features S with an instance \mathbf{x} , in the same way by a classifier represented by a Boolean function f . For $\delta \in (0, 1]$, the term t_S is said to be a δ -probable reason for \mathbf{x} given f if $\frac{h_{f,\mathbf{x}}(S)}{2^{n-|S|}} \geq \delta$. Furthermore, for any $\ell \in t_S$, this condition is not satisfied. If $\delta = 1$, t_S is a prime implicant for \mathbf{x} given f .

A **binary decision tree** on X_n is a binary tree T , where each internal node is labeled with one of the n Boolean input variables from X_n , and each leaf is labeled with either 0 or 1. Each variable is assumed to appear at most once on any path from the root to a leaf (read-once property). The value $T(x) \in \{0, 1\}$ of T for an input instance \mathbf{x} is determined by the label of the leaf reached from the root node.

A **random forest** on X_n is a set $F = \{T_1, \dots, T_m\}$, where each T_i ($i \in [m]$) is a decision tree on X_n , and the value $F(\mathbf{x})$ is given by

$$F(\mathbf{x}) = \begin{cases} 1 & \text{if } \frac{1}{m} \sum_{i=1}^m T_i(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

The *size* of F is given by $|F| = \sum_{i=1}^m |T_i|$, where $|T_i|$ is the number of nodes present in T_i . The class of decision trees on X_n is denoted by DT_n , and the class of random forests with at most m decision trees (for $m \geq 1$) on DT_n is denoted by $\text{RF}_{n,m}$. Finally, $\text{RF}_n = \bigcup_{m \geq 1} \text{RF}_{n,m}$ and $\text{RF} = \bigcup_{n \geq 1} \text{RF}_n$. It is well known that any decision tree $T \in \text{DT}_n$ can be transformed in linear time into an equivalent DNF (or an equivalent CNF). This DNF is an orthogonal DNF (Audemard et al., 2022b). However, when moving to random forests, the situation is quite different. Any formula in CNF or DNF can be converted in linear time into an equivalent random forest, but there is no polynomial

space conversion from a random forest to CNF or DNF (Audemard et al., 2022c).

Example 1. The random forest $F = \{T_1, T_2, T_3\}$ presented in Figure 1 consists of three trees. It classifies bank loans using the features $\{x_1, x_2, x_3, x_4\}$.

Consider the instance $\mathbf{x} = (1, 1, 1, 1)$. Since $F(\mathbf{x}) = 1$, the client \mathbf{x} is granted a bank loan. The direct reason for \mathbf{x} , given by F , is $P_{\mathbf{x}}^F = x_1 \wedge x_2 \wedge x_3 \wedge x_4$. Now consider the instance $\mathbf{x}' = (0, 0, 1, 0)$, which is recognized as a loan rejection since $F(\mathbf{x}') = 0$. The direct reason for \mathbf{x}' and F is $P_{\mathbf{x}'}^F = \bar{x}_1 \wedge \bar{x}_2 \wedge x_3 \wedge \bar{x}_4$.

We conclude this section by recalling some important properties and definitions for what follows. The first property concerns the fact that the evaluation of the function $h_{f,\mathbf{x}}(S)$ in formula 1 is a #SAT problem when considering a classifier represented by a Boolean function f (for example, a random forest). However, for a binary decision tree T , $h_{x,T}(S)$ can be rewritten in the form $h_{x,T}(S) = w(\text{DNF}(T) | t_S)$, where $\text{DNF}(T)$ is the disjunctive normal form representation of the tree T , and $w(\text{DNF}(T) | t_S)$ is the number of models of the formula $\text{DNF}(T) | t_S$. This formula is also an orthogonal DNF (Darwiche, 1999), and therefore $w(\text{DNF}(T) | t_S)$ can be evaluated in linear time (Bounia and Koriche, 2023).

A central result to recall in this work is the encoding of a random forest into a CNF formula, as well as the ability to perform an implication test via a call to a SAT oracle, as demonstrated in (Audemard et al., 2022c).

Proposition 1. Let $F = \{T_1, \dots, T_m\}$ be a random forest from $\text{RF}_{n,m}$, and let t be a term over X_n and $k \in \mathbb{N}$. Let H be the following CNF formula:

$$H = \{(y_i \vee c) : i \in [m], c \in \text{CNF}(\neg T_i)\} \cup \text{CNF}\left(\sum_{i=1}^m y_i > k\right)$$

where $\{y_1, \dots, y_m\}$ are new variables and $\text{CNF}(\sum_{i=1}^m y_i > k)$ is the CNF encoding of the cardinality constraint $\sum_{i=1}^m y_i > k$. For $k = \frac{m}{2}$, t is an implicant of F (an implicant of the strict majority of the trees) if and only if $H \wedge t$ is unsatisfiable.

Based on such encoding, explanations by prime implicants (or sufficient reasons) for an instance \mathbf{x} given a random forest F can be characterized in terms of **MUS** (minimal unsatisfiable subsets (Liffiton and Sakallah, 2008)). However, their very high computational cost (the DP-complete problem) makes their derivation challenging to perform. A natural question arises: are there minimal abductive explanations for inclusion that can be computed in polynomial time? The answer is yes, with the *majoritary reasons* (Audemard et al., 2022c), which are abductive explanations. For an instance \mathbf{x} , a majority reason is an implicant of a strict majority of the trees in the forest F .

