Symbolic Explanations for Multi-Label Classification

Ryma Boumazouza[a], Fahima Cheikh-Alili[b], Bertrand Mazure[c] and Karim Tabia[d]
CRIL, Univ. Artois and CNRS, F62300 Lens, France

Keywords: Explainable AI, Multi-Label Classification, Factual and Counterfactual Explanations, SAT Solving.

Abstract: This paper proposes an agnostic and declarative approach to provide different types of symbolic explanations for multi-label classifiers. More precisely, in addition to global sufficient reason and counterfactual explanations, our approach makes it possible to generate explanations at different levels of granularity in addition to structural relationships between labels. Our approach is declarative and allows to take advantage of the strengths of modern SAT-based oracles and solvers. Our experimental study provides promising results on many multi-label datasets.

1 INTRODUCTION

In many fields such as the medical field, it is sensitive and critical to understand how and why a model makes a given prediction. This is reinforced by laws and regulations in several parts of the world (such as the GDPR in Europe) aiming to ensure that AI-based systems are ethical, transparent and make interpretable decisions for users. There are currently many explanation approaches (such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), ANCHORS (Ribeiro et al., 2018)) to explain ML models but they most often address the multi-class classification problem (where a data instance is associated with a single class). Unfortunately, very few studies have focused on explaining multi-label classifiers (where a data instance is associated with a subset of labels). This work proposes a new approach to explain the predictions of a multi-label classifier. This approach overcomes several challenges of multi-label classification. Among the main characteristics of our approach, we mention the following: - *Symbolic*: The symbolic explanations that we propose answer the question *Why a model predicted certain labels (sufficient reasons) ?* or *What is enough to change in an input instance to have a different prediction (counterfactuals) ?* This contrasts with the majority of existing approaches which are numerical and which answer the question *To what extent does a feature influence the prediction of the classifier?* Moreover, the approach provides both feature and label-based explanations. - *Agnostic* : Thanks to using surrogate models, our approach can be used to explain any multi-label classifier, regardless of the used technique and implementation.

- *Declarative*: Our approach to generate symbolic explanations is based on modeling the problem in the form of variants of the propositional satisfiability problem (SAT[1]) in the spirit of the symbolic explainer ASTERYX (Boumazouza et al., 2021). This makes it possible to exploit SAT-based oracles for the enumeration of explanations without implementing dedicated programs.

2 REVIEW OF RELATED WORKS

A lot of current works focus on binary and multi-class classification problems compared to the multi-label ones. The majority of explainability approaches are posthoc and allow to provide essentially two types of explanations: (1) symbolic explanations (e.g. (Shih et al., 2018), (Ignatiev et al., 2019b), (Reiter, 1987)) or (2) numerical ones (e.g. SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016)). It is important to emphasize that these two main categories attempt to answer two different types of questions: While numerical approaches attempt to quantify the influence

[a] https://orcid.org/0000-0002-3940-8578
[b] https://orcid.org/0000-0002-4543-625X
[c] https://orcid.org/0000-0002-3508-123X
[d] https://orcid.org/0000-0002-8632-3980

[1]Boolean satisfiability problem (SAT) is the decision. problem, which, given a propositional logic formula often encoded in CNF, determines whether there is an assignment of propositional variables that makes the formula true

Boumazouza, R., Cheikh-Alili, F., Mazure, B. and Tabia, K.
Symbolic Explanations for Multi-Label Classification.
DOI: 10.5220/0011668700003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 342-349
ISBN: 978-989-758-623-1; ISSN: 2184-433X

of each feature on the prediction, symbolic explanations aim at justifying why a model predicted a given label for an instance through identifying causes (or sufficient reasons) or listing what should be modified in an input instance to have an alternative decision (counterfactuals).
Explanation approaches in multi-label classification can mainly be categorized into feature importance explanations and decision rule explanations. In (Panigutti et al., 2019), the authors propose "MARLENA", a model-agnostic method to explain multi-label black-box decisions. It generates a synthetic neighborhood around the sample to be explained and learns a multi-label decision tree on it. The explanations are simply the decision rules derived from the decision trees. In (Ciravegna et al., 2020), the authors propose an approach to explain neural network-based systems by learning first-order logic rules from the outputs of the multi-label model. This approach completely ignores the features when providing explanations. In (Singla and Biswas, 2021), the authors focus on multi-label model explainability and propose a method to merge multiple feature importance explanations corresponding to each label into a single list of feature contributions. The aggregation of the feature weights is simply the average feature weights over the k labels. The same idea is used in (Chen, 2021) except that they compute Shapley values over the dataset using kernel SHAP and then compute a global feature importance per label. Such methods are limited when it comes to the explanation types they provide. For instance, one can not identify which part of the features is responsible for a given part of the multi-label prediction.

