# An Exploration into A Fortiori Case-Based Argumentation as an Interpretable Classification Approach

Joeri Peters[1,2,*], Floris Bex[1] and Henry Prakken[1]

[1]Utrecht University, Utrecht, The Netherlands
[2]Netherlands National Police, Driebergen, The Netherlands

**Abstract**

This paper is aimed at investigating whether the XAI approach 'AF-CBA', which produces model-agnostic justifications for binary classifications using case-based argumentation, can function effectively as a classification approach. Its performance is evaluated across multiple datasets. The results indicate that AF-CBA can be used as a classifier comparable to decision trees for some tasks, albeit one that may rely on a fallback mechanism. This may make it a viable option for specific high-risk domains, such as law (enforcement), in which interpretability is paramount.

**Keywords**

Machine Learning, Classification, Argumentation, XAI Case-Based Reasoning,

## 1. Introduction

The decisions of a black-box model are opaque to its users. The post-hoc, model-agnostic explainable artificial intelligence (XAI [1]) approach called '*a fortiori* case-based argumentation' (AF-CBA) by Prakken and Ratsma [2] can justify the binary predictions of any black-box [3, 4] machine learning (ML) classifier, thereby allowing for their interpretation in terms users can understand. Transparency and interpretability are essential for establishing trust, especially in domains where decisions have significant consequences. Transparency/interpretability is often a legal requirement [5]. AF-CBA is built on the theory of precedential constraint [6], referencing cases from a case base and presenting a case-based argumentation framework's dialogue to the user as a justification of the predicted class label. This approach thus follows legal patterns of reasoning, in which precedents are key. Extensions to AF-CBA have been suggested [7, 8, 9, 10] as a way of improving AF-CBA's ability to address the trade-off between performance and interpretability [11] with its justifications. However, this trade-off can at times be an acceptable price to pay [12, 13]. The trade-off can be problematic in certain high-risk domains, such as justice, law enforcement and counter-terrorism, medicine and the military. Some would argue that it is preferable to improve interpretable models, rather than to rely on black-box models [14]. The question that can therefore be raised is whether the approach of AF-CBA could be used to form an interpretable classifier. As an example in this paper, we adopt the fictional scenario, based on our own application domain, of officials using a classifier to label incidents as acts of terrorism.

The rest of this paper is structured as follows. The AF-CBA classifier and how it relates to the original justification formalism [2] is described in Section 2. Classification experiments are presented in Section 3, where the performance of AF-CBA as a classifier is compared to decision trees and random forests. Section 4 is concerned with the interpretability of this approach. Section 5 describes related work and Section 6 discusses limitations and future work directions.



*RELAF'26: Reasoning with Evidence in Law Enforcement and Forensics, Workshop at ICAIL 2026, Singapore.*
*Corresponding author.
j.g.t.peters@uu.nl (J. Peters); f.j.bex@uu.nl (F. Bex); h.prakken@uu.nl (H. Prakken)
0009-0009-1493-9872 (J. Peters); 0000-0002-5699-9656 (F. Bex); 0000-0002-3431-7757 (H. Prakken)

## 2. Preliminaries

Binary class predictions are analogous to court rulings based on legal precedent. Prakken and Ratsma [2] introduced the 'top-level model' that was later named 'AF-CBA', drawing on AI & law research and influenced by Horty's theory of *precedential constraint* [15], CATO [16] and Čyras et al. [17, 18]. AF-CBA generates post-hoc and model-agnostic justifications, meaning it explains classification predictions without having access to the black-box model. It does, however, require access to a repository of labelled instances – the *case base*. This can be either the original training data or an archive of earlier classifications [9], but the former is used in this paper. A *focus case* is a new, individual instance for which the classifier predicts an *outcome*. AF-CBA conducts an argument game between a *proponent* and an *opponent* of that prediction. Similar cases are *cited* as meaningful precedents, provided that differences with the focus case only serve to reinforce the outcome for the latter (precedential constraint).

### 2.1. Precedential Constraint

A *case* within a case base (CB) consists of an outcome and a *fact situation*. The outcome is a binary label, $o$ or $o'$. The variables $s$ and $\bar{s}$ represent the two sides, with $s = o$ if $\bar{s} = o'$, and vice versa. The fact situation contains values of *dimensions*, each a tuple $d = (V, \leq_o, \leq_{o'})$. This includes a value set $V$ and two partial orderings on $V$, $\leq_o$ and $\leq_{o'}$, such that $v \leq_o v'$ if and only if $v' \leq_{o'} v$, where $v, v' \in V$. Each dimension has a *tendency*, with a positive tendency indicating that higher values are associated with one outcome and the inverse tendency with the opposite outcome. The tendency may be given explicitly as $d_i^+$ or $d_i^-$ (with regards to outcome $o$). A dimension is thus a feature with a tendency. A value assignment, $(d, v)$, specifies the value $x$ of dimension $d$ in case $c \in CB$, noted as $v(d, c) = x$. The set of value assignments for all dimensions $D$ (non-empty) constitutes a fact situation denoted as $F$. It is assumed that two fact situations refer to the same $D$. A case is denoted as $c = (F, outcome(c))$, with $outcome(c) \in \{o, o'\}$. The fact situation $c$ is denoted as $F(c)$. When two fact situations are compared, one case is determined to be 'stronger' or 'better' for a particular outcome than the other (i.e. it is considered positively correlated). For instance, Table 1 shows precedent $c$ with a better value for dimension $d^-_{weapon}$ than the focus case $f$, but a worse value for $d^+_{casualties}$. The focus case's outcome is considered *forced* if there exists a precedent with the same outcome for which all differences with the focus case only reinforce the focus case's strength for that outcome [6].

**Table 1**
Precedent case $c$ and focus case $f$.

| Case | $d^+_{casualties}$ | $d^-_{weapon}$ | ... | Outcome |
|---|---|---|---|---|
| $c$ | 5 | low | ... | True |
| $f$ | 10 | high | ... | True |

Given two fact situations $F$ and $F'$, there exists a *preference relation* $F \leq_s F'$ between fact situations if $v \leq_s v'$ for all $(d, v) \in F$ and $(d, v') \in F'$. Deciding $F$ for $s$ is forced if the case base contains a case $c = (F', s)$ such that $F' \leq_s F$ (precedential constraint). A case base is consistent if it does not contain two cases $c = (F, s)$ and $c' = (F', \bar{s})$ such that $F \leq_s F'$. Otherwise, it is inconsistent. A *best precedent* (Definition 2) is defined as having the same outcome as the focus case, with as few *relevant differences* (Definition 1) as possible. More than one case can meet these criteria.