A limitation of majority explanations, including those of minimum-size, is that they involve only a strict majority of trees, which can make them too large to be

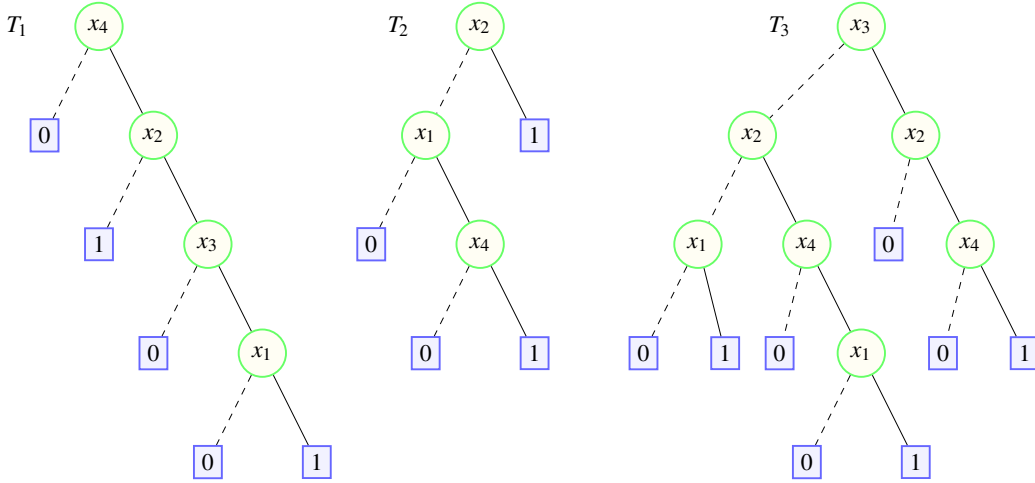


Figure 1: A random forest for bank loan allocation based on the features $\{x_1, x_2, x_3, x_4\}$.

interpretable. In this work, we proposed k -majoritary explanations (of minimum-size), which, although difficult to compute, better explain the decision made by the forest, thus improving upon existing majoritary explanations. Furthermore, to provide more concise and interpretable explanations, we introduced δ -probable majoritary reasons, which are δ -probable explanations for the strict majority of trees, thereby offering an alternative to traditional probabilistic explanations.

3 FOR BETTER EXPLANATIONS FOR RANDOM FORESTS

The concepts of δ -probable majoritary reasons and probabilistic explanations, as defined in (Wäldchen, 2022), do not coincide. A probabilistic explanation is a δ -probable reason for \mathbf{x} given forest F , while a δ -probable majoritary reason is a δ -probable reason for a strict majority of decision trees in F , with the additional condition that t is minimal for inclusion for at least one tree. Majoritary δ -probable reasons can be considered simplified versions of probabilistic explanations, potentially including irrelevant features. Now, consider a random forest $F \in \text{RF}_{n,m}$ and an instance \mathbf{x} . The set $F_c = \{T_i \mid T_i(\mathbf{x}) = F(\mathbf{x})\}$ represents the trees classifying \mathbf{x} in the same way as the forest F , with $\frac{m}{2} \leq |F_c| \leq m$.

3.1 k -Majoritary Reason

Definition 1 (k -Majoritary Reason). *Let $F = \{T_1, \dots, T_m\}$ be a random forest in $\text{RF}_{n,m}$ and $\mathbf{x} \in \{0, 1\}^n$ an instance, with $k \in \mathbb{N}$ such that $\lfloor \frac{m}{2} \rfloor + 1 \leq k \leq |F_c|$. A k -majoritary reason (k -MAJ) for \mathbf{x} given F is a term t covering \mathbf{x} , which is an implicant of at least k trees. Furthermore, for each literal $l \in t$, $t \setminus \{l\}$ no longer satisfies this condition. A minimum-size*

k -majoritary reason (k -minMAJ) is one that contains the minimal number of literals.

Lemma 1. *Let F be a random forest in $\text{RF}_{n,m}$, an instance $\mathbf{x} \in \{0, 1\}^n$, and $k \geq \lfloor \frac{m}{2} \rfloor + 1$. We can always derive a majoritary reason (MAJ) from a k -majoritary reason (k -MAJ) using a greedy algorithm.*

Example 2. *Based on example 1, the minMAJ reasons for \mathbf{x} given F are $x_1 \wedge x_2 \wedge x_4$, $x_1 \wedge x_3 \wedge x_4$, and $x_2 \wedge x_3 \wedge x_4$. Each of these explanations is more concise than the direct reason $P_{\mathbf{x}}^F$, but none of these explanations is a 3-MAJreason, as they only involve 2 trees.*

In contrast, for the instance \mathbf{x}' , $\bar{x}_1 \wedge \bar{x}_2 \wedge \bar{x}_4$ and $\bar{x}_2 \wedge x_3 \wedge \bar{x}_4$ are minMAJ reasons, but $\bar{x}_1 \wedge \bar{x}_2 \wedge \bar{x}_4$ is the unique 3-minMAJ reason for \mathbf{x}' given F .

The k -MAJ reasons ensure coverage by a larger number of decision trees, making them more robust and reliable (as shown in Example 2). Naturally, the user will be interested in deriving concise reasons, particularly the k -minMAJ reasons. However, it is important to remember the inherent complexity in deriving them.

