Recourse under Model Multiplicity via Argumentative Ensembling

Paper #293

ABSTRACT

Model Multiplicity (MM) arises when multiple, equally performing machine learning models can be trained to solve the same prediction task. Recent studies show that models obtained under MM may produce inconsistent predictions for the same input. When this occurs, it becomes challenging to provide counterfactual explanations (CEs), a common means for offering recourse recommendations to individuals negatively affected by models' predictions. In this paper, we formalise this problem, which we name *recourse-aware ensembling*, and identify several desirable properties which methods for solving it should satisfy. We demonstrate that existing ensembling methods, naturally extended in different ways to provide CEs, fail to satisfy these properties. We then introduce *argumentative ensembling*, deploying computational argumentation as a means to guarantee robustness of CEs to MM, while also accommodating customisable user preferences. We show theoretically and experimentally that argumentative ensembling is able to satisfy properties which the existing methods lack, and that the trade-offs are minimal wrt the ensemble's accuracy.

KEYWORDS

Bipolar Argumentation, Model Multiplicity, Counterfactual Explanations

ACM Reference Format:

Paper #293. 2024. Recourse under Model Multiplicity via Argumentative Ensembling. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 10 pages.

1 INTRODUCTION

Model Multiplicity (MM), also known as predictive multiplicity or the Rashomon Effect, refers to a scenario where multiple, equally performing machine learning (ML) models may be trained to solve a prediction task [5, 7, 41]. While the existence of multiple models that achieve the same accuracy is not a problem per se, recent literature [5, 41] has drawn attention to the fact that these models may differ greatly in their internals. and might thus produce inconsistent predictions when deployed. Consider the commonly used scenario of a loan application, where an individual modelled by input $\mathbf{x}$ with features *unemployed* status, *33* years of age and *low* credit rating applies for a loan. Assume the bank has trained a set of ML models $\mathcal{M} = \{M_1, M_2, M_3\}$ to predict whether the loan should be granted or not. Even though each M_i may exhibit good performance overall, their internal differences may lead to conflicts, e.g. if $M_1(\mathbf{x}) = M_2(\mathbf{x}) = 0$ (i.e., reject), while $M_3(\mathbf{x}) = 1$ (i.e., accept).

Ensembling techniques are commonly used to deal with MM scenarios [4, 5]. A standard such technique is *naive ensembling* [4], where the predictions of several models are aggregated to produce a single outcome that reflects the opinion of the majority of models. For instance, naive ensembling applied to our running example would result in a rejection, as a majority of the models agree that the applicant is not creditworthy. While ensembling methods have been shown to be effective in practice, their application to consequential decision-making tasks raises some important challenges.

Indeed, these methods tend to ignore the need to provide avenues for recourse to users negatively impacted by the models' outputs, which the ML literature typically achieves via the provision of *counterfactual explanations* (CEs) for the predictions (see [30, 43] for recent overviews). Dealing with MM while also taking CEs into account is non-trivial. Indeed, standard algorithms designed to generate CEs for single models typically fail to produce recourse recommendations that are valid across equally performing models [38, 47]. This phenomenon may have troubling implications as a lack of robustness may lead users to question whether a CE is actually explaining the underlying decision-making task and is not just an artefact of a (subset of) model(s).

Another challenge is that naive ensembling ignores other meta-evaluation aspects of models like fairness, robustness, and interpretability, while it has been shown that models under MM can demonstrate substantial differences in these regards [13, 16, 52] and users may have strong preferences for some of these aspects, e.g. they may prefer model fairness to accuracy.

In this paper, we frame the recourse problem under MM formally and propose different approaches to accommodate recourse as well as user preferences over models. After covering related work (§2) and the necessary preliminaries (§3), we make the following contributions. In §4, we purpose a formalisation of the problem and a number of desirable properties for ensembling methods for recourse under MM. We also consider two natural extensions of naive ensembling to accommodate generation of CEs and show that they may violate some of the properties we define. We then propose *argumentative ensembling* in §5, a novel technique rooted in computational argumentation (see [2, 3] for overviews). We show that it is able to solve the recourse problem effectively while also naturally incorporating user preferences, a problem in which the conflicts between different objectives quickly become cognitively intractable. We present extensive experiments in §6, where we show that our framework always provides valid recourse under MM, showing better evaluations on a number of metrics and without compromising the individual prediction accuracies. We also demonstrate the usefulness of specifying user preferences in our framework. We then conclude and look ahead to the various avenues of future work in §7.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

2 RELATED WORK

Model Multiplicity. MM has been shown to affect several important dimensions of trustworthy ML. In particular, among equally accurate models, there could exist different fairness characteristics [13, 23, 50, 61], levels of model interpretability [12, 52, 53], model robustness evaluations [16] and even inconsistent explanations [6, 20, 28, 38–40, 42].

Recent attempts have been made to address the MM problem. Black et al. [5] suggested candidate models should be evaluated across additional dimensions other than accuracy (e.g., robustness or fairness evaluation thresholds). They provided a potential solution based on applying meta-rules to filter out undesirable models, then use ensemble methods to aggregate them, or randomly select one of them. Extending these ideas, the selective ensembling method of [4] embeds statistical testing into the ensembling process, such that when the numbers of candidate models predicting an input to the top two classes are close or equal, which could happen under naive ensembling strategies like majority voting, an abstention signal can be flagged for relevant stakeholders. Xin et al. [62] looked at decision trees and proposed an algorithm to enumerate all models obtainable under MM; Roth et al. [51] instead proposed a model reconciling procedure to resolve the conflicts between two disagreeing models. Meanwhile, a number of works [33, 41, 53, 60] propose metrics to quantify the extent of MM in prediction tasks.

Counterfactual Explanations and MM. A CE for a prediction of an input by an ML model is typically defined as another data point that minimally modifies the input such that the ML would yield a desired classification [55, 59]. CEs are often advocated as a means to provide algorithmic recourse for data subjects, and as such, modern algorithms have been extended to enforce additional desirable properties such as *actionability* [57], *plausiblity* [18] and *diversity* [44]. We refer to [30] for a recent overview.

More central to the MM problem, Pawelczyk et al. [47] pointed out that CEs on data manifold are more likely to be robust under MM than minimum-distance CEs. Leofante et al. [38] proposed an algorithm for a given ensemble of feed-forward neural networks to compute robust CEs that are provably valid for all models in the ensemble. Also related to MM is the line of work that focuses on the robustness of CEs against model changes, e.g., parameter updates due to model retraining on the same or slightly shifted data distribution [6, 9, 22, 26, 31, 35, 36, 45, 56]. These studies usually aim to generate CEs that are robust across retrained versions of the same model, which is different from MM where several (potentially structurally different) models are targeted together.

Computational Argumentation. Argumentation, as understood in AI and inspired by the seminal paper by Dung [21], is a set of formalisms for representing and reasoning about conflicting information. These formalisms' strengths in managing and resolving these conflicts have been demonstrated in numerous application areas, e.g. online debate [10], scheduling [14] and judgmental forecasting [34]. There have also been a broad range of works (see [15, 58] for overviews) demonstrating argumentation's capability for explaining the outputs of AI models, e.g. neural networks [17, 48], Bayesian classifiers [54] and random forests [49]. To the best of our knowledge the only work in which argumentation has been applied to MM specifically is [1], where the introduced method extracts and selects winning rules from an ensemble of classifiers, leveraging on argumentation's strengths in providing explainability for this process. This differs from our method as we consider pre-existing CEs computed for single models, rather than using rules as explanations, and we consider preferences in the ensembling.

3 PRELIMINARIES

Given a set of classification *labels* $\mathcal{L}$, a *model* is a mapping $M : \mathbb{R}^n \to \mathcal{L}$; we denote that M classifies an *input* $\mathbf{x} \in \mathbb{R}^n$ as ℓ iff $M(\mathbf{x}) = \ell$. In this paper, we focus on binary classification, i.e. $\mathcal{L} = \{0, 1\}$, for simplicity, though our approach is extensible to multi-class problems. Then, a *counterfactual explanation* (CE) for $\mathbf{x}$, given M, is some $\mathbf{c} \in \mathbb{R}^n \setminus \{\mathbf{x}\}$ such that $M(\mathbf{c}) \neq M(\mathbf{x})$, which may be optimised by some distance metric between the inputs.

A *bipolar argumentation framework* (BAF) [11] is a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$, where $\mathcal{X}$ is a set of *arguments*, $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ is a directed relation of *direct attack* and $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{X}$ is a directed relation of *direct support*. Given a BAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$, for any $\alpha_1 \in \mathcal{X}$, we refer to $\mathcal{A}(\alpha_1) = \{\alpha_2 | (\alpha_2, \alpha_1) \in \mathcal{A}\}$ as α_1*'s direct attackers* and to $\mathcal{S}(\alpha_1) = \{\alpha_2 | (\alpha_2, \alpha_1) \in \mathcal{S}\}$ as α_1*'s direct supporters*. Then, an *indirect attack* from α_x on α_y is a sequence $\alpha_1, r_1, \ldots, r_{n-1}, \alpha_n$, where $n \geq 3$, $\alpha_1 = \alpha_x$, $\alpha_n = \alpha_y$, $r_1 \in \mathcal{A}$ and $r_i \in \mathcal{S}$ $\forall i \in \{2, \ldots, n-1\}$. Similarly, a *supported attack* from α_x on α_y is a sequence $\alpha_1, r_1, \ldots, r_{n-1}, \alpha_n$, where $n \geq 3$, $\alpha_1 = \alpha_x$, $\alpha_n = \alpha_y$, $r_{n-1} \in \mathcal{A}$ and $r_i \in \mathcal{S}$ $\forall i \in \{1, \ldots, n-2\}$. Straightforwardly, a supported attack on an argument implies there is a direct attack on an argument.

