

# Trading Complexity for Conciseness in Random Forest Explanations

Tracking Number: #12693

## Abstract

Random forests have long been considered as powerful model ensembles in machine learning. By training multiple decision trees, whose diversity is fostered through data and feature subsampling, the resulting random forest can lead to more stable and reliable predictions than a single decision tree. This however comes at the cost of decreased interpretability: while decision trees are often easily interpretable, the predictions made by random forests are much more difficult to understand, as they involve a majority vote over multiple decision trees. In this paper, we examine different types of *reasons* that explain “why” an input instance is classified as positive or negative by a Boolean random forest. Notably, as an alternative to *sufficient reasons* taking the form of prime implicants of the random forest, we introduce *majoritary reasons* which are prime implicants of a strict majority of decision trees. For these abductive explanations, the tractability of the generation problem (finding one reason) and the minimization problem (finding one shortest reason) are investigated. Experiments conducted on various datasets reveal the existence of a trade-off between runtime complexity and conciseness. Sufficient reasons - for which the identification problem is DP-complete - are slightly larger than majority reasons that can be generated using a simple linear-time greedy algorithm, and significantly larger than *minimal* majority reasons that can be approached using an anytime PARTIAL MAXSAT algorithm.

## Introduction

Over the past two decades, rapid progress in statistical machine learning has led to the deployment of models endowed with remarkable predictive capabilities. Yet, as the spectrum of applications using statistical learning models becomes increasingly large, explanations for why a model is making certain predictions are ever more critical. For example, in medical diagnosis, if some model predicts that an image is malignant, then the doctor may need to know which features in the image have led to this classification. Similarly, in the banking sector, if some model predicts that a customer is a fraud, then the banker might want to know why. Therefore, having explanations for why certain predictions are made is essential for securing user confidence in machine learning technologies (Miller 2019; Molnar 2019).

This paper focuses on classifications made by *random forests*, a popular ensemble learning method that constructs multiple randomized decision trees during the training phase, and predicts by taking a majority vote over the base classifiers (Breiman 2001). Since decision tree randomization is achieved by essentially coupling data subsampling (or bagging) and feature subsampling, random forests are fast and easy to implement, with few tuning parameters. Furthermore, they often make accurate and robust predictions in practice, even for small data samples and high-dimensional feature spaces (Biau 2012). For these reasons, random forests have been used in various applications including, among others, computer vision (Criminisi and Shotton 2013), crime prediction (Bogomolov et al. 2014), ecology (Cutler et al. 2007), genomics (Chen and Ishwaran 2012), and medical diagnosis (Azar et al. 2014).

Despite their success, random forests are much less interpretable than decision trees. Indeed, the prediction made by a decision tree on a given data instance can be easily interpreted by reading the unique root-to-leaf path that covers the instance. By contrast, there is no such *direct reason* in a random forest, since the prediction is derived from a majority vote over multiple decision trees. So, a key issue in random forests is to infer *abductive* explanations, that is, to capture in concise terms why a data instance is classified as positive or negative by the model ensemble.

**Related Work.** Explaining random forest predictions has received increasing attention in recent years (Bénard et al. 2021; Choi et al. 2020; Izza and Marques-Silva 2021). Notably, in the classification setting, (Choi et al. 2020; Izza and Marques-Silva 2021) have focused on *sufficient reasons*, which are abductive explanations involving only relevant features (Darwiche and Hirth 2020). Informally, if we view any random forest classifier as a Boolean function  $f$ , then a sufficient reason for classifying a data instance  $x$  as positive by  $f$  is a *prime implicant*  $t$  of  $f$  covering  $x$ . By construction, removing any feature from a sufficient reason  $t$  would question the fact that  $t$  explains the way  $x$  is classified by  $f$ . Note that if  $f$  is described by a single decision tree, then generating a sufficient reason for any input instance  $x$  can be done in linear time. Yet, in the general case where  $f$  is represented by an arbitrary number of decision trees, the problem of identifying a sufficient reason has recently shown to be



DP-complete (Izza and Marques-Silva 2021). Despite this intractability statement, the empirical results reported by the authors indicate that a MUS-based algorithm for computing sufficient reasons proves quite efficient in practice.

In addition to *model-based* explanations described above, *model-agnostic* explanations can be applied to random forests. Notably, the LIME method (Ribeiro, Singh, and Guestrin 2016) extrapolates a linear threshold function  $g$  from the behavior of the random forest  $f$  around an input instance  $x$ . For the ANCHOR method (Ribeiro, Singh, and Guestrin 2018), the extrapolated model  $g$  takes the form of a decision rule. Yet, even if in both cases a prime implicant of  $g$  can be easily computed, the resulting explanation is *not* guaranteed abductive since  $g$  is only an approximation of  $f$ .

**Contributions.** In this paper, we introduce several new notions of abductive explanations: *direct reasons*, which extend to the case of random forests the corresponding notion defined primarily for decision trees, and *majority reasons*, which are abductive explanations taking into account the averaging rule of random forests. Informally, a majoritary reason for classifying an instance  $x$  as positive by some random forest  $f$  is a prime implicant  $t$  of a majority of decision trees in  $f$  that covers  $x$ . What make direct and majoritary reasons valuable is the possibility of inferring them in a tractable way, whilst there is no similar tractability result when dealing with sufficient reasons, unless  $P = NP$ .

More specifically, we examine in this study the tractability of both the generation (finding one explanation) and the minimization (finding one shortest explanation) problems for direct reasons and majoritary reasons. As far as we know, all complexity results related to random forest explanations are new, if we make an exception for the intractability of generating sufficient reasons, which was recently established in (Izza and Marques-Silva 2021). Notably, direct reasons and majoritary reasons can be derived in time polynomial in the size of the input (the instance and the random forest used to classify it). By contrast, the identification of minimal majoritary reasons is NP-complete, and the identification of minimal sufficient reasons is  $\Sigma_2^P$ -complete.

Based on these results, we provide algorithms for deriving random forest explanations, which open the way for an empirical comparison. Our experiments made on standard benchmarks reveal the existence of a trade-off between the runtime complexity of finding (possibly minimal) abductive explanations and the conciseness of such explanations. In a nutshell, majoritary reasons and minimal majoritary reasons offer interesting compromises in comparison to, respectively, sufficient reasons and minimal sufficient reasons. Indeed, the size of majoritary reasons and the computational effort required to generate them are generally smaller than those obtained for sufficient reasons. Furthermore, minimal majoritary reasons outperform minimal sufficient reasons, since the latter are too computationally demanding. In fact, using an *anytime* PARTIAL MAXSAT solver for minimizing majoritary reasons, we derive concise explanations which are typically much shorter than all other forms of abductive explanations. Proofs and additional empirical results are provided as supplementary material.

## Preliminaries

For an integer  $n$ , let  $[n] = \{1, \dots, n\}$ . By  $\mathcal{F}_n$  we denote the class of all Boolean functions from  $\{0, 1\}^n$  to  $\{0, 1\}$ , and we use  $X_n = \{x_1, \dots, x_n\}$  to denote the set of input Boolean variables. Any Boolean vector  $x \in \{0, 1\}^n$  is called an *instance*. For any function  $f \in \mathcal{F}_n$ , an instance  $x \in \{0, 1\}^n$  is called a *positive example* of  $f$  if  $f(x) = 1$ , and a *negative example* if  $f(x) = 0$ .

We refer to  $f$  as a propositional formula when it is described using the Boolean connectives  $\wedge$  (conjunction),  $\vee$  (disjunction) and  $\neg$  (negation), together with the constants 1 (true) and 0 (false). As usual, a *literal*  $l_i$  is a variable  $x_i$  or its negation  $\neg x_i$ , also denoted  $\bar{x}_i$ . A *term*  $t$  is a conjunction of literals, and a *clause*  $c$  is a disjunction of literals. A *DNF formula* is a disjunction of terms and a *CNF formula* is a conjunction of clauses. The set of variables occurring in a formula  $f$  is denoted  $\text{Var}(f)$ . In the following, we shall often treat instances as terms, and terms as sets of literals. For an assignment  $z \in \{0, 1\}^n$ , the corresponding term is

$$t_z = \bigwedge_{i=1}^n x_i^{z_i} \text{ where } x_i^0 = \bar{x}_i \text{ and } x_i^1 = x_i$$

A term  $t$  *covers* an assignment  $z$  if  $t \subseteq t_z$ . An *implicant* of a Boolean function  $f$  is a term that implies  $f$ , that is, a term  $t$  such that  $f(z) = 1$  for every assignment  $z$  covered by  $t$ . A *prime implicant* of  $f$  is an implicant  $t$  of  $f$  such that no proper subset of  $t$  is an implicant of  $f$ .