Proposition 2. *Let $F \in \text{RF}_{n,m}$, $\mathbf{x} \in \{0, 1\}^n$, $k \geq \lfloor \frac{m}{2} \rfloor + 1$, and $p \in \mathbb{N}$. Deciding whether there exists a k -minMAJ reason t for \mathbf{x} given the random forest F such that t contains at most p features is an NP-complete problem.*

A common approach to resolve NP optimization problems is to rely on modern SAT solvers. In this perspective, recall that a PARTIAL MAXSAT problem consists of a pair $(C_{\text{soft}}, C_{\text{hard}})$ where C_{soft} and C_{hard} are (finite) sets of clauses. The goal is to find a Boolean assignment that maximizes the number of clauses c in C_{soft} that are satisfied, while satisfying all clauses in C_{hard} .

Proposition 3. *Let $F = \{T_1, \dots, T_m\}$ ($F \in \text{RF}_{n,m}$), $\mathbf{x} \in \{0, 1\}^n$ and $k \geq \frac{m}{2}$. Let $(C_{\text{soft}}, C_{\text{hard}})$ be an instance of the PARTIAL MAXSAT problem such that:*

$$\begin{aligned}
C_{\text{soft}} &= \{\bar{x}_i : x_i \in t_x\} \cup \{x_i : \bar{x}_i \in t_x\} \\
C_{\text{hard}} &= \{(\bar{y}_i \vee c_{i|x}) : i \in [m], c \in \text{CNF}(T_i^\pm)\} \\
&\cup \text{CNF}\left(\sum_{i=1}^m y_i > k\right)
\end{aligned}$$

where $c_{i|x} = c \cap t_x$ is the restriction of c to the literals in t_x , $\{y_1, \dots, y_m\}$ are auxiliary variables, $T_i^\pm = T_i$ ($i \in [m]$) if $F(\mathbf{x}) = 1$, $T_i = \neg T_i$ if $F(\mathbf{x}) = 0$, and $\text{CNF}(\sum_{i=1}^m y_i > k)$ is the CNF encoding of the cardinality constraint $\sum_{i=1}^m y_i > k$. Let \mathbf{z}^* be an optimal solution of $(C_{\text{soft}}, C_{\text{hard}})$. Then, $t_x \cap t_{\mathbf{z}^*}$ is a k -minMAJ reason for \mathbf{x} given F .

Thanks to this characterization result, it is possible to leverage the many algorithms that have been developed so far for PARTIAL MAXSAT (see, for example, (Ansótegui et al., 2013; Saikko et al., 2016; Ignatiev, 2019)) to compute k -minMAJ reasons.

3.2 Majoritary Probabilistic Explanation (δ -Probable Majoritary Reason)

The concepts of majoritary reason and prime implicant explanation are considered natural explanation concepts for random forests. However, these reasons are often large in size, making them difficult to interpret (Audemard et al., 2022a; Izza and Marques-Silva, 2021; Audemard et al., 2022c). To mitigate this limitation, a probabilistic generalization of explanations was proposed by (Wäldchen et al., 2021). However, deriving probabilistic explanations is NP^{PP}-hard. Therefore, we propose a more easily derivable alternative: δ -probable majoritary reasons.

Definition 2 (δ -Probable Majoritary Reason). *Let $F = \{T_1, \dots, T_m\}$ be a random forest in $\text{RF}_{n,m}$ and $\mathbf{x} \in \{0, 1\}^n$, with $\delta \in (0, 1]$. A majoritary δ -probable reason for \mathbf{x} given F is a subset of features S (or its term t_S that covers \mathbf{x}) that satisfies:*

$$\sum_{T_i \in F_c} \mathbb{1}\left[\frac{h_{\mathbf{x}, T_i(S)}}{2^{n-|S|}} \geq \delta\right] \geq \frac{m}{2}$$

And for each $l \in S$, $S \setminus \{l\}$ does not satisfy this last condition.

A minimum-size δ -probable majoritary reason for \mathbf{x} given F is a δ -probable majoritary reason for \mathbf{x} given F that contains a minimal number of features.

The definition of a δ -probable majoritary reason aims to improve the interpretability of explanations for random forests. By specifying a subset of features that meet a probabilistic threshold δ , this definition allows for explanations that are representative of the model's decision-making process while ensuring a significant level of confidence. The requirement for support by a majority of trees ensures that these explanations are based on collective judgment, thereby increasing their

reliability. Furthermore, the constraint that removing a feature compromises support highlights the importance of each attribute in the decision. This approach thus balances concise explanations with a probabilistic framework, enhancing user confidence and understanding in the face of complex predictions.

Example 3. *Based on example 1 and for the instance \mathbf{x} , given F , $S = \{x_1, x_4\}$ (the term associated with S is $t_S = x_1 \wedge x_4$) is a 0.75-probable majoritary reason for \mathbf{x} given F .*

We have :

- $\frac{h_{\mathbf{x}, T_1}(\{x_1, x_4\})}{4} = 0.75$
- $\frac{h_{\mathbf{x}, T_2}(\{x_1, x_4\})}{4} = 1$
- $\frac{h_{\mathbf{x}, T_3}(\{x_1, x_4\})}{4} = 0.75$

Thus, t_S is a 0.75-probable reason for a strict majority of trees and is smaller than all of minMAJ reasons for \mathbf{x} given F .

Deriving a δ -probable explanation for an instance \mathbf{x} given a decision tree T is generally an NP-complete problem (Arenas et al., 2022). Given that a decision tree is a particular random forest (composed of a single tree) and that a δ -probable majoritary reason is equivalent to a δ -probable reason, we suggest that calculating a δ -probable majoritary reason for a random forest F is also NP-hard.

