Toward Defeasible Machine Learning: When Learned Rule Sets Become Defeasible Theories

Ruvarashe Madzime[1], Thomas Meyer[2]

[1]University of the Western Cape and CAIR, Cape Town, South Africa
[2]University of Cape Town and CAIR, Cape Town, South Africa

Abstract

Some interpretable classifiers produce conjunctive if-then rules that look like defaults and exceptions, but looking defeasible is not the same as being defeasible. Interpretable models show a user which rule produced a prediction, yet they do not guarantee that the default-and-exception relationships humans naturally read between rules are logically justified. We identify a single structural condition on learned conjunctive rule sets, called strict global exception closure, under which every prediction the classifier makes is provably a defeasible entailment. We translate the rule set into a knowledge base in the KLM framework and prove that the classifier's Most Specific Wins prediction, which selects the label given by the most specific applicable rule, agrees exactly with defeasible entailment under Rational Closure, Lexicographic Closure, and System W simultaneously. We further show that strict global exception closure is the only condition that needs to be checked directly: one auxiliary property follows from it logically, and another is restored by prediction-preserving pruning. We analyse what breaks when the condition fails, revealing insights for both machine learning and nonmonotonic reasoning. Experiments on benchmark datasets confirm that the condition arises naturally from decision tree extraction, producing rule sets with nontrivial exception hierarchies at competitive accuracy.

Keywords

defeasible reasoning, nonmonotonic reasoning, rational closure, lexicographic closure, System W, interpretable machine learning, exception-closed rule sets

1. Introduction

Interpretable machine learning methods such as decision trees, RIPPER [1], and optimal rule lists [2] produce models that a person can read directly [3]. When the features are Boolean, the output is often a set of conjunctive rules, each mapping a conjunction of feature literals to a predicted class label. For instance, a rule learner trained on a medical dataset might produce:

$$a \Rightarrow 1, \qquad a \wedge b \Rightarrow 0, \qquad a \wedge b \wedge c \Rightarrow 1. \tag{1}$$

Here each rule says: if the features in the antecedent are present, predict the label on the right. A human reader does not treat these as three independent predictions. Instead, they are read as a structured pattern of defaults and exceptions: feature a normally supports label 1, except when b is also present, unless c is also present. This kind of reasoning, where a general rule is overridden by a more specific one, appears throughout real-world classification. A pathologist may reason that uniform cell size normally indicates a benign tumour, unless prominent nucleoli are also present. A toxicologist may reason that a certain cap shape normally indicates an edible mushroom, unless the odour profile is foul. In each case, the exception overrides the default rather than sitting alongside it as an independent claim. This pattern of default and exception is what defeasible reasoning formalises. A natural question then arises: is the resemblance between a learned rule set and a defeasible theory superficial, or can it be made formally exact? The distinction matters because it marks a gap between two levels of interpretability. Interpretable classifiers address what black-box models lack: they show a user which rule produced a prediction. But showing which rule fired is not the same as showing that the relationships between rules are logically coherent. When a doctor reads two rules and interprets the

24th International Workshop on Nonmonotonic Reasoning, July 17–19, 2026, Lisbon, Portugal
madzimeruvarashe@gmail.com (R. Madzime); tmeyer@cair.org.za (T. Meyer)
0009-0002-9307-5572 (R. Madzime); 0000-0003-2204-6969 (T. Meyer)

more specific one as an exception to the more general one, the doctor is imposing a defeasible reasoning narrative onto the classifier's output. That narrative may be correct, or it may be a projection onto rules that simply conflict without any principled resolution. Without a formal guarantee, there is no way to tell the difference.

This paper works within the KLM framework [4], which formalises defeasible conditionals of the form $\alpha \mathrel{|\!\sim} \beta$ ("typically, if α then β"). Given a knowledge base of such conditionals, several closure operators determine what the knowledge base defeasibly entails. We consider three: Rational Closure (RC) [5], Lexicographic Closure (LC) [6], and System W [7]. These three are known to disagree on general knowledge bases [8], so showing that all three agree with the classifier's predictions confirms that the correspondence is a consequence of the classifier's structure, not an artefact of any one formalism's design.

This paper identifies a single structural condition on learned conjunctive rule sets, called *strict global exception closure*, under which the resemblance between a classifier and a defeasible theory becomes a provable correspondence. Every pair of rules that could apply to the same instance and disagree on the label must be related by specificity: one must be a strict extension of the other, placing them in a default/exception relationship rather than leaving them as independent conflicting claims. Under this condition, we translate the rule set into a defeasible knowledge base and prove that the classifier's prediction on every fully observed instance agrees exactly with defeasible entailment under RC, LC, and System W.

Establishing this correspondence is not straightforward. Translating a rule set into a knowledge base does not by itself guarantee that the classifier's predictions match the defeasible entailments. All three formalisms rely, directly or indirectly, on a hierarchy that ranks formulas by how exceptional they are. For the correspondence to hold, the exception hierarchy that the classifier encodes in its rules must match the hierarchy that the formalisms construct from the translated knowledge base. The structural condition is precisely what forces this match, and showing that it does requires non-trivial proof. Two separate bridge results are needed: one connecting the classifier's exception depth to the ranking used by RC and LC, and another connecting the classifier's depth partition to the tolerance partition used by System W.

We further show that strict global exception closure is the only condition that needs to be imposed directly. Two additional conditions used in the initial statement of the theorems, *sanity* and *no total override*, turn out to be not needed: the first follows logically from strict global exception closure, and the second is restored by a prediction-preserving pruning algorithm. This pruning result shows that a rule set can be broken as a KLM theory while working correctly as a classifier, and that removing the offending rules fixes the theory without changing any prediction. The paper also analyses what breaks when the condition fails, showing that dead rules can collapse the defeasible theory while the classifier works correctly, and conversely that a classifier can predict correctly from a semantically inconsistent knowledge base.

A predecessor of this work [9] introduced a classifier whose rules were structured as Horn defaults paired with exceptions, learned via Answer Set Programming. That work showed empirically that default-and-exception rule structures can match interpretable baselines in accuracy, but it provided no formal semantic guarantees. The present paper fills that gap. To our knowledge, no prior work has proved that a learned classifier's predictions are exactly defeasible entailments under any KLM-style formalism; the result here holds under three simultaneously. We validate the results in practice by learning Exception-Closed Conjunctive Rule Sets (ECCRS) from decision trees on three benchmark datasets [10], producing compact rule sets with nontrivial exception hierarchies at competitive accuracy.

A natural question is why we prove the correspondence rather than simply learning in a defeasible target language from the start. The answer is that whether a learned rule set is a defeasible theory is a question about what the rules mean, not about how they were produced. Standard algorithms such as decision tree induction already produce high-quality conjunctive rule sets. A rule set that happens to satisfy strict global exception closure carries a formal defeasible guarantee regardless of the algorithm that produced it, and our result makes this precise without requiring any change to the learning pipeline.