With these basic notions in hand, a (Boolean) *decision tree* on  $X_n$  is a binary tree  $T$ , each of whose internal nodes is labeled with one of  $n$  input variables, and whose leaves are labeled 0 or 1. Without loss of generality, every variable is supposed to occur at most once on any root-to-leaf path. The value  $T(x)$  of  $T$  on an input instance  $x$  is given by the label of the leaf reached from the root as follows: at each node go to the left or right child depending on whether the input value of the corresponding variable is 0 or 1, respectively. A (Boolean) *random forest* on  $X_n$  is an ensemble  $F = \{T_1, \dots, T_m\}$ , where each  $T_i$  ( $i \in [m]$ ) is a decision tree on  $X_n$ , and such that the value  $F(x)$  is given by

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

The size of  $F$  is given by  $|F| = \sum_{i=1}^m |T_i|$ , where  $|T_i|$  is the number of nodes occurring in  $T_i$ . The class of decision trees on  $X_n$  is denoted  $\text{DT}_n$ , and the class of random forests with at most  $m$  decision trees (with  $m \geq 1$ ) over  $\text{DT}_n$  is denoted  $\text{RF}_{n,m}$ . Finally,  $\text{RF}_n$  is the union of all  $\text{RF}_{n,m}$  for  $m \geq 1$ .

**Example 1.** The random forest  $F = \{T_1, T_2, T_3\}$  in Figure 1 is composed of three decision trees. It separates *Cattleya orchids* from other orchids using the following features:  $x_1$ : “has fragrant flowers”,  $x_2$ : “has one or two leaves”,  $x_3$ : “has large flowers”, and  $x_4$ : “is sympodial”.

We conclude this section with two important properties of random forests. The first property is related to the fact that any decision tree  $T$  can be transformed into its negation  $\neg T \in \text{DT}_n$ , by simply reverting the label of leaves. Negating a random forest can also be achieved in polynomial time:



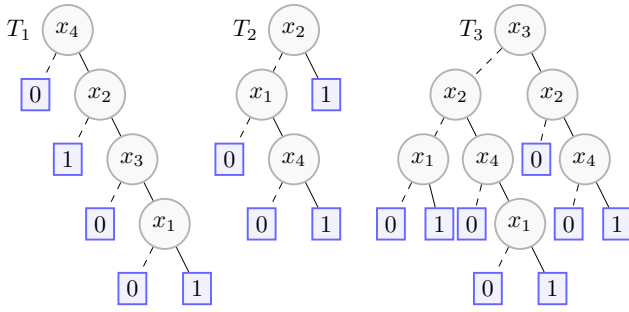


Figure 1: A random forest for recognizing *Cattleya* orchids. The left (resp. right) child of any decision node labelled by  $x_i$  corresponds to the assignment of  $x_i$  to 0 (resp. 1).

**Proposition 1.** *There exists a linear-time algorithm that computes a random forest  $\neg F \in \text{RF}_{n,m}$  equivalent to the negation of a given random forest  $F \in \text{RF}_{n,m}$ .*

For the second property, it is well-known that any decision tree  $T$  can be encoded in linear time into an equivalent disjunction of terms  $\text{DNF}(T)$ , where each term coincides with a 1-path (i.e., a path from the root to a leaf labeled with 1), or a conjunction of clauses  $\text{CNF}(T)$ , where each clause is the negation of term describing a 0-path. Yet, when switching to random forests, the picture is quite different:

**Proposition 2.** *Any CNF or DNF formula can be converted in linear time into an equivalent random forest, but there is no polynomial-space translation from  $\text{RF}$  to CNF or to DNF.*

## Random Forest Explanations

The key focus of this study is to explain *why* a random forest classifies some data instance as positive or negative. This calls for a notion of abductive explanation<sup>1</sup>. Specifically, an *abductive explanation* for an instance  $\mathbf{x} \in \{0, 1\}^n$  given a Boolean function  $f \in \mathcal{F}_n$  is an implicant  $t$  of  $f$  (resp.  $\neg f$ ) if  $f(\mathbf{x}) = 1$  (resp.  $f(\mathbf{x}) = 0$ ) that covers  $\mathbf{x}$ . Such an abductive explanation always exists, since  $t = t_{\mathbf{x}}$  is such a (trivial) explanation. So, in the rest of this section, we shall mainly concentrate on *concise* forms of abductive explanations.

## Direct Reasons

For a decision tree  $T \in \text{DT}_n$  and a data instance  $\mathbf{x} \in \{0, 1\}^n$ , the *direct reason* of  $\mathbf{x}$  given  $T$  is the term  $t_{\mathbf{x}}^T$  corresponding to the unique root-to-leaf path of  $T$  that covers  $\mathbf{x}$ . This simple form of abductive explanation can be extended to random forests as follows:

**Definition 1.** *Let  $F = \{T_1, \dots, T_m\}$  be a random forest in  $\text{RF}_{n,m}$ , and  $\mathbf{x} \in \{0, 1\}^n$  be an instance. Then, the direct reason for  $\mathbf{x}$  given  $F$  is the term  $t_{\mathbf{x}}^F$  defined by*

$$t_{\mathbf{x}}^F = \begin{cases} \bigwedge_{T_i \in F: T_i(\mathbf{x})=1} t_{\mathbf{x}}^{T_i} & \text{if } F(\mathbf{x}) = 1 \\ \bigwedge_{T_i \in F: T_i(\mathbf{x})=0} t_{\mathbf{x}}^{T_i} & \text{if } F(\mathbf{x}) = 0 \end{cases}$$

<sup>1</sup>Unlike (Ignatiev, Narodytska, and Marques-Silva 2019), we do not require those explanations to be minimal w.r.t. set inclusion, in order to keep the concept distinct (and actually more general) than the one of sufficient reasons.

By construction,  $t_{\mathbf{x}}^F$  is an abductive explanation that can be computed in  $\mathcal{O}(|F|)$  time.

**Example 2.** *Based on Example 1, consider the instance  $\mathbf{x} = (1, 1, 1, 1)$ . Since  $F(\mathbf{x}) = 1$ , it is recognized as a *Cattleya* orchid. The direct reason for  $\mathbf{x}$  given  $F$  is  $t_{\mathbf{x}}^F = x_1 \wedge x_2 \wedge x_3 \wedge x_4$ . Consider now the instance  $\mathbf{x}' = (0, 1, 0, 0)$ , which is not recognized as a *Cattleya* orchid, since  $F(\mathbf{x}') = 0$ . The direct reason for  $\mathbf{x}'$  given  $F$  is  $t_{\mathbf{x}'}^F = x_2 \wedge \bar{x}_3 \wedge \bar{x}_4$ .*

## Sufficient Reasons

Another valuable notion of abductive explanation is the one of *sufficient reason*<sup>2</sup>, defined for any Boolean classifier (Darwiche and Hirth 2020). In the setting of random forests, such explanations can be defined as follows:

**Definition 2.** *Let  $F \in \text{RF}_n$  be a random forest and  $\mathbf{x} \in \{0, 1\}^n$  be an instance. A sufficient reason for  $\mathbf{x}$  given  $F$  is a prime implicant  $t$  of  $F$  (resp.  $\neg F$ ) if  $F(\mathbf{x}) = 1$  (resp.  $F(\mathbf{x}) = 0$ ) that covers  $\mathbf{x}$ .*

**Example 3.** *For our running example,  $x_2 \wedge x_3 \wedge x_4$  and  $x_1 \wedge x_4$  are the sufficient reasons for  $\mathbf{x}$  given  $F$ .  $\bar{x}_4$  and  $\bar{x}_1 \wedge \bar{x}_3$  are the sufficient reasons for  $\mathbf{x}'$  given  $F$ .*

Importantly, all features occurring in a sufficient reason  $t$  are *relevant*. Indeed, removing any literal from  $t$  would question the fact that  $t$  implies  $F$ . Note that the direct reason  $t_{\mathbf{x}}^F$  for  $\mathbf{x}$  given  $F$  may contain arbitrarily many more features than a sufficient reason for  $\mathbf{x}$  given  $F$ , since this is already known in the case where  $F$  consists in a single decision tree (Izza, Ignatiev, and Marques-Silva 2020).