**Definition 1** (Differences between cases)
Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases. The set $D(c, f)$ of relevant differences between $c$ and $f$ is defined as:

$$D(c, f) = \{d \in D \mid v(d, c) \not\leq_{outcome(c)} v(d, f)\}.$$

**Definition 2** (Best precedent)
Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases. $c$ is a best precedent for $f$ iff:

- $outcome(c) = outcome(f)$ and
- there is no $c' \in CB$ such that $outcome(c') = outcome(c)$ and $D(c', f) \subset D(c, f)$.

## 2.2. Argument Game

The argument game of Prakken and Ratsma [2] for explaining outcomes is based on Prakken's game [19] for grounded semantics of abstract argumentation frameworks [20]. An abstract argumentation framework (AF), introduced by Dung [20], is a pair $AF = \langle A,\ attack \rangle$, where $A$ is a set of arguments and $attack$ is a binary relation on $A$. A subset $B$ of $A$ is considered *conflict-free* if no argument in $B$ attacks any other argument in $B$. It is considered *admissible* if it is both conflict-free and able to defend itself against attacks. More specifically, if an argument $A_1$ belongs to $B$, and some argument $A_2$ in $A$ attacks $A_1$, there must then be an argument in $B$ that counters $A_2$.

In the grounded game, an application of the argument game that is sound and complete for grounded semantics, two players debate about any differences between the focus case and cited precedents. It is normally played for the outcome predicted by a classifier. The first player, the proponent, argues in favour of that game's target outcome. The opponent challenges that outcome. The game starts with the proponent citing a best precedent. The opponent attempts to respond either by citing a *counterexample* or by playing a *distinguishing move*, $Worse(c, x)$, which indicates that the focus case is weaker than the precedent $c$ for the dimension-value pairs $x$. This move can then be attacked by the proponent's *downplaying move*, $Compensates(c, y, x)$, with dimension-value pairs $y$ offsetting the shortcomings of dimension-value pairs $x$. In response to a cited counterexample, the *transformation move*, $Transformed(c, c')$, states that the initially cited case can be transformed into one with $D(c', f) = \emptyset$ using a chain of downplaying moves. This back-and-forth continues until a player is unable to make further moves. Note that if the proponent's argument is of the form $Transformed(c, c')$, the case $c$ in question refers to the proponent's initial citation argument – the best precedent. A player wins a dialogue if the other player cannot move.

Note that in Prakken and Ratsma [2], the compensating set $y$ in $Compensates(c, y, x)$ can be the empty set. This guarantees a winning strategy for the proponent, thus making AF-CBA 'explanation complete' [2]. The argument game is based on the following abstract argumentation framework:

**Definition 3** (Case-based argumentation framework)
Given a case base $CB$, a focus case $f \notin CB$, and definitions of compensation $sc$, an abstract argumentation framework AF is a pair $\langle \mathcal{A}, attack \rangle$, where:

- $\mathcal{A} = CB \cup M$,
  with $M = \{Worse(c, x) \mid c \in CB, x \neq \emptyset$ and
  $x = \{(d, v) \in F(f) \mid v(d, f) <_{outcome(f)} v(d, c)\}\} \cup$
  $\{Compensates(c, y, x) \mid c \in CB, y \subseteq \{(d, v) \in F(f) \mid v(d, c) <_{outcome(f)} v(d, f)\}$,
  $x = \{(d, v) \in F(f) \mid v(d, f) <_{outcome(f)} v(d, c)\}$ and $y$ compensates $x$ according to $sc\} \cup$
  $\{Transformed(c, c') \mid c \in CB$ and $c$ can be transformed into $c'$ and $D(c', f) = \emptyset\}$
- $A$ attacks $B$ iff:
  - $A, B \in CB$ and $outcome(A) \neq outcome(B)$ and $D(B, f) \not\subset D(A, f)$;
  - $B \in CB$ with $outcome(B) = outcome(f)$ and $A$ is of the form $Worse(B, x)$;
  - $B$ is of the form $Worse(c, x)$ and $A$ is of the form $Compensates(c, y, x)$;
  - $B \in CB$ and $outcome(B) \neq outcome(f)$ and $A$ is of the form $Transformed(c, c')$.

In the grounded game, the proponent and opponent take turns to argue about the outcome, which forms *dispute trees*. A dispute tree that branches after the proponent's moves and contains all the opponent's legal replies, is a strategy for the proponent. If the proponent wins all disputes in that strategy and can thus win the argument, it is a *winning strategy*. This is presented as a justification for the predicted outcome as an argument graph. This dialogue flow is graphically depicted in Figure 1.

For example, consider a focus case $f$ involving a possible terrorist incident with three casualties and low weapon sophistication. The case base contains a terrorist precedent $c$ with more casualties

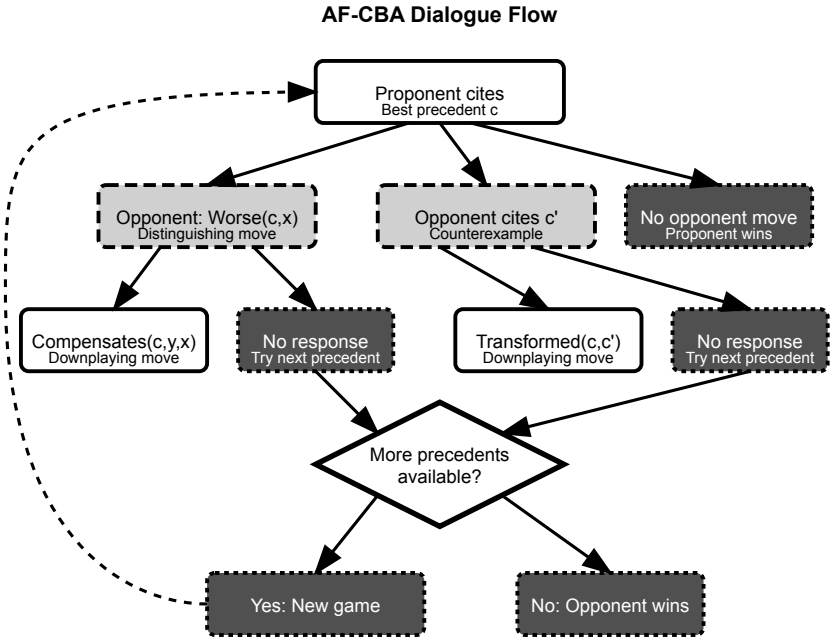


**Figure 1:** A schematic representation of the AF-CBA dialogue process, showing how proponent and opponent moves create a tree of possible dialogue moves.

(15, higher values being associated with acts of terrorism) but higher weapon sophistication ('high', not associated with acts of terrorism but instead with organised crime). Assuming that the model predicts $outcome(f) =$ terrorism, that the opponent can point at a worse dimension, as well as a counterexample, and also that both of these can be countered by the proponent in the way described above, AF-CBA could produce a justification through the argument game as shown in Figure 2.