The main contributions of this paper are as follows: (1) we identify strict global exception closure as a single structural condition under which a learned conjunctive rule classifier is a defeasible theory in the KLM sense, with predictions matching defeasible entailment under RC, LC, and System W simultaneously; (2) we establish two independent bridge results connecting the classifier's exception hierarchy to the ranking constructions used by the formalisms, one for RC and LC and another for System W; (3) we show that the two auxiliary conditions used in the theorems are not independent, since one follows logically from closure and the other is restored by prediction-preserving pruning; and (4) we validate the approach experimentally on benchmark datasets, showing that the condition arises naturally from decision tree extraction. The paper proceeds as follows. Section 2 gives the necessary background from both machine learning and nonmonotonic reasoning, with a running example woven through each definition. Section 3 introduces a running example. Section 4 presents the translation and establishes the depth-equals-BaseRank result. Section 5 proves that the classifier's predictions are defeasible entailments. Section 6 looks at what happens when the structural condition fails and shows that one assumption is enough. Section 7 describes how ECCRS are learned in practice. Section 8 reports experiments. Section 9 discusses related work and concludes.

2. Background

2.1. Rule-based classification

In a typical classification setting, a learner is given a dataset of instances, each described by a vector of Boolean feature values $(a_1, \ldots, a_n) \in \{0,1\}^n$ and labelled with a class $\ell \in \{0,1\}$. The goal is to learn a model that, given a new feature vector, predicts the correct label. We focus on models whose output is a set of conjunctive rules, where each rule maps a conjunction of feature conditions to a predicted label.

Let $A = \{a_1, \ldots, a_n\}$ be a finite set of propositional atoms representing features. A body B is a consistent conjunction of literals over A. It describes a partial pattern of feature values: for example, $a_1 \wedge \neg a_3$ says that feature a_1 is true and feature a_3 is false, without saying anything about the other features. A complete feature pattern F is a complete assignment over A, identified with the conjunction of all its literals, so that every feature is either asserted or negated. We write $B \subset C$ when C strictly extends B, meaning that C contains all the literals of B plus at least one more. Equivalently, $C \models B$ and $B \not\models C$. Two bodies are compatible if $B \wedge C$ is satisfiable, that is, they do not impose contradictory requirements on any feature.

Definition 2.1 (Rules and applicability)**.** A rule is a pair (B, ℓ) where B is a body and $\ell \in \{0,1\}$ is the predicted label. A rule (B, ℓ) is applicable to a complete feature pattern F if $F \models B$, meaning that F satisfies every literal in B.

When multiple rules apply to the same instance, a conflict resolution strategy is needed. The natural strategy for conjunctive rule sets is to prefer more specific rules over more general ones, since a more specific rule describes a more refined pattern in the data and therefore carries more targeted information.

Definition 2.2 (Most Specific Wins)**.** For a rule set $\mathcal{R}$ and a complete feature pattern F, let $\mathrm{A}(F) := \{(B, \ell) \in \mathcal{R} : F \models B\}$ be the set of applicable rules, and let $\mathrm{M}(F)$ be the inclusion-maximal applicable rules, that is, those rules in $\mathrm{A}(F)$ whose body is not strictly extended by the body of any other applicable rule. The Most Specific Wins (MSW) decision is:

$$\mathrm{MSW}_{\mathcal{R}}(F) := \begin{cases} \ell & \text{if } \mathrm{M}(F) \neq \emptyset \text{ and all rules in } \mathrm{M}(F) \text{ share label } \ell, \\ ? & \text{if } \mathrm{M}(F) = \emptyset. \end{cases}$$

Here ? denotes abstention: the classifier makes no prediction when no applicable rule exists.

The key structural condition that makes the bridge to defeasible reasoning possible is the following.

Definition 2.3 (Strict global exception closure)**.** A rule set $\mathcal{R}$ satisfies strict global exception closure if for all rules (B, ℓ) and $(C, 1-\ell)$ in $\mathcal{R}$, whenever B and C are compatible, either $B \subset C$ or $C \subset B$.

In plain terms: any two rules that could both apply to some instance and disagree on the label must be related by specificity. One must be a strict extension of the other, placing them in a default/exception relationship rather than leaving them as independent conflicting claims. We call a rule set satisfying this condition an Exception-Closed Conjunctive Rule Set (ECCRS).

To see why this matters, consider the rule set $\{(a, 1), (b, 0)\}$. Here both rules could apply to a pattern $a \wedge b$, they disagree on the label, and neither body extends the other: this rule set fails closure. The condition rules out exactly this kind of unstructured conflict. Any two compatible opposite-label rules must lie in a specificity chain, so that the more specific one always resolves the conflict.

As a concrete example, suppose a medical classifier learns two rules: "fever predicts infection" and "rash predicts allergy." A patient who presents with both fever and rash triggers both rules, which disagree, yet neither rule refines the other. The classifier must break the tie by some ad hoc mechanism, and there is no principled default/exception reading. Strict global exception closure prevents this situation. If the two rules had instead been "fever predicts infection" and "fever $\wedge$ rash predicts allergy," the second rule strictly extends the first, and the conflict is resolved by specificity: fever normally predicts infection, except when a rash is also present.

Two further properties are used in the initial statement of the theorems.

Definition 2.4 (Sanity)**.** A rule set $\mathcal{R}$ is sane if no body appears with both labels: there is no B such that both $(B, 0) \in \mathcal{R}$ and $(B, 1) \in \mathcal{R}$.

Definition 2.5 (No total override)**.** A rule set $\mathcal{R}$ satisfies no total override if every rule $(B, \ell) \in \mathcal{R}$ has at least one witness completion: a complete feature pattern F with $F \models B$ such that no stricter opposite-label rule applies to F. In other words, every rule must have at least one situation where it actually determines the prediction.

Section 6 shows that both conditions are not needed under strict global exception closure.

Lemma 2.6. *If $\mathcal{R}$ satisfies strict global exception closure, then for every F with $\mathrm{M}(F) \neq \emptyset$, all rules in $\mathrm{M}(F)$ share the same label.*

Proof. Suppose $(B, 0), (C, 1) \in \mathrm{M}(F)$. Then $F \models B \wedge C$, so B and C are compatible. Closure forces $B \subset C$ or $C \subset B$, contradicting maximality of both.

2.2. Defeasible reasoning in the KLM framework

Nonmonotonic reasoning formalises the intuition that general rules can have exceptions. Classical logic cannot capture this: if $\alpha \to \beta$ holds classically, then $\alpha \wedge \gamma \to \beta$ must also hold, so there is no way for additional information γ to defeat the conclusion β. The KLM framework [4], named after Kraus, Lehmann, and Magidor, addresses this by introducing a weaker kind of conditional. A defeasible conditional $\alpha \mathrel{|\!\sim} \beta$ is read as "typically, if α then β." Unlike a classical implication, a defeasible conditional can be overridden by more specific information: one can consistently hold both $\alpha \mathrel{|\!\sim} \beta$ and $\alpha \wedge \gamma \mathrel{|\!\sim} \neg\beta$, meaning that α typically implies β except in the presence of γ. A defeasible knowledge base $\mathcal{K}$ is a finite set of such conditionals. The central question is what $\mathcal{K}$ defeasibly entails. Several closure operators have been developed to answer this question. All three we consider rely, directly or indirectly, on a way of ranking formulas by how typical or exceptional they are. The BaseRank construction [5, 8] provides this ranking. A formula with a low BaseRank is typical and uncontroversial; a high BaseRank means the formula lives in an unusual corner of the space.