Remark 1 (Complexity of δ -probable majoritary reasons). *In this work, we did not define the exact complexity of the problem of computing δ -probable majoritary reasons, but we know that this problem belongs to the complexity class NP. By drawing an analogy with the fact that computing a probabilistic explanation for decision trees is an NP-hard problem, we suggest that the problem of computing a probabilistic majoritary explanation is also an NP-hard problem.*

Based on the previous remark and the results presented in (Arenas et al., 2022; Bounia and Koriche, 2023; Izza et al., 2024; Louenas, 2023), which clearly show that deriving a δ -probable reason is out of reach when the classifier in question is a complex tree and the size of the input instance is high-dimensional, we suggest that deriving δ -probable majoritary reasons is also out of reach when the classifier is a random forest. Based on these results, we therefore propose a greedy algorithm to derive minimum-size majoritary probabilistic reasons for inclusion, although they are not necessarily minimal in size.

Greedy Algorithm. In the following, we propose an algorithm to derive δ -probable majoritary reasons from a random forest. This algorithm (see Algorithm 1) aims to identify a reason that meets a given confidence threshold. It takes as input a random forest composed of decision trees, an input instance, and a confidence threshold δ . The algorithm examines the literals of a term and adjusts the set of literals based on the classifications of the

decision trees, thereby ensuring that the resulting reason is predominantly compelling according to the specified threshold. Here is the algorithm:

```

Input: Random forest  $F = \{T_1, \dots, T_m\}$ ,
         instance  $\mathbf{x} \in \{0, 1\}^n$ , confidence
         threshold  $\delta \in (0, 1]$ 
Output: A  $\delta$ -probable majority reason

 $S \leftarrow t$  /* is a term (a set of
    literals) */
 $F_c = \{T_i : F(\mathbf{x}) = T_i(\mathbf{x})\}$  /*The trees that
    classify  $\mathbf{x}$  as  $F$  */
for  $l \in t$  do
    if  $\sum_{T_i \in F_c} 1 \left[ \frac{h_{\mathbf{x}, T_i}(S)}{2^{n-|S|}} \geq \delta \right] \geq \frac{m}{2}$  then
        if  $h_{\mathbf{x}, T_i}(S) \geq \delta$  then
             $F_c \leftarrow F_c - \{T_i\}$ 
        end
         $S \leftarrow S - \{l\}$  /*Remove literal  $l$ 
            from  $S$  */
    end
end
return  $S$ 

```

Algorithm 1: δ -Probable Majority Reason.

Proposition 4. *The algorithm 1 runs in time $O(n \times |F|)$. In the case where $F(\mathbf{x}) = 1$, the algorithm 1 starts with $S = t_x$, where t_x represents the initial set of literals corresponding to the input instance \mathbf{x} . The algorithm then proceeds to iterate over the literals l of S . For each literal l , it checks whether the set S without this literal (noted $S - \{l\}$) constitutes a δ -probable reason for at least $\lfloor \frac{m}{2} \rfloor + 1$ decision trees in the forest F . If the condition is satisfied, it means that the elimination of the literal l from S does not affect the validity of the reason, and therefore, l is removed from S . The algorithm then moves to the next literal, repeating this process until it has examined all the literals of the initial set S .*

At the end of this iteration, the final term S is, by construction, a δ -probable reason for the majority of the decision trees in F . This ensures that, despite the removals made, the remaining set of literals continues to satisfy the specified confidence threshold, thus representing a robust explanation for the given classification. This greedy algorithm runs in time $O(n \times |F|)$, where n is the number of literals in the initial set S and $|F|$ is the total number of trees in the forest. Indeed, checking whether S is an implicant for each tree T_i (for each $i \in [m]$) requires time $O(n \times |T_i|)$, which is justified by the approach used in verifying the implications of the literals on the decision trees (Audemard et al., 2022b; Izza et al., 2020).

4 EXPERIMENTS

We conducted several experiments to evaluate the performance of our approaches, aiming to measure the size difference between minMAJ reasons and k -minMAJ reasons by randomly drawing k between $\lfloor \frac{m}{2} \rfloor + 1$ and $|F_c|$, while also assessing the computation time required to derive them. This study highlights the utility of k -MAJ reasons, which serve as an improved version of MAJ reasons. Additionally, we aim to evaluate the improvement in intelligibility achieved through the reduction in size of the δ -probable majority reasons (calculated with algorithm 1) compared to that of direct reasons and minMAJ reasons.

4.1 Experimental Protocol

We used $B = 20$ standard binary classification datasets from the sites **Kaggle**¹, **OpenML**², and **UCI**³. Categorical attributes were treated as integers, while numerical attributes were binarized using the learning algorithm employed for constructing the decision trees of the forest. The classification performance for F_b was evaluated by measuring the average accuracy on a test set of over 150 instances. The learning of the forest F_b was performed using the CART algorithm, utilizing the implementation from the Scikit-Learn library (Pedregosa et al., 2011), with default hyperparameters, except for the parameter (**nb_estimator**) that controls the number of trees in the forest, which was chosen to limit this number to prevent an explosion of our encodings while maintaining good accuracy.