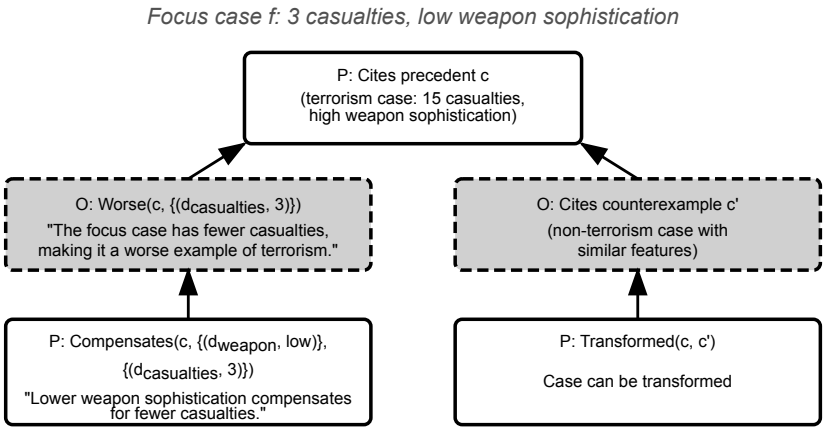


**Figure 2:** An example of a winning strategy for the proponent.

The set $sc$ (Definition 3) serves as the cornerstone of AF-CBA's reasoning process, determining which dimensions can compensate for deficiencies in which other dimensions. The set $sc$ was intentionally not specified by Prakken and Ratsma [2], but it is used by Peters et al. [10] to incorporate domain knowledge into the framework as preference relations between dimension sets, e.g. $D_w \prec D_b$, which they present as optionally discoverable from data. Preference relations motivate the downplaying moves ($Compensates(c, y, x)$), in the sense that any poor values of focus case $f$ in relation to precedent $c$ for dimension set $D_w$ can thus be compensated by the stronger values of $f$ in dimension set $D_b$. The intuition is that stronger dimensions are regarded as more significant for the outcome. Before preference relation discovery can commence, however, dimension tendencies must be determined, that is, the

direction of their correlation with the outcome. The coefficients from logistic regression are used to this end, after Van Woerkom et al. [21].

## 2.3. Preference relation discovery

In Peters et al. [10], it is shown that the feature importance scores of the black-box classifier whose predictions AF-CBA is meant to justify can be combined with a rule-based classifier to infer conditional preference relations. Specifically, a rule-based classifier (RIPPER [22, 23]) is used to induce a set of decision rules, each of the form IF $\Psi_k$ THEN outcome $= s_k$. For each rule context $\Psi_k$, conditional preference relations are derived by comparing aggregated SHAP importance scores of dimension sets, computed from the black-box classifier over the subset of cases satisfying $\Psi_k$. A preference relation $D'_j \prec D'_i$ is recognised when the relative difference in aggregated importance exceeds a threshold $\delta$. This produces context-specific preference relations that only apply when conditions $\Psi_k$ are satisfied. The set of all such instantiations is written as $\mathcal{P} = \bigcup_{k=1}^{K} \{ \, (\Psi_k, \; D'_j \prec D'_i) \, \}$.

Because first inducing rules and then determining preference relations is arguably rather complex, it is worth considering a simplified alternative. In addition to this conditional approach, a more straightforward approach is therefore adopted as an alternative. This *global preference relation discovery* is depicted in Definition 4. This alternative approach likewise uses SHAP [24], globally (unconditionally) rather than locally (conditionally).

**Definition 4** (Global preference relation discovery)
Let $CB$ be a case base, $D$ its set of dimensions, $\delta$ the importance difference threshold, and $C_b$ a statistical classifier that makes predictions on cases from $CB$. For any dimension $d \in D$, let $\phi(d)$ denote the SHAP importance of dimension $d$ as computed from $C_b$. For any set of dimensions $D' \subseteq D$, define the aggregated importance of $D'$ as:

$$\Phi(D') = \sum_{d \in D'} |\phi(d)|.$$

A preference relation $D'_j \prec D'_i$ is discovered if and only if $D'_i \cap D'_j = \emptyset$ and

$$\frac{\Phi(D'_i) - \Phi(D'_j)}{\Phi(D'_j)} > \delta,$$

where both $D'_i$ and $D'_j$ are non-empty sets. [1]

For both methods, conditional [10] as well as global (Definition 4) preference relation discovery, a threshold $\delta$ is used to determine when the importance difference between dimension sets is significant enough to justify a preference relation. For instance, a preference relation $\{d_{casualties}\} \prec \{d_{location}\}$ is produced if $(\Phi((\{d_{location}\})) - \Phi((\{d_{casualties}\})))/\Phi((\{d_{casualties}\})) > \delta$, with $\delta = 2.0$ used throughout this paper. This simplified method of deriving global preference relations directly based on aggregated SHAP importance scores over the entire case base is less computationally expensive and simpler to implement than the original conditional method of Peters et al. [10]. Its simplicity may arguably improve the interpretability of the overall approach. However, it could prove to be less effective due to feature importances varying locally, whereas the context-specificity of Peters et al. [10] would then lead to better classification performance.

In AF-CBA's original role as a post-hoc XAI approach, any ML prediction is justifiable. This *explanation completeness* [2] is guaranteed by allowing the proponent to play Compensates$(c, \emptyset, x)$, meaning the precedent's outcome is adopted regardless of any dimensions in which the focus case is worse. When AF-CBA is used as a classifier, this guarantee is neither necessary nor desirable – compensation is only applied when worse dimensions are genuinely outweighed by better dimensions according to a preference relation, and there is no predetermined outcome for which a winning strategy must be found. In terms of Definition 3, there is no need to explicitly state $y \neq \emptyset$ in the downplaying move.

[1] A slightly better formulation of this condition has been kindly suggested: $\Phi(D'_i) > \delta' \Phi(D'_j)$ with $\delta' = \delta + 1$.

Consequently, some focus cases may receive no justified outcome, in which case a default outcome can be used. The default outcome not only operates as a fallback for when AF-CBA fails to find a winning strategy for either outcome; it operates as a tiebreaker when it finds one for both. Thus, the *single justification fraction* is the relative number of cases for which AF-CBA finds a winning strategy for only one of the two outcomes and thus does not rely on the default heuristic and the *double justification fraction* represents the fraction of cases for which the heuristic was applied as a tiebreaker.