Given $\mathcal{K}$, let $E_0 := \overrightarrow{\mathcal{K}}$ be the materialisation, obtained by replacing each defeasible conditional $\alpha \mathrel{|\!\sim} \beta$ with the classical implication $\alpha \to \beta$. Then define $E_{i+1} := \{\alpha \to \beta \in E_i : E_i \models \neg\alpha\}$: at each stage, keep only those implications whose antecedents are exceptional, meaning they are inconsistent with the current set. The BaseRank of a formula α is $\mathrm{br}_{\mathcal{K}}(\alpha) := \min\{i : E_i \not\models \neg\alpha\}$, or ∞ if no such i exists. Rational Closure (RC) [5] uses BaseRank to decide entailment. It accepts $\alpha \mathrel{|\!\sim} \beta$ if and only if $\mathrm{br}(\alpha) < \mathrm{br}(\alpha \wedge \neg\beta)$, that is, the antecedent becomes strictly more exceptional when the consequent is denied. RC is viewed as the canonical form of defeasible entailment satisfying the

KLM rationality postulates. It does, however, suffer from a well-known limitation called the drowning problem [6]. If an antecedent is made exceptional by one unusual property, RC may withdraw all of its other defaults, even those completely unrelated to the unusual property. For example, if penguins are exceptional birds because they cannot fly, RC may also stop concluding that penguins have beaks, even though beak-having has nothing to do with flight. Lexicographic Closure (LC) [6, 8] addresses the drowning problem [6] by comparing possible worlds more carefully. Rather than a single BaseRank comparison, LC counts how many defaults each world violates at each rank level. For each world u and rank i, let $N_i(u)$ be the number of defaults of BaseRank i that u violates. The seriousness tuple $N(u) := \langle N_p(u), \ldots, N_0(u) \rangle$ is compared from the highest rank down. A world with fewer violations at the highest rank where the two worlds differ is preferred. LC entails $\alpha \mid\!\sim \beta$ when every most-preferred world satisfying α also satisfies β.

System W [7] takes a structurally different approach that does not use BaseRank at all. Instead, it splits the conditionals into layers using a notion of tolerance: a conditional is tolerated by a set S when there exists some world that makes the conditional true without making anything in S false. Intuitively, System W asks which world is more normal by looking at which defaults it violates, layer by layer from the most exceptional down. The world that violates fewer defaults at the most exceptional layer where the two worlds differ is preferred. Formally, the ordered tolerance partition $\mathrm{OP}(\mathcal{K}) = (D_0, \ldots, D_p)$ is built by repeatedly extracting the tolerated conditionals: D_0 contains the conditionals tolerated by the full knowledge base, D_1 contains those tolerated by what remains after removing D_0, and so on. For each world u, let $\xi_i(u)$ denote the set of conditionals in layer D_i that u falsifies. System W prefers world u to world v when, at the highest layer where the two worlds differ in what they falsify, u falsifies a strict subset of what v falsifies. Because System W does not build on BaseRank, connecting it to the classifier requires a separate bridge argument from those used for RC and LC.

3. Running Example

Throughout this paper we use the three-rule chain from Equation (1), with atoms $A = \{a, b, c\}$ and rule set $\mathcal{R}_0 = \{(a, 1),\ (a \wedge b, 0),\ (a \wedge b \wedge c, 1)\}$. This rule set satisfies strict global exception closure: the only compatible opposite-label pairs are $(a, 1)$ with $(a \wedge b, 0)$, and $(a \wedge b, 0)$ with $(a \wedge b \wedge c, 1)$, and in each case one body strictly extends the other. The exception structure is captured by the exception depth of each rule, which we now define informally (the formal definition appears in Definition 4.2). Let $\mathrm{Opp}(B)$ be the set of opposite-label predecessors of B: rule bodies with the opposite label that are strictly less specific than B. Rule $(a, 1)$ has $\mathrm{Opp}(a) = \emptyset$, so $\mathrm{depth}(a) = 0$: it is a plain default with no opposite-label predecessor. Rule $(a \wedge b, 0)$ has $\mathrm{Opp}(a \wedge b) = \{a\}$, since $(a, 1)$ has the opposite label and $a \subset a \wedge b$, giving $\mathrm{depth}(a \wedge b) = 1$: it is an exception to the default. Rule $(a \wedge b \wedge c, 1)$ has $\mathrm{Opp}(a \wedge b \wedge c) = \{a \wedge b\}$, giving $\mathrm{depth}(a \wedge b \wedge c) = 2$: it is a counter-exception that overrides the exception.

Now consider three complete patterns over A. For $F_1 = a \wedge \neg b \wedge c$, only $(a, 1)$ applies, so $\mathrm{MSW}(F_1) = 1$. For $F_2 = a \wedge b \wedge \neg c$, both $(a, 1)$ and $(a \wedge b, 0)$ apply, but the latter is more specific, so $\mathrm{MSW}(F_2) = 0$. For $F_3 = a \wedge b \wedge c$, all three rules apply, and the most specific is $(a \wedge b \wedge c, 1)$, giving $\mathrm{MSW}(F_3) = 1$: the counter-exception overrides the exception.

Now we translate to a defeasible knowledge base. Using a fresh atom l to represent label 1, the translation gives $\mathcal{T}(\mathcal{R}_0) = \{a \mid\!\sim l,\ a \wedge b \mid\!\sim \neg l,\ a \wedge b \wedge c \mid\!\sim l\}$, with materialisation $E_0 = \{a \to l,\ a \wedge b \to \neg l,\ a \wedge b \wedge c \to l\}$. The BaseRank sequence (as defined in Section 2) peels off one depth layer at a time. At stage E_0: the antecedent a is consistent with E_0 (the valuation $a, \neg b, \neg c, l$ satisfies all three implications), so $\mathrm{br}(a) = 0$. But $a \wedge b$ triggers a label contradiction: both $a \to l$ and $a \wedge b \to \neg l$ fire, forcing l and $\neg l$, so $a \wedge b$ is exceptional and survives into E_1. Similarly, $a \wedge b \wedge c$ also survives. This gives $E_1 = \{a \wedge b \to \neg l,\ a \wedge b \wedge c \to l\}$. At stage E_1: the antecedent $a \wedge b$ is now consistent (with $\neg c$, only $a \wedge b \to \neg l$ fires), so $\mathrm{br}(a \wedge b) = 1$. But $a \wedge b \wedge c$ again triggers a contradiction within E_1, giving $E_2 = \{a \wedge b \wedge c \to l\}$. At stage E_2: $a \wedge b \wedge c$ is consistent, so $\mathrm{br}(a \wedge b \wedge c) = 2$ and $E_3 = \emptyset$.

The BaseRanks match the exception depths exactly: $\mathrm{br}(a) = \mathrm{depth}(a) = 0$, $\mathrm{br}(a \wedge b) = \mathrm{depth}(a \wedge$

$b) = 1$, and $\mathrm{br}(a \wedge b \wedge c) = \mathrm{depth}(a \wedge b \wedge c) = 2$. This match between exception depth and BaseRank is not a coincidence. It is the central structural result of Section 4. Once that match is established, the MSW predictions on F_1, F_2, F_3 will be shown in Section 5 to be not just reasonable classifier outputs but provable defeasible entailments under RC, LC, and System W at the same time. The LC and System W calculations appear in Sections 5.2 and 5.3.

4. The Bridge: Translation and Depth = BaseRank

The first step toward showing that the classifier is a defeasible theory is to establish that its internal hierarchy matches the hierarchy used by the defeasible theory.