3 SYMBOLIC EXPLANATIONS FOR MULTI-LABEL CLASSIFICATION

This section presents the main types of symbolic explanations for multi-label classification. Explanations are distinguished according to the associated semantics (sufficient reasons or counterfactuals), the elements composing an explanation and the level of granularity of the explanations (the whole prediction or parts of the prediction).

3.1 Multi-Label Classification

A multi-label classification problem is formally defined by a set of feature variables X=$\{X_1,..,X_n\}$ and a set of label (binary) variables Y=$\{Y_1,..,Y_k\}$. A dataset in multi-label classification is a collection of couples $<x,y>$ where x is an instance of X and y an instance of Y encoding the true labels associated with x. Let us first formally recall some definitions used in this paper. For the sake of simplicity, the presentation is limited to classifiers with binary features.

Definition 1 (Multi-label classifier)**.** *A multi-label classifier is a function mapping each input data instance x to a multi-label prediction y. Each input x is a vector of n values assigned to X. Each output is a vector y of k binary values assigned to Y. Given the prediction* y=$f(x)$*, the instance x is classified by f in the label* Y_j *if* Y_j=1 *in the prediction y.*

3.2 Features-Based Explanations

A feature-based explanation involves only features. It can be associated with different semantics and different granularity levels. We focus on two complementary types of feature-based explanations that are the *sufficient reasons* and *counterfactuals*. *Sufficient reason* explanations correspond to the minimal part of the input data that is sufficient to trigger the current prediction while *counterfactual* explanations refer to the minimal changes needed to make in the input data to get an alternative, possibly desired target.
Depending on the problem under study, it may be relevant to have different types of explanations. Assume that we have a MLC problem with a large output set (eg. hundreds). It may be irrelevant to provide an explanation for the entire outcome of the model, especially for datasets with very low density. This is true especially since in most cases, the user is interested in the few classes predicted positively. For example, in document categorization tasks, a user may want to understand why a document is classified in such or such classes. Why this document was not classified in all the remaining classes may be irrelevant. Based on this observation, our approach provides explanations for both the entire prediction and explanations for parts of the prediction that are of interest to the user. We summarize in Table 1 the different cases we distinguish for feature-based explanations:

Table 1: The symbolic-based multi-label explanations.

	Entire-outcome	**Fine-grained**
Sufficient Reasons (Which features cause the current prediction)	*Why* $f(x)$=y ?	*What causes a subset of labels to be predicted by f ?*
Counterfactuals (Which features modify to have an alternative prediction)	*Which* x' *st.* $f(x')$=y' ?	*Which* x' *st. to force f to make a desired partial prediction ?*

In order to illustrate the different concepts, let us use the following example:

Example 1 (Running example: Classifying Yelp reviews into 5 categories)**.** *The "yelp reviews classification" is a categorization problem of reviews to know whether a review positively comments on certain aspects such as food, service, ambiance, deals and worthiness. The dataset contains more than 10000 reviews from food and restaurant areas.*

Input raw data is first pre-processed and two types of features are extracted that are i) textual features consisting of unigrams, bigrams and trigrams and ii) binary features representing rating 1-2 stars, 3 stars, and 4-5 stars respectively. The classes are : F (Food), S (Service), A (Ambience), D (Deals), W (Worthiness). Assume now that we are considering the following review "We went out with friends to have Mexican food, the quesadillas was delicious and came with a lot of cheese. We find the place a little boring but the dining room seemed nice" accompanied with a 4 stars rating. Assume also that we are given the multi-label classifier f depicted in Fig. 1 and consisting in a Binary Relevance classifier using decision trees as base classifiers. The predicted outcome for this review x is $f(x)=(1,0,0,0,0)$.