Not being able to justify either outcome can be viewed as a strength: it signals that the decision cannot be grounded sceptically in precedential constraint, which is itself informative. In many application domains, there would still be an interest in classifying such cases by default. Possible defaults include risk-based heuristics (favouring the outcome with more acceptable worst-case consequences), or expert-specified ones reflecting professional standards or legal principles (e.g. presumption of innocence). The *majority-class heuristic* is arguably the most straightforward. It minimises overall classification errors under the assumption that the sample distribution reflects the true population distribution. The majority class is therefore the heuristic used throughout the experiments in this paper.

## 3. Classification Experiments

To assess whether AF-CBA is viable as a classifier, the approach is implemented – using a Python version of the grounded game originally implemented in Java by Visser [25] after Vreeswijk [26] – and its performance is evaluated using three traditional ML metrics: accuracy, macro F1-score and MCC. SHAP values are determined using HistGradientBoosting [27] on default settings.

Model performance is evaluated across four datasets: Graduate Admissions (applicants to a master's programme and whether they were accepted, see Acharya et al. [28]), Telco Customer Churn (customers of a telecommunications provider and whether they switched providers, see IBM [29]), the Global Terrorism Dataset (GTD, a dataset with terrorism-related incident data [30, 31], taking all incidents post-1997), and COMPAS (convicted criminals and whether they are recidivists, see Angwin et al. [32] and Ofer [33]). The GTD can act as a proxy for terrorism-related datasets used in law enforcement practice, relevant to our own domain, which are typically not publically available due to privacy concerns and other sensitivities. It contains many descriptors such as attack, weapon and target types, with whether the incident is a suicide attack being the class label in this paper.

The following preprocessing steps are applied: standardisation of continuous features, one-hot-encoding of categorical features and the correction or removal of instances with missing or erroneous values. The exact steps undertaken per dataset are available as part of the online code repository (https://github.com/JGTP/AF-CBA). Across all four, the maximum (stratified) sample size is set to 7000, with Admissions already being smaller than that (500).

AF-CBA is compared with Scikit-learn's [27] default decision tree and random forest classifiers, using default parameters. All approaches are evaluated using 5-fold stratified cross-validation (CV) with a 20% holdout test set. The CV scores provide a robust estimate for comparing models, whilst the holdout scores offer an independent check on whether performance generalises beyond the data used. When the two scores are similar, the comparison between models is stable; when they diverge, it may indicate overfitting or sensitivity to the particular data split.

The datasets exhibit some class imbalance, which can affect both performance and interpretability. This is particularly pertinent to metrics like accuracy, which may be misleading when dominated by the majority class. Macro F1-scoring and especially MCC are more robust in this regard [34]. The class distributions resulting from stratified sampling (again, see the code repository) are reported in Table 2, with cross-validation and holdout sets maintaining approximately the same distributions.

Real-world case bases often contain inconsistent cases – precedents that support contradictory outcomes despite being similar to one another. As discussed in Peters et al. [9], such inconsistencies can undermine the quality of AF-CBA's justifications. Because there is no definitive outcome to justify when AF-CBA is used as a classifier, this is less relevant to the purposes of this paper. Adopting the recommendation of Peters et al. [9] of filtering best precedents (Definition 2) based on case base

**Table 2**
Class distribution for the datasets used in the experiments when sampled to the target size. The CV folds and holdout sets are stratified samples of these datasets.

| | Majority class $s$ | Minority class $\bar{s}$ | Total |
|---|---|---|---|
| ADMISSIONS | 461 (92.2%) | 39 (7.8%) | 500 |
| CHURN | 5140 (73.4%) | 1860 (26.6%) | 7,000 |
| GTD | 6639 (94.8%) | 361 (5.2%) | 7,000 |
| COMPAS | 4595 (65.6%) | 2405 (34.4%) | 7,000 |

inconsistency in order to avoid inconsistent forcing here, in this paper, merely has the effect of reducing the number of best precedents that may be considered before reaching the conclusion that no winning strategy exists.

Both preference relation discovery methods are tested, conditional [10] as well as global (Definition 4). These are compared with the decision tree and random forest. The results of the classification experiments, using the data in Table 2, are presented in Table 3.

**Table 3**
Classification performance on the four datasets comparing AF-CBA (with C. for conditional and G. for global preference discovery, respectively) with traditional machine learning models, using accuracy, macro F1-score, and MCC. Cross-validation results are reported as $\mu \pm \sigma$; holdout performance is parenthesised.

| Model | Accuracy | F1 | MCC |
|---|---|---|---|
| | ADMISSIONS | | |
| Decision Tree | .89±0.01 (.94) | .64±0.09 (.80) | .28±0.17 (.59) |
| Random Forest | .92±0.03 (.94) | .62±0.13 (.77) | .27±0.27 (.55) |
| AF-CBA (C.) | .92±0.03 (.89) | .72±0.10 (.71) | .46±0.20 (.43) |
| AF-CBA (G.) | **.93**±0.02 (.92) | **.73**±0.10 (.73) | **.51**±0.18 (.46) |
| | CHURN | | |
| Decision Tree | .73±0.01 (.71) | .65±0.01 (.64) | .31±0.02 (.28) |
| Random Forest | **.79**±0.00 (.78) | **.71**±0.00 (.70) | **.44**±0.01 (.41) |
| AF-CBA (C.) | .77±0.00 (.77) | .67±0.01 (.65) | .36±0.01 (.33) |
| AF-CBA (G.) | .76±0.00 (.76) | .54±0.01 (.52) | .27±0.02 (.24) |
| | GTD | | |
| Decision Tree | **.98**±0.00 (.98) | .90±0.02 (.91) | .80±0.04 (.82) |
| Random Forest | **.98**±0.00 (.98) | **.91**±0.02 (.90) | **.82**±0.05 (.81) |
| AF-CBA (C.) | .95±0.01 (.94) | .67±0.08 (.53) | .38±0.16 (.11) |
| AF-CBA (G.) | .95±0.00 (.95) | .56±0.03 (.54) | .20±0.05 (.18) |
| | COMPAS | | |
| Decision Tree | .64±0.00 (.60) | .59±0.00 (.56) | .18±0.01 (.11) |
| Random Forest | .66±0.00 (.63) | **.61**±0.01 (.57) | **.22**±0.02 (.14) |
| AF-CBA (C.) | **.67**±0.01 (.66) | .48±0.03 (.46) | .15±0.04 (.10) |
| AF-CBA (G.) | .66±0.00 (.66) | .42±0.00 (.41) | .07±0.03 (.04) |

In terms of accuracy, AF-CBA (in combination with the majority-class heuristic) performs somewhat comparably to the baselines across three of the four datasets, typically falling within a few percentage points of the decision tree and random forest classifiers. For Admissions, a very small dataset, AF-CBA even outperforms the rest. The difference in macro F1-score is larger, especially for the GTD and COMPAS, with AF-CBA performing worse than the ML models. The difference is even more pronounced with MCC, which may indicate a greater sensitivity due to class imbalance, possibly as a result of the majority-class default. All models struggle with COMPAS and global AF-CBA's MCC scores are particularly poor for that dataset. When comparing the two preference relation discovery approaches, the conditional approach appears to perform better on all datasets except Admissions. In fact, conditional AF-CBA is better than the decision tree for Churn. This may suggest that the context-specificity of the conditional approach [10] is beneficial for all datasets but Admissions, that the more straightforward comparison of feature importance scores of the global approach (Definition 4)

is more suited for that small dataset – likely due to the conditions becoming poorly generalisable for such a small dataset. The holdout results broadly mirror the trends observed under CV, suggesting that the CV results provide a reasonable characterisation of the generalised model behaviour. The most notable possible exception is the GTD with the conditional approach to AF-CBA, with $0.38$ versus $0.11$ for MCC, suggesting that those CV folds lead to some small degree of overfitting under specific circumstances.