Definition 4.1 (Translation)**.** Given a rule set $\mathcal{R}$, fix a fresh atom l for the positive class and define the defeasible knowledge base $\mathcal{T}(\mathcal{R}) := \{B \mathrel{|\!\sim} l : (B, 1) \in \mathcal{R}\} \cup \{B \mathrel{|\!\sim} \neg l : (B, 0) \in \mathcal{R}\}$, with materialisation $\overrightarrow{\mathcal{T}(\mathcal{R})} := \{B \to l : (B, 1) \in \mathcal{R}\} \cup \{B \to \neg l : (B, 0) \in \mathcal{R}\}$.

A key feature of this translation is that feature atoms appear only in antecedents and the label atom l appears only in consequents. As a result, the only way a body B can become inconsistent with the materialised theory is through a label contradiction: assuming B forces both l and $\neg l$ at the same time, because B triggers two implications with opposite label consequents (Lemma A.1 in the appendix).

Definition 4.2 (Exception depth)**.** For $(B, \ell) \in \mathcal{R}$, let $\mathrm{Opp}(B) := \{C : (C, 1-\ell) \in \mathcal{R},\ C \subset B\}$ be the set of opposite-label rule bodies strictly less specific than B. Define $\mathrm{depth}(B) := 0$ if $\mathrm{Opp}(B) = \emptyset$, and $\mathrm{depth}(B) := 1 + \max\{\mathrm{depth}(C) : C \in \mathrm{Opp}(B)\}$ otherwise. For a complete feature pattern F, set $\mathrm{depth}(F) := \max\{\mathrm{depth}(B) : (B, \ell) \in \mathcal{R},\ F \models B\}$ if $\mathrm{A}(F) \neq \emptyset$, and 0 otherwise.

Depth 0 means the rule is a plain default with no opposite-label predecessor. Depth 1 means it overrides some default. Depth 2 means it overrides an exception, and so on. Three supporting lemmas capture how depth behaves under strict global exception closure. Their proofs appear in the appendix.

Lemma 4.3 (Opposite predecessors force depth increase)**.** *If $\mathcal{R}$ satisfies strict global exception closure and $(B, \ell), (C, 1-\ell) \in \mathcal{R}$ with $C \subset B$, then* $\mathrm{depth}(B) \geq 1 + \mathrm{depth}(C)$.

Lemma 4.4 (Same-label extension preserves depth)**.** *If $\mathcal{R}$ satisfies strict global exception closure and $(B, \ell), (D, \ell) \in \mathcal{R}$ with $B \subset D$, then* $\mathrm{depth}(D) \geq \mathrm{depth}(B)$.

Lemma 4.5 (Depth-maximal rule in $\mathrm{M}(F)$)**.** *If $\mathcal{R}$ satisfies strict global exception closure and $\mathrm{A}(F) \neq \emptyset$, then there exists $(B, \ell) \in \mathrm{M}(F)$ with* $\mathrm{depth}(B) = \mathrm{depth}(F)$.

The central result of this section establishes that the classifier's exception hierarchy and the BaseRank hierarchy [5] used by the defeasible formalisms are the same. This match is what will force the predictions to agree in Section 5.

Proposition 4.6 (Depth = BaseRank)**.** *Assume $\mathcal{R}$ satisfies strict global exception closure, sanity, and no total override, and let $\mathcal{K} := \mathcal{T}(\mathcal{R})$. For every rule body B and every complete feature pattern F,* $\mathrm{depth}(B) = \mathrm{br}_{\mathcal{K}}(B)$ *and* $\mathrm{depth}(F) = \mathrm{br}_{\mathcal{K}}(F)$.

Proof sketch. The argument rests on a precise description of the BaseRank sequence (Lemma A.2 in the appendix): at every stage i, E_i contains exactly the implications whose exception depth is at least i. The BaseRank construction peels off one depth layer per stage.

If $\mathrm{depth}(B) \geq i{+}1$, some opposite-label predecessor C with $\mathrm{depth}(C) \geq i$ survives in E_i. Since $C \subset B$, assuming B forces both l and $\neg l$, so B stays exceptional. If $\mathrm{depth}(B) = i$, the witness completion provides a valuation satisfying B and all of E_i without contradiction: strict closure ensures that any opposite-label implication triggered would either be a predecessor of B (needing depth $\geq i$, contradicting the inductive description) or a strict extension (blocked by the witness). So B drops out at stage $i{+}1$. Together these give $\mathrm{br}_{\mathcal{K}}(B) = \mathrm{depth}(B)$. The argument for F uses Lemma 4.5.

The classifier and the defeasible theory have therefore been built on the same structure: the exception hierarchy that the rule set encodes is the BaseRank hierarchy that all three KLM formalisms use to resolve conflicts. The next section shows that this shared structure forces the predictions to agree.

5. The Classifier as a Defeasible Theory

We now prove the main result: under strict global exception closure, every prediction the classifier makes is a defeasible entailment of the translated knowledge base. The intuition is as follows. Under strict global exception closure, all compatible cross-label conflicts lie in specificity chains. For any query F, the decisive rule sits at a single depth k, and above k no applicable rule body exists. At depth k, one of the two possible label worlds violates the decisive default and the other does not. This separation is sufficient for all three formalisms, each of which resolves it the same way despite using different mechanisms. Two separate bridge results underlie the correspondence: the depth = BaseRank identity (Proposition 4.6) grounds the argument for RC and LC, while a separate depth-partition-equals-tolerance-partition result (Lemma A.3) grounds the argument for System W. To make the proof sketches uniform, we fix notation used throughout this section. For a complete feature pattern F with $\mathrm{A}(F) \neq \emptyset$, let β denote the label literal corresponding to $\mathrm{MSW}(F)$, let $k := \mathrm{depth}(F)$, and define two candidate worlds: $u^+ := F \wedge \beta$, $u^- := F \wedge \neg\beta$. The world u^+ assigns F the MSW-predicted label; u^- assigns the opposite. All three proof sketches reduce to showing that the relevant formalism prefers u^+ over u^-. By Lemma 4.5, there exists a rule $(B, \ell^*) \in \mathrm{M}(F)$ with $\mathrm{depth}(B) = k$ and label literal β. This rule serves as the decisive default in every argument below.

5.1. Rational Closure

Theorem 5.1 (RC correspondence)**.** *Assume $\mathcal{R}$ satisfies strict global exception closure, sanity, and no total override, and let $\mathcal{K} := \mathcal{T}(\mathcal{R})$. For every complete feature pattern F: (1) if $\mathrm{A}(F) = \emptyset$, neither $F \mathrel{|\!\sim} l$ nor $F \mathrel{|\!\sim} \neg l$ belongs to the rational closure of $\mathcal{K}$; (2) if $\mathrm{A}(F) \neq \emptyset$ and $\mathrm{MSW}_{\mathcal{R}}(F) = 1$, then $\mathcal{K} \mid\!\approx_{\mathrm{RC}} F \mathrel{|\!\sim} l$ and $\mathcal{K} \not\mid\!\approx_{\mathrm{RC}} F \mathrel{|\!\sim} \neg l$; (3) symmetrically for $\mathrm{MSW}_{\mathcal{R}}(F) = 0$.*

Proof sketch. If no rule applies, the materialisation puts no label constraint on F, so the BaseRanks of F, $F \wedge l$, and $F \wedge \neg l$ are all 0 and the strict inequality test fails for both labels. If some rule applies, Proposition 4.6 gives $\mathrm{br}(F) = k$. The decisive rule (B, ℓ^*) contributes $B \to \beta$ to E_k by Lemma A.2. Since $F \models B$, classical entailment gives $E_k \cup \{F\} \models \beta$, and therefore $E_k \models \neg(F \wedge \neg\beta)$, giving $\mathrm{br}(u^-) \geq k + 1$. At stage $k{+}1$, no applicable antecedent of depth $\geq k{+}1$ is satisfied by F (by maximality of k), so $\mathrm{br}(u^-) = k + 1$, and the RC test $k < k + 1$ succeeds. For the opposite label, $E_k \cup \{F \wedge \beta\}$ is satisfiable, so $\mathrm{br}(u^+) = k$ and the test $k < k$ fails.