We will also use notions of acceptability of sets of arguments in BAFs [11]. A set of arguments $X \subseteq \mathcal{X}$, also called an *extension*, is said to *set-attack* any $\alpha_1 \in \mathcal{X}$ iff there exists an attack (whether direct, indirect or supported) from some $\alpha_2 \in X$ on α_1. Meanwhile, X is said to *set-support* any $\alpha_1 \in \mathcal{X}$ iff there exists a direct support from some $\alpha_2 \in X$ on α_1.[1] Then, a set $X \subseteq \mathcal{X}$ *defends* any $\alpha_1 \in \mathcal{X}$ iff $\forall \alpha_2 \in \mathcal{X}$, if $\{\alpha_2\}$ set-attacks α_1 then $\exists \alpha_3 \in X$ such that $\{\alpha_3\}$ set-attacks α_2. Any set $X \subseteq \mathcal{X}$ is then said to be *conflict-free* iff $\nexists \alpha_1, \alpha_2 \in X$ such that $\{\alpha_1\}$ set-attacks α_2, and *safe* iff $\nexists \alpha_3 \in \mathcal{X}$ such that X set-attacks α_3 and either: X set-supports α_3; or $\alpha_3 \in X$. (Note that a safe set is guaranteed to be conflict-free.) The notion of a set $X \subseteq \mathcal{X}$ being *d-admissible* (based on *admissibility* in [21]) is then introduced, requiring that X is conflict-free and defends all of its elements. This notion is extended to account for safe sets: X is said to be *s-admissible* iff it is safe and defends all of its elements. (Note that an s-admissible set is guaranteed to be d-admissible.) It is also extended for the notion of *external coherence* [11]: X is said to be *c-admissible* iff it is conflict-free, closed for $\mathcal{S}$ and defends all of its elements. Finally, a set $X \subseteq \mathcal{X}$ is said to be *d-preferred* (respectively, *s-preferred*, *c-preferred*) iff it is d-admissible (respectively, s-admissible, c-preferred) and maximal wrt set-inclusion.

4 RECOURSE UNDER MODEL MULTIPLICITY

As mentioned in §1, a common way to deal with MM in practice is to employ ensembling techniques, where the prediction outcomes of several models are aggregated to produce a single outcome. Aggregation can be performed in different ways, as discussed in [4, 5].

[1] Note that in [11], set-supports are defined based on sequences of supports, but we do not use them here.

In the following, we formalise a notion of *naive ensembling*, adopted in [5], and also known as *majority voting*, which will serve as a baseline for our analysis.[2]

Definition 4.1. Given an input $\mathbf{x}$, a set of models $\mathcal{M}$ and a set of labels $\mathcal{L}$, we define the set of *top labels* $\mathcal{L}_{max} \subseteq \mathcal{L}$ as:

$$\mathcal{L}_{max} = argmax_{\ell_i \in \mathcal{L}} |\{M_j \in \mathcal{M} | M_j(\mathbf{x}) = \ell_i\}|.$$

Then we then use $\mathcal{M}^n(\mathbf{x}) \in \mathcal{L}_{max}$ to denote the aggregated classification by *naive ensembling*. In the cases where $|\mathcal{L}_{max}| > 1$, we select $\mathcal{M}^n(\mathbf{x})$ from $\mathcal{L}_{max}$ randomly. With an abuse of notation, we also let $\mathcal{M}^n = \{M_j \in \mathcal{M} | M_j(\mathbf{x}) = \mathcal{M}^n(\mathbf{x})\}$ denote the set of models that agree on the aggregated classification.

Coming back to our loan example where M_1 and M_2 reject the loan ($M_1(\mathbf{x}) = M_2(\mathbf{x}) = 0$) while M_3 accepts it ($M_3(\mathbf{x}) = 1$), we obtain $\mathcal{M}^n(\mathbf{x}) = 0$ and $\mathcal{M}^n = \{M_1, M_2\}$. Naive ensembling is known to be an effective strategy to mediate conflicts between models and is routinely used in practical applications. However, in this paper, we take an additional step and aim to generate CEs providing recourse for a user that has been impacted by $\mathcal{M}^n(\mathbf{x})$. Recent work by [38] has shown that standard algorithms designed to generate CEs for single models typically fail to produce recourse recommendations that are robust across $\mathcal{M}^n$. One natural idea to address this would be to extend naive ensembling to account for CEs. Next, we formalise this idea in terms of several properties that are important in this setting. We then discuss two concrete methods extending naive ensembling and analyse them in terms of the properties we define.

4.1 Problem Statement and Desirable Properties

Consider a non-empty set of models $\mathcal{M}=\{M_1, \ldots, M_m\}$ and, for an input $\mathbf{x}$, assume a set $\mathcal{C}(\mathbf{x}) = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ where each $\mathbf{c}_i \in \mathcal{C}(\mathbf{x})$ is a CE for $\mathbf{x}$, given M_i. In the rest of the paper, wherever it is clear that we refer to a given $\mathbf{x}$, we use $\mathcal{C}$ and omit its dependency on $\mathbf{x}$ for readability. Our aim is to solve the problem outlined below.

Problem: Recourse-Aware Ensembling (RAE)
Input: input $\mathbf{x}$, set $\mathcal{M}$ of models, set $\mathcal{C}$ of CEs
Output: "optimal" set $S \subseteq \mathcal{M} \cup \mathcal{C}$ of models and CEs.

To characterise optimality, we propose a number of desirable properties for the outputs of ensembling methods. We refer to these outputs as *solutions* for RAE. The most basic requirement requires that both models and CEs in the output are non-empty.

Definition 4.2. An ensembling method satisfies *non-emptiness* iff for any given input $\mathbf{x}$, set $\mathcal{M}$ of models and set $\mathcal{C}$ of CEs, any solution $S \subseteq \mathcal{M} \cup \mathcal{C}$ is such that $S \cap \mathcal{M} \neq \emptyset$ and $S \cap \mathcal{C} \neq \emptyset$.

Specifically, non-emptiness ensures that the RAE method returns some models and some CEs. We then look to ensure that the ensembling method returns a non-trivial set of models, as formalised by the next property.

Definition 4.3. An ensembling method satisfies *non-triviality* iff for any given input $\mathbf{x}$, set $\mathcal{M}$ of models and set $\mathcal{C}$ of CEs, any solution $S \subseteq \mathcal{M} \cup \mathcal{C}$ is such that $|S \cap \mathcal{M}| > 1$.

[2]It should be noted that in [5], the case where there is no majority is not discussed.

Clearly, the returned models should not disagree amongst themselves on the classification, which leads to our next requirement.

Definition 4.4. An ensembling method satisfies *model agreement* iff for any given input $\mathbf{x}$, set $\mathcal{M}$ of models and set $\mathcal{C}$ of CEs, any solution $S \subseteq \mathcal{M} \cup \mathcal{C}$ is such that $\forall M_i, M_j \in S \cap \mathcal{M}$, $M_i(\mathbf{x}) = M_j(\mathbf{x})$.

The next property, which itself requires model agreement to be satisfied, checks whether the set of returned models is among the largest of the agreeing sets of models, a motivating property of naive ensembling.

Definition 4.5. An ensembling method satisfies *majority vote* iff it satisfies model agreement and for any given input $\mathbf{x}$, set $\mathcal{M}$ of models, set $\mathcal{C}$ of CEs and set $\mathcal{L}$ of labels, any solution $S \subseteq \mathcal{M} \cup \mathcal{C}$ is such that, letting $\ell_i = M_j(\mathbf{x})$ $\forall M_j \in S \cap \mathcal{M}$, $\nexists \ell_k \in \mathcal{L} \setminus \{\ell_i\}$ such that $|\{M_l \in \mathcal{M} | M_l(\mathbf{x}) = \ell_k\}| > |\{M_l \in \mathcal{M} | M_l(\mathbf{x}) = \ell_i\}|$.

Next, we consider the robustness of recourse. Previous work [38] considered a very conservative notion of robustness whereby explanations are required to be valid for all models in $\mathcal{M}$. While this might be desirable in some cases, we highlight that satisfying this property may not always be feasible in practice. We therefore propose a relaxed notion of robustness, which requires that CEs are valid only for the models that support them.

Definition 4.6. An ensembling method satisfies *counterfactual validity* iff for any given input $\mathbf{x}$, set $\mathcal{M}$ of models and set $\mathcal{C}$ of CEs, any solution $S \subseteq \mathcal{M} \cup \mathcal{C}$ is such that $\forall M_i \in S \cap \mathcal{M}$ and $\forall \mathbf{c}_j \in S \cap \mathcal{C}$, $M_i(\mathbf{c}_j) \neq M_i(\mathbf{x})$.

While counterfactual validity is a fundamental requirement for any sound ensembling method, one also needs to ensure that the solutions it generates are coherent, as formalised below.