The above suggests that AF-CBA can be a reasonably well-performing classifier for some datasets. It is, however, important to consider the justification fractions. If both are low, it is not AF-CBA itself that is an acceptable classifier, but the majority-class heuristic.

**Table 4**
The single and double justification fractions for the AF-CBA results in Table 3. Cross-validation results are reported as $\mu \pm \sigma$; holdout performance is parenthesised.

| | Approach | Single | Double |
|---|---|---|---|
| ADMISSIONS | Conditional | .86±0.03 (.87) | .12±0.02 (.11) |
| | Global | .75±0.04 (.72) | .25±0.04 (.28) |
| CHURN | Conditional | 71±0.02 (.70) | .16±0.01 (.18) |
| | Global | .19±0.02 (.16) | .81±0.02 (.84) |
| GTD | Conditional | .76±0.02 (.74) | .12±0.03 (.15) |
| | Global | .75±0.05 (.72) | .25±0.05 (.28) |
| COMPAS | Conditional | .14±0.02 (.14) | .86±0.03 (.86) |
| | Global | .03±0.00 (.03) | .97±0.00 (.97) |

The justification fractions associated with the results in Table 3 are presented in Table 4. Across all datasets except COMPAS, the conditional approach favours single justifications. The global approach tends to produce a larger share of double justifications, but still favours single justifications with Admissions and the GTD. The sum of the two fractions approaches 1.0 in most instances, with the exception of Churn and GTD (both conditional). The dataset where AF-CBA performs the worst, COMPAS, is also the one with the lowest single justification fractions and the highest double justification fractions, possibly suggesting a useful diagnostic for when AF-CBA is likely to underperform. Holdout performance approximates CV performance to a reasonable extent. Finally, the single justification fraction for the global approach with the GTD is notably high. Although the conditional approach performs better in Table 3, that difference if not reflected in this table for the GTD in terms of the single justification fraction.

The lower single justification fractions for Churn (global) and COMPAS (both) raise an important question: when AF-CBA finds a winning strategy, how well does it then classify specifically? Evaluating performance metrics only on singly justified cases – those for which AF-CBA produces an unambiguous decision rather than defaulting to the majority class or justifying both outcomes – isolates the classifier's reasoning capability from its fallback behaviour. Because exactly which cases are singly justified can vary as a result of the difference between the conditional and global preference relation discovery approaches, these are presented separately in Tables 5 and 6, respectively. Note that these tables do not represent new experiments and the ML models are not retrained; the performance metrics are simply recalculated by limiting the test set to singly justified cases only. The size of each test sample in these two tables is therefore equal to the size of the CV folds' test sets (for example, $7000 \cdot 0.8 \cdot 0.2$) or holdout test sets ($7000 \cdot 0.2$) used in Table 3 multiplied by the corresponding single justification fraction in Table 4, as listed in Tables 5 and 6 underneath the dataset names.

Restricting evaluation to singly justified cases only changes the picture somewhat. In terms of both F1 and MCC, and frequently accuracy also, AF-CBA outperforms the ML models for all datasets except the GTD – although there the difference in performance is reduced as well. This is the case for both the conditional and the global approach. What this seems to suggest is that *if* AF-CBA finds a winning strategy for a single outcome only, that winning strategy is for the correct outcome more often than the ML models manage to predict.

**Table 5**
Classification performance when only counting singly justified cases, using conditional preference relation discovery. Test sizes are shown as: 'approximate CV mean (holdout)'.

| | Model | Accuracy | F1 | MCC |
|---|---|---|---|---|
| ADMISSIONS ∼69 (87) | Decision Tree | .90±0.01 (.95) | .64±0.09 (.82) | .30±0.19 (.64) |
| | Random Forest | **.94**±0.02 (.95) | .65±0.15 (.82) | .32±0.29 (.64) |
| | AF-CBA (C.) | .93±0.04 (.90) | **.75**±0.11 (.73) | **.53**±0.22 (.52) |
| CHURN ∼795 (980) | Decision Tree | .77±0.01 (.76) | .68±0.02 (.68) | .36±0.04 (.35) |
| | Random Forest | .82±0.00 (.83) | .73±0.01 (.74) | .48±0.01 (.49) |
| | AF-CBA (C.) | .82±0.01 (.82) | **.75**±0.01 (.75) | **.50**±0.01 (.49) |
| GTD ∼851 (1036) | Decision Tree | **.99**±0.00 (1.00) | .87±0.07 (.72) | .75±0.13 (.44) |
| | Random Forest | **.99**±0.00 (1.00) | **.89**±0.06 (.75) | **.80**±0.10 (.51) |
| | AF-CBA (C.) | .98±0.00 (.99) | .82±0.08 (.71) | .65±0.15 (.47) |
| COMPAS ∼157 (196) | Decision Tree | .66±0.03 (.66) | .63±0.02 (.62) | .25±0.03 (.24) |
| | Random Forest | .71±0.03 (.73) | .67±0.01 (.69) | .36±0.03 (.38) |
| | AF-CBA (C.) | **.75**±0.02 (.73) | **.71**±0.02 (.70) | **.43**±0.03 (.40) |

**Table 6**
Classification performance when only counting singly justified cases, using global preference relation discovery. Test sizes are shown as: 'approximate CV mean (holdout)'.