The problem of finding a sufficient reason  $t$  for an input instance  $\mathbf{x} \in \{0, 1\}^n$  given random forest  $F \in \text{RF}_n$ , has recently been shown  $\text{DP}$ -complete (Izza and Marques-Silva 2021). In fact, even the apparently simple task of *checking* whether  $t$  is an implicant of  $F$  is already hard:

**Proposition 3.** *Let  $F$  be a random forest in  $\text{RF}_n$  and  $t$  be a term over  $X_n$ . Then, deciding whether  $t$  is an implicant of  $F$  is  $\text{coNP}$ -complete.*

The above result is in stark contrast with the computational complexity of checking whether a term  $t$  is an implicant of a decision tree  $T$ . This task can be solved in polynomial time, using the fact that  $T$  can be converted (in linear time) into its clausal form  $\text{CNF}(T)$ , together with the fact that testing whether  $t$  implies  $\text{CNF}(T)$  can be done in  $\mathcal{O}(|T|)$  time. That mentioned, in the case of random forests, the implicant test can be achieved via a call to a SAT oracle:

**Proposition 4.** *Let  $F = \{T_1, \dots, T_m\}$  be a random forest of  $\text{RF}_{n,m}$ , and  $t$  be a (satisfiable) term over  $X_n$ . Let  $H$  be the CNF formula*

$$\{(\bar{y}_i \vee c) : i \in [m], c \in \text{CNF}(\neg T_i)\} \cup \text{CNF}\left(\sum_{i=1}^m y_i > \frac{m}{2}\right)$$

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

<sup>2</sup>Sufficient reasons are also known as prime-implicant explanations (Shih, Choi, and Darwiche 2018).



Based on such an encoding, the sufficient reasons for an instance  $x$  given a random forest  $F$  can be characterized in terms of MUS (minimal unsatisfiable subsets), as suggested in (Izza and Marques-Silva 2021). This characterization is useful because many SAT-based algorithms for computing a MUS (or even all MUSes) of a CNF formula have been pointed out for the past decade (Audemard, Lagniez, and Simon 2013; Liffiton et al. 2016; Marques-Silva, Janota, and Mencía 2017), and hence, one can take advantage of them for computing sufficient reasons.

Going one step further, a natural way for improving the clarity of sufficient reasons is to focus on those of minimal size. Specifically, given  $F \in \text{RF}_n$  and  $x \in \{0, 1\}^n$ , a *minimal sufficient reason* for  $x$  with respect to  $F$  is a sufficient reason for  $x$  given  $F$  of minimal size.<sup>3</sup>

**Example 4.** For our running example,  $x_1 \wedge x_4$  is the unique minimal sufficient reason for  $x$  given  $F$ , and  $\bar{x}_4$  is the unique minimal reason for  $x'$  given  $F$ .

As a by-product of the characterization of a sufficient reason in terms of MUS (Izza and Marques-Silva 2021), a minimal sufficient reason for  $x$  given  $f$  can be viewed as a *minimal* MUS. Thus, we can exploit algorithms for computing minimal MUSes (see e.g., (Ignatiev et al. 2015)) in order to infer minimal sufficient reasons. However, deriving a minimal sufficient reason is computationally harder than deriving a sufficient reason:

**Proposition 5.** Let  $F \in \text{RF}_n$ ,  $x \in \{0, 1\}^n$ , and  $k \in \mathbb{N}$ . Then, deciding whether there exists a minimal sufficient reason  $t$  for  $x$  given  $F$  containing at most  $k$  features is  $\Sigma_2^P$ -complete.

## Majoritary Reasons

Based on the above considerations, a natural question arises: does there exist a middle ground between direct reasons, which can include many irrelevant features but are easy to calculate, and sufficient reasons, which only contain relevant features but are potentially much harder to infer? Inspired by the way prime implicants can be computed when dealing with decision trees, we can reply in the affirmative using the notion of *majoritary reasons*, defined as follows.

**Definition 3.** Let  $F = \{T_1, \dots, T_m\}$  be a random forest in  $\text{RF}_{n,m}$  and  $x \in \{0, 1\}^n$  be an instance. Then, a *majoritary reason* for  $x$  given  $F$  is a term  $t$  covering  $x$ , such that  $t$  is an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees  $T_i$  (resp.  $\neg T_i$ ) if  $F(x) = 1$  (resp.  $F(x) = 0$ ), and for every  $l \in t$ ,  $t \setminus \{l\}$  does not satisfy this last condition.

**Example 5.** Based on our running example, the *majoritary reasons* for  $x$  given  $F$  are  $x_1 \wedge x_2 \wedge x_4$ ,  $x_1 \wedge x_3 \wedge x_4$ , and  $x_2 \wedge x_3 \wedge x_4$ . All these explanations are more concise than the direct reason  $t_x^F$ . For  $x'$ , the *majoritary reasons* given  $F$  are  $\bar{x}_1 \wedge \bar{x}_4$ ,  $x_2 \wedge \bar{x}_4$ , and  $\bar{x}_1 \wedge x_2 \wedge \bar{x}_3$ . Note that the *majoritary reasons*  $x_1 \wedge x_2 \wedge x_4$  and  $x_1 \wedge x_3 \wedge x_4$  for  $x$  given  $F$  include an irrelevant variable for the task of classifying  $x$  using  $F$ .

<sup>3</sup>Minimal sufficient reasons should not to be confused with *minimum-cardinality explanations* (Shih, Choi, and Darwiche 2018), where the minimality condition bears on the features set to 1 in the data instance  $x$ .

since  $x_1 \wedge x_4$  is a sufficient reason for  $x$  given  $F$ . Similarly, all *majoritary reasons* for  $x'$  given  $F$  contain an irrelevant variable for the task of classifying  $x'$  using  $F$ .

In general, *majoritary reasons* and *sufficient reasons* do not coincide. Indeed, a sufficient reason is a prime implicant (covering  $x$ ) of the forest  $F$ , while a *majoritary reason* is an implicant  $t$  (covering  $x$ ) of a majority of decision trees in the forest  $F$ , satisfying the additional condition that  $t$  is a prime implicant of at least one of these decision trees.

A key observation justifying this difference is that even if every implicant of a Boolean function  $f$  is an implicant of the function  $f \vee g$ , it is not always the case that every *prime* implicant of  $f$  is a prime implicant of  $f \vee g$ . To this point, consider our running example and take the term  $t = x_1 \wedge x_3 \wedge x_4$ . Here,  $t$  is a *majoritary reason* for  $x = (1, 1, 1, 1)$  given  $F$ , since it covers  $x$ , it is a prime implicant of  $T_1$ , and it is an implicant of  $T_2$ . Thus,  $t$  is an implicant of  $f = T_1 \wedge T_2$  (it is a prime one), and hence an implicant of  $F$ , using the fact that  $F$  is logically equivalent to  $(T_1 \wedge T_2) \vee (T_1 \wedge T_3) \vee (T_2 \wedge T_3)$ . However,  $t$  is *not* a prime implicant of  $F$ . Indeed, the sub-term  $x_1 \wedge x_4$  is a sufficient reason for  $x$  given  $F$ , since it is a prime implicant of  $F$  that covers  $x$ .

Viewing *majoritary reasons* as “weak” forms of sufficient reasons, they can include irrelevant features:

**Proposition 6.** Let  $F = \{T_1, \dots, T_m\}$  be a random forest of  $\text{RF}_{n,m}$  and  $x \in \{0, 1\}^n$  such that  $F(x) = 1$ . Unless  $m < 3$ , it can be the case that every *majoritary reason* for  $x$  given  $F$  contains arbitrarily many more features than any sufficient reason for  $x$  given  $F$ .

What makes *majoritary reasons* valuable is that they are abductive and can be generated in linear time. The evidence that any *majoritary reason*  $t$  for  $x$  given  $F$  is an abductive explanation comes directly from the fact that if  $t$  implies a majority of decision trees in  $F$ , then it is an implicant of  $F$  (note that the converse implication does not hold in general).

The tractability of generating *majoritary reasons* lies in the fact that they can be found using a simple greedy algorithm. For the case where  $F(x) = 1$ , start with  $t = t_x$ , and iterate over the literals  $l$  of  $t$  by checking whether  $t$  deprived of  $l$  is an implicant of at least  $\lfloor \frac{m}{2} \rfloor + 1$  decision trees of  $F$ . If so, remove  $l$  from  $t$  and proceed to the next literal. Once all literals in  $t_x$  have been examined, the final term  $t$  is by construction an implicant of a majority of decision trees in  $F$ , such that removing any literal from it would lead to a term that is no longer an implicant of this majority. So,  $t$  is by construction a *majoritary reason*. The case where  $F(x) = 0$  is similar, by simply replacing each  $T_i$  with its negation in  $F$ . This greedy algorithm runs in  $\mathcal{O}(n|F|)$  time, using the fact that, on each iteration, checking whether  $t$  is an implicant of  $T_i$  (for each  $i \in [m]$ ) can be done in  $\mathcal{O}(|T_i|)$  time.

By analogy with *minimal sufficient reasons*, a natural way of improving the quality of *majoritary reasons* is to seek for shortest ones. Formally, a *minimal majority reason* for an instance  $x \in \{0, 1\}^n$  and a random forest  $F \in \text{RF}_n$  is a minimal-size *majoritary reason* for  $x$  given  $F$ .