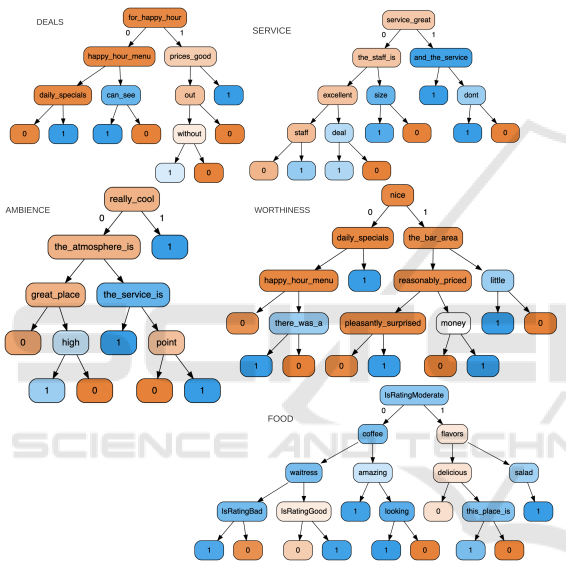

Figure 1: Binary Relevance based on decision trees on Yelp.

3.2.1 Entire-Outcome Explanations

An entire-outcome explanation explains all the predicted labels simultaneously. Our feature-based explanations are based on the definition of sufficient reason explanations and counterfactuals proposed initially for the multi-class case (Shih et al., 2018; Ignatiev et al., 2019a; Ignatiev et al., 2019b; Boumazouza et al., 2021).

Entire-Outcome Sufficient Reasons Explanations

An entire-outcome explanation (*SR* for short) identifies the minimal part of a data sample x (namely, the subset of features) capable to trigger the current multi-label outcome. Formally,

Definition 2 (*SR* Explanations)**.** *Let x be a data instance and $y=f(x)$ be its prediction by the multi-label classifier f. An entire-outcome sufficient reason explanation $\tilde{x}$ is such that:*

1. *$\tilde{x} \subseteq x$ ($\tilde{x}$ is a part of x),*
2. *$\forall \acute{x}, \tilde{x} \subset \acute{x} : f(\acute{x})=f(x)$ ($\tilde{x}$ suffices to trigger $f(x)$),*
3. *There is no $\hat{x} \subset \tilde{x}$ satisfying 1 and 2 (minimality).*

While the two fist conditions in Definition 2 search for parts of x allowing to fire the same prediction, the minimality one allows to find parsimonious explanations (in terms of the number of features involved in the explanation).

Example 2 (Example 1 Continued)**.** ***SR explanations:*** *In order to explain the prediction $y=(1,0,0,0,0)$ for the review in hand, an example of sufficient reason is ['IsRatingBad:0', 'waitress:0', 'looking:0', 'this_place_is:0', 'delicious:1', 'the_staff_is:0', 'staff:0', 'excellent:0', 'service_great:0', 'great_place:0', 'really_cool:0', 'the_atmosphere_is:0', 'daily_specials:0', 'happy_hour_menu:0', 'prices_good:0', 'for_happy_hour:0', 'the_bar_area:0', 'pleasantly_surprised:0', 'reasonably_priced:0']. One can check that this SR forces the five decisions trees to predict $y=(1,0,0,0,0)$ (see Table 2).*

Table 2: An *SR* for the prediction $y=(1,0,0,0,0)$.

Labels	y	Sufficient Reason explanations
Food	1	['IsRatingBad : 0', 'waitress : 0', 'looking : 0', 'this_place_is : 0', 'delicious : 1']
Service	0	['the_staff_is : 0', 'staff : 0', 'excellent : 0', 'service_great : 0']
Ambience	0	['great_place : 0', 'really_cool : 0', 'the_atmosphere_is : 0']
Deals	0	['daily_specials : 0', 'happy_hour_menu : 0', 'for_happy_hour : 0']
Worthiness	0	['daily_specials : 0', 'happy_hour_menu : 0', 'the_bar_area : 0', 'pleasantly_surprised : 0', 'reasonably_priced : 0']

Entire-Outcome Counterfactual Explanations

Another important type of explanations that are actionable are the ones of counterfactuals. Given a target outcome $\acute{y}$, an entire-outcome explanation (*CF* for short) is the minimal changes to be done in x in order to obtain $\acute{y}$. In other words, if for some reason, one wants to force the classifier to predict $\acute{y}$, then a counterfactual explanation is those minimal changes $\acute{x}$ needed to make on x such that $f(x[\acute{x}])$=$\acute{y}$ (the notation $x[\acute{x}]$ denotes the instance x where the variables involved in $\acute{x}$ are inverted).