| | Model | Accuracy | F1 | MCC |
|---|---|---|---|---|
| ADMISSIONS ∼60 (72) | Decision Tree | .95±0.03 (.96) | .75±0.16 (.82) | .50±0.33 (.65) |
| | Random Forest | **.96**±0.02 (.97) | .68±0.16 (.89) | .37±0.33 (.79) |
| | AF-CBA (G.) | **.96**±0.04 (.93) | **.88**±0.12 (.79) | **.77**±0.24 (.60) |
| CHURN ∼213 (224) | Decision Tree | .91±0.01 (.93) | .86±0.02 (.89) | .71±0.04 (.78) |
| | Random Forest | .94±0.01 (.95) | .91±0.02 (.92) | .81±0.04 (.84) |
| | AF-CBA (G.) | **.95**±0.02 (.95) | **.92**±0.03 (.92) | **.85**±0.06 (.85) |
| GTD ∼840 (1008) | Decision Tree | .99±0.00 (1.00) | .83±0.05 (.93) | .66±0.10 (.87) |
| | Random Forest | **1**±0.00 (1.00) | **.88**±0.06 (.93) | **.78**±0.10 (.87) |
| | AF-CBA (G.) | .99±0.00 (1.00) | .80±0.05 (.90) | .62±0.09 (.82) |
| COMPAS ∼34 (42) | Decision Tree | .69±0.04 (.61) | .67±0.03 (.58) | .34±0.05 (.17) |
| | Random Forest | .68±0.05 (.61) | .65±0.08 (.59) | .32±0.15 (.22) |
| | AF-CBA (G.) | **.74**±0.08 (.72) | **.73**±0.09 (.71) | **.47**±0.18 (.46) |

Overall, AF-CBA performs reasonably well for some datasets – Admissions, GTD, and Churn to some extent – and more so for the singly justified focus cases. The conditional approach is slightly better, but also risks a small degree of overfitting due to its specificity to local conditions. These results suggest that AF-CBA can be a suitable classifier for datasets where its justification fractions are sufficiently high. Whether this implies that AF-CBA can be considered to be a viable alternative to traditional ML models depends on its relation to the performance-interpretability trade-off, as considered in the next section.

## 4. Interpretability

Interpretability is not an intrinsic property of a model but a subjective quality that varies with the expertise of the interpreter, the complexity of the model, and the domain in question. The question is therefore not whether AF-CBA is interpretable in absolute terms, but where it sits relative to other approaches and under what conditions its winning strategies are sufficiently interpretable.

Random forests are typically not considered interpretable. Decision trees and AF-CBA's dispute trees both could be, but this requires qualification. Decision trees produce threshold-based rules motivated by statistical correlations, potentially spanning many branches. AF-CBA's winning strategies instead take the form: 'the focus case receives outcome 1 because it is similar to precedent $c_i$, despite differences on dimensions $D_2$ which are outweighed by $D_3$'. This precedential structure arguably aligns more naturally with how domain experts reason – through attack patterns, diagnostic cases, or legal precedents.

Both approaches nonetheless share three fundamental challenges. First, interpretability degrades

with complexity regardless of architecture: a dispute tree citing hundreds of precedents is no more intelligible than a decision tree with 1,000 leaves. Second, both presuppose cognitive prerequisites that may not be universally accessible: decision trees require familiarity with feature-based partitioning, whilst AF-CBA requires some grasp of precedential constraint and preference relations. Which is more intuitive to domain experts remains empirically unclear. Third, the source of domain knowledge differs: decision trees derive feature importance from training data through metrics like Gini impurity reduction, whereas AF-CBA makes preference relations explicit and open to contestation. However, when those preferences are automatically discovered, a black-box component remains in the pipeline. The conditional approach [10] is arguably more interpretable for relying on a rule-based classifier, though also more complex.

These observations suggest that interpretability claims for both decision trees and AF-CBA are conditional rather than absolute. Both approaches can produce interpretable explanations for simple cases, but both degrade as complexity increases. The key question is not whether AF-CBA is interpretable per se, but whether it produces explanations of manageable complexity for a substantial proportion of cases. This can be addressed in part through structural analysis of tree complexity.

Whilst interpretability ultimately depends on user comprehension, structural complexity metrics provide useful proxies for cognitive accessibility. Decision tree interpretability degrades with depth, with depths exceeding perhaps five or six levels becoming difficult to interpret. The same can be said of tree breadth, where wider trees are more difficult to interpret. The structural complexity of a tree is due to some combination of depth and breadth, with high numbers for both producing a particularly complex structure. There is no obvious threshold beyond which interpretability degrades excessively. Some practitioners might use 'printability' on a page of size A4 as a practical criterion.

Depth refers to the path length from root to leaf. For decision trees, the maximum depth is the longest possible path, whilst mean depth is the weighted average across all leaves, where the number of samples reaching each leaf is used as a weight. This represents the expected number of decisions traversed during classification. Breadth is the total number of leaves. Since there is only one decision tree, mean and maximum breadth are necessarily identical. For AF-CBA, a separate dispute tree is constructed for each focus case, so mean and maximum breadth can differ.

**Table 7**
Structural complexity comparison between AF-CBA dispute trees and decision trees on the four datasets (holdout). Metrics include average and maximum depth and breadth.

| Model | Avg Depth | Max Depth | Avg Breadth | Max Breadth |
|---|---|---|---|---|
| ADMISSIONS | | | | |
| Decision Tree | 4.88 | 9 | 38 | 38 |
| AF-CBA (C.) | 1.22 | 3 | 1.02 | 3 |
| AF-CBA (G.) | 1.36 | 3 | 1.40 | 11 |
| CHURN | | | | |
| Decision Tree | 11.09 | 24 | 1091 | 1091 |
| AF-CBA (C.) | 1.41 | 3 | 2.44 | 121 |
| AF-CBA (G.) | 1.86 | 3 | 5.69 | 160 |
| GTD | | | | |
| Decision Tree | 6.58 | 15 | 118 | 118 |
| AF-CBA (C.) | 1.34 | 3 | 1.86 | 56 |
| AF-CBA (G.) | 1.54 | 3 | 3.88 | 56 |
| COMPAS | | | | |
| Decision Tree | 14.58 | 35 | 1993 | 1993 |
| AF-CBA (C.) | 2.72 | 3 | 71.87 | 1093 |
| AF-CBA (G.) | 2.79 | 3 | 79.47 | 975 |

Table 7 reports breadth and depth metrics for AF-CBA and decision trees trained using Scikit-learn's default implementation across the datasets used in Section 3. This table considers the winning strategy for each singly justified focus case as well as the selected (default) winning strategy for every doubly

justified focus case. AF-CBA's winning strategies are consistently shallower and narrower than decision trees across all datasets, with conditional preference relation discovery consistently producing the simplest justifications. Decision trees exhibit average depths ranging from 4.88 to 14.58, with maximum depths reaching 35 levels. AF-CBA's average depths, by contrast, remain between 1.22 and 2.79, with an observed maximum of 3. The breadth comparison is even starker: decision trees produce between 38 and 1993 leaf nodes, whilst AF-CBA's average breadth varies between 1.02 and 79.47 (with a maximum of 1093 for COMPAS). Overall, these results suggest that the winning strategies produced by AF-CBA are less complex – and so arguably more interpretable – than decisions trees. For those datasets where AF-CBA's performance is at least somewhat comparable to that of decisions trees, it can therefore be said to be a viable alternative as an interpretable classification approach. And in high-risk domains like law (enforcement), an increased interpretability in exchange for a slightly reduced performance – compared to, say, random forests – may be warranted.