Definition 4.7. An ensembling method satisfies *counterfactual coherence* iff for any given input $\mathbf{x}$, set $\mathcal{M}$ of models and set $\mathcal{C}$ of CEs, any solution $S \subseteq \mathcal{M} \cup \mathcal{C}$, where $\mathcal{M} = \{M_1, \ldots, M_m\}$ and $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$, is such that $\forall i \in \{1, \ldots, m\}$, $M_i \in S$ iff $\mathbf{c}_i \in S$.

Intuitively coherence requires that *(i)* a CE is returned only if it is supported by a model and *(ii)* when the CE is chosen, then its corresponding model must be part of the support. This ultimately guarantees strong justification as to why a given recourse is suggested since selected models and their reasoning (represented by their CEs) are assessed in tandem.

The properties defined above may not be all satisfiable at the same time in practice, therefore heuristic approaches might be needed. Next, we discuss two such heuristics to extend naive ensembling towards solving RAE.

4.2 Extending Naive Ensembling for Recourse

We now present two strategies that leverage naive ensembling to solve RAE. In particular, we use the relationship between models in the ensemble $\mathcal{M}^n$ and their corresponding CEs as follows.

Definition 4.8. Consider an input $\mathbf{x}$, a set $\mathcal{M}$ of models and a set $\mathcal{C}$ of CEs. Let $\mathcal{M}^n \subseteq \mathcal{M}$ be the set of models obtained by naive ensembling. We define the set of *naive CEs* as:

$$\mathcal{C}^n = \{\mathbf{c}_i \in \mathcal{C} \mid M_i \in \mathcal{M}^n\};$$

and the set of *valid CEs* as:

$$\mathcal{C}^v = \{\mathbf{c}_i \in \mathcal{C} \mid M_i \in \mathcal{M}^n \wedge \forall M_j \in \mathcal{M}^n, M_j(\mathbf{c}_i) \neq M_j(\mathbf{x})\}.$$

Then, two possible solutions to RAE are $S^n = \mathcal{M}^n \cup \mathcal{C}^n$, named *augmented ensembling*, or $S^v = \mathcal{M}^n \cup \mathcal{C}^v$, named *robust ensembling*.

Intuitively, augmented ensembling suggests taking all the CEs in $\mathcal{C}$ that correspond to the models in $\mathcal{M}^n$. Meanwhile, robust ensembling extends augmented ensembling by enforcing the additional constraint that CEs are selected only if they are valid for all models in $\mathcal{M}^n$. We now provide an illustrative example to clarify the results produced by the two strategies.

Example 4.9. Consider $\mathcal{M} = \{M_1, M_2, M_3, M_4, M_5\}$ and an input $\mathbf{x}$ such that $M_1(\mathbf{x}) = M_2(\mathbf{x}) = M_3(\mathbf{x}) = 0$ and $M_4(\mathbf{x}) = M_5(\mathbf{x}) = 1$. Let $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}_5\}$ be the set of CEs generated for $\mathbf{x}$, i.e. $M_1(\mathbf{c}_1) = M_2(\mathbf{c}_2) = M_3(\mathbf{c}_3) = 1$, while $M_4(\mathbf{c}_4) = M_5(\mathbf{c}_5) = 0$. Applying naive ensembling to $\mathcal{M}$ yields $\mathcal{M}^n = \{M_1, M_2, M_3\}$ and $\mathcal{M}^n(\mathbf{x}) = 0$. Then, the set of naive CEs is $\mathcal{C}^n = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$, and thus augmented ensembling gives $S^n = \{M_1, M_2, M_3, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$. Now, assume that $\mathbf{c}_1$ is invalid for M_2 (i.e. $M_2(\mathbf{c}_1) = 0$), $\mathbf{c}_2$ is invalid for M_1, $\mathbf{c}_3$ is invalid for M_2, and all three CEs are otherwise valid for all models in $\mathcal{M}^n$. Then, then the set of valid CEs is empty, i.e. $\mathcal{C}^v = \varnothing$, and thus robust ensembling gives $S^v = \{M_1, M_2, M_3\}$.

This example shows that both methods host major drawbacks: augmented ensembling may produce CEs which are invalid and thus it is not robust to MM, while robust ensembling is prone to returning no CEs. We now present a theoretical analysis to assess the extent to which augmented and robust ensembling are able to satisfy the properties given in Definitions 4.2 to 4.7.

Theorem 4.10. *Augmented ensembling satisfies non-emptiness, model agreement, majority vote and counterfactual coherence. It satisfies non-triviality if $|\mathcal{M}| > 2$. It does not satisfy counterfactual validity.*

Proof. By Definition 4.1, it can be seen by inspection that $|\mathcal{M}^n| > 0$. Thus, non-emptiness is satisfied. It can also be seen by inspection of the same definition that $\forall M_i, M_j \in \mathcal{M}^n$, $M_i(\mathbf{x}) = M_j(\mathbf{x})$. Thus, model agreement is satisfied. We can also see that $\nexists \ell_i \in \mathcal{L} \setminus \{\mathcal{M}^n(\mathbf{x})\}$ such that $|\{M_j \in \mathcal{M} | M_j(\mathbf{x}) = \ell_i\}| > |\mathcal{M}^n|$. Thus, majority vote is satisfied. By Definitions 4.1 and 4.8, it can be seen that $\forall M_i \in \mathcal{M}$ and $\forall \mathbf{c}_i \in \mathcal{C}$, $M_i \in S$ iff $\mathbf{c}_i \in S$. Thus, counterfactual coherence is satisfied.

Example 4.9 provides a counterexample which shows that counterfactual validity is not satisfied.

Finally, the partial satisfaction of non-triviality can be proven by contradiction: assume $|\mathcal{M}| = n$, $n > 2$ but $|\mathcal{M}^n| = 1$. By Definition 4.1, $\mathcal{M}^n$ is the largest subset of $\mathcal{M}$ containing models with the same classification outcome. However, for binary classification, this implies that the remaining $n-1$ all agree on the opposite classification, i.e. $|\mathcal{M} \setminus \mathcal{M}^n| > |\mathcal{M}^n|$, which leads to a contradiction. □

Theorem 4.11. *Robust ensembling satisfies model agreement, majority vote and counterfactual validity. It satisfies non-triviality if $|\mathcal{M}| > 2$. It does not satisfy non-emptiness or counterfactual coherence.*

Proof. The proofs for model agreement, majority vote and non-triviality are analogous to those in Theorem 4.10 and so are omitted.

It can be seen by inspection of Definition 4.8 that $\forall M_i \in \mathcal{M}^n$ and $\forall \mathbf{c}_j \in \mathcal{C}^v$, $M_i(\mathbf{c}_j) \neq M_i(\mathbf{x})$. Thus, counterfactual validity is satisfied.

	$\mathcal{M}^n \cup \mathcal{C}^n$	$\mathcal{M}^n \cup \mathcal{C}^v$	$\mathcal{M}^a \cup \mathcal{C}^a$
non-emptiness	✓		✓
non-triviality	✓*	✓*	✓*
model agreement	✓	✓	✓
majority vote	✓	✓	
counterfactual validity		✓	✓
counterfactual coherence	✓		✓

Table 1: Augmented ($\mathcal{M}^n \cup \mathcal{C}^n$) and robust ($\mathcal{M}^n \cup \mathcal{C}^v$) ensembling, as well as our approach ($\mathcal{M}^a \cup \mathcal{C}^a$, defined in §5), assessed against the desirable properties defined in §4.1. Satisfaction of a property is indicated with ✓, while partial satisfaction under given conditions is indicated with ✓*.

Example 4.9 provides a counterexample which shows that non-emptiness and counterfactual coherence are not satisfied. □

These results, summarised in Table 1, demonstrate that there may exist cases in which both augmented and robust ensembling fail to solve RAE satisfactorily. This has strong implications on the quality of the results obtained in practice, as we will show experimentally in §6. Further, these methods provide no way to take into account users' preferences over the models. As previously mentioned (see §1 and §2), there could be different characteristics among the models in $\mathcal{M}$ in terms of meta-evaluation aspects, e.g. a model's fairness, robustness, and simplicity. Depending on the task, a model's fairness might be specified as being more important than its robustness or simplicity. In such cases, it would be desirable to have a principled way to rank models according to the preference specification, as promoted in [5]. The combination of these deficiencies motivates the need for a richer ensembling framework to solve RAE while incorporating user preferences, given next.

5 ARGUMENTATIVE ENSEMBLING

We now present our method for ensembling models and CEs which inherently supports specifying preferences over models; then we undertake a theoretical analysis of its properties.

5.1 Definition

We start by formalising the ways to incorporate the previously mentioned preferences over models. These preferences could be obtained by any information, e.g. meta-rules over the models as suggested in [5], but we will assume that they are extracted wrt a set of properties of the models, e.g. their accuracy or (a metric representing) their simplicity.

Definition 5.1. Given a set $\mathcal{M}$ of models, a set $\mathcal{P}$ of *properties* is such that $\forall \pi \in \mathcal{P}$, $\pi : \mathcal{M} \to \mathbb{R}$ is a total function.

We then define a preference over the properties such that users can impose a ranking of priority over them. Here and onwards, for simplicity we use total orders, denoted $\preceq$, over any set S such that, as usual, for any $s_i, s_j \in S$, $s_i \prec s_j$ iff $s_i \preceq s_j$ and $s_i \not\succeq s_j$. Also as usual, we say that $s_i \simeq s_j$ iff $s_i \preceq s_j$ and $s_i \succeq s_j$.