**Example 6.** For our running example, the three *majoritary reasons* for  $x$  given  $F$  are *minimal ones*. Contrastingly,



among the majoritary reasons for  $\mathbf{x}'$  given  $F$ , only  $\bar{x}_1 \wedge \bar{x}_4$  and  $x_2 \wedge \bar{x}_4$  are minimal.

Unsurprisingly, the optimization task for majoritary reasons is more demanding than the generation task. Still, minimal majoritary reasons are easier to find than minimal sufficient reasons. In more formal terms:

**Proposition 7.** *Let  $F \in \text{RF}_n$ ,  $\mathbf{x} \in \{0, 1\}^n$ , and  $k \in \mathbb{N}$ . Then, deciding whether there exists a minimal majoritary reason  $t$  for  $\mathbf{x}$  given  $F$  containing at most  $k$  features is NP-complete.*

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

**Proposition 8.** *Let  $F \in \text{RF}_{n,m}$  and  $\mathbf{x} \in \{0, 1\}^n$  be an instance such that  $F(\mathbf{x}) = 1$ . Let  $(C_{\text{soft}}, C_{\text{hard}})$  be an instance of the PARTIAL MAXSAT problem such that:*

$$\begin{aligned} C_{\text{soft}} &= \{\bar{x}_i : x_i \in t_{\mathbf{x}}\} \cup \{x_i : \bar{x}_i \in t_{\mathbf{x}}\} \\ C_{\text{hard}} &= \{(\bar{y}_i \vee c_{|\mathbf{x}}) : i \in [m], c \in \text{CNF}(T_i)\} \\ &\cup \text{CNF}\left(\sum_{i=1}^m y_i > \frac{m}{2}\right) \end{aligned}$$

where  $c_{|\mathbf{x}} = c \cap t_{\mathbf{x}}$  is the restriction of  $c$  to the literals in  $t_{\mathbf{x}}$ ,  $\{y_1, \dots, y_m\}$  are fresh variables, and  $\text{CNF}(\sum_{i=1}^m y_i > \frac{m}{2})$  is a CNF encoding of the constraint  $\sum_{i=1}^m y_i > \frac{m}{2}$ . Let  $\mathbf{z}^*$  be an optimal solution of  $(C_{\text{soft}}, C_{\text{hard}})$ . Then, the intersection of  $t_{\mathbf{x}}$  with  $t_{\mathbf{z}^*}$  is a minimal majoritary reason for  $\mathbf{x}$  given  $F$ .

Clearly, in the case where  $F(\mathbf{x}) = 0$ , it is enough to consider the same instance of PARTIAL MAXSAT as above, except that  $C_{\text{hard}} = \{(\bar{y}_i \vee c_{|\mathbf{x}}) : i \in [m], c \in \text{CNF}(\neg T_i)\} \cup \text{CNF}(\sum_{i=1}^m y_i > \frac{m}{2})$ .

Thanks to this characterization result, one can leverage the numerous algorithms that have been developed so far for PARTIAL MAXSAT (see e.g. (Ansótegui, Bonet, and Levy 2013; Morgado, Ignatiev, and Marques-Silva 2014; Narodytska and Bacchus 2014; Saikko, Berg, and Järvisalo 2016)) in order to compute minimal majoritary reasons.

## Experiments

**Experimental Setup.** The empirical protocol was as follows. We have considered 15 datasets for binary classification, which are standard benchmarks from the repositories Kaggle (www.kaggle.com), OpenML (www.openml.org), and UCI (archive.ics.uci.edu/ml/). These datasets are *compas*, *placement*, *recidivism*, *adult*, *ad-data*, *mnist38*, *mnist49*, *gisette*, *dexter*, *dorothea*, *farm-ads*, *higgs-boson*, *christine*, *gina*, and *bank*. *mnist38* and *mnist49* are subsets of the *mnist* dataset, restricted to the instances of 3 and 8 (resp. 4 and 9) digits. Due to space constraints, additional information about the datasets (especially the numbers and types of features, the number of instances), and about the

random forests that have been trained (especially, the number of Boolean features used, the number of trees, the depth of the trees, the mean accuracy) are reported as a supplementary material.

Categorical features have been treated as arbitrary numbers (the scale is nominal). As to numeric features, no data preprocessing has taken place: these features have been binarized on-the-fly by the random forest learning algorithm. For this learner, we have used the version 0.23.2 of the Scikit-Learn library (Pedregosa et al. 2011). The maximal depth of any decision tree in a forest has been bounded at 8. All other hyper-parameters of the learning algorithm have been set to their default value except the number of trees. We made some preliminary tests for tuning this parameter in order to ensure that the accuracy is good enough.

For every benchmark  $b$ , a 10-fold cross validation process has been achieved: a set of 10 random forests have been computed and evaluated from the labelled instances of  $b$ , partitioned into 10 parts. One part was used as the test set and the remaining 9 parts as the training set for generating a forest. The classification performance on  $b$  was measured using the mean accuracy obtained over the 10 random forests. For each benchmark  $b$ , each random forest  $F$ , and a pool of 25 instances  $\mathbf{x}$  drawn at random from the test set (leading to 250 instances per dataset), we have run the algorithms described in the previous section for deriving the direct reason for  $\mathbf{x}$  given  $F$ , a sufficient reason for  $\mathbf{x}$  given  $F$ , a majoritary reason  $\mathbf{x}$  given  $F$ , a minimal majoritary reason for  $\mathbf{x}$  given  $F$ , and a minimal sufficient reason for  $\mathbf{x}$  given  $F$ .

For computing sufficient reasons and minimal majoritary reasons, we took advantage of the Pysat library (Ignatiev, Morgado, and Marques-Silva 2018) (version 0.1.6.dev15) which provides the implementation of the RC2 PARTIAL MAXSAT solver and an interface to MUSER (Belov and Marques-Silva 2012). For majority reasons, we picked uniformly at random 50 permutations of the literals describing the instance and tried to eliminate those literals (within the greedy algorithm) following the ordering corresponding to the permutation. We kept a smallest explanation among those that have been derived (of course, the corresponding runtime that has been measured is the cumulated time over the 50 tries). Sufficient reasons have been computed as MUSes, as explained before.

We also derived a “LIME explanation” for each instance. Such an explanation has been inferred as follows. Given an input instance  $\mathbf{x}$  under consideration, we first used LIME (Ribeiro, Singh, and Guestrin 2016) to generate a linear zero-threshold function  $w_{\mathbf{x}} \in \mathbb{R}^n$ . In other words, the value  $w_{\mathbf{x}}(z)$  of  $w_{\mathbf{x}}$  on any instance  $z$  is given by  $w_{\mathbf{x}}(z) = 1$  if  $w_{\mathbf{x}}^T z > 0$ , and  $w_{\mathbf{x}}(z) = 0$  otherwise. Now, if  $\mathbf{x}$  is classified positively by  $w_{\mathbf{x}}$ , then in order to derive an explanation, it is enough to sum in a decreasing way the positive weights occurring in  $w_{\mathbf{x}}$  until this sum exceeds (the opposite of) the sum of all the negative weights occurring in  $w_{\mathbf{x}}$ . The term  $t$  composed of the variables  $x_i$  associated with the positive weights which have been selected is, by construction, a minimal sufficient reason for  $\mathbf{x}$  given  $w_{\mathbf{x}}$  since for every  $\mathbf{x}'$  cov-



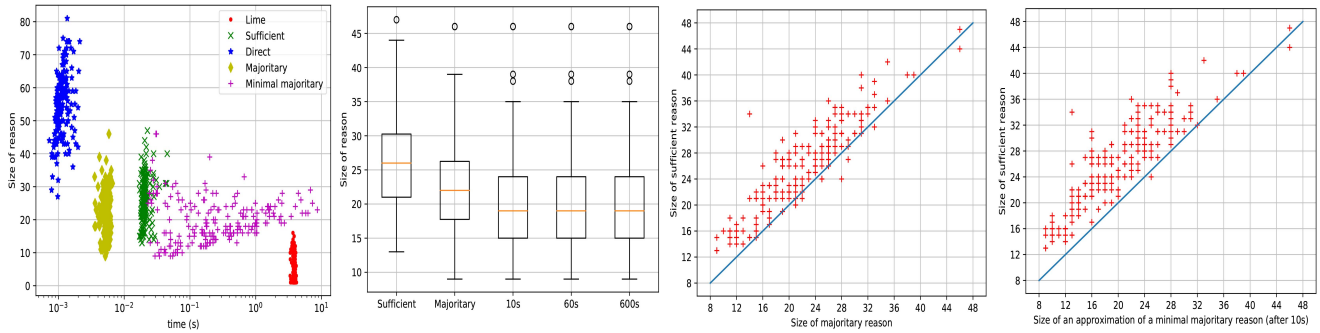


Figure 2: Empirical results for the *placement* dataset.

ered by  $t$ , the inequality  $w_x^\top x' > 0$  holds.<sup>4</sup> Instances that are classified negatively can be handled in a similar way.