Definition 3 (*CF* Explanations)**.** *Let x be a complete data instance and y=$f(x)$ be its prediction by the multi-label classifier f. Given a target outcome $\acute{y}$, an entire-outcome counterfactual explanation $\tilde{x}$ of x is such that:*

1. *$\tilde{x} \subseteq x$ ($\tilde{x}$ is part of x),*
2. *$f(x[\tilde{x}]) = \acute{y}$ (fire target prediction),*
3. *There is no $\hat{x} \subset \tilde{x}$ such that $f(x[\hat{x}])$=$f(x[\tilde{x}])$ (minimality).*

Example 3 (Example 2 Continued)**.** *Assume that the initial prediction y is $(1,0,0,0,0)$ and that the target prediction $\acute{y}$ is $(0,1,1,1,1)$. An example of entire-outcome counterfactual is : ['delicious:1', 'IsRatingModerate:0' 'staff:0', 'great_place:0', 'daily_specials:0', 'little:1', 'the_bar_area:0'] . Table 3 shows how this CF forces each decision tree to trigger the target outcome $\acute{y}$.*

Table 3: Example of entire-outcome *CF* explanation.

Labels	y	$\acute{y}$	**Counterfactual explanations**
Food	1	0	['delicious : 1', 'IsRatingModerate : 0']
Service	0	1	['staff : 0']
Ambience	0	1	['great_place : 0']
Deals	0	1	['daily_specials : 0']
Worthiness	0	1	'little : 1', 'the_bar_area : 0']

3.2.2 Fine-Grained Explanations

In practice, it can be more useful to get explanations about a label or a subset of labels of interest rather than an explanation for the entire prediction (a vector of k labels). We say that the Y_j label is positively predicted if $Y_j = 1$, and negatively predicted if $Y_j = 0$.

Fine-Grained Sufficient Reasons Explanations

Similar to the definition of sufficient reasons for the entire-outcome, a fine-grained sufficient reason is limited to explaining the part of y that is of interest to the user.

Definition 4 (SR$_y$ Explanations)**.** *Let x be a data instance, y=$f(x)$ be its multi-label (entire) prediction by the classifier f and $\tilde{y}$ a subset of y representing the labels of interest ($\tilde{y}$ can involve labels that are predicted positively of negatively). A fine-grained sufficient reason explanation $\tilde{x}$ of x is such that:*

1. *$\tilde{x} \subseteq x$ ($\tilde{x}$ is a part of x),*
2. *$\forall \acute{x}$, $\tilde{x} \subset \acute{x}$: $f(\acute{x})$=$\tilde{y}$ ($\tilde{x}$ suffices to trigger $\tilde{y}$),*
3. *There is no $\hat{x} \subset \tilde{x}$ satisfying 1 and 2 (minimality).*

Example 4 (Example 7 Continued)**.** *Assume we want to explain the predictions regarding labels "Food", "Service" and "Ambience" (i.e ($Y_1 = 1, Y_2 = 0, Y_3 = 0$)). The following is an example of a fine-grained SR$_y$ explanation : ['IsRatingBad:0', 'waitress:0', 'looking:0', 'this_place_is:0', 'delicious:1', 'the_staff_is:0', 'staff:0', 'excellent:0', 'service_great:0', 'great_place:0', 'really_cool:0', 'the_atmosphere_is:0'].*

Fine-Grained Counterfactual Explanations

Definition 5 (CF$_y$ Explanations)**.** *Let x be a data instance, y=$f(x)$ be its multi-label prediction by the classifier f. Let $\tilde{y}$ a subset of y representing the labels of interest (namely, the labels to flip). A fine-grained counterfactual explanation $\tilde{x}$ of x is such that:*

1. *$\tilde{x} \subseteq x$ ($\tilde{x}$ is a part of x),*
2. *$f(x[\tilde{x}]) = y[\tilde{y}]$ (inversion of labels into $\tilde{y}$),*
3. *There is no $\hat{x} \subset \tilde{x}$ such that, $f(x[\hat{x}])$=$f(x[\tilde{x}])$ (minimality)*

The term $y[\tilde{y}]$ denotes the prediction y where labels included in $\tilde{y}$ are inverted (set to the target outcome).