Definition 5.2. Given a set $\mathcal{P}$ of properties, a *property preference* $\preceq_{\mathcal{P}}$ is a total order over $\mathcal{P}$.

Model preferences can be defined using property preferences. In the following example, we define one way for doing so.

Example 5.3. Consider the same set of models as in Example 4.9 and a set of properties $\mathcal{P}=\{\pi_1, \pi_2\}$ where π_1, π_2 represent model accuracy and simplicity, respectively, with a property preference $\preceq_{\mathcal{P}}$ such that $\pi_1 \succ_{\mathcal{P}} \pi_2$ and values for the properties as follows.

	M_1	M_2	M_3	M_4	M_5
π_1 (accuracy)	0.85	0.87	0.86	0.86	0.87
π_2 (simplicity)	0	0.75	1	0.5	0.75

A simple model preference $\preceq_{\mathcal{M}}$ over $\mathcal{M}$ may be such that, for any $M_i, M_j \in \mathcal{M}$, $M_i \succ_{\mathcal{M}} M_j$ iff: (i) $\pi_1(M_i) > \pi_1(M_j)$; or (ii) $\pi_1(M_i) = \pi_1(M_j)$ and $\pi_2(M_i) > \pi_2(M_j)$. This $\preceq_{\mathcal{M}}$ is a total order over $\mathcal{M}$ and results in $M_2 \simeq_{\mathcal{M}} M_5 \succ_{\mathcal{M}} M_3 \succ_{\mathcal{M}} M_4 \succ_{\mathcal{M}} M_1$.

Other ways to define preferences over models from preferences over properties of models can be defined, e.g. based on more sophisticated notions of dominance. We will assume some given notion of total preference over models, as follows, ignoring how it is obtained.

Definition 5.4. Given a set $\mathcal{M}$ of models, a *model preference* $\preceq_{\mathcal{M}}$ over $\mathcal{M}$ is a total order over $\mathcal{M}$.

How can we incorporate these model preferences into the ensembling, while still satisfying the properties defined in §4.1? To tackle this problem, we use argumentation as follows.

Definition 5.5. The BAF corresponding to an input $\mathbf{x}$, a set $\mathcal{M}$ of models, a set $\mathcal{C}$ of CEs and a model preference $\preceq_{\mathcal{M}}$ is $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ with:

- $\mathcal{X} = \mathcal{M} \cup \mathcal{C}$;
- $\mathcal{A} \subseteq (\mathcal{M} \times \mathcal{M}) \cup (\mathcal{M} \times \mathcal{C})$ where:
  - for any $M_i, M_j \in \mathcal{M}$, $(M_i, M_j) \in \mathcal{A}$ iff $M_i(\mathbf{x}) \neq M_j(\mathbf{x})$ and $M_i \succeq_{\mathcal{M}} M_j$;
  - for any $M_i \in \mathcal{M}$ and $\mathbf{c}_j \in \mathcal{C}$ where $M_i(\mathbf{c}_j) = M_i(\mathbf{x})$, $(M_i, \mathbf{c}_j) \in \mathcal{A}$ iff $M_i \succeq_{\mathcal{M}} M_j$ and $(\mathbf{c}_j, M_i) \in \mathcal{A}$ iff $M_j \succeq_{\mathcal{M}} M_i$;
- $\mathcal{S} \subseteq (\mathcal{M} \times \mathcal{C}) \cup (\mathcal{C} \times \mathcal{M})$ where for any $M_i \in \mathcal{M}$ and $\mathbf{c}_j \in \mathcal{C}$, $(M_i, \mathbf{c}_j), (\mathbf{c}_j, M_i) \in \mathcal{S}$ iff $i = j$.

Here, a model attacks another model if they disagree on the prediction and the latter is not strictly preferred to the former. This means that models which are outperformed with regards to preferences must be defended by more-preferred, agreeing models in order to be considered acceptable. Models and CEs are treated similarly, with the CEs inheriting the preferences from the models by which they were generated, and attacks being present between them when the model considers the CE invalid. This, along with the fact that models and their CEs support one another, ensures that the models are inherently linked to their reasoning, in the form of their CEs, and conflicts are drawn not only when two models' predictions differ, but also when their reasoning differs.

Argumentative ensembling makes use of the set of s-preferred sets of arguments, referred to as P_s, in the corresponding BAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$, in order to resolve the MM problem.

Definition 5.6. Consider an input $\mathbf{x}$, a set $\mathcal{M}$ of models, a set $\mathcal{C}$ of CEs, a set $\mathcal{L}$ of labels and a model preference $\preceq_{\mathcal{M}}$. Let the largest s-preferred sets for the corresponding BAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ be defined as:

$$X_{max} = argmax_{X \in P_s} |X|.$$

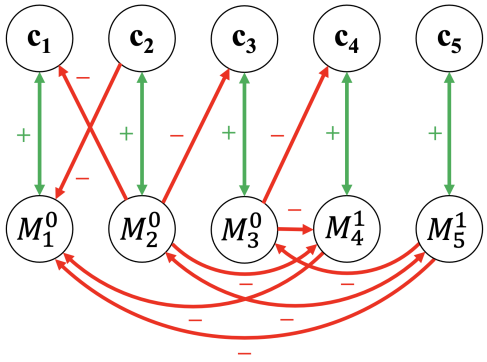


Figure 1: BAF for Example 5.7 where: models' predictions for the input x are given as superscripts, e.g. $M_1(\mathbf{x}) = 0$ but $M_4(\mathbf{x}) = 1$; reciprocal supports are represented by dual-headed green arrows labelled with + and standard (reciprocal) attacks are represented by single-headed (dual-headed, respectively) red arrows labelled with −.

Then, the solution to RAE by *argumentative ensembling* is defined as $S^a \in X_{max}$, where $\mathcal{M}^a = S^a \cap \mathcal{M}$ and $\mathcal{C}^a = S^a \cap \mathcal{C}$ and in the case of $|X_{max}| > 1$, we select S^a from X_{max} randomly. We also let $\mathcal{M}^a(\mathbf{x}) = \ell_i \in \mathcal{L}$ where $M_j(\mathbf{x}) = \ell_i \ \forall M_j \in \mathcal{M}^a$.

Note that, alternatively, when $|X_{max}| > 1$, we could choose to report all viable resulting ensembles in X_{max} rather than a random one as defined above, so that more informed decisions could be made by relevant stakeholders. We leave this to future work.

The following example demonstrates how quickly the problem, when preferences are included, can become complex. This is the case even with only five models, far fewer than usual in MM.

Example 5.7. The BAF corresponding to input, models and CEs as in Example 5.3 is $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ with (see Fig. 1): $\mathcal{X} = \{M_1, M_2, M_3, M_4, M_5, \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}_5\}$; $\mathcal{A} = \{(M_2, M_4), (M_2, M_5), (M_2, \mathbf{c}_1), (M_2, \mathbf{c}_3), (M_3, M_4), (M_3, \mathbf{c}_4), (M_4, M_1), (M_5, M_1), (M_5, M_2), (M_5, M_3), (\mathbf{c}_2, M_1)\}$; and $\mathcal{S} = \{(M_1, \mathbf{c}_1), (M_2, \mathbf{c}_2), (M_3, \mathbf{c}_3), (M_4, \mathbf{c}_4), (M_5, \mathbf{c}_5), (\mathbf{c}_1, M_1), (\mathbf{c}_2, M_2), (\mathbf{c}_3, M_3), (\mathbf{c}_4, M_4), (\mathbf{c}_5, M_5)\}$. This leads to $P_s = \{\{M_2, \mathbf{c}_2\}, \{M_4, M_5, \mathbf{c}_4, \mathbf{c}_5\}\}$, and thus $S^a = \{M_4, M_5, \mathbf{c}_4, \mathbf{c}_5\}$ and $\mathcal{M}^a(\mathbf{x}) = 1$.

This example shows how the use of CEs in the ensembling directly results in the prediction being reversed, relative to the other ensembling methods, violating majority vote. This is due to the fact that, while the preferences over the two sets of models are roughly similar, the validity of the CEs for the models selected by the augmented and robust ensemblings is very poor. This means that when a CE which is valid for all models is required, some compromise must be made on the model selection, as we demonstrated.

5.2 Theoretical Analysis

We will now undertake a theoretical analysis of argumentative ensembling, demonstrating some of the desirable behaviours thereof via properties. First, we consider the properties introduced in §4.1.

Theorem 5.8. *Argumentative ensembling satisfies non-emptiness, model agreement, counterfactual validity and counterfactual coherence. It satisfies non-triviality if for some $M_i \in \mathcal{M}$, where $\nexists M_j \in \mathcal{M} \setminus$*

$\{M_i\}$ *such that* $M_j \succ_{\mathcal{M}} M_i$, $\exists M_k \in \mathcal{M}$ *such that* $M_k(\mathbf{x}) = M_i(\mathbf{x})$, $M_k(\mathbf{c}_i) \neq M_k(\mathbf{x})$ *and* $M_i(\mathbf{c}_k) \neq M_i(\mathbf{x})$. *It does not satisfy majority vote.*

Proof. Let us first prove counterfactual coherence. It can be seen from Definition 5.5 that $\forall M_i \in \mathcal{M}$ and $\forall \mathbf{c}_j \in \mathcal{C}$, $(M_i, \mathbf{c}_j), (\mathbf{c}_j, M_i) \in \mathcal{S}$. Thus, there exists an indirect attack on any $M_i \in \mathcal{M}$ iff there exists a direct attack on $\mathbf{c}_i \in \mathcal{C}$. Likewise, there exists an indirect attack on any $\mathbf{c}_i \in \mathcal{C}$ iff there exists a direct attack on $M_i \in \mathcal{M}$. Then, letting $\mathcal{M} = \{M_1, \ldots, M_m\}$ and $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$, since any $X \in P_s$ is maximal wrt $\mathcal{X}$ by Definition 5.6, X must be such that $\forall i \in \{1, \ldots, m\}$, $M_i \in S$ iff $\mathbf{c}_i \in S$.