All the experiments have been conducted on a computer equipped with Intel(R) XEON E5-2637 CPU @ 3.5 GHz and 128 Gib of memory. A time-out (TO) of 600s has been considered for each instance and each type of explanation, except LIME explanations.

**Experimental Results.** A first conclusion that can be drawn from our experiments is the intractability of computing in practice minimal sufficient reasons (this is not surprising, since this coheres with the complexity result given by Proposition 5). Indeed, we have been able to compute within the time limit of 600s a minimal reason for only 10 instances and a single dataset (*compas*).

Due to space limitations, we report hereafter empirical results about three datasets only, namely *placement*, *gisette* and *dorothea*. The results obtained on the other datasets are similar and available as a supplementary material. *placement* is a small dataset about the placement of 215 students in a campus; students are described using 13 features, related to their curricula, the type and work experience, and the salary. An instance is labelled as positive when the student gets a job. The random forests consist of 25 trees, and their mean accuracy is 97.6%. *gisette* is much larger, including 5000 features and 7000 examples. Each feature is a pixel, and the task is to separate the digits 4 and 9. The random forests consist of 85 trees, and their mean accuracy is 96%. Finally, *dorothea* is a high-dimensional dataset, with 100,000 features and 1950 examples. Each instance is an organic molecule, and the goal is to discriminate binding compounds from non-binding ones. Here, the random forest consists of 71 trees, with a mean accuracy of 93%.

Figure 2 provides the results obtained for *placement*, using four plots. Each dot represents an instance. The first plot shows the time needed to compute a reason on the x-axis, and the size of this reason on the y-axis. On this plot, there are no dots for minimal sufficient reasons, because their computation did not terminate before the time-out. The plot also highlights that all other reasons have been computed

within the time limit, and in general using a small amount of time. In particular, it shows that the direct reason can be quite large, that the computation of LIME explanations is usually more expensive than the ones of the other explanations, and that LIME explanations can be very short.<sup>5</sup> A box plot about the sizes of all the explanations is reported (the LIME ones and the direct reasons are not presented for the sake of readability). The figure also provides two scatter plots, aiming to compare the size of majority reasons with the size of sufficient reasons, as well as the size of the minimal majority reasons with the size of sufficient reasons. These plots clearly show the benefits w.r.t. size reduction that can be offered by considering majority reasons and minimal majority reasons instead of sufficient reasons. At first sight, these empirical observations may look surprising since, by construction, for any majority reason  $t$  for  $x$  given  $f$  (including the minimal ones) there exists at least one sufficient reason for  $x$  given  $f$  that is implied by  $t$  (hence that cannot be larger). As to majority reasons, one must keep in mind that the result that is reported is a shortest reason out of a set of 50 majority reasons that are computed for each  $x$  (so to say, we leverage the tractability of computing such reasons to tackle the size issue). For minimal majority reasons, the PARTIAL MAXSAT algorithm used to compute them aims at minimizing the size of the reason that is derived, while MUS algorithms for computing sufficient reasons do not focus on the size (computing minimal MUSes is much harder, as explained previously).

Figures 3 and 4 synthesize the results obtained for *gisette* and *dorothea*, respectively. Conclusions similar to those drawn for *placement* can be derived for *gisette* and *dorothea*, with some exceptions. First of all, there are here no dots for minimal majority reasons because their computation did not terminate before the time-out. Furthermore, LIME explanations are much longer. This can be partly explained by the fact that the computation achieved by LIME relies on a binary representation of the instance that is quite different (and possibly much larger) than the one considered in the representation of the random forest. Indeed, each decision tree in the forest focuses only on a subset of most impor-

<sup>4</sup>Indeed, the inequality  $w_x^\top x' > 0$  holds in the worst situation where all the variables associated with a positive weight in  $w_x$  and not belonging to  $t$  are set to 0, whilst all the variables associated with a negative weight in  $w_x$  are set to 1.

<sup>5</sup>Recall that LIME explanations are not guaranteed to be abductive. See also (Narodytska et al. 2019) that reports some experiments about ANCHOR explanations.



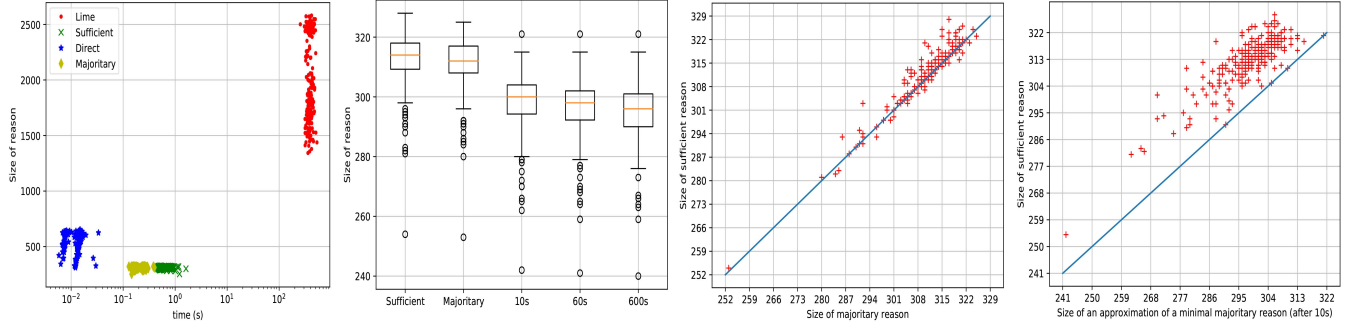


Figure 3: Empirical results for the *gisette* dataset.

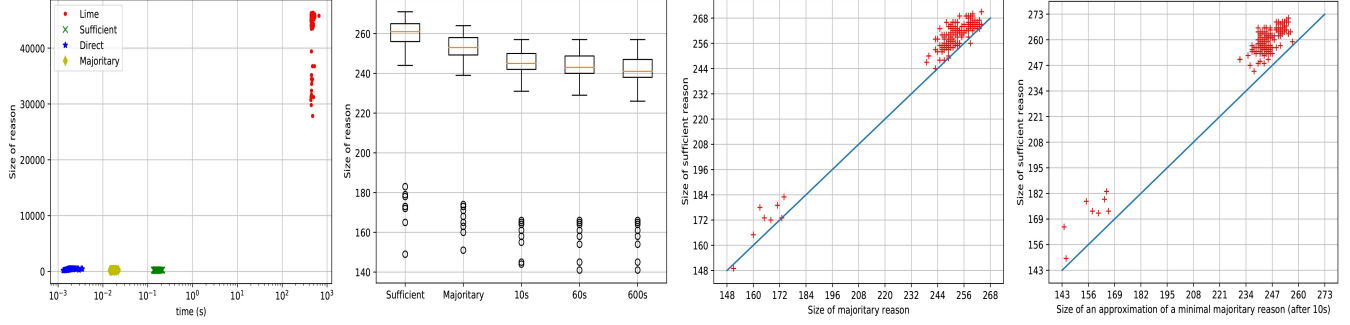


Figure 4: Empirical results for the *dorothea* dataset.

tant features (in the sense of Gini criterion) found during the learning phase. In our experiments, the size of LIME explanations was typically large for high-dimensional datasets.

When minimal majority reasons are difficult to calculate (as it is the case for *gisette* and *dorothea*), a natural approach is to approximate them. From this perspective, we can take advantage of an incremental PARTIAL MAXSAT algorithm, like LMHS (Saikko, Berg, and Järvisalo 2016), to do the job. Specifically, the result given in Proposition 8 provides a way to derive abductive explanations for an instance  $x$  given a random forest  $F$  in an *anytime* fashion. Basically, using LMHS, a Boolean assignment  $z$  satisfying all the hard constraints of  $C_{\text{hard}}$  and a given number, say  $k$ , of soft constraints from  $C_{\text{soft}}$  is looked for ( $k$  is set to 0 at start). If such an assignment is found, then one looks for an assignment satisfying  $k + 1$  soft constraint, and so on, until an optimal solution is found or a preset time bound is reached. In many cases, the most demanding step from a computational standpoint is the one for which  $k$  is the optimal value (but one ignores it); we look for an assignment that satisfies  $k + 1$  soft constraint (and such an assignment does not exist). By construction, every  $z$  that is generated that way is such that  $t_x \cap t_z$  is an implicant of  $F$  that covers  $x$  (and hence, an abductive explanation). The approximation  $z$  of a minimal majority reason for  $x$  given  $F$ , which is obtained when the time limit is met, can be significantly shorter than the sufficient reason for  $x$  given  $F$  that has been derived.

In our experiments, we used three time limits: 10s, 60s, and 600s. The results are reported in the box plots and the scatter plots in Figures 2, 3, and 4. As illustrated by the box

plots, the sizes of the approximations  $z$  which are derived gently decrease with time. The scatter plots indicate that significant size savings can be achieved even for the smallest time bound of 10s that has been considered.