For the running example, consider $F_2 = a \wedge b \wedge \neg c$ with $\mathrm{MSW}(F_2) = 0$: $k = 1$, the set $E_1 \cup \{F_2\}$ entails $\neg l$ since $F_2 \models a \wedge b$ and $a \wedge b \to \neg l$ is in E_1, so $\mathrm{br}(u^-) = \mathrm{br}(F_2 \wedge l) = 2$, and $1 < 2$ confirms that RC entails $F_2 \mathrel{|\!\sim} \neg l$.

5.2. Lexicographic Closure

Theorem 5.2 (LC correspondence)**.** *Assume $\mathcal{R}$ satisfies strict global exception closure, sanity, and no total override, and let $\mathcal{K} := \mathcal{T}(\mathcal{R})$. For every complete F, the lexicographic closure of $\mathcal{K}$ entails the same label as $\mathrm{MSW}_{\mathcal{R}}(F)$.*

Proof sketch. If no rule applies, all violation counts are zero, both label worlds tie, and neither label is entailed. Otherwise, at rank k the world u^- violates $B \mathrel{|\!\sim} \beta$ (it satisfies B but not β), giving $N_k(u^-) \geq 1$. The world u^+ violates no rank-k default: any such violation would require a compatible opposite-label rule of depth k applicable to F, but closure and the maximality of B in $\mathrm{M}(F)$ rule this out via Lemma 4.3, giving $N_k(u^+) = 0$. Above rank k, no body of depth greater than k is satisfied by F, so no rank-$(i > k)$ default fires under F at all, giving $N_i(u^+) = N_i(u^-) = 0$ for all $i > k$. The seriousness tuple of u^+ is therefore lexicographically smaller when read from the highest rank down, and LC entails β.

For $F_3 = a \wedge b \wedge c$ with $\mathrm{MSW}(F_3) = 1$: the seriousness tuples are $N(u^+) = \langle 0, 1, 0\rangle$ and $N(u^-) = \langle 1, 0, 1\rangle$. At rank 2, $0 < 1$ settles the comparison. LC entails $F_3 \mathrel{|\!\sim} l$.

5.3. System W

Theorem 5.3 (System W correspondence)**.** *Assume $\mathcal{R}$ satisfies strict global exception closure, sanity, and no total override, and let $\mathcal{K} := \mathcal{T}(\mathcal{R})$. For every complete F, System W entails the same label as* $\mathrm{MSW}_{\mathcal{R}}(F)$.

Proof sketch. This argument requires a bridge result specific to System W and independent of BaseRank. The depth partition $D_i := \{B \mid\sim \gamma \in \mathcal{K} : \mathrm{depth}(B) = i\}$ matches the ordered tolerance partition of $\mathcal{K}$ (Lemma A.3 in the appendix). The proof of this identity shows two things: first, that each depth-i conditional is tolerated within the residual knowledge base $\mathcal{K}_{\geq i}$, using the witness completion to construct a world that verifies the conditional without falsifying anything in $\mathcal{K}_{\geq i}$; and second, that no conditional of depth $d > i$ can be tolerated within $\mathcal{K}_{\geq i}$, because any world verifying it must also satisfy a shallower opposite-label predecessor and thereby falsify something in $\mathcal{K}_{\geq i}$. This is genuinely different from the BaseRank argument, since it works with tolerance rather than exceptionality.

Once the partition identity is established, the comparison at layer k follows a pattern parallel to the LC case, but the mechanism differs. LC compares violation counts numerically ($N_k(u^+) = 0 < 1 \leq N_k(u^-)$). System W instead compares the falsified sets ξ_i by strict subset inclusion: u^+ is preferred to u^- at the highest layer where their falsified sets differ. Specifically, $\xi_k(u^+) = \emptyset$ (by the same closure reasoning: any falsified conditional in D_k would require a compatible opposite-label rule of depth k, contradicting maximality of B via Lemma 4.3), while $\xi_k(u^-) \ni B \mid\sim \beta$. For all layers above k, $\xi_i(u^+) = \xi_i(u^-) = \emptyset$ since no body of depth greater than k is satisfied by F. Since $\emptyset \subsetneq \xi_k(u^-)$, we have $u^+ \prec_W u^-$, and System W entails β.

For $\mathcal{R}_0$, the tolerance partition is $D_0 = \{a \mid\sim l\}$, $D_1 = \{a \wedge b \mid\sim \neg l\}$, $D_2 = \{a \wedge b \wedge c \mid\sim l\}$. For $F_3 = a \wedge b \wedge c$: at layer 2, $\xi_2(u^+) = \emptyset$ and $\xi_2(u^-) = \{a \wedge b \wedge c \mid\sim l\}$. Since $\emptyset \subsetneq \{a \wedge b \wedge c \mid\sim l\}$, System W entails $F_3 \mid\sim l$.

5.4. Summary

Corollary 5.4 (MSW = defeasible entailment)**.** *Assume $\mathcal{R}$ satisfies strict global exception closure, sanity, and no total override, and let $\mathcal{K} := \mathcal{T}(\mathcal{R})$. For every complete feature pattern F with $\mathrm{A}(F) \neq \emptyset$:*

$$\mathrm{MSW}_{\mathcal{R}}(F) = 1 \iff \mathcal{K} \mid\approx_{\mathrm{RC}} F \mid\sim l \iff \mathcal{K} \mid\approx_{\mathrm{LC}} F \mid\sim l \iff \mathcal{K} \mid\approx_{\mathrm{W}} F \mid\sim l,$$

and symmetrically for label 0*. If $\mathrm{A}(F) = \emptyset$, all three leave the label unresolved, matching abstention.*

The predictions of a learned conjunctive rule classifier satisfying strict global exception closure are formally identical to the defeasible entailments of the translated knowledge base, and this holds under RC, LC, and System W simultaneously. Because these three formalisms are known to disagree in general, the fact that all three agree here confirms that the correspondence is a consequence of the classifier's structure rather than an accident of any one formalism's design.

6. When the Condition Fails, and Why One Is Enough

The correspondence requires three assumptions: strict global exception closure (A1), sanity (A2), and no total override (A3). We examine what breaks when each fails, and then show that A1 alone suffices.

Sanity (A2) cannot fail under A1, and so needs no separate analysis. Having both $(B, 0)$ and $(B, 1)$ in $\mathcal{R}$ would require $B \subset B$ by closure, but strict inclusion is irreflexive.