Let us prove non-emptiness by contradiction. Assume that $\exists X \in P_s$ such that $X \cap \mathcal{M} = \varnothing$ or $X \cap \mathcal{C} = \varnothing$. We know from the above proof that, $\forall X \in P_s$, $X \cap \mathcal{M} \neq \varnothing$ iff $X \cap \mathcal{C} \neq \varnothing$. Then, by the definition of s-preferred extensions (see §3), it must be the case that $\forall X \in P_s$, $X = \varnothing$. Based on the fact that, by Definition 5.4, $\preceq_{\mathcal{M}}$ is a total ordering and thus transitive, this is not possible as it will always be the case that $\exists M_i \in \mathcal{M}$ such that $\nexists M_j \in \mathcal{M}$ where $M_j \succ_{\mathcal{M}} M_i$. Thus, by Definition 5.5, either $\mathcal{A}(M_i) \cup \mathcal{A}(\mathbf{c}_i) = \varnothing$ or $\forall \alpha_k \in \mathcal{A}(M_i) \cup \mathcal{A}(\mathbf{c}_i)$, $M_i \in \mathcal{A}(\alpha_k)$ or $\mathbf{c}_i \in \mathcal{A}(\alpha_k)$, meaning $\{M_i, \mathbf{c}_i\}$ is either unattacked or is able to defend itself, therefore would be acceptable in at least one s-preferred extension, and we have the contradiction.

Let us prove model agreement by contradiction. Assume that $\exists M_i, M_j \in \mathcal{M}$ such that $M_i(\mathbf{x}) \neq M_j(\mathbf{x})$ and $\exists X \in P_s$ such that $M_i, M_j \in X$. By Definition 5.5, it follows that $\exists (M_i, M_j) \in \mathcal{A}$ or $\exists (M_j, M_i) \in \mathcal{A}$, which cannot be the case in an s-preferred set, which must be conflict-free (see §3), and so we have the contradiction.

Let us prove counterfactual validity by contradiction. Assume that $\exists M_i \in \mathcal{M}$ and $\exists \mathbf{c}_j \in \mathcal{C}$ such that $M_i(\mathbf{c}_j) = M_i(\mathbf{x})$ and $\exists X \in P_s$ such that $M_i, \mathbf{c}_j \in X$. By Definition 5.5, it can be seen that $\exists (M_i, \mathbf{c}_j) \in \mathcal{A}$ or $\exists (\mathbf{c}_j, M_i) \in \mathcal{A}$, which cannot be the case in an s-preferred set, which, again, must be conflict-free, and so we have the contradiction.

Let us prove the partial satisfaction of non-triviality by contradiction. It can be seen from Definition 5.6 and the proof for counterfactual coherence above that for $|\mathcal{M}^a| = 1$, it must be the case that $\forall X \in P_s$, $|X \cap \mathcal{M}| = 1$. However, from the above assumptions we can see that for some $M_i \in \mathcal{M}$, where $\nexists M_j \in \mathcal{M} \setminus \{M_i\}$ such that $M_j \succ_{\mathcal{M}} M_i$, $\exists M_k \in \mathcal{M}$ such that $M_k(\mathbf{x}) = M_i(\mathbf{x})$, $M_k(\mathbf{c}_i) \neq M_k(\mathbf{x})$ and $M_i(\mathbf{c}_k) \neq M_i(\mathbf{x})$. Then, to avoid a contradiction, it must be the case that $\exists \alpha_l \in \mathcal{M} \cup \mathcal{C}$ such that $(\alpha_l, M_k) \in \mathcal{A}$ or $(\alpha_l, \mathbf{c}_k) \in \mathcal{A}$, and $(M_i, \alpha_l), (\mathbf{c}_i, \alpha_l) \notin \mathcal{A}$. If $\alpha_l \in \mathcal{M}$, and thus $M_l(\mathbf{x}) \neq M_k(\mathbf{x}) = M_i(\mathbf{x})$, then it must be the case that $(M_i, M_l) \in \mathcal{A}$ (a contradiction, by Definition 5.5, since $M_i \succeq_{\mathcal{M}} M_l$. If $\alpha_l \in \mathcal{C}$ and $M_l(\mathbf{x}) \neq M_k(\mathbf{x}) = M_i(\mathbf{x})$, then $(M_i, M_l) \in \mathcal{A}$ (a contradiction by the same reasoning). Finally, if $\alpha_l \in \mathcal{C}$ and $M_l(\mathbf{x}) = M_k(\mathbf{x}) = M_i(\mathbf{x})$, then either: if $M_i(\mathbf{c}_l) = M_i(\mathbf{x})$ or $M_l(\mathbf{c}_i) = M_l(\mathbf{x})$, then $(M_i, \mathbf{c}_l) \in \mathcal{A}$ (a contradiction; or if this is not the case, then $M_l \in X'$ (which can be checked by repeating the steps completed for M_k, for M_l instead). It therefore must be the case that $\exists X' \in P_s$ where $|X' \cap \mathcal{M}| > 1$, and we have the contradiction in all cases.

Finally, Example 5.7 provides a counterexample which shows that majority vote is not satisfied. □

These results contrast with those for augmented and robust ensembling, as shown in Table 1. Argumentative ensembling avoids the pitfalls of augmented and robust ensembling by satisfying non-emptiness, counterfactual validity and counterfactual coherence. In order to achieve this behaviour, majority vote is sacrificed as a guarantee. In §6 we will assess the impact of not guaranteeing majority vote on argumentative ensembling's accuracy, along with other metrics.

However, first, we consider the relationship between the different ensembling methods. For the remainder of the section, we assume as given an input $\mathbf{x}$, a set $\mathcal{M}$ of models, a set $\mathcal{C}$ of CEs, a model preference $\preceq_{\mathcal{M}}$ and a corresponding BAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ with P_s the set of all its s-preferred sets.

Theorem 5.9. *If* $\forall M_i, M_j \in \mathcal{M}$, $M_i \simeq_{\mathcal{M}} M_j$, *and* $\forall M_k \in \mathcal{M}$ *and* $\mathbf{c}_l \in \mathcal{C}$, *where* $M_k(\mathbf{x}) = M_l(\mathbf{x})$, $M_k(\mathbf{c}_l) \neq M_k(\mathbf{c}_l)$, *then augmented, robust and argumentative ensembling are equivalent.*

Proof. If $\forall M_i, M_j \in \mathcal{M}$, $M_i \simeq_{\mathcal{M}} M_j$, then it can be seen from Definition 5.5 that $(M_i, M_j), (M_j, M_i) \in \mathcal{A}$ iff $M_i(\mathbf{x}) \neq M_j(\mathbf{x})$. Note that, by Definition 4.8, augmented and robust ensembling are equivalent since $\forall M_k \in \mathcal{M}$ and $\mathbf{c}_l \in \mathcal{C}$, where $M_k(\mathbf{x}) = M_l(\mathbf{x})$, $M_k(\mathbf{c}_l) \neq M_k(\mathbf{c}_l)$. Also by the assumptions, it can be seen from Definition 5.5 that $\forall M_k \in \mathcal{M}$ and $\forall \mathbf{c}_l \in \mathcal{C}$, $(M_k, \mathbf{c}_l), (\mathbf{c}_l, M_k) \in \mathcal{A}$ iff $M_k(\mathbf{c}_l) = M_k(\mathbf{x})$ and $M_k(\mathbf{x}) \neq M_l(\mathbf{x})$ due to the assumptions in the theorem, meaning any attack is reciprocated and all arguments defend themselves. Then, $\forall M_m, M_n \in \mathcal{M}$ such that $M_m(\mathbf{x}) = M_n(\mathbf{x})$, $(M_m, M_n) \notin \mathcal{A}$. Thus, $P_s = \{\{M_o \in \mathcal{M}, \mathbf{c}_o \in \mathcal{C} | M_o(\mathbf{x}) = 0\}, \{M_p \in \mathcal{M}, \mathbf{c}_p \in \mathcal{C} | M_p(\mathbf{x}) = 1\}$. By Definitions 4.1, 4.8 and 5.6, all forms of ensembling select from the same two sets of models and CEs in the same manner and are thus equivalent. □

We also provide a number of theoretical results concerning the behaviour of argumentative ensembling, first relating to the preferences. The first result demonstrates how a completely dominant model wrt the preferences will be present in all s-preferred sets.

Proposition 5.10. *If* $\exists M_i \in \mathcal{M}$ *such that* $\forall M_j \in \mathcal{M}$, $M_i \succ_{\mathcal{M}} M_j$, *then* $\forall X \in P_s$, $M_i \in X$.