## Conclusion

In this paper, we have introduced, analyzed and evaluated some new notions of abductive explanations suited to random forest classifiers, namely majority reasons and minimal majority reasons. Our investigation reveals the existence of a trade-off between conciseness and runtime complexity for abductive explanations. Unlike sufficient reasons, majority reasons and minimal majority reasons may contain irrelevant features. Nevertheless, despite this evidence, majority reasons and minimal majority reasons appear as valuable alternative to sufficient reasons. Indeed, majority reasons can be computed in polynomial time while sufficient reasons cannot (unless  $P = NP$ ). In addition, in most of our experiments, majority reasons slightly smaller than sufficient reasons can be computed thanks to a simple greedy algorithm with random permutations of literals. Minimal majority reasons can be looked for when majority reasons are too large, but this is at the cost of an extra computation time that can be important, and even prohibitive in some cases. However, minimal majority reasons can be approximated using an *anytime* PARTIAL MAXSAT algorithm. Empirically, approximations can be derived within a small amount of time and their sizes are significantly smaller than the ones of sufficient reasons.



## References

- Ansótegui, C.; Bonet, M. L.; and Levy, J. 2013. SAT-based MaxSAT algorithms. *Artificial Intelligence*, 196: 77–105.
- Audemard, G.; Lagniez, J.-M.; and Simon, L. 2013. Improving Glucose for Incremental SAT Solving with Assumptions: Application to MUS Extraction. In *Proceedings of the 16th International Conference on Theory and Applications of Satisfiability Testing (SAT’13)*, 309–317.
- Azar, A. T.; Elshazly, H. I.; Hassanien, A. E.; and Elkorany, A. M. 2014. A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, 113(2): 465–473.
- Belov, A.; and Marques-Silva, J. 2012. MUSer2: An Efficient MUS Extractor. *J. Satisf. Boolean Model. Comput.*, 8(3/4): 123–128.
- Bénard, 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*, 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, ICM’14*, 427–434. ACM.
- Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.
- Chen, X.; and Ishwaran, H. 2012. Random forests for genomic data analysis. *Genomics*, 99(6): 323–329.
- Choi, A.; Shih, A.; Goyanka, A.; and Darwiche, A. 2020. On Symbolically Encoding the Behavior of Random Forests. In *Proceedings of the 3rd Workshop on Formal Methods for ML-Enabled Autonomous Systems (FoMLAS)*.
- Criminisi, A.; and Shotton, J. 2013. *Decision Forests for Computer Vision and Medical Image Analysis*. Advances in Computer Vision and Pattern Recognition. Springer.
- Cutler, R.; Thomas, C. E. J.; Beard, K. H.; Cutler, A.; Hess, K. T.; Gibson, J.; and Lawler, J. J. 2007. Random Forests for Classification in Ecology. *Ecology*, 88(11): 2783–2792.
- Darwiche, A.; and Hirth, A. 2020. On the Reasons Behind Decisions. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI’20)*, 712–720.
- Darwiche, A.; and Marquis, P. 2002. A knowledge compilation map. *Journal of Artificial Intelligence Research*, 17: 229–264.
- Ignatiev, A.; Morgado, A.; and Marques-Silva, J. 2018. PySAT: A Python Toolkit for Prototyping with SAT Oracles. In *Proceedings of the 21st International Conference on Theory and Applications of Satisfiability Testing (SAT’2018)*, 428–437.
- Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-Based Explanations for Machine Learning Models. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI’19)*, 1511–1519.
- Ignatiev, A.; Previti, A.; Liffiton, M.; and Marques-Silva, J. 2015. Smallest MUS Extraction with Minimal Hitting Set Dualization. In *Proceedings of the 21st International Conference on Principles and Practice of Constraint Programming (CP’15)*, 173–182.
- Izza, Y.; Ignatiev, A.; and Marques-Silva, J. 2020. On Explaining Decision Trees. *CoRR*, abs/2010.11034.
- Izza, Y.; and Marques-Silva, J. 2021. On Explaining Random Forests with SAT. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI’21)*, 2584–2591.
- Karp, R. 1972. *Reducibility among combinatorial problems*, chapter Complexity of Computer Computations, 85–103. New York: Plenum Press.
- Lang, J.; Liberatore, P.; and Marquis, P. 2003. Propositional Independence: Formula-Variable Independence and Forgetting. *Journal of Artificial Intelligence Research*, 18: 391–443.
- Liberatore, P. 2005. Redundancy in logic I: CNF propositional formulae. *Artificial Intelligence*, 163(2): 203–232.
- Liffiton, M.; Previti, A.; Malik, A.; and Marques-Silva, J. 2016. Fast, flexible MUS enumeration. *Constraints An Int. J.*, 21(2): 223–250.
- Marques-Silva, J.; Janota, M.; and Mencía, C. 2017. Minimal sets on propositional formulae. Problems and reductions. *Artificial Intelligence*, 252: 22–50.
- 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.
- Morgado, A.; Ignatiev, A.; and Marques-Silva, J. 2014. MSCG: Robust Core-Guided MaxSAT Solving. *J. Satisf. Boolean Model. Comput.*, 9(1): 129–134.
- Narodytska, N.; and Bacchus, F. 2014. Maximum Satisfiability Using Core-Guided MaxSAT Resolution. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2717–2723.
- Narodytska, N.; Shrotri, A.; Meel, K.; Ignatiev, A.; and Marques-Silva, J. 2019. Assessing Heuristic Machine Learning Explanations with Model Counting. In *Proceedings of 22nd International Conference on the Theory and Applications of Satisfiability Testing (SAT’19)*, 267–278.
- 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 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.
- Ribeiro, M.-T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In McIlraith,



S. A.; and Weinberger, K. Q., eds., *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 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)*, 539–546.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A Symbolic Approach to Explaining Bayesian Network Classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI'18)*, 5103–5111.



## Proofs

For every  $f, g \in \mathcal{F}_n$ , we note  $f \models g$  when for every  $x \in \{0, 1\}^n$ ,  $f(x) = 1$  implies that  $g(x) = 1$ .

### Proof of Proposition 1

*Proof.* By definition, an instance  $x$  is a model of the negation of a given random forest  $F = \{T_1, \dots, T_m\}$  if and only if it is a model of at most  $\frac{m}{2}$  trees among those of  $T$ . Let us state (w.l.o.g.) that  $x$  is a model of  $T_1, \dots, T_k$  and a counter-model of  $T_{k+1}, \dots, T_m$ , with  $k \leq \frac{m}{2}$ . Equivalently, we have that  $x$  is a counter-model of  $T'_1, \dots, T'_k$  and a model of  $T'_{k+1}, \dots, T'_m$  where each  $T'_i$  ( $i \in \{1, \dots, m\}$ ) is a decision tree equivalent to the negation of  $T_i$ . This precisely means that  $x$  is a model of  $\neg F = \{T'_1, \dots, T'_m\}$ . Since each  $T'_i$  ( $i \in \{1, \dots, m\}$ ) can be computed in time linear in  $|T_i|$ , the result follows.  $\square$

### Proof of Proposition 2

*Proof.* Let  $G = c_1 \wedge \dots \wedge c_p$  be a CNF formula with  $p > 0$  clauses. Each  $c_i$  can be transformed into a decision tree  $T_i$  using the following linear-time recursive algorithm. For the base cases, if  $c_i$  is empty, then  $T_i = 0$  and if  $c_i$  is a tautology, then  $T_i = 1$ . For the inductive case, suppose that  $c_i = l_j \vee c'_i$  and let  $T'_j$  the decision tree encoding  $c'_i$ . If  $l_j = x_j$  (resp.  $l_j = \bar{x}_j$ ), then  $T_i$  is the decision tree rooted at  $x_j$  with right (resp. left) child labeled by the leaf 1 (resp. 0) and with left (resp. right) child encoding  $T'_j$ . Now, let  $F = \{T_1, \dots, T_p, T_{p+1}, \dots, T_q\}$ , where  $T_{p+1}, \dots, T_q$  are decision trees rooted at 0, and  $q = 2p - 1$ . For any input instance  $x \in \{0, 1\}^n$ , we have  $G(x) = 1$  iff  $T_i(x) = 1$  for every  $i \in [p]$ . Since for any positive integer  $p$ , the function

$$\phi_p(z) = \frac{p - z}{2p - 1}$$

satisfies  $\phi_p(0) > 1/2$  and  $\phi_p(z) < 1/2$  for  $z \in [p]$ , it follows that  $G(x) = 1$  iff  $F(x) > 1/2$ .

The case for DNF formulas is dual: given  $G = t_1 \vee \dots \vee t_p$ , compute in linear time a CNF formula equivalent to  $\neg G$ , then turn it in linear time into an equivalent random forest using the transformation above, and finally negate in linear time the resulting random forest by taking advantage of Proposition 1.