Example 5 (Example 4 Continued)**.** *Let us assume that we want to invert the prediction of the labels "Service" and "Ambience" (i.e $\tilde{y} = (Y_2 = 1, Y_3 = 1)$). The following is an example of fine-grained CF$_y$ explanation: ['staff:0', 'great_place:0'].*

3.3 Label-Based Explanations

Up to now, we explain the predictions of a classifier only using the features of the input data. Relying solely on features to form symbolic explanations can be problematic in terms of the clarity and relevance of explanations to the user. As shown in the figures of the example 6, explaining a complex concept or label based solely on features can be difficult for the user to understand. In some cases, this aspect can be greatly improved by exploiting relationships or structures between the labels. For instance, if a label Y_i is subsumed by a label Y_j according to the multi-label classifier f, then clearly sufficient reasons of Y_i are also sufficient reasons for Y_j. Other examples of relations that can be easily extracted and exploited are label equivalence and disjointedness.

The main advantage is that we will have a parsimonious explanation which will be easier for a user to understand, and by reducing the number of the explanations generated, it will simplify their presentation.

Example 6. *Let us consider the example of digits classification using an augmented version of the MNIST dataset with labels "Odd", "Even" and "Prime". The existing labels $Y_{i \in 0...9}$ indicate whether the input image x is recognized as an i-digit while the new labels Y_{ODD}, Y_{EVEN} and Y_{PRIME} correspond respectively to the labels "Odd", "Even" and "Prime". Assume an input image x, and its multi-label prediction $f(x)=(0,0,0,0,0,0,0,0,0,1,1,0,0)$ (namely, x is predicted to be the digit "9" and "Odd"), we have the following explanations:*

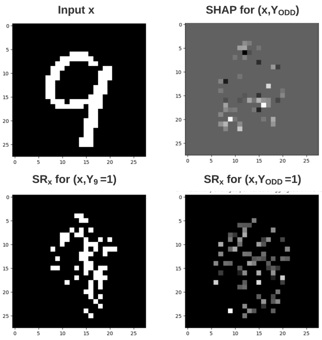

Figure 2: Feature-based explanations for a sample from augmented MNIST dataset.

4 A MODEL-AGNOSTIC SAT-based APPROACH

As mentioned in the introduction, our approach for providing symbolic explanations is agnostic and declarative. It is based on modeling the multi-label classifier and our explanation enumeration problems as variants of the propositional satisfiability problem (SAT) exactly in the spirit of ASTERYX (Boumazouza et al., 2021). The modeling goes through two steps: a first step for encoding the multi-label classifier in an "equivalent" (or "faithful" in case of using a surrogate model) canonical symbolic representation then a second step for enumerating the explanations. Before diving into more details, Fig. 3 depicts a general overview of our approach:

4.1 Step 1: Classifier Modeling

The aim of this step is to associate the multi-label classifier with a symbolic an equivalent/faithful symbolic representation that can be processed by a SAT-based oracle to enumerate our symbolic explanations. As shown in Fig. 3, two cases are considered:

- Direct Encoding : Some machine learning models have direct encoding in conjunctive normal form

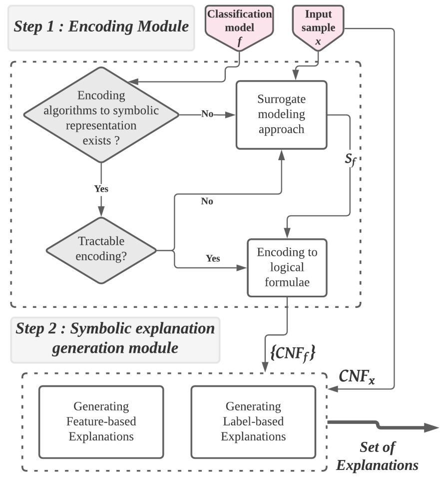

Figure 3: Overview of the proposed approach.