For each dataset $b \in [B]$, each random forest F_b , and each instance \mathbf{x} from the corresponding test set, we compared the average size of the MAJ reasons and the k -MAJ reasons, as well as the corresponding minimum-sizes (minMAJ and k -minMAJ), in order to highlight the advantage of the k -MAJ reasons compared to the standard MAJ reasons. We also measured the average time required to compute the minMAJ and k -minMAJ reasons (see Table 1). The derivation of the reasons MAJ and k -MAJ was performed using the greedy algorithm described in (Audemard et al., 2022c), with the direct reason P_x^F as input, while the derivation of the minimum-size reasons minMAJ and k -minMAJ was carried out using the PARTIAL MAXSAT solver via the RC2 interface (Ignatiev, 2019), configured with *Glucose* using its implementation with the **PySAT** library⁴.

In the second phase, in order to evaluate the improvement in intelligibility related to the reduction in the size of explanations, we reported the sizes of the direct reasons, the MAJ reasons, as well as the minMAJ

¹<https://www.kaggle.com/datasets>

²<https://www.openml.org>

³<https://archive.ics.uci.edu/datasets>

⁴<https://pysathq.github.io/>

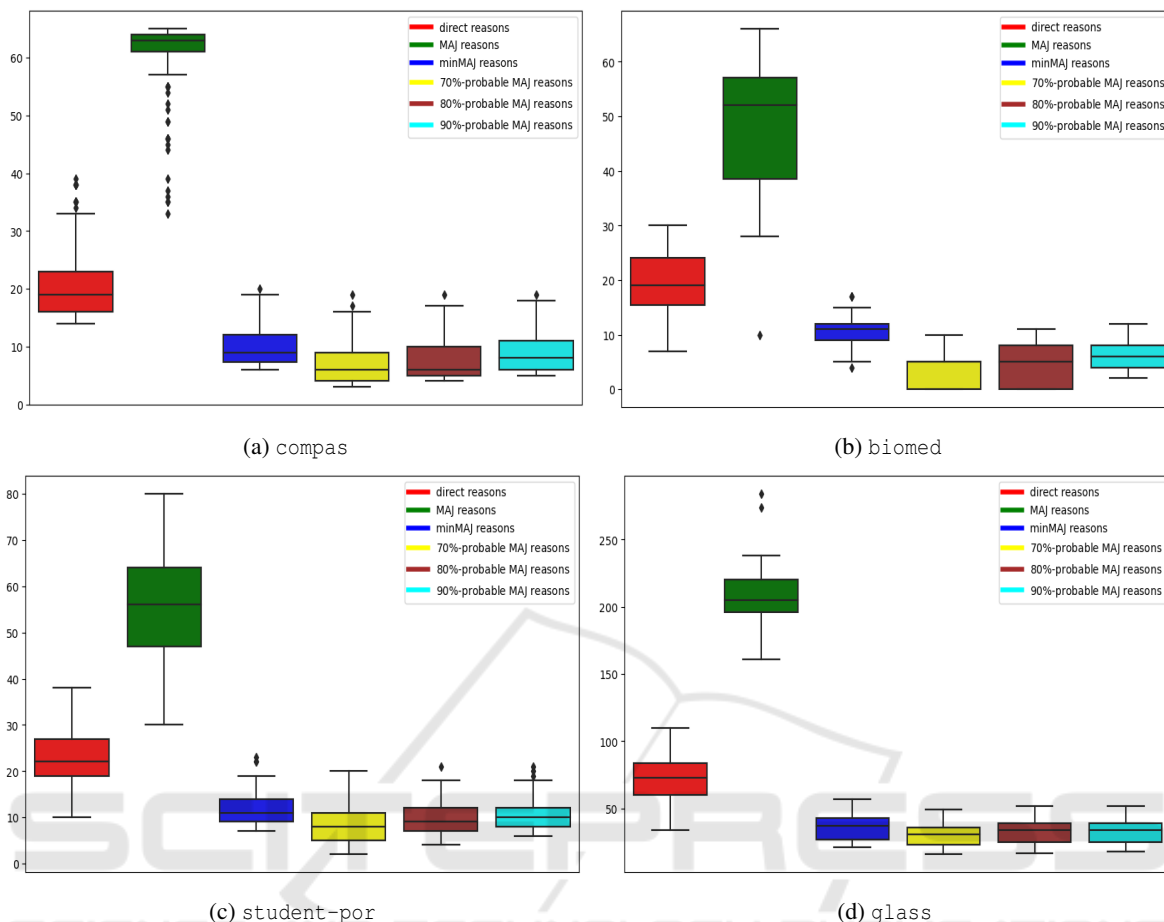


Figure 2: The boxplots illustrating the sizes of direct reasons, MAJ reasons, minMAJ reasons, as well as the {70%, 80%, 90%}-probable majority reasons.

reasons calculated from the given instance x for the forest F_b . The calculation of the direct reasons P_x^F was performed using the **PyXAI** library (Audemard et al., 2023), while the computation of the δ -probable majority reasons was carried out using the algorithm 1 for values of $\delta = \{0.7, 0.8, 0.9\}$. A boxplot visualization was then produced (see 3). To show the reduction obtained in terms of the size of the δ -probable majority reasons compared to the minMAJ reasons, we used **a minMAJ reason as input for the algorithm 1** and varied δ from 0.3 to 1 (that is, until reaching the minMAJ reason).

All experiments were conducted using Python. They were executed on a computer equipped with an Intel(R) Core(TM) i9-9900 processor, operating at a base clock speed of 3.10 GHz. This high-performance processor has 8 cores and 16 threads, allowing for efficient multitasking and parallel processing, which is particularly beneficial for compute-intensive tasks commonly found in machine learning applications. Additionally, the system is equipped with 64 GiB of memory (RAM), providing ample resources to handle large datasets and high-complexity algorithms without significant slowdowns.