## 5. Related Work

The work presented in this paper builds upon several ideas from argumentation theory, CBR and explainable AI. For example, Čyras et al. [35] employ abstract argumentation frameworks to support CBR in their 'AA-CBR' approach. They represent cases by sets of features – 'factors', but not in the sense of pro- and con-factors as in Prakken and Ratsma [2] – with binary outcomes. This involves a dialogue between a proponent who is in favour of a default outcome and an opponent who argues against it. An important property of AA-CBR is the concept of 'nearest cases', which are cases from the case base that are most similar to the new case. When multiple nearest cases exist with contradictory outcomes, AA-CBR uses grounded semantics to determine an outcome. This is determined by means of three essential criteria: outcome contradiction, specificity (more specific cases attack less specific ones) and conciseness (attacks should be as similar to the attacked case as possible). Additionally, AA-CBR can explain decisions through dispute trees. AA-CBR was a direct inspiration during the initial development of AF-CBA as a justification approach. In a continuation of this work, Fungwacharakorn et al. [36] present an approach to enable factor-based reasoning over non-factorised case bases. Their framework preserves the semantics of AA-CBR, but delegates case coverage and factor extraction to LLM agents. This means that the contents of precedential cases do not have to be unnecessarily disclosed (maintaining privacy) or preprocessed (improving flexibility). Their approach outperforms single-prompt performance as the number of features of a case increases.

Carstens and Toni [37] integrate supervised ML with structured argumentation. ML classifiers are trained normally, but their predictions are evaluated using an argumentation framework. Their arguments formalise prototypical knowledge, with the intention of identifying and correcting misclassifications. This can result in performance gains. The key distinction with the work in this paper is that they treat argumentation as a post-processing refinement in order to correct an ML classifier, whereas AF-CBA employs CBA to either justify classifications or act as the actual classification mechanism.

Nouwens et al. [38] consider the use of CBR for decision support in Dutch administrative law. They also emphasise the importance of transparency and explainability over accuracy maximisation. Their work focuses on medical assessments for the purpose of driving certification. They compare traditional CBR, AI and law models of precedential constraint and a hybrid approach. Testing was performed on a case base of over 30,000 instances. Their results suggest that traditional CBR achieves the best accuracy, and that both it and the hybrid model outperform the approaches that rely purely on precedential constraint. In contrast with AF-CBA, their hybrid approach relies on a (relatively complex) similarity measure to express differences between cases when unforced and does not use AF-CBA's notions of compensation and transformation.

Van Woerkom et al. [21] connect the theory of precedential constraint [15] to order theory and many-sorted logic. Their work builds upon that by Prakken and Ratsma [2] and extends this with concepts of landmarks (cases that set new precedents) and completeness. They characterise the kinds of datasets that typically adhere to precedential constraint and how consistency affects linear separability.

Furthermore, Odekerken et al. [39] use precedential constraint to construct a human-in-the-loop classification approach, focusing on notions of stability and relevance in light of incomplete fact situations. If the justification status of a focus case – that is, whether is forced – does not change regardless of its incomplete elements, than that focus case is considered stable. Relevance reflects which dimension-values would affect this justification status.

Finally, an augmented version of Horty's result model [15] that estimates precedents' 'authoritativeness' is used for classification by Morello et al. [40].

## 6. Discussion and Conclusion

The results suggest that AF-CBA can be used as an interpretable classifier, albeit an imperfect one. Any reduction in performance compared to traditional models has to be weighed against its interpretability. The fact that AF-CBA's singly justified outcomes perform relatively well suggests a possible complementary relation with a less interpretable but better performing ML classifier, where AF-CBA is the preferred model and the other classifier acts as a backup and tiebreaker – whose predictions can then be justified by AF-CBA regardless. This complementary setup would possibly be of interest when the difference in interpretability between AF-CBA and the ML model is particularly pronounced.

The majority-class heuristic currently plays a double role: as a default outcome to fall back on when neither outcome can be justified, and as a tiebreaker when both outcomes are justifiable. The fact that AF-CBA may produce double outcomes is not necessarily problematic (it is still interpretable), and the final decision could be resolved in other ways. One option is to modify the framework itself to include additional constraints or conflicts, thereby preventing double justifications from occurring. Another option is to consider the degree to which an outcome is satisfactorily justified. For example, the number of winning strategies for one outcome could be compared with the number for the other outcome, or case base inconsistency could be penalised. It is not evident how such a tiebreaker mechanism can be optimised for the classification task, making it an interesting possibility for future work.

More fundamental than any small improvement in performance is the question of interpretability. AF-CBA currently produces winning strategies represented as graphs, which are assumed to be more interpretable than most ML models. Interpretability could be enhanced by automatically generating textual explanations of the form: 'the focus case receives outcome 1 because it is similar to precedent $i$, despite its worse values for dimensions $x$, which are compensated by dimensions $y$...' Combining such textual descriptions with graphical depictions would arguably increase accessibility. At the same time, one might argue that a full understanding of these strategies presupposes knowledge of the underlying grounded game and preference discovery mechanisms. Yet this difficulty is not unique to AF-CBA: many classifiers described as 'interpretable' are so only to ML specialists rather than end users. A systematic user study is therefore needed to determine under which conditions approaches such as AF-CBA are considered sufficiently interpretable in practice.

In conclusion, this paper has demonstrated that AF-CBA can be used as a classifier under some circumstances and as a complementary one under other ones. Depending on the dataset, it may depend on a default heuristic – itself not part of AF-CBA. With some datasets, AF-CBA proves to be a notably poorer classifier than its alternatives. When restricted to singly justified cases, however, AF-CBA's performance is competitive with traditional, out-of-the-box ML models. Where its performance is only slightly worse, this can be viewed as a trade-off in exchange for interpretability. For datasets where AF-CBA's performance is significantly worse than traditional ML models, be it in terms of performance or of fallback behaviour, the results from this paper can instead be interpreted as supporting the initial understanding of AF-CBA as a post-hoc XAI approach.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their feedback and suggestions.

## Declaration on Generative AI

During the preparation of this work, the authors used Claude in order to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] E. S. Ortigossa, T. Gonçalves, L. G. Nonato, EXplainable Artificial Intelligence (XAI)—From theory to methods and applications, IEEE Access 12 (2024) 80799–80846.

[2] H. Prakken, R. Ratsma, A top-level model of case-based argumentation for explanation: Formalisation and experiments, Argument & Computation 13 (2022) 159–194.