Since every CNF formula  $G$  can be turned in linear time into an equivalent random forest  $G'$  (as we have just proved it), the size of  $G'$  is polynomial in the size of  $G$  for a fixed polynomial (independent of  $G$ ). Then, exploiting Proposition 1, one can negate  $G'$  in linear time. The resulting random forest  $G''$  is equivalent to the negation of  $G$  and its size is also polynomial in the size of  $G$  for a fixed polynomial.

Finally, suppose, towards a contradiction, that a polynomial-space translation from RF to CNF exists. If so, one could compute a CNF formula  $G'''$  equivalent to  $G''$  and having a size polynomial in the size of  $G''$  for a fixed polynomial. Thus, the CNF formula  $G'''$  would have a size polynomial in the size of  $G$  for a fixed polynomial. This CNF formula could be negated in linear time into a DNF formula  $G''''$  by applying De Morgan's laws. By construction,  $G''''$  would be a DNF formula equivalent to  $G$ , and its size

would be polynomial in the size of  $G$  for a fixed polynomial. This conflicts with the fact that there is no polynomial-space translation from CNF to DNF, see e.g., (Darwiche and Marquis 2002). Using duality, we prove similarly that there is no polynomial-space translation from RF to DNF.  $\square$

### Proof of Proposition 3

*Proof.*

- **Membership to coNP:** we show that the complementary problem, i.e., the problem of deciding whether a term  $t$  is not an implicant of a random forest  $F$ , is in NP. This is direct given the characterization of the implicants of  $F$  provided by Proposition 4: it is enough to compute in time polynomial in the size of  $F$  the CNF formula  $H$  given in Proposition 4, and to exploit the fact that  $t$  is not an implicant of  $F$  if and only if  $t \wedge H$  is satisfiable. Finally, deciding whether  $t \wedge H$  is satisfiable can be easily achieved by a non-deterministic algorithm running in time polynomial in the size of the input (just guess a truth assignment over the variables occurring in  $t \wedge H$  and check in polynomial time that this assignment is a model of  $t \wedge H$ ).
- **coNP-hardness:** by reduction from VAL, the validity problem for DNF formulae. Let  $G = d_1 \vee \dots \vee d_p$  be a DNF formula over  $X_n$ . We can associate with  $G$  in polynomial time an equivalent random forest  $F$  using Proposition 2. Consider now the term  $t = \top$ .  $t$  is an implicant of  $F$  if and only if  $G$  is valid.  $\square$

### Proof of Proposition 4

*Proof.* We have  $t \not\models F$  if and only if  $t \wedge \neg F$  is satisfiable. From  $F$ , exploiting Proposition 1 one can generate in polynomial time a random forest  $F' = \{\neg T_1, \dots, \neg T_m\}$  equivalent to  $\neg F$ . Each  $\neg T_i$  is the decision tree obtained by replacing every 1-leaf in  $T_i$  by a 0-leaf, and vice-versa. We thus have  $t \not\models F$  if and only if  $t \wedge F'$  is satisfiable. Then  $F'$  can be associated in polynomial time with the following Boolean quantified formula  $\exists Y.H$  when  $Y = \{y_i : i \in [m]\} \cup A$  is a set of new variables and  $H$  is a CNF formula which is the conjunction of the clauses of

$$\{(\bar{y}_i \vee c) : i \in [m], c \in \text{CNF}(T'_i)\}$$

with a CNF encoding of the cardinality constraint

$$\sum_{i=1}^m y_i > \frac{m}{2}$$

using auxiliary variables in  $A$ .  $F'$  is equivalent to  $\exists Y.H$ , therefore  $t \wedge F'$  is satisfiable if and only if  $t \wedge \exists Y.H$  is satisfiable. Since the variables of  $Y$  do not occur in  $t$ ,  $t \wedge \exists Y.H$  is equivalent to  $\exists Y.(t \wedge H)$ . Since  $\exists Y.(t \wedge H)$  is satisfiable if and only if  $t \wedge H$  is satisfiable, we get that  $t \wedge \exists Y.H$  is satisfiable if and only if  $t \wedge H$  is satisfiable.  $\square$

### Proof of Proposition 5

*Proof.*



- Membership to  $\Sigma_2^P$ : if there exists a minimal reason  $t$  for  $x$  given  $f$  such that  $t$  contains at most  $k$  features, then one can guess  $t$  using a nondeterministic algorithm running in polynomial time (the size of  $t$  is bounded by the size of  $x$ ), then check in polynomial time that  $t$  is a sufficient reason for  $x$  given  $f$  using an NP-oracle (this comes directly from the fact that this problem belongs to DP), and finally check in polynomial time that the size of  $t$  is upper bounded by  $k$ .
- $\Sigma_2^P$ -hardness: in (Liberatore 2005) (Theorem 2), it is shown that the problem of deciding whether a CNF formula  $\Pi = \bigwedge_{i=1}^p c_i$  has an irredundant equivalent subset of size at most  $k$  is  $\Sigma_2^P$ -complete, and that the problem is  $\Sigma_2^P$ -hard even in the case when  $\Pi$  is unsatisfiable. Whenever  $\Pi$  is unsatisfiable, an irredundant equivalent subset of  $\Pi$  precisely is a MUS of  $\Pi$  (every clause being considered as a soft clause). Accordingly, there exists an irredundant equivalent subset  $E$  of an unsatisfiable CNF formula  $\Pi$  such  $E$  is of size at most  $k$  if and only if there exists a MUS  $I = \{y_i : c_i \in E\}$  of  $S = \{y_i : c_i \in \Pi\}$  given  $H = \{\bar{y}_i \vee c_i : c_i \in \Pi\}$  such that  $I$  is of size at most  $k$ . Because of this equivalence, the problem of deciding whether  $S$  has a MUS of size at most  $k$  given  $H$  has the same complexity as the problem of deciding whether  $\Pi$  has an irredundant equivalent subset of size at most  $k$ , so it is  $\Sigma_2^P$ -hard. Finally, we reduce this latter problem to the one of deciding whether a term is a minimal reason for an instance given a random forest. The reduction is as follows. With  $(H, S)$  where  $S$  is satisfiable and  $H \cup S$  is unsatisfiable (as obtained from the previous reduction), one associates in polynomial time the pair  $(x, F)$  where  $x$  is any interpretation that extends  $S$  and  $F$  is a random forest from  $\text{RF}_{n,m}$  equivalent to  $\neg H$  (since  $H$  is a CNF formula, a DNF formula equivalent to  $\neg H$  can be computed in linear time from  $H$  and turned in linear time into an equivalent random forest  $F$  as shown by Proposition 2). Since  $H \cup S$  is unsatisfiable, we have  $S \models \neg H$  showing that  $x \models F$ . Now,  $I$  is a MUS of  $S$  given  $H$  if and only if  $I \cup H$  is unsatisfiable and for every  $l \in I$ ,  $(I \setminus \{l\}) \cup H$  is satisfiable. Taking  $t = I$ , this is equivalent to state that  $t \wedge \neg F$  is unsatisfiable and for every  $l \in t$ ,  $(t \setminus \{l\}) \wedge \neg F$  is satisfiable. Equivalently,  $t \models F$  and for every  $l \in t$ ,  $(t \setminus \{l\}) \not\models F$ , or stated otherwise  $t$  is a prime implicant of  $F$ . Since  $t = I$  and  $I \subseteq S$ , we also have  $S \models t$ , hence  $x \models t$ . Thus  $t$  is a sufficient reason for  $x$  given  $F$ . Since  $|I| = |t|$ , a MUS  $I$  of  $S$  given  $H$  such that  $|I| \leq k$  exists if and only if a sufficient reason  $t$  for  $x$  given  $F$  such that  $|t| \leq k$  exists. This completes the proof.  $\square$

### Proof of Proposition 6

*Proof.* Again, we focus only on the case when  $F(x) = 1$  (if  $F(x) = 0$ , it is enough to consider the random forest  $\neg F$  instead of  $F$ ).

If  $F$  contains at most 2 trees, then  $F$  is equivalent to the conjunction of its elements. In this case, testing whether a

term  $t$  implied by  $x$  is an implicant of  $F$  boils down to testing that  $t$  is an implicant of every tree in  $F$ , so that the sufficient reasons for  $x$  given  $F$  are precisely the majoritary reasons for  $x$  given  $F$ .