Proof. If $\forall M_j \in \mathcal{M}, M_i \succ_{\mathcal{M}} M_j$ then, by Definition 5.5, $\mathcal{A}(M_i) = \mathcal{A}(\mathbf{c}_i) = \varnothing$ and $\forall M_j \in \mathcal{M}$ where $M_j(\mathbf{x}) \neq M_i(\mathbf{x})$, $M_i \in \mathcal{A}(M_j)$. Similarly, $\forall \mathbf{c}_k \in \mathcal{C}$ where $M_i(\mathbf{c}_k) = M_i(\mathbf{x})$, $M_i \in \mathcal{A}(\mathbf{c}_k)$, and so M_i indirectly attacks M_k. Then, $\nexists X \in P_s$ such that $M_j \in X$ or $M_k \in X$ and thus, $\forall X \in P_s$, $M_i \in X$. □

We also show how, for any two s-preferred sets, there exists some trade-off between their models wrt the preferences.

Lemma 5.11. *Given two s-preferred sets* $X, X' \in P_s$, $\nexists M_i \in (X \cap \mathcal{M}) \setminus X'$ *such that* $M_i \succ_{\mathcal{M}} M_j$ $\forall M_j \in X'$.

Meanwhile, in the non-strict case, a model which is not outperformed by any other model wrt the preferences will be present in at least one s-preferred set.

Proposition 5.12. *If* $\exists M_i \in \mathcal{M}$ *such that* $\forall M_j \in \mathcal{M}$, $M_i \succeq_{\mathcal{M}} M_j$, *then* $\exists X \in P_s$ *such that* $M_i \in X$.

Proof. If $\forall M_j \in \mathcal{M}, M_i \succeq_{\mathcal{M}} M_j$ then, by Definition 5.5, $\nexists \alpha_k \in \mathcal{A}(M_i) \cup \mathcal{A}(\mathbf{c}_i)$ such that $M_i \notin \mathcal{A}(\alpha_k)$ and $\mathbf{c}_i \notin \mathcal{A}(\alpha_k)$. Thus, it must be the case that $\exists X \in P_s$ such that $M_i \in X$. □

We now show that if a model is outperformed wrt the preferences by all other models, then the outperformed model cannot exist in an s-preferred set unless it is defended by a more preferred model.

Proposition 5.13. *For any $M_i \in \mathcal{M}$, if $M_j \succ_{\mathcal{M}} M_i$ $\forall M_j \in \mathcal{M} \setminus \{M_i\}$ and $\exists X \in P_s$ such that $M_i \in X$, then, $\forall M_k \in \mathcal{M}$ where $M_k(\mathbf{x}) \neq M_i(\mathbf{x})$, $\exists M_l \in X$ such that $M_l \succeq_{\mathcal{M}} M_k$.*

Proof. By Definition 5.5, $M_k \in \mathcal{A}(M_i)$ and $\{M_m \in \mathcal{M} | M_i \in \mathcal{A}(M_m) \vee \mathbf{c}_i \in \mathcal{A}(M_m)\} = \emptyset$. Then, given that $\exists X \in P_s$ such that $M_i \in X$, we know that, $\forall M_k \in \mathcal{A}(M_i)$ $\exists M_l \in X$ where $M_l(\mathbf{x}) = M_i(\mathbf{x})$ and $M_l \in \mathcal{A}(M_k)$. Thus, by Definition 5.5, $M_l \succeq_{\mathcal{M}} M_k \succ_{\mathcal{M}} M_i$. □

We also consider the behaviour of argumentative ensembling wrt the selected CEs, demonstrating that those from disagreeing models are guaranteed not to be included in any s-preferred set.

Proposition 5.14. *Any s-preferred set $X \in P_s$ is such that $\nexists \mathbf{c}_i, \mathbf{c}_j \in X \cap \mathcal{C}$ where $M_i(\mathbf{x}) \neq M_j(\mathbf{x})$.*

Proof. Let us prove by contradiction, assuming that $\exists \mathbf{c}_i, \mathbf{c}_j \in X \cap \mathcal{C}$. Counterfactual coherence (Theorem 5.8) requires that $M_i, M_j \in X$. However, by Definition 5.5, $M_i(\mathbf{x}) \neq M_j(\mathbf{x})$ requires that $M_i \in \mathcal{A}(M_j)$ or $M_j \in \mathcal{A}(M_i)$, and so we have the contradiction. □

Finally, we show that s-preferred sets of corresponding BAFs in our setting satisfy all the forms of admissibility for BAFs in [11].

Proposition 5.15. *Any s-preferred set $X \in P_s$ is d-admissible, s-admissible and c-admissible.*

Proof. Trivially, any s-preferred set is s-admissible and thus also d-admissible (see §3). Then, it can be seen from Definition 5.5 that $\mathcal{S} \subseteq (\mathcal{M} \times \mathcal{C}) \cup (\mathcal{C} \times \mathcal{M})$, $i = j$ $\forall (M_i, \mathbf{c}_j) \in \mathcal{S} \cap (\mathcal{M} \times \mathcal{C})$ and $k = l$ $\forall (\mathbf{c}_k, M_l) \in \mathcal{S} \cap (\mathcal{C} \times \mathcal{M})$. By counterfactual coherence (Theorem 5.8), $\forall M_i \in \mathcal{M}$ and $\forall \mathbf{c}_i \in \mathcal{C}$, $M_i \in X$ iff $\mathbf{c}_i \in X$. Then, since P_s contains the sets of $\mathcal{X}$ which are maximal wrt set-inclusion, any $X' \in P_s$ must be closed for $\mathcal{S}$ and thus c-admissible. □

6 EMPIRICAL EVALUATION

We now examine the effectiveness of our approach using three real-world datasets. Specifically, we empirically evaluate the extent to which each of the ensembling methods introduced in §4.2 and §5 satisfy the desirable properties defined in §4.1. We also instantiate three variations of argumentative ensembling by including two different types of model properties $\mathcal{P}$ and demonstrate the usefulness of incorporating model preferences into ensembling methods.

6.1 Experiment Setup

We apply all ensembling methods on three datasets in the legal and financial contexts: loan approval (heloc) [27], recidivism prediction (compas) [37], and credit risk (credit) [32]. Due to neural networks' sensitivity to randomness at training time, they suffer severely from MM and are frequently targeted when investigating this research topic (as discussed in §2). Therefore, even though our method is model-agnostic, we focus on neural networks for the experiments.

For each dataset, we train-test 150 classifiers with five different hidden layer sizes using 80% of the dataset (this 80% is train-test split for training each model; see Appendix A.1 for dataset and training details). The 150 neural networks are trained using different random seeds for parameter initialisation and different train-test splits (within the train-test 80% of the dataset), forming a pool of possible models under MM from which we sample multiple sets $\mathcal{M}$ of models to which we apply our ensembling methods. We use the remaining 20% of each dataset as test inputs for the ensembling methods (limited to 500 inputs test set if larger than this).

At each run, we randomly sample, from the model pool, sets $\mathcal{M}$ with 10, 20, or 30 models, then we feed each input to the models to receive their predicted labels and generate one CE from each model using the nearest neighbour CEs approach of [8], and finally apply the ensembling methods. For each size (10, 20, 30), we perform five different choices of model sets $\mathcal{M}$, and we record the mean and standard deviation of the results (over the five choices of model sets for each size).

As concerns model preferences, we focus on accuracy of the trained classifiers over the (20%) test inputs and model structure simplicity. For the latter, we assign the models, from the most complex to the simplest (depending on the number of neurons in the hidden layers, see see Appendix A.1), scores of {0, 0.25, 0.5, 0.75, 1} such that higher values imply simpler models. Note that multiple models in $\mathcal{M}$ may have the same simplicity scores because we adopt only five different model structures to obtain the 150 neural networks for each dataset. Models in $\mathcal{M}$ may have any (near-optimal) test accuracy: in the experiments each model in $\mathcal{M}$ has a different accuracy.

Evaluation metrics. Each ensembling method is evaluated against the following metrics: prediction accuracy over the test set (*acc.*), average model simplicity in the ensemble (*simp.*), average size of models and CEs in the ensemble, measured as percentages of $|\mathcal{M}|$ (*size M/C*), average validities of ensembled CEs over the ensembled models (**c** *val.*). Also, we report the percentage of test inputs for which a method fails to produce CEs (fail). The results are averaged over all the test inputs except for the failure cases. We also report the average test set accuracies of the models in $\mathcal{M}$. Note that the model agreement property (Property 4.4) is omitted as it is satisfied by every compared method. To understand how the violation of the majority vote property affects our method, we measure the proportion of test inputs for which the predicted label using our method is the same as that of naive ensembling (*mv*).

Ensembling Methods. We use augmented and robust ensembling as our baselines. For argumentative ensembling, we use four variations with different preferences: S^a ($\mathcal{P} = \emptyset$), S^a-A ($\mathcal{P} = \{accuracy\}$), S^a-S ($\mathcal{P} = \{simplicity\}$) and S^a-AS ($\mathcal{P} = \{accuracy, simplicity\}$, where $accuracy \simeq_{\mathcal{P}} simplicity$). In our implementation for argumentative ensembling, when $|X_{max}| > 1$ (see Definition 5.6), we return S^a, which has the same prediction label as naive ensembling. We additionally report the percentage of test inputs with $|X_{max} > 1|$ in Table 3 in Appendix A.2.