4.2 Results

Table 1 presents an excerpt of our results for 20 datasets. The column $\#I$ represents the number of instances, $\#F$ the number of binary attributes, and $\%A$ the accuracy of the forest F_b . The column |Explanation| shows the average size of the computed explanations: P_x^F (direct reason), MAJ (majority reason), k -MAJ (k -majority reason), minMAJ (minimum-size majority reason), and k -minMAJ (k -minimum-size majority reason). The column |Times| indicates the average computation times for generating the minMAJ and k -minMAJ reasons using the PARTIAL MAXSAT solver. K : average size of k .

We observe that the average sizes of the MAJ and k -MAJ reasons are remarkably similar, with a small difference of only 4. This indicates that adding the parameter k does not substantially affect the size of the derived majority reasons. However, the gap decreases further when considering the minMAJ and k -minMAJ reasons, going from 4 to 1.8. This trend suggests that the parameter k influences the minMAJ reasons less than the MAJ reasons, leading to more compact representations.

In analyzing computation times, we find that the av-

erage time required to compute the minMAJ and $k\text{-minMAJ}$ reasons is nearly identical, showing a negligible difference. For example, in the *startup* dataset, the average computation time for a minMAJ reason is about 6.5 seconds, while the $k\text{-minMAJ}$ reasons take about 7 seconds. This small difference indicates that incorporating the parameter k does not have a significant impact on computational efficiency. Overall, these results imply that using $k\text{-minMAJ}$ reasons, which include the minMAJ reasons, does not lead to a notable increase in the size of the reasons or computation time. This observation highlights the potential benefits of $k\text{-minMAJ}$ reasons in terms of enhancing explainability and robustness without significantly compromising computation time and reason size.

To illustrate the improvement in intelligibility achieved by moving from abductive explanations (such as direct reasons, MAJ reasons, and minMAJ reasons) to δ -probable majoritary reasons, we created several boxplots based on 150 instances from four datasets: *compas* and *student-por* (shown on the right), and *biomed* and *glass* (shown on the left) (see 4.1). These diagrams visualize the transition from minMAJ and MAJ reasons to probabilistic majoritary reasons for thresholds of 70%, 80%, and 90%. Examining these boxplots reveals a clear trend towards a significant reduction in the number of attributes used in both direct and majoritary reasons when adopting a 0.7-probable majoritary reason. This demonstrates the effectiveness of δ -probable majoritary reasons in simplifying explanations while maintaining an adequate level of confidence in the model's decision-making process. In other words, while traditional explanations may involve a broader set of attributes, shifting to δ -probable majoritary reasons allows for a focus on the most relevant features, thereby facilitating user understanding. This simplification of explanations not only enhances interpretability but also promotes better comprehension of the predictions provided by the model, making the results more accessible and actionable for the human user.

In our study, we observed a significant reduction in the size of explanations achieved with 70% δ -probable majoritary reasons compared to traditional explanations based on direct reasons and MAJ reasons. To further investigate this phenomenon, we conducted additional experiments assessing the impact of varying δ values from 0.3 to 1 across three datasets: *employee*, which includes attributes related to employee performance and demographics, *compas*, focusing on characteristics associated with criminal recidivism; and *backache*, encompassing health information on patients with back issues. By calculating the sizes of the δ -probable majoritary reasons for various instances at each δ value, we found that the average size of these explanations gradually increased with higher δ values, indicating that greater confidence leads to more detailed explanations. However, this size eventually stabilized when the algorithm effectively captured the minMAJ reason, highlighting a threshold where

increasing δ no longer significantly enhances the explanations. These results demonstrate the balance between explanation size and intelligibility, as adjusting δ allows for tailoring the level of detail in explanations, thereby enhancing their interpretability without adding unnecessary complexity.

Remark 2. *Although we have not conducted specific experiments to directly compare the $k\text{-minMAJ}$ explanations with the δ -probabilistic majoritary explanations, we can draw relevant conclusions from the results obtained for each of these approaches. The $k\text{-minMAJ}$ explanations, while offering additional robustness by involving a greater number of trees in the forest, tend to generate larger explanations than the minMAJ explanations. However, the minMAJ explanations are generally larger than the δ -probabilistic majoritary explanations, as confirmed by our experimental results. Therefore, we can deduce, without the need for further experiments, that a further size reduction will occur if the algorithm's input is a $k\text{-minMAJ}$ explanation.*