As to the case  $m \geq 3$ , whatever  $n \geq 1$ , let  $T$  be a decision tree equivalent to the parity function  $\oplus_{i=1}^n x_i$ . Consider the random forest  $F$  containing  $\lfloor \frac{m}{2} \rfloor$  copies of  $T$ ,  $\lfloor \frac{m}{2} \rfloor$  copies of the decision tree  $\neg T$ , and a decision tree reduced to a 1-leaf. By construction,  $F$  is valid. Indeed, among the subsets of  $F$  containing a strict majority of decision trees, one can find the one containing all the  $\lfloor \frac{m}{2} \rfloor$  copies of  $T$  plus the 1-leaf (their conjunction is thus equivalent to  $T$ ) and the one containing all the  $\lfloor \frac{m}{2} \rfloor$  copies of  $\neg T$  plus the 1-leaf (their conjunction is thus equivalent to  $\neg T$ ). Their disjunction is thus valid. As a consequence, whatever  $x$ , we have  $F(x) = 1$ . Indeed, we have either  $T(x) = 1$  or  $\neg T(x) = 1$  (and obviously,  $1(x) = 1$ ). Thus,  $t_x$  is an implicant of a (strict) majority of decision trees of  $F$ . Now, consider any literal  $l$  of  $t_x$ . The term  $t_x \setminus \{l\}$  is not an implicant of  $T$  nor an implicant of  $\neg T$  since the implicants of the parity function  $\oplus_{i=1}^n x_i$  (or of its negation) depend on every variable  $x_i$  ( $i \in \{1, \dots, n\}$ ). Therefore,  $t_x$  is the unique majoritary reason for  $x$  given  $F$  and it contains  $n$  characteristics. But since  $F$  is valid,  $\top$  is the unique sufficient reason for  $x$  given  $F$ .  $\square$

### Proof of Proposition 7

*Proof.*

- Membership to NP: if there exists a minimal majoritary reason  $t$  for  $x$  given  $F$  such that  $t$  contains at most  $k$  features, then one can guess  $t$  using a nondeterministic algorithm running in polynomial time (the size of  $t$  is bounded by the size of  $x$ ), then check in polynomial time whether  $t$  is a sufficient reason for  $x$  given  $T$  for a majority of trees  $T \in F$ , and finally check in polynomial time that the size of  $t$  is upper bounded by  $k$ .
- NP-hardness: in the following, we focus only on the case when  $F(x) = 1$  (if  $F(x) = 0$ , it is enough to consider the random forest  $\neg F$  instead of  $F$ ; this is harmless given that  $\neg F$  can be computed in time linear in the size of  $F$ , see Proposition 1). We assume that  $m = 1$ , i.e.,  $F$  consists of a single decision tree  $T \in \text{DT}_n$ .

We call MINIMAL SUFFICIENT REASON the problem that asks, given  $T \in \text{DT}_n$ ,  $x \in \{0, 1\}^n$  with  $T(x) = 1$  and  $k \in \mathbb{N}$ , whether there is an implicant  $t$  of  $T$  of size at most  $k$  that covers  $x$ .

Our objective is to prove that MINIMAL SUFFICIENT REASON is NP-hard. To this end, let us first recall that a *vertex cover* of an undirected graph  $G = (X, E)$  is a subset  $V \subseteq X$  of vertices such that  $\{y, z\} \cap V \neq \emptyset$  for every edge  $e = \{y, z\}$  in  $E$ . In the MIN VERTEX COVER problem, we are given a graph  $G$  together with an integer  $k \in \mathbb{N}$ , and the task is to find a vertex cover  $V$  of  $G$  of size at most  $k$ . MIN VERTEX COVER is a well-known NP-hard problem (Karp 1972), and we now show that it can be reduced in polynomial time to MINIMAL SUFFICIENT REASON.

Suppose that we are given a graph  $G = (X, E)$  and assume, without loss of generality, that  $G$  does not include



isolated vertices. For any  $y \in X$ , let  $E_y = \{e \in E : y \in e\}$  denote the set of edges in  $G$  that are adjacent to  $y$ , and let  $N_y = \{z \in X : \{y, z\} \in E\}$  denote the set of neighbors of  $y$  in  $G$ . By  $G \setminus y$ , we denote the deletion of  $y$  from  $G$ , obtained by removing  $y$  and its adjacent edges, i.e.,  $G \setminus y = (X \setminus \{y\}, E \setminus E_y)$ . We associate with  $G$  a decision tree  $T(G)$  over  $X_n = X$  using the following recursive algorithm. If  $G$  is the empty graph (i.e.  $E = \emptyset$ ), then return the decision tree rooted at a 1-leaf. Otherwise, pick a node  $y \in X$  and generate a decision tree  $T(G)$  such that:

- (1) the root is labeled by  $y$ ;
- (2) the left child is the decision tree encoding the monomial  $\bigwedge N_y$ ;
- (3) the right child is the decision tree  $T(G')$  returned by calling the algorithm on  $G' = G \setminus y$ .

By construction,  $T(G)$  is a complete backtrack search tree of the formula  $\text{CNF}(E) = \bigwedge \{(y \vee z) : \{y, z\} \in E\}$ , which implies that  $T(G)$  and  $\text{CNF}(E)$  are logically equivalent. Furthermore,  $T(G)$  is a comb-shaped tree since recursion only on the rightmost branch. In particular, the algorithm runs in  $\mathcal{O}(n|E|)$  time, since step (1) takes  $\mathcal{O}(1)$  time, step (2) takes  $\mathcal{O}(n)$  time, and step (3) is called at most  $|E|$  times.

Now, with an instance  $P_1 = (G, k)$  of MIN VERTEX COVER, we associate the instance  $P_2 = (T(G), x, k)$  of MINIMAL SUFFICIENT REASON, where  $x = (1, \dots, 1)$ . Based on the above algorithm,  $P_2$  can be constructed in time polynomial in the size of  $P_1$ .

Let  $V$  be a solution of  $P_1$ . Since  $V$  is a vertex cover of  $G$ , the term  $t_V = \bigwedge V$  is an implicant of the formula  $\text{CNF}(E)$ . Since  $t_V \subseteq t_x$  and  $|t_V| \leq k$ , it follows from the fact that  $\text{CNF}(E)$  and  $T(G)$  are logically equivalent that  $t_V$  is a solution of  $P_2$ .

Conversely, let  $t$  be a solution of  $P_2$ . Since  $t$  is an implicant of  $T(G)$ , it follows that  $t$  is an implicant of  $\text{CNF}(E)$ . This together with the fact that  $t \subseteq t_x$  implies that the subset of vertices  $V \subseteq X_n$ , satisfying  $\bigwedge V = t$ , is a vertex cover of  $G$ . Since  $|V| \leq k$ , it is therefore a solution of  $P_1$ .

□

## Proof of Proposition 8

*Proof.* Let us first recall that the forgetting  $\exists V.f$  of a set of variables  $V$  in a formula  $f$  denotes a formula that is a most general consequence of  $f$  that is independent of  $V$  (in the sense that it is equivalent to a formula where no variable from  $V$  occurs) (Lang, Liberatore, and Marquis 2003).

Let  $z^*$  be any optimal solution of  $(C_{\text{soft}}, C_{\text{hard}})$ . On the one hand,  $z^*$  is a model of  $C_{\text{hard}}$ . Let  $V$  be the set of variables occurring in  $C_{\text{hard}}$  but not in  $X_n$ . Since  $z^* \models C_{\text{hard}}$ , we have that  $\exists V.z^* \models \exists V.C_{\text{hard}}$  (see (Lang, Liberatore, and Marquis 2003)). Stated otherwise, the projection  $\exists V.z^*$  of  $z^*$  on  $X_n$  implies the projection of  $C_{\text{hard}}$  on  $X_n$ .

On the other hand, by construction, a consistent term  $t$  over  $X_n$  implies the projection of  $C_{\text{hard}}$  on  $X_n$  if and only if  $t$  is an implicant of more than  $\frac{m}{2}$  decision trees of  $F$ . Thus,

the term  $\exists V.z^*$  is an implicant of more than  $\frac{m}{2}$  decision trees of  $F$ .

Finally, if  $z^*$  is an optimal solution of  $(C_{\text{soft}}, C_{\text{hard}})$ , then  $z^*$  satisfies a maximal number of soft clauses from  $C_{\text{soft}}$ . Since those soft clauses are precisely the negations of the literals occurring in  $t_x$ , the term  $t_{z^*} \cap t_x$  obtained from  $\exists V.z^*$  by removing every literal that coincides with a soft clause is still an implicant of more than  $\frac{m}{2}$  decision trees of  $F$ . Indeed,  $C_{\text{hard}}$  is monotone on  $X_n$  and the polarity of every variable of  $X_n$  in  $C_{\text{hard}}$  is the same as its polarity in  $t_x$ . Since  $z^*$  satisfies a maximal number of soft clauses,  $t_{z^*} \cap t_x$  contains a minimal number of literals. As  $t_{z^*} \cap t_x \subseteq t_x$ ,  $t_{z^*} \cap t_x$  is a minimal majoritary reason for  $x$  given  $F$ . □