6.2 Results and Analyses

We report the results for all experiments in Table 2 (the standard deviations are presented in Tables 4 to 6 in Appendix A.3).

Usefulness of preferences. With test accuracy specified as model preference, S^a-A shows the best accuracy in all experiments. This validates Proposition 5.10, because, assuming that the accuracy for every model in $\mathcal{M}$ is different, for S^a-A, there exists a model in $\mathcal{M}$

	acc.	simp.	size M/C	c val. (fail)	mv	acc.	simp.	size M/C	c val. (fail)	mv	acc.	simp.	size M/C	c val. (fail)	mv
	heloc					compas					credit				
$\|\mathcal{M}\| = 10$	.709±.003					.856±.001					.664±.008				
S^n	.709	.495	.943/.943	.657 (.00)	1.00	.858	.464	.980/.980	.572 (.00)	1.00	.697	.588	.817/.817	.757 (.00)	1.00
S^v	.709	.495	.943/.309	1.00 (.34)	1.00	.858	.464	.980/.174	1.00 (.51)	1.00	.697	.588	.817/.457	1.00 (.44)	1.00
S^a	.712	.504	.499/.499	1.00 (.00)	.983	.859	.463	.369/.369	1.00 (.00)	.994	.694	.600	.580/.580	1.00 (.00)	.953
S^a-**A**	.726	.485	.357/.357	1.00 (.00)	.943	.864	.430	.295/.295	1.00 (.00)	.988	.710	.593	.486/.486	1.00 (.00)	.825
S^a-**S**	.710	.608	.462/.462	1.00 (.00)	.967	.860	.657	.306/.306	1.00 (.00)	.987	.689	.626	.565/.565	1.00 (.00)	.925
S^a-**AS**	.712	.528	.493/.493	1.00 (.00)	.980	.860	.501	.360/.360	1.00 (.00)	.994	.696	.607	.578/.578	1.00 (.00)	.946
$\|\mathcal{M}\| = 20$	.710±.003					.855±.001					.663±.004				
S^n	.717	.488	.940/.940	.626 (.00)	1.00	.859	.538	.978/.978	.544 (.00)	1.00	.708	.571	.810/.810	.734 (.00)	1.00
S^v	.717	.388	.940/.230	1.00 (.37)	1.00	.859	.538	.978/.111	1.00 (.60)	1.00	.708	.571	.810/.351	1.00 (.62)	1.00
S^a	.716	.466	.460/.460	1.00 (.00)	.984	.859	.514	.331/.331	1.00 (.00)	.992	.691	.580	.557/.557	1.00 (.00)	.961
S^a-**A**	.728	.432	.361/.361	1.00 (.00)	.950	.866	.541	.235/.235	1.00 (.00)	.982	.709	.586	.481/.481	1.00 (.00)	.862
S^a-**S**	.711	.551	.420/.420	1.00 (.00)	.966	.857	.609	.304/.304	1.00 (.00)	.987	.684	.590	.549/.549	1.00 (.00)	.947
S^a-**AS**	.715	.473	.459/.459	1.00 (.00)	.984	.859	.555	.324/.324	1.00 (.00)	.990	.693	.581	.556/.556	1.00 (.00)	.959
$\|\mathcal{M}\| = 30$	.710±.003					.855±.001					.663±.004				
S^n	.718	.512	.940/.940	.620 (.00)	1.00	.859	.527	.976/.976	.530 (.00)	1.00	.710	.540	.807/.807	.727 (.00)	1.00
S^v	.718	.512	.940/.205	1.00 (.41)	1.00	.859	.527	.976/.087	1.00 (.56)	1.00	.710	.540	.807/.311	1.00 (.70)	1.00
S^a	.716	.499	.456/.456	1.00 (.00)	.982	.862	.519	.308/.308	1.00 (.00)	.990	.683	.549	.551/.551	1.00 (.00)	.943
S^a-**A**	.729	.406	.353/.353	1.00 (.00)	.946	.865	.532	.225/.225	1.00 (.00)	.983	.711	.552	.441/.441	1.00 (.00)	.850
S^a-**S**	.712	.518	.445/.445	1.00 (.00)	.978	.861	.567	.294/.294	1.00 (.00)	.988	.684	.555	.546/.546	1.00 (.00)	.940
S^a-**AS**	.716	.500	.456/.456	1.00 (.00)	.982	.862	.543	.303/.303	1.00 (.00)	.988	.683	.549	.551/.551	1.00 (.00)	.944

Table 2: Quantitative evaluations of ensembling methods on three datasets, heloc, compas, and credit. $|\mathcal{M}| = \{10, 20, 30\}$ stands for results for different model set sizes, the acc. entries in the rows starting with $|\mathcal{M}|$ are the average single model accuracies.

that is the most preferred and is included in the ensemble. Similarly, S^a-S shows the best simp. scores in all experiments. However, since simplicity scores are not unique for each model, usually a single most preferred model does not exist, therefore an optimal simp. evaluation is not guaranteed. When specifying both properties as model preferences (S^a-AS), at least one of the two metrics is improved compared with S^a.

Desirable properties of ensembling methods. For up to 70% of test inputs, robust ensembling does not find any CEs ($S^v \cap \mathcal{C} = \emptyset$), confirming its violation of non-emptiness. As $|\mathcal{M}|$ increases, S^v would require finding CEs which are valid for more models, and the number of CEs found would drop as shown by the results for the size C evaluations. In contrast, the remaining methods, including argumentative ensembling, always find non-empty ensembles, the sizes of which are also not affected by the model set sizes.

S^n demonstrates low c Val. scores, showing that, on average, a CE from a model in the ensemble is only valid for 53.0% to 75.7% of other agreeing models. Therefore, the violation of counterfactual validity has a significant impact on results in practice. S^v produces valid CEs over models in the ensemble, however, as previously mentioned, they do not always exist.

For S^a, we observe the same number of models and CEs in the solution set and 100% counterfactual validity, confirming the behaviour predicted by the satisfaction of counterfactual coherence and counterfactual validity (Theorem 5.8). Argumentative ensembling shows model and CE ensemble sizes of 22.5% (when $|\mathcal{M}| = 30$) to 58.0% of $|\mathcal{M}|$, meaning that it is non-trivially more selective than S^n and S^v, only accepting the largest set of models with similar reasoning local to the test input (validated by agreement on CEs as required by counterfactual coherence). This results in comparable test accuracies as S^n and S^v with guaranteed CE validity. In fact, in most cases, the S^a-S option has the lowest agreement rate with majority vote prediction result (mv), however, it is more accurate than the baselines which use naive ensembling. When no preference is specified, argumentative ensembling has higher accuracy than majority vote for the heloc dataset when $|\mathcal{M}| = 10$ and for the compas dataset when $|\mathcal{M}| = 10, 30$. This indicates we do not necessarily lose accuracy in satisfying all desirable properties besides majority vote.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a formal study of the problem of providing recourse under MM. We defined several properties which are desirable in methods for solving this problem, highlighting deficiencies in methods extending conservatively the standard naive ensembling used for MM without recourse. We have then introduced argumentative ensembling, a novel method for providing recourse under MM, which leverages computational argumentation to incorporate robustness guarantees and user preferences over models. We show, by means of a theoretical analysis, that argumentative ensembling hosts advantages over the other methods, notably in non-emptiness of solutions and validity of CEs, notwithstanding its ability to handle user preferences. This is, however, achieved by sacrificing the satisfaction of the property of majority vote, and so we conducted an empirical analysis with three real-world datasets to examine the effects of this sacrifice. Our results demonstrate that argumentative ensembling always finds valid CEs without compromising prediction accuracy, and shows the usefulness of specifying preferences over models.

This paper opens up several potentially fruitful directions for future work. First, it would be interesting to examine the extent to which considering attacks to or from *sets* of arguments, rather than single arguments, as in [19, 24, 29, 46], may help in MM given the conflicts between agreeing sets of models. We would also like to exploit the explanatory potential of argumentation to support explainable ensembling, e.g. using sub-graphs as in [25, 63].

REFERENCES

[1] Nadia Abchiche-Mimouni, Leila Amgoud, and Farida Zehraoui. 2023. Explainable Ensemble Classification Model based on Argumentation. In *AAMAS 2023*. 2367–2369. https://doi.org/10.5555/3545946.3598936

[2] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. 2017. Towards Artificial Argumentation. *AI Magazine* 38, 3 (2017), 25–36.

[3] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre (Eds.). 2018. *Handbook of Formal Argumentation*. College Publications.