Lemma 6.1. *Strict global exception closure implies sanity.*

Dropping A3 (no total override): consider $\mathcal{R}_1 = \{(a, 1),\ (a \wedge b, 0),\ (a \wedge \neg b, 0)\}$. The rule $(a, 1)$ has no witness: every pattern satisfying a satisfies either $a \wedge b$ or $a \wedge \neg b$, both with the opposite label. The classifier works correctly, since MSW always returns label 0 for patterns satisfying a. On the defeasible reasoning side, however, the tautology $b \vee \neg b$ forces a label contradiction under a at every stage, giving

Algorithm 1 Pruning to the witness fixed point

Require: Rule set $\mathcal{R}$ satisfying strict global exception closure
Ensure: Pruning fixed point $\mathcal{R}'$

$S \leftarrow \mathcal{R}$
repeat
  $\mathcal{D} \leftarrow \{(B, \ell) \in S : (B, \ell)$ is totally overridden in $S\}$
  $S \leftarrow S \setminus \mathcal{D}$
until $\mathcal{D} = \emptyset$
return S

$\mathrm{br}(a) = \infty$, which then cascades to all bodies and breaks the entire BaseRank construction. A single dead rule, one that never determines any classifier prediction, can collapse an entire defeasible theory while the classifier based on the same rules works perfectly.

Dropping A1 (strict global exception closure): consider $\mathcal{R}_2 = \{(a, 1),\ (b, 0)\}$ and $F = a \wedge b$. Both rules apply, neither is more specific, and they disagree. MSW returns ?. RC also entails neither label, but the agreement is coincidental. Without closure, no general guarantee holds.

Lemma 6.1 already removes the need for A2. The witness condition (A3) is recovered by pruning.

Definition 6.2 (Totally overridden)**.** A rule (B, ℓ) is totally overridden in a rule set S if $B \models \bigvee\{C : (C, 1-\ell) \in S,\ B \subset C\}$. That is, every completion of B satisfies some stricter opposite-label rule in S. Whether a rule is totally overridden can be determined by inspecting the rule bodies alone and can be checked efficiently.

Theorem 6.3 (Pruning fixed point)**.** *Let $\mathcal{R}$ satisfy strict global exception closure and let $\mathcal{R}'$ be the output of Algorithm 1. Then: (1) $\mathcal{R}'$ satisfies strict global exception closure; (2) $\mathcal{R}'$ satisfies no total override; (3)* $\mathrm{MSW}_{\mathcal{R}}(F) = \mathrm{MSW}_{\mathcal{R}'}(F)$ *for every complete feature pattern F.*

Proof. (1) holds because $\mathcal{R}' \subseteq \mathcal{R}$, so any pair of compatible opposite-label rules in $\mathcal{R}'$ is also a pair in $\mathcal{R}$, where the specificity ordering already holds. (2) holds because the algorithm stops only when no rule in S is totally overridden. (3) holds because a totally overridden rule is never inclusion-maximal in $\mathrm{A}(F)$ for any F: by definition, every completion of its body satisfies some stricter opposite-label rule, so there is always a more specific rule applicable. Removing it therefore does not change $\mathrm{M}(F)$ or the MSW decision.

Corollary 6.4. *Let $\mathcal{R}$ satisfy strict global exception closure and let $\mathcal{R}'$ be its pruning fixed point. Then Corollary 5.4 holds for $\mathcal{R}'$ and $\mathcal{T}(\mathcal{R}')$, and* $\mathrm{MSW}_{\mathcal{R}}(F) = \mathrm{MSW}_{\mathcal{R}'}(F)$ *for every F.*

In summary, sanity is not an independent requirement since it follows logically from strict global exception closure. The witness condition is not an independent requirement either since it holds by construction on the pruning fixed point, which is uniquely determined by the rule set, computable by inspecting the rule bodies alone, and makes identical MSW predictions to the original rule set. The formal grounding of the classifier as a defeasible theory therefore rests on a single, efficiently checkable structural property of the learned rule set.

7. Learning ECCRS in Practice

The theoretical results above hold for any rule set satisfying strict global exception closure. We now describe a practical method for learning such rule sets from data.

A decision tree is a classifier that makes predictions by asking a sequence of yes/no questions about the features of an instance. Starting from the root, each internal node tests one feature, and the answer determines which child node to visit next. This continues until a leaf node is reached, and the leaf's label is the prediction. The path from the root to any node can be read as a conjunction of feature

conditions: for example, if the first split tests whether feature a is true and the second tests whether feature b is true, then the node reached after both tests corresponds to the conjunction $a \wedge b$.

Definition 7.1 (Decision tree)**.** A decision tree over a feature set $A = \{a_1, \ldots, a_n\}$ and label set $\{0, 1\}$ is a rooted binary tree in which every internal node is labelled with a feature $a_j \in A$ and has two children corresponding to $a_j = 1$ (true) and $a_j = 0$ (false), and every leaf node is labelled with a class $\ell \in \{0, 1\}$. Given a complete feature pattern F, the tree classifies F by following the unique root-to-leaf path determined by the feature values in F. The body of a node is the conjunction of all feature literals along the path from the root to that node.

At each node in the tree, the training instances that reach that node have a majority label: the label held by more than half of those instances. The majority label can change as the tree grows deeper. For example, the root node might have majority label 1, but after splitting on feature b, the child node corresponding to $a \wedge b$ might have majority label 0. When this happens, the deeper node naturally plays the role of an exception: feature a normally supports label 1, except when b is also present.

The learning procedure works as follows. First, a standard decision tree is trained on the data. Then, rules are extracted not only from the leaves of the tree but from every node where the majority label differs from the label of the node's parent. Each such node produces a rule whose body is the conjunction of feature tests along the path from the root to that node, and whose label is the majority label at that node. Leaf nodes always produce a rule regardless of whether their label differs from their parent's, since they represent the tree's final predictions. After extraction, duplicate bodies are removed (keeping the higher-accuracy rule) and totally overridden rules are pruned via Algorithm 1.

Extracting rules from all levels of the tree, not just the leaves, is what produces the default-and-exception structure. Strict global exception closure holds by construction: two rules from the same root-to-leaf path are related by specificity (the deeper node adds more feature tests), while two rules from different branches have incompatible bodies (they disagree on the splitting feature) and so the closure condition does not apply. The MSW prediction of the resulting ECCRS matches the tree's own prediction by construction.

8. Experiments

Can the learning procedure of Section 7 produce rule sets with nontrivial defeasible structure while remaining competitive in predictive performance? We test this on three benchmark datasets from the UCI Machine Learning Repository [10]: Breast Cancer Wisconsin (699 instances, 9 features, 65.5%/34.5% class split), Mushroom (8124 instances, 22 features, nearly balanced), and Bank Marketing (45211 instances, 16 features, 88.3%/11.7% class split). Features are Boolean-encoded via one-hot encoding. All experiments use 10-fold stratified cross-validation with scikit-learn [11]. Because accuracy alone can be misleading on imbalanced data, we also report macro-averaged precision, recall, and F1 score. ECCRS hyperparameters (maximum depth, minimum leaf size, minimum support) are set per dataset to produce compact rule sets with nontrivial exception structure.

The baselines include three symbolic classifiers and two subsymbolic classifiers. CART [12] with unrestricted depth is the canonical decision tree learner. RIPPER [1] is a widely used rule induction system, the closest existing family to ECCRS in output format. OneR [13] is a single-rule baseline. On the subsymbolic side, we include a Random Forest (100 trees) and a neural network (two hidden layers of 64 and 32 units).

Figure 1 and Table 1 report the results. On Cancer, ECCRS achieves 0.946 accuracy with 13 rules and 8 exception pairs, within two percentage points of the best system (Random Forest at 0.959) and ahead of the neural network (0.940). On Mushroom, CART, RIPPER, and Random Forest all achieve perfect accuracy, and ECCRS is close behind at 0.999 with 14 rules and 8 exception pairs.