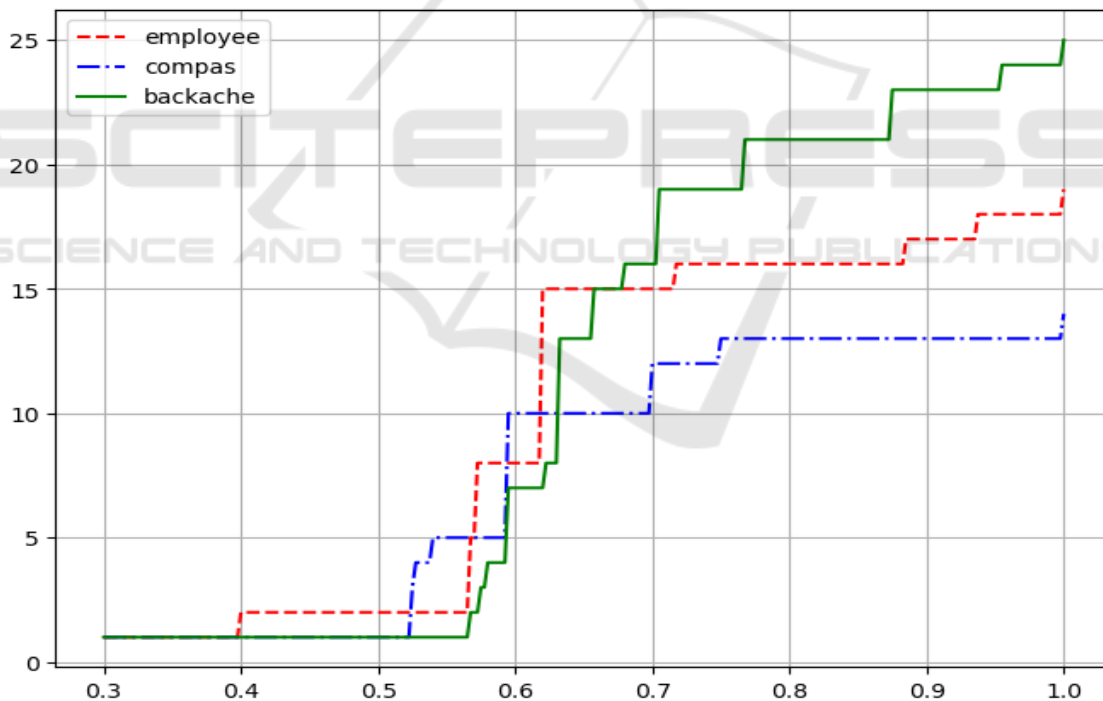
5 CONCLUSION AND FUTURE WORK

Conclusion. In this paper, we introduced k -majoritary reasons and δ -probable majoritary reasons as innovative extensions of traditional majoritary reasons for random forests, aiming to enhance the robustness and conciseness of explanations provided by these models. The $k\text{-MAJ}$ reasons differ from standard majoritary reasons by leveraging a larger subset of trees in the forest for decision-making. This approach not only strengthens the forest's overall decision-making capability but also yields more compact and robust explanations. Importantly, our findings suggest that employing $k\text{-MAJ}$ reasons does not incur a significant increase in computation time, making them a practical choice for real-world applications. However, we also observed that minMAJ and $k\text{-minMAJ}$ explanations can become excessively large, which may hinder their interpretability due to cognitive limitations faced by users. To address this issue, we proposed the concept of δ -probable majoritary reasons. By establishing a probabilistic threshold, δ , this method allows for the generation of more concise and interpretable explanations, ultimately improving user experience. Our experiments demonstrated that these novel approaches are not only effective but also flexible, striking a balance between intelligibility and the robustness of explanations for random forests.

Future Work. Looking ahead, we plan to extend the concepts of k -majoritary reasons and δ -probable majoritary reasons to other tree-based models, such as boosted trees, to further explore their applicability across various machine learning paradigms. This will involve defining and formalizing the complexity of these new expla-

Table 1: Evaluation of k -Majority Explanations: Dataset Statistics and Computation Times.

dataset / random forest				Explanation					Times		
name	#I	#F	%A	P_x^F	MAJ	k -MAJ	minMAJ	k -minMAJ	K	minMAJ	k -minMAJ
australian	690	641	89.89	73.35	68.7	70.9	33.49	34.73	14.41	0.658	0.8003
horse	299	361	86.56	77.04	69.63	74.13	35.82	36.97	19.41	10.877	11.9031
titanic	623	553	79.07	64.12	58.95	61.05	28.05	29.92	17.09	2.6108	2.8402
gina	3153	3802	91.65	183.07	165.31	174.59	97.03	100.73	15.18	3.0707	3.2105
student-por	649	149	90.26	58.38	54.85	56.43	20.61	21.32	26.14	21.8167	23.7152
anneal	898	202	99.26	63.31	57.39	60.91	21.45	22.12	23.55	4.9525	5.02314
startup	923	1704	79.26	129.09	122.3	125.62	74.23	75.34	15.81	6.5699	7.0585
heart	303	386	84.62	61.23	58.35	59.41	25.45	25.83	23.05	2.1122	2.2051
cars	406	484	94.62	80.97	71.43	76.67	32.14	33.74	23.75	5.3147	5.9147
hungarian	294	297	78.65	62.65	57.3	59.73	26.94	28.14	21.73	6.9276	6.9708
vote	434	16	95.42	15.83	11.6	12.86	5.58	5.72	35.85	0.0227	0.0241
soybean	683	81	96.1	43.69	30.62	35.13	11.42	12.97	29.18	0.3899	0.4177
hepatitis	142	188	88.37	59.16	50.19	55.42	23.05	24.87	27.33	16.7504	18.1012
haberman	306	153	65.22	56.77	52.6	54.55	26.39	27.54	22.47	3.501	3.7391
divorce	170	49	96.08	26.78	14.2	19.53	9.06	11.02	20.0	0.0702	0.0712
appendicitis	106	210	93.75	64.38	50.31	56.97	25.12	28.22	24.16	1.8912	1.9521
balance	625	28	84.57	17.19	14.53	15.13	7.86	8.15	34.69	0.0522	0.05432
ecoli	336	175	96.04	43.66	33.68	38.88	16.45	17.64	14.29	0.1095	0.1104
yeast	2417	301	97.93	90.54	80.51	83.13	33.07	35.86	11.29	0.0366	0.0382
student-mat	395	155	87.39	58.02	53.73	55.15	22.13	22.91	20.92	3.0831	3.2341

Figure 3: Size of δ -probable majoritary reasons when δ varies from 0.3 to 1 (minMAJ reason) for the datasets *employee*, *compas*, and *backache*.

nations to better understand their implications in different contexts. Furthermore, another promising direction for future research is to optimize the underlying algorithms used to generate these explanations. We aim to enhance their reliability and conciseness while preserving the quality of insights provided to users. This en-

tails exploring novel computational strategies that improve efficiency without sacrificing the clarity and utility of the explanations. Ultimately, our goal is to ensure that these explanations remain not only comprehensible but also actionable for users, thereby facilitating better decision-making processes in machine learning applica-

tions.