[4] Emily Black, Klas Leino, and Matt Fredrikson. 2022. Selective Ensembles for Consistent Predictions. In *ICLR 2022*. https://openreview.net/forum?id=HfUyCRBeQc

[5] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *FAccT 2022*. 850–863. https://doi.org/10.1145/3531146.3533149

[6] Emily Black, Zifan Wang, and Matt Fredrikson. 2022. Consistent Counterfactuals for Deep Models. In *ICLR 2022*. https://openreview.net/forum?id=St6eyiTEHnG

[7] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[8] Dieter Brughmans, Pieter Leyman, and David Martens. 2023. NICE: An Algorithm for Nearest Instance Counterfactual Explanations. *Data Mining and Knowledge Discovery* (2023), 1–39. https://doi.org/10.1007/s10618-023-00930-y

[9] Ngoc Bui, Duy Nguyen, and Viet Anh Nguyen. 2022. Counterfactual Plans under Distributional Ambiguity. In *ICLR 2022*. https://openreview.net/forum?id=noaG7SrPVK0

[10] Elena Cabrio and Serena Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions†. *Argument Comput.* 4, 3 (2013), 209–230. https://doi.org/10.1080/19462166.2013.862303

[11] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *ECSQARU 2005*. 378–389. https://doi.org/10.1007/11518655_33

[12] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An Interpretable Model with Globally Consistent Explanations for Credit Risk. *CoRR* abs/1811.12615 (2018). arXiv:1811.12615 http://arxiv.org/abs/1811.12615

[13] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. In *ICML 2021*. 2144–2155. http://proceedings.mlr.press/v139/coston21a.html

[14] Kristijonas Cyras, Dimitrios Letsios, Ruth Misener, and Francesca Toni. 2019. Argumentation for Explainable Scheduling. In *AAAI 2019*. 2752–2759. https://doi.org/10.1609/aaai.v33i01.33012752

[15] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: A Survey. In *IJCAI 2021*. 4392–4399. https://doi.org/10.24963/ijcai.2021/600

[16] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *JMLR* 23, 1 (2022), 10237–10297.

[17] Adam Dejl, Chloe He, Pranav Mangal, Hasan Mohsin, Bogdan Surdu, Eduard Voinea, Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. 2021. Argflow: A Toolkit for Deep Argumentative Explanations for Neural Networks. In *AAMAS 2021*. 1761–1763. https://doi.org/10.5555/3463952.3464229

[18] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *NeurIPS 2018*. 590–601. https://proceedings.neurips.cc/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html

[19] Yannis Dimopoulos, Wolfgang Dvorák, Matthias König, Anna Rapberger, Markus Ulbricht, and Stefan Woltran. 2023. Sets Attacking Sets in Abstract Argumentation. In *NMR 2023*. 22–31. https://ceur-ws.org/Vol-3464/paper3.pdf

[20] Jiayun Dong and Cynthia Rudin. 2019. Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. *CoRR* abs/1901.03209 (2019). arXiv:1901.03209 http://arxiv.org/abs/1901.03209

[21] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artif. Intell.* 77, 2 (1995), 321–358. https://doi.org/10.1016/0004-3702(94)00041-X

[22] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. 2022. Robust Counterfactual Explanations for Tree-Based Ensembles. In *ICML 2022*. 5742–5756. https://proceedings.mlr.press/v162/dutta22a.html

[23] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. 2020. Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. In *ICML 2020*. 2803–2813. http://proceedings.mlr.press/v119/dutta20a.html

[24] Wolfgang Dvorák, Matthias König, Markus Ulbricht, and Stefan Woltran. 2022. Rediscovering Argumentation Principles Utilizing Collective Attacks. In *KR 2022*. 122–131. https://proceedings.kr.org/2022/13/

[25] Xiuyi Fan and Francesca Toni. 2014. On Computing Explanations in Abstract Argumentation. In *ECAI 2014*. 1005–1006. https://doi.org/10.3233/978-1-61499-419-0-1005

[26] Andrea Ferrario and Michele Loi. 2022. The Robustness of Counterfactual Explanations Over Time. *IEEE Access* 10 (2022), 82736–82750. https://doi.org/10.1109/ACCESS.2022.3196917

[27] FICO. 2018. Explainable Machine Learning Challenge. https://community.fico.com/s/explainable-machine-learning-challenge

[28] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* 20 (2019), 177:1–177:81. http://jmlr.org/papers/v20/18-760.html

[29] Giorgos Flouris and Antonis Bikakis. 2019. A comprehensive study of argumentation frameworks with sets of attacking arguments. *Int. J. Approx. Reason.* 109 (2019), 55–86. https://doi.org/10.1016/j.ijar.2019.03.006

[30] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (2022), 1–55. https://doi.org/10.1007/s10618-022-00831-6

[31] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. 2023. Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees. In *ICML 2023*. 12351–12367. https://proceedings.mlr.press/v202/hamman23a.html

[32] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77

[33] Hsiang Hsu and Flávio P. Calmon. 2022. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. In *NeurIPS 2023*. 28988–29000. http://papers.nips.cc/paper_files/paper/2022/hash/ba4caa85ecdcafbf9102ab8ec384182d-Abstract-Conference.html

[34] Benjamin Irwin, Antonio Rago, and Francesca Toni. 2022. Forecasting Argumentation Frameworks. In *KR 2022*. 533–543. https://proceedings.kr.org/2022/55/

[35] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. 2023. Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. *CoRR* abs/2309.12545 (2023). https://doi.org/10.48550/arXiv.2309.12545 arXiv:2309.12545

[36] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2023. Formalising the Robustness of Counterfactual Explanations for Neural Networks. In *AAAI 2023*. 14901–14909. https://doi.org/10.1609/aaai.v37i12.26740

[37] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. 2016. There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

[38] Francesco Leofante, Elena Botoeva, and Vineet Rajani. 2023. Counterfactual Explanations and Model Multiplicity: a Relational Verification View. In *KR 2023*. 763–768. https://doi.org/10.24963/kr.2023/78

[39] Dan Ley, Leonard Tang, Matthew Nazari, Hongjin Lin, Suraj Srinivas, and Himabindu Lakkaraju. 2023. Consistent Explanations in the Face of Model Indeterminacy via Ensembling. *CoRR* abs/2306.06193 (2023). https://doi.org/10.48550/arXiv.2306.06193 arXiv:2306.06193

[40] Charles Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. 2023. But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI. In *AISTATS 2023*. 7375–7391. https://proceedings.mlr.press/v206/marx23a.html

[41] Charles T. Marx, Flávio P. Calmon, and Berk Ustun. 2020. Predictive Multiplicity in Classification. In *ICML 2020*. 6765–6774. http://proceedings.mlr.press/v119/marx20a.html

[42] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. 2020. Individual differences among deep neural network models. *Nature communications* 11, 1 (2020), 5725. https://doi.org/10.1038/s41467-020-19632-w

[43] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. 2021. A Survey on the Robustness of Feature Importance and Counterfactual Explanations. *CoRR* abs/2111.00358 (2021). arXiv:2111.00358 https://arxiv.org/abs/2111.00358

[44] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT 2020*. 607–617. https://doi.org/10.1145/3351095.3372850

[45] Tuan-Duy H. Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. 2022. Robust Bayesian recourse. In *UAI 2022*. 1498–1508. https://proceedings.mlr.press/v180/nguyen22a.html

[46] Søren Holbech Nielsen and Simon Parsons. 2006. A Generalization of Dung's Abstract Framework for Argumentation: Arguing with Sets of Attacking Arguments. In *ArgMAS 2006*. 54–73. https://doi.org/10.1007/978-3-540-75526-5_4

[47] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *UAI 2020*. 809–818. http://proceedings.mlr.press/v124/pawelczyk20a.html

[48] Nico Potyka. 2021. Interpreting Neural Networks as Quantitative Argumentation Frameworks. In *AAAI 2021*. 6463–6470. https://doi.org/10.1609/aaai.v35i7.16801

[49] Nico Potyka, Xiang Yin, and Francesca Toni. 2023. Explaining Random Forests Using Bipolar Argumentation and Markov Networks. In *AAAI 2023*. 9453–9460. https://doi.org/10.1609/aaai.v37i8.26132

[50] Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. *Nat. Mach. Intell.* 3, 10 (2021), 896–904. https://doi.org/10.1038/s42256-021-00396-x

[51] Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2023. Reconciling Individual Probability Forecasts. In *FAccT 2023*. 101–110. https://doi.org/10.1145/3593013.3593980

[52] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[53] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the Existence of Simpler Machine Learning Models. In *FAccT 2022*. 1827–1858. https://doi.org/10.1145/3531146.3533232

[54] Sjoerd T. Timmer, John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. 2015. Explaining Bayesian Networks Using Argumentation. In *ECSQARU 2015*. 83–92. https://doi.org/10.1007/978-3-319-20807-7_8

[55] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *KDD 2017*. 465–474. https://doi.org/10.1145/3097983.3098039

[56] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. 2021. Towards Robust and Reliable Algorithmic Recourse. In *NeurIPS 2021*. 16926–16937. https://proceedings.neurips.cc/paper/2021/hash/8ccfb1140664a5fa63177fb6e07352f0-Abstract.html

[57] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *FAT 2019*. 10–19. https://doi.org/10.1145/3287560.3287566

[58] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.* 36 (2021), e5. https://doi.org/10.1017/S0269888921000011

[59] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

[60] Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. 2023. Predictive Multiplicity in Probabilistic Classification. In *AAAI 2023*. 10306–10314. https://doi.org/10.1609/aaai.v37i9.26227

[61] Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *NeurIPS 2019*. 8780–8789. https://proceedings.neurips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html

[62] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo I. Seltzer, and Cynthia Rudin. 2022. Exploring the Whole Rashomon Set of Sparse Decision Trees. In *NeurIPS 2022*. 14071–14084. http://papers.nips.cc/paper_files/paper/2022/hash/5afaa8b4dd18eb1eed055d2d821b58ae-Abstract-Conference.html

[63] Zhiwei Zeng, Chunyan Miao, Cyril Leung, Zhiqi Shen, and Jing Jih Chin. 2019. Computing Argumentative Explanations in Bipolar Argumentation Frameworks. In *AAAI 2019*. 10079–10080. https://doi.org/10.1609/aaai.v33i01.330110079