[3] Z. Lipton, The mythos of model interpretability, Communications of the ACM 61 (2016) 96–100.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 93:1–93:42.

[5] A. Bibal, M. Lognoul, A. De Streel, B. Frénay, Legal requirements on explainability in machine learning, Artificial Intelligence and Law 29 (2021) 149–169.

[6] J. F. Horty, Rules and reasons in the theory of precedent, Legal Theory 17 (2011) 1–33.

[7] W. Van Woerkom, D. Grossi, H. Prakken, B. Verheij, Landmarks in case-based reasoning, in: S. Schlobach, M. Perez-Ortiz, M. Tielman (Eds.), HHAI2022: Augmenting Human Intellect., IOS Press, 2022, pp. 212–224.

[8] W. Van Woerkom, D. Grossi, H. Prakken, B. Verheij, Hierarchical precedential constraint, in: Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023), ACM, 2023, pp. 333–342.

[9] J. G. T. Peters, F. J. Bex, H. Prakken, Model- and data-agnostic justifications with A Fortiori Case-Based Argumentation, in: Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023), ACM, 2023, pp. 207–216.

[10] J. G. T. Peters, F. J. Bex, H. Prakken, Justifying black-box predictions with domain knowledge, in: J. Maranhão (Ed.), Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025), ACM, 2025, pp. 102–111.

[11] A. Assis, J. Dantas, E. Andrade, The performance-interpretability trade-off: A comparative study of machine learning models, Journal of Reliable Intelligent Environments 11 (2024). doi:10.1007/s40860-024-00240-0.

[12] U. Johansson, C. Sönströd, U. Norinder, H. Boström, Trade-off between accuracy and interpretability for predictive in silico modeling, Future Medicinal Chemistry 3 (2011) 647–663.

[13] A. A. Freitas, Automated machine learning for studying the trade-off between predictive accuracy and interpretability, in: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl (Eds.), International cross-domain conference for machine learning and knowledge extraction, Springer International Publishing, 2019, pp. 48–66.

[14] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[15] J. F. Horty, Reasoning with dimensions and magnitudes, Artificial Intelligence and Law 27 (2019) 309–345.

[16] V. Aleven, Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment, Artificial Intelligence 150 (2003) 183–237.

[17] K. Čyras, K. Satoh, F. Toni, Explanation for Case-Based Reasoning via Abstract Argumentation, in: P. Baroni, T. F. Gordon, T. Scheffler, M. Stede (Eds.), Computational Models of Argument, Proceedings of COMMA 2016, IOS Press, 2016, pp. 243–254.

[18] K. Čyras, D. Birch, Y. Guo, F. Toni, R. Dulay, S. Turvey, D. Greenberg, T. Hapuarachchi, Explanations by arbitrated argumentative dispute, Expert Systems with Applications 127 (2019) 141–156.

[19] H. Prakken, Dialectical proof theory for defeasible argumentation with defeasible priorities

(preliminary report), in: J.-J. C. Meyer, P.-Y. Schobbens (Eds.), Formal Models of Agents: ESPRIT Project ModelAge Final Workshop, Selected Papers, Springer Lecture Notes in AI 1760, Springer, 1999, pp. 202–215.
[20] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–357.
[21] W. Van Woerkom, D. Grossi, H. Prakken, B. Verheij, A Fortiori Case-Based Reasoning: From theory to data, Journal of Artificial Intelligence Research 81 (2024) 401–441.
[22] W. W. Cohen, Fast effective rule induction, in: A. Prieditis, S. Russell (Eds.), Proceedings of the twelfth international conference on machine learning, Morgan Kaufmann, 1995, pp. 115–123.
[23] I. Moscovitz, Wittgenstein, 2025. URL: https://github.com/imoscovitz/wittgenstein.
[24] S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions (2017). ArXiv:1705.07874.
[25] W. Visser, Implementation of Argument-Based Practical Reasoning, 2008.
[26] G. A. W. Vreeswijk, An algorithm to compute minimally grounded and admissible defence sets in argument systems, in: P. E. Dunne, T. Bench-Capon (Eds.), Computational Models of Argument, Proceedings of COMMA 2006, IOS Press, 2006, pp. 109–120.
[27] O. Kramer, Scikit-Learn, in: O. Kramer (Ed.), Machine Learning for Evolution Strategies, Springer International Publishing, 2016, pp. 45–53.
[28] M. Acharya, A. Armaan, A. Antony, A comparison of regression models for prediction of graduate admissions, in: 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), IEEE, 2019, pp. 1–5.
[29] IBM, Telco Customer Churn, 2018. URL: https://www.kaggle.com/blastchar/telco-customer-churn.
[30] National Consortium for the Study of Terrorism and Responses to Terrorism (START), Global Terrorism Database, 2018. URL: https://www.start.umd.edu/gtd/.
[31] G. LaFree, L. Dugan, Introducing the Global Terrorism Database, Terrorism and Political Violence 19 (2007) 181–204.
[32] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine Bias: There's software used across the country to predict future criminals and it's biased against blacks, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[33] D. Ofer, COMPAS Recidivism Racial Bias, 2017. URL: https://www.kaggle.com/datasets/danofer/compass.
[34] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020). doi:10.1186/s12864-019-6413-7.
[35] K. Čyras, K. Satoh, F. Toni, Abstract argumentation for case-based reasoning, in: Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning, AAAI Press, 2016, pp. 549–552.
[36] W. Fungwacharakorn, M. M. Zin, H.-T. Nguyen, Y. Kong, K. Satoh, Argumentative reasoning with language models on non-factorized case bases (2025). ArXiv:2512.12656.
[37] L. Carstens, F. Toni, Using argumentation to improve classification in natural language problems, ACM Transactions on Internet Technology 17 (2017) 30:1–30:23.
[38] J. Nouwens, A. Borg, H. Prakken, An application of case-based reasoning to decision-making in Dutch administrative law, in: J. Savelka, J. Harasta, T. Novotna, J. Misek (Eds.), Legal Knowledge and Information Systems, JURIX 2024: The Thirty-seventh Annual Conference, IOS Press, 2024, pp. 131–141.
[39] D. Odekerken, F. J. Bex, H. Prakken, Precedent-based reasoning with incomplete information for human-in-the-loop decision support, Artificial Intelligence and Law (2024). doi:10.1007/s10506-024-09421-x.
[40] Y. Morello, A. Ciabattoni, M. Gray, The result model under inconsistent knowledge: Theory and experiments, in: R. Markovich, L. Di Caro, A. Rapp, C. Schifanella (Eds.), Legal Knowledge and Information Systems, JURIX 2025: The Thirty-eighth Annual Conference, IOS Press, ????, pp. 133–144.