On Bank Marketing, the heavy class imbalance makes accuracy alone uninformative: the top five systems span less than one percentage point (0.893 to 0.900). ECCRS achieves the highest precision (0.817) of any system, meaning that when it predicts the minority class it is correct more often than any

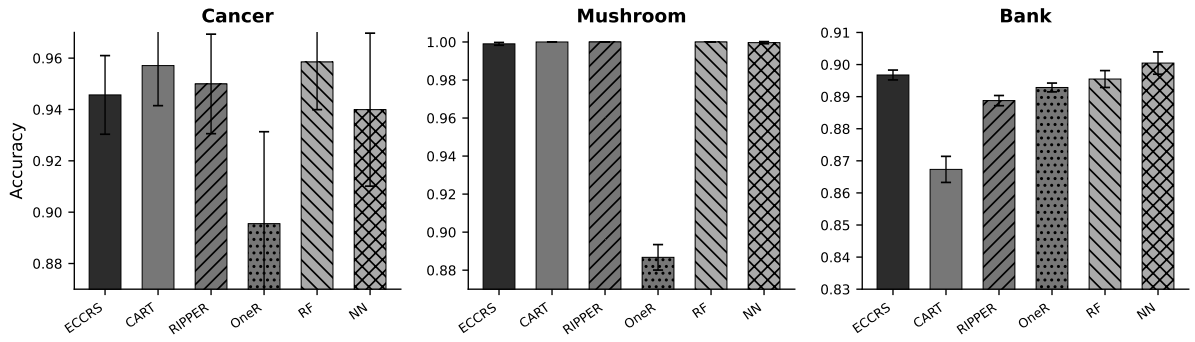

Figure 1: Mean accuracy (± std) over 10-fold CV on three UCI datasets. Y-axes are scaled per dataset. RF = Random Forest, NN = neural network.

Table 1
Mean (± std) over 10-fold CV. Best per metric in bold. Systems above each dashed line are symbolic; those below are subsymbolic. ECCRS is the only system producing exception pairs.

Dataset	System	Accuracy	F1	Precision	Recall	Rules / Exc. pairs
Cancer	ECCRS	.946±.015	.941±.016	.935±.019	.952±.014	13 / 8
	CART	.957±.016	.953±.017	.948±.018	.961±.016	34 / 0
	RIPPER	.950±.019	.946±.021	.940±.023	.955±.018	6 / 0
	OneR	.896±.036	.887±.037	.883±.038	.897±.033	2 / 0
	RF	**.959±.019**	**.955±.020**	**.950±.023**	**.963±.017**	–
	NN	.940±.030	.933±.034	.936±.034	.934±.036	–
Mushroom	ECCRS	.999±.001	.999±.001	.999±.001	.999±.001	14 / 8
	CART	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	17 / 0
	RIPPER	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	8 / 0
	OneR	.887±.007	.886±.007	.896±.005	.890±.007	2 / 0
	RF	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	–
	NN	.9996±.001	.9996±.001	.9996±.001	.9996±.001	–
Bank	ECCRS	.897±.002	.620±.011	**.817±.016**	.589±.008	38 / 6
	CART	.867±.004	.679±.011	.679±.010	.679±.013	6017 / 0
	RIPPER	.889±.002	.565±.013	.769±.022	.552±.008	24 / 0
	OneR	.893±.001	.615±.009	.775±.010	.586±.007	2 / 0
	RF	.895±.003	.670±.011	.762±.012	.635±.010	–
	NN	**.900±.003**	**.696±.021**	.779±.014	**.660±.024**	–

other classifier. Its recall is lower, giving a lower F1 (0.620) than the Random Forest (0.670) or CART (0.679), which captures more minority-class instances at the cost of more false positives using over 6000 rules. ECCRS achieves its performance with just 38 rules and 6 exception pairs. Neither CART nor RIPPER produce any exception pairs: CART extracts rules only from leaves so no rule can ever strictly extend another, and RIPPER uses a priority ordering rather than nested specificity chains.

9. Discussion and Conclusion

The central question this paper asks is whether a learned rule set that looks defeasible can be shown to be defeasible in a formal sense. The answer is yes, under a single structural condition. A conjunctive rule classifier satisfying strict global exception closure is, in the precise KLM sense, a defeasible theory: every prediction it makes is a defeasible entailment of the translated knowledge base, and this holds under Rational Closure, Lexicographic Closure, and System W simultaneously. Because these formalisms disagree on general knowledge bases, their agreement here confirms that the correspondence follows

from the classifier's structure, not from any one formalism's design. The pruning result sharpens the picture: dead rules can collapse the BaseRank construction even though the classifier behaves correctly, but pruning removes them without changing any prediction, after which strict global exception closure alone suffices. Experiments on three benchmark datasets confirm that this structure arises naturally from decision tree extraction and produces compact rule sets competitive with both symbolic and subsymbolic alternatives.

In terms of related work, this paper connects two research traditions that have largely developed independently. On the nonmonotonic reasoning side, the KLM framework [4] provides the formal foundation, with Rational Closure [5], Lexicographic Closure [6, 8], and System W [7] as the three closure operators we consider. Prior work has studied when RC and LC agree on certain knowledge bases [8]; our focus is different, showing that a class of learned classifiers forms a defeasible theory under all three at once. On the machine learning side, Rudin [3] and Freitas [14] argue for interpretable-by-design models. Rule learners such as RIPPER [1], optimal rule lists [2], and ILP [15] produce readable rule sets but none guarantees correspondence to defeasible knowledge. Within ILP, Bain and Muggleton [16] studied learning rules with their exceptions, but the connection to KLM entailment was not made. A predecessor to ECCRS [9] showed empirical competitiveness without formal grounding.

A growing body of work uses argumentation as a bridge between learning and reasoning (see Čyras et al. [17] for a survey). Recent contributions structure learned rules as arguments [18], learn argumentation frameworks from data [19], combine argumentation with case-based reasoning [20], and connect belief change to classifier explanation [21]. Schwind et al. [22] formalise iterated belief change as learning. None of these establishes that a learned classifier's predictions are exactly the defeasible entailments of a KLM knowledge base. The most closely related work is Rienstra [23], who uses KLM preferential models to improve post-hoc explanations of boolean classifiers. The direction is reversed: Rienstra explains an existing classifier using KLM formalisms from the outside, while we show that the classifier itself is a KLM defeasible theory from the inside. On the explainability side, Ribeiro et al. [24] introduce Anchors, conjunctive rules that locally explain black-box predictions. Anchors are post-hoc and carry no global semantic guarantee, whereas ECCRS rules collectively form a provable defeasible theory.

Several open questions remain. The results hold only for complete feature patterns; when features are missing, ECCRS falls back to the most specific applicable rule, but this does not in general match RC, LC, or System W since each handles incomplete information differently. A second direction concerns fairness: the correspondence depends only on strict global exception closure, not on how the rules were learned, so a fairness-constrained algorithm could produce an ECCRS in which protected attributes never appear in exception bodies while preserving the formal guarantees.