REFERENCES

- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.
- Ansótegui, C., Bonet, M. L., and Levy, J. (2013). SAT-based MaxSAT algorithms. *Artificial Intelligence*, 196:77–105.
- Arenas, M., Barceló, P., Romero Orth, M., and Subercaseaux, B. (2022). On computing probabilistic explanations for decision trees. *Advances in Neural Information Processing Systems*, 35:28695–28707.
- Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J., and Marquis, P. (2021). On the computational intelligibility of boolean classifiers. In *Proc. of KR'21*, pages 74–86.
- Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J., and Marquis, P. (2022a). On preferred abductive explanations for decision trees and random forests. In *Proc. of IJCAI'22*.
- Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J.-M., and Marquis, P. (2022b). On the explanatory power of boolean decision trees. *Data & Knowledge Engineering*, 142:102088.
- Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J.-M., and Marquis, P. (2022c). Trading complexity for sparsity in random forest explanations. In *Proc. of AAAI'22*.
- Audemard, G., Bellart, S., Bounia, L., Lagniez, J.-M., Marquis, P., and Szczepanski, N. (2023). Pyxai : calculer des explications pour des modèles d'apprentissage supervisé. EGC.
- Audemard, G., Koriche, F., and Marquis, P. (2020). On tractable XAI queries based on compiled representations. In *Proc. of KR'20*, pages 838–849.
- Azar, A. T., Elshazly, H. I., Hassanien, A. E., and Elkorary, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, 113.
- Bénaud, C., Biau, G., Veiga, S. D., and Scornet, E. (2021). Interpretable random forests via rule extraction. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, AISTATS'21*, pages 937–945.
- Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095.
- Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI'14*, pages 427–434. ACM.
- Bounia, L. and Koriche, F. (2023). Approximating probabilistic explanations via supermodular minimization (corrected version). In *Uncertainty in Artificial Intelligence (UAI 2023)*, volume 216, pages 216–225.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Choi, A., Shih, A., Goyanka, A., and Darwiche, A. (2020). On symbolically encoding the behavior of random forests. In *Proc. of FoMLAS'20, 3rd Workshop on Formal Methods for ML-Enabled Autonomous Systems, Workshop at CAV'20*.
- Criminisi, A. and Shotton, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*. Advances in Computer Vision and Pattern Recognition. Springer.
- Darwiche, A. (1999). Compiling devices into decomposable negation normal form. pages 284–289.
- Darwiche, A. and Hirth, A. (2020). On the reasons behind decisions. In *Proc. of ECAI'20*.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42.
- Ignatiev, A. (2019). Rc2: an efficient maxsat solver. *J. Satisf. Boolean Model. Comput.*
- Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019). Abduction-based explanations for machine learning models. In *Proc. of AAAI'19*, pages 1511–1519.
- Izza, Y., Ignatiev, A., and Marques-Silva, J. (2020). On explaining decision trees. *ArXiv*, abs/2010.11034.
- Izza, Y. and Marques-Silva, J. (2021). On explaining random forests with SAT. In *Proc. of IJCAI'21*, pages 2584–2591.
- Izza, Y., Meel, K. S., and Marques-Silva, J. (2024). Locally-minimal probabilistic explanations. *ArXiv*.
- Liffiton, M. H. and Sakallah, K. A. (2008). Algorithms for computing minimal unsatisfiable subsets of constraints. *Journal of Automated Reasoning*, 40:1–33.
- Louenas, B. (2023). *Modèles formels pour l'IA explicable: des explications pour les arbres de décision*. PhD thesis, Université d'Artois.
- Louenas, B. (2024). Enhancing the Intelligibility of Boolean Decision Trees with Concise and Reliable Probabilistic Explanations. In *20th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Lisboa, Portugal.
- Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting m(ijcaidel predictions). In *Proc. of NIPS'17*, pages 4765–4774.
- Marques-Silva, J. (2023). Logic-based explainability in machine learning. *ArXiv*, abs/2211.00541.
- Marques-Silva, J. and Huang, X. (2023). Explainability is not a game. *Communications of the ACM*, 67:66–75.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.
- Molnar, C. (2019). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*. Leanpub.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proc. of SIGKDD'16*, pages 1135–1144.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proc. of AAAI'18*, pages 1527–1535.
- Saikko, P., Berg, J., and Jarvisalo, M. (2016). LMHS: A SAT-IP hybrid MaxSAT solver. In *Proceedings of the 19th International Conference of Theory and Applications of Satisfiability Testing (SAT'16)*, pages 539–546.
- Wäldchen, S. (2022). Towards explainable artificial intelligence: interpreting neural network classifiers with probabilistic prime implicants.
- Wäldchen, S., Macdonald, J., Hauch, S., and Kutyniok, G. (2021). The computational complexity of understanding binary classifier decisions. *J. Artif. Intell. Res.*, 70:351–387.