(CNF). For instance, the authors in (Narodytska et al., 2018) proposed a CNF encoding for Binarized Neural Networks (BNNs) for verification purposes. The authors in (Shih et al., 2019) proposed algorithms for compiling Naive and Latent-Tree Bayesian network classifiers into decision graphs. Hence, in some cases, a multi-label classifier can be directly and equivalently encoded in CNF. For instance, the Binary Relevance classifier using decision trees as base classifiers can be equivalently encoded in CNF as illustrated in our running example (same thing holds of random forests, Binarized Neural Networks and some Bayesian network classifiers). The idea is to associate a CNF Σ_i with each base classifier f_i such that the binary prediction of f_i for a data instance x is captured by the truth value or consistency of Σ_i and Σ_x (Σ_x stands for the CNF encoding of the data instance x). Formally, f_i is said to be equivalent to Σ_i iff for any data instance x :

$$\Sigma_i \wedge \Sigma_x = \begin{cases} \top & \text{if } f_i(x) = 1 \\ \bot & \text{otherwise.} \end{cases} \quad (1)$$

Where $\top$ means that the conjunction of Σ_i and Σ_x is satisfiable, corresponding to a positive prediction. Similarly, $\bot$ means that the conjunction of Σ_i and Σ_x is unsatisfiable (in case of negative prediction).

- Surrogate Modeling : In case the multi-label classifier cannot be directly encoded in CNF or in case the encoding is intractable, our approach proceeds by associating with the multi-label classifier a faithful surrogate model that can be encoded in CNF. In addition to allowing the handling of any multi-label classifier, the surrogate modeling offer another useful advantage that is providing local explanations. Indeed, it is challenging to explain a model's prediction over the whole dataset where the decision boundary may not be easily distinguished. The surrogate model built locally will make it possible to provide explanations in the neighborhood of x. Our approach associates a surrogate model s_i for each label Y_i. The surrogate model s_i is trained on the vicinity of the data sample x using the original training instances with the predictions from the MLC model as targets or generated data through perturbing the input instance x. A good surrogate model is the one able to ensure a good trade-off between a high faithfulness to the initial model and tractability of the CNF encoding.

Example 7. *Let us continue our running example. The encoding of the decision trees of Fig. 1 into CNF is direct as shown in the following (encoding a decision tree in CNF comes down to encoding the paths leading to leaves labeled 0).*

Food $y_1 \Leftrightarrow$ (*IsRatingModerate* $\lor$ *coffee* $\lor$ *waitress* $\lor$ $\neg$*IsRatingBad*) $\land$
(*IsRatingModerate* $\lor$ *coffee* $\lor$ $\neg$*waitress* $\lor$ *IsRatingGood*) $\land$
(*IsRatingModerate* $\lor$ $\neg$*coffee* $\lor$ $\neg$*amazing*$\neg$*looking*) $\land$
($\neg$*IsRatingModerate* $\lor$ *flavors* $\lor$ *delicious*) $\land$
($\neg$*IsRatingModerate* $\lor$ *flavors* $\lor$ $\neg$*delicious* $\lor$ $\neg$*this_place_is*)

Service $y_2 \Leftrightarrow$ (*service_great* $\lor$ *the_staff_is* $\lor$ *excellent* $\lor$ *staff*) $\land$
(*service_great* $\lor$ *the_staff_is* $\lor$ $\neg$*excellent* $\lor$ $\neg$*deal*) $\land$
(*service_great* $\lor$ $\neg$*the_staff_is* $\lor$ $\neg$*size*) $\land$
($\neg$*service_great* $\lor$ $\neg$*and_the_service* $\lor$ $\neg$*dont*)

Ambien $y_3 \Leftrightarrow$ (*really_cool* $\lor$ *the_atmosphere_is* $\lor$ *great_place*) $\land$
(*really_cool* $\lor$ *the_atmosphere_is* $\lor$ $\neg$*great_place* $\lor$ $\neg$*high*) $\land$
(*really_cool* $\lor$ $\neg$*the_atmosphere_is* $\lor$ $\neg$*the_service_is* $\lor$ *point*)

Deals $y_4 \Leftrightarrow$ (*for_happy_hour* $\lor$ *happy_hour_menu* $\lor$ *daily_specials*) $\land$
(*for_happy_hour* $\lor$ $\neg$*happy_hour_menu*$\neg$*can_see*) $\land$
($\neg$*for_happy_hour* $\lor$ *prices_good* $\lor$ $\neg$*out*) $\land$
($\neg$*for_happy_hour* $\lor$ *prices_good* $\lor$ *out* $\lor$ $\neg$*without*)