In conclusion, whether a learned rule set is a genuine defeasible theory depends on its structure, not just on how the rules are written. Strict global exception closure is enough to guarantee full correspondence with RC, LC, and System W, while dropping it removes any systematic guarantee. Interpretable classifiers show a user which rule fired; this paper asks whether the relationships between rules are logically sound. Strict global exception closure answers that question: when it holds, the default-and-exception reading that humans naturally impose on a rule set is provably correct, not merely a narrative projected onto the output. The condition is efficiently checkable, letting practitioners distinguish formally grounded exception structures from ad hoc conflict resolution. More broadly, this suggests a direction for defeasible machine learning: not only asking whether a classifier produces readable rules, but whether those rules satisfy structural conditions that make their meaning well-defined, rather than just simple to read. The appendix contains proof sketches for the main technical lemmas; full proofs are available in a technical report [25].

10. Acknowledgements

This work is based on the research supported in part by the National Research Foundation of South Africa (REFERENCE NO: SAI240823262612). This work was also supported in part by funding from the

Mastercard Foundation Scholars Program.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: Paraphrase and reword, Improve writing style. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. W. Cohen, Fast effective rule induction, in: Proceedings of the 12th International Conference on Machine Learning (ICML), Morgan Kaufmann, 1995, pp. 115–123.

[2] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning certifiably optimal rule lists for categorical data, Journal of Machine Learning Research 18 (2018) 1–78.

[3] 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. doi:10.1038/s42256-019-0048-x.

[4] S. Kraus, D. Lehmann, M. Magidor, Nonmonotonic reasoning, preferential models and cumulative logics, Artificial Intelligence 44 (1990) 167–207. doi:10.1016/0004-3702(90)90101-5.

[5] D. Lehmann, M. Magidor, What does a conditional knowledge base entail?, Artificial Intelligence 55 (1992) 1–60. doi:10.1016/0004-3702(92)90041-U.

[6] D. Lehmann, Another perspective on default reasoning, Annals of Mathematics and Artificial Intelligence 15 (1995) 61–82. doi:10.1007/BF01535841.

[7] C. Komo, C. Beierle, Nonmonotonic reasoning from conditional knowledge bases with System W, Annals of Mathematics and Artificial Intelligence 90 (2022) 231–268. doi:10.1007/s10472-021-09777-9.

[8] A. Kaliski, An Overview of KLM-Style Defeasible Entailment, Master's thesis, University of Cape Town, 2020. URL: http://hdl.handle.net/11427/32743.

[9] R. S. Madzime, L. Leenen, T. Meyer, An override-aware classifier for transparent AI, in: Proceedings of the Southern African Conference for Artificial Intelligence Research (SACAIR 2025), 2025. Online Proceedings, Vol. II. ISBN: 978-1-0370-5280-4.

[10] D. Dua, C. Graff, UCI machine learning repository, 2019. URL: https://archive.ics.uci.edu/ml.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[12] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[13] R. C. Holte, Very simple classification rules perform well on most commonly used datasets, Machine Learning 11 (1993) 63–91. doi:10.1023/A:1022631118932.

[14] A. A. Freitas, Comprehensible classification models: A position paper, ACM SIGKDD Explorations Newsletter 15 (2014) 1–10. doi:10.1145/2594473.2594475.

[15] S. Muggleton, L. De Raedt, Inductive logic programming: Theory and methods, Journal of Logic Programming 19–20 (1994) 629–679. doi:10.1016/0743-1066(94)90035-3.

[16] M. Bain, S. Muggleton, Non-monotonic learning, in: D. Michie (Ed.), Machine Intelligence 12, Oxford University Press, 1991, pp. 105–120.

[17] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI 2021), 2021, pp. 4392–4399.

[18] N. Prentzas, A. Kakas, C. S. Pattichis, Explainable machine learning via argumentation, in: Proceedings of the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), volume 14318 of *LNCS*, Springer, 2023, pp. 371–385.

[19] C. Tirsi, M. Proietti, F. Toni, ABALearn: An automated logic-based learning system for ABA frameworks, in: Proceedings of the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), volume 14318 of *LNCS*, Springer, 2023, pp. 3–17.
[20] A. Gould, G. Paulino-Passos, S. Dadhania, M. Williams, F. Toni, Preference-based abstract argumentation for case-based reasoning, in: Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR 2024), 2024, pp. 394–404. doi:10.24963/kr.2024/37.
[21] A. Rago, M. V. Martinez, Advancing interactive explainable AI via belief change theory, in: Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR 2024), 2024, pp. 931–937. doi:10.24963/kr.2024/87.
[22] N. Schwind, K. Inoue, S. Konieczny, P. Marquis, Iterated belief change as learning, in: Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI 2025), 2025, pp. 4669–4677. doi:10.24963/ijcai.2025/520.
[23] T. Rienstra, Explaining boolean classifiers with non-monotonic background theories, in: F. A. Oliehoek, M. Kok, S. Verwer (Eds.), Artificial Intelligence and Machine Learning, Springer Nature Switzerland, Cham, 2025, pp. 174–188.
[24] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI), 2018, pp. 1527–1535.
[25] R. Madzime, T. Meyer, Toward defeasible machine learning (full proofs), 2026. URL: https://doi.org/10.5281/zenodo.20700274. doi:10.5281/zenodo.20700274, technical report containing full proofs.

A. Proof Sketches

The lemma statements and proof sketches below provide the key ideas behind the technical results. Full proofs are available in [25].

Lemma A.1 (Label contradiction mechanism)**.** *Let E be any set of implications whose antecedents are feature formulas and whose consequents are in $\{l, \neg l\}$. Let α be a body or feature pattern (hence satisfiable over the feature atoms). Then $E \models \neg\alpha$ if and only if $E \cup \{\alpha\} \models l$ and $E \cup \{\alpha\} \models \neg l$.*

The proof is immediate: α is satisfiable over feature atoms, so the only source of unsatisfiability is forcing both l and $\neg l$. The depth lemmas (Lemmas 4.3, 4.4, 4.5) follow from the definition of Opp and the finiteness of the rule set; each proof is a short calculation given in [25].

Lemma A.2 (Description of E_i)**.** *Assume $\mathcal{R}$ satisfies strict global exception closure, sanity, and no total override. For every $i \geq 0$, $E_i = \{B \to \gamma : (B, \ell) \in \mathcal{R},\ \gamma = l \text{ iff } \ell = 1,\ \mathrm{depth}(B) \geq i\}$.*

Sketch. By induction on i. The forward direction of the inductive step uses closure: an opponent at depth $\geq i$ produces a label contradiction via Lemma A.1. The reverse direction constructs a satisfying world from a witness completion, using closure to show no opposite-label implication in E_i can apply. *Sketch of Proposition 4.6.* For rule bodies the result follows directly from Lemma A.2. For complete patterns, the lower bound uses applicable opponents to force label contradictions at each stage below k; the upper bound argues by contradiction using closure and Lemma 4.3.

Lemma A.3. *Assume $\mathcal{R}$ satisfies strict global exception closure, sanity, and no total override. Let $D_i := \{B \mid\sim \gamma \in \mathcal{T}(\mathcal{R}) : \mathrm{depth}(B) = i\}$ and $p := \max\{\mathrm{depth}(B) : (B, \ell) \in \mathcal{R}\}$. Then the ordered tolerance partition of $\mathcal{T}(\mathcal{R})$ is $(D_0, D_1, \ldots, D_p)$.*

Sketch. A conditional at depth i is tolerated in $D_i \cup \cdots \cup D_p$ because a witness completion provides a verifying world that falsifies nothing in the remaining set. A conditional at depth $d > i$ is not tolerated because every verifying world falsifies its opponent at depth $d - 1$.