Worth $y_5 \Leftrightarrow$ (*nice* $\lor$ *daily_specials* $\lor$ *happy_hour_menu*) $\land$
(*nice* $\lor$ *daily_specials* $\lor$ $\neg$*happy_hour_menu* $\lor$ $\neg$*there_was_a*) $\land$
($\neg$*nice* $\lor$ *the_bar_area* $\lor$ *reasonably_priced* $\lor$ *pleasantly_surprised*) $\land$
($\neg$*nice* $\lor$ *the_bar_area* $\lor$ $\neg$*reasonably_priced* $\lor$ *money*) $\land$
($\neg$*nice* $\lor$ $\neg$*the_bar_area* $\lor$ $\neg$*little*)

Once the encoding step is achieved, we can rely on SAT-based oracles to provide explanations as follows:

4.2 Step 2: Explanation Enumeration

Recall that in Step 2 we are given a set of CNFs $\Sigma_1,..,\Sigma_k$ encoding the MLC f and a data instance x encoded in CNF and denoted Σ_x. The aim is to explain the prediction $y{=}f(x)$. Recall also that in order to provide sufficient reasons or counterfactuals for a given label Y_i, we rely on a SAT oracle on Σ_i and Σ_x. In the following, let $SR(x,s_i)$ (resp. $CR(x,s_i)$) denote the set of sufficient reasons (resp. counterfactuals) to explain individual prediction $s_i(x)$. Such explanations are obtained thanks to a SAT-based oracle (see for instance (Boumazouza et al., 2021) how one can use a SAT oracle to provide sufficient reasons and counterfactuals for binary classifiers).

4.2.1 Feature-Based Explanations

Depending on the type of explanations to provide, our approach proceeds as follows:
- Entire-Outcome Sufficient Reasons ***SR*****:** Since we can provide sufficient reasons for each label Y_i, then it suffices to combine (join) an *SR* from each classifier S_i to form an *SR* for the whole outcome as shown in the example of Table 2.
- Entire-Outcome Counterfactuals CF**:** Similar to sufficient reasons, one can form entire-outcome counterfactual CF_x as far as we have counterfactuals CF_i for each label Y_i. More precisely, let the MLC f predict y for x (namely, $f(x){=}y$). Let us assume that the user wants to force the prediction to y'. Then, an entire-outcome CF is formed by joining a counterfactual from each CF_i (see example in Table 3).
- Fine-Grained Sufficient Reasons SR_y**:** For fine-grained explanations, we proceed in a similar way while restricting to the part $y'{\subseteq}y$ of interest to the user. Namely, given sufficient reasons for each label $y_i{\in}y'$, then joining an SR_i from each classifier f_i with $y_i{\in}y'$ is enough to form an SR_y for the partial outcome y' as shown in Example 4.
- Fine-Grained Counterfactuals CF_y**:** Given counterfactuals for each label $y_i{\in}y'$, then joining an CF_i from each classifier f_i such that $y_i{\in}y'$ allows to build an CF_y allowing to obtain y' as in Example 5.

4.2.2 Label-Based Explanations

Recall that label-based explanations denote structural relationships between labels. In order to extract some relationships, one can also rely on a SAT-based oracle since each individual labels Y_i is associated with a CNF Σ_i. Hence, checking whether some relationships hold between subsets of labels comes down to checking the corresponding logical relationships between CNF formulas.

For instance, assume we are given an input x and the we want to check whether $Y_1{\equiv}Y_2$ (label equivalence relation) in the vicinity of x. We can easily check if the CNF Σ_1 is logically equivalent to Σ_2 in which case they must share the same models. Another simple method consists simply in checking if for any prediction $y'{=}f(x')$ such that x' is an instance from the

Table 5: Evaluating the CNF encoding over different datasets.

Dataset	radius	avg RF's accuracy	min CNF size	avg CNF size	max CNF size	min enc_runtime(s)	avg enc_runtime(s)	max enc_runtime(s)
YELP Review Analysis	60	92.67%	96/232	4827/13004	13732/36864	0.48	3.29	13.73
	180	92.73%	4625/12416	6812/18395	15963/428941	2.97	4.64	15.32
Augmented MNIST	150	93.97%	509/1268	12095/32353	14308/38344	0.68	12.58	16.13
	250	96.27	423/1119	9556/25455	15105/40530	1.35	7.93	14.41
IMDB Movie Genre Pred	30	99.53%	863/2344	1282/3533	3149/8558	0.82	1.09	2.73
Patient Characteristics (NYS15)	63	96.73%	2446/6615	7887/21370	11305/30594	1.91	6.73	10.12

neighborhood of x that Y_1=1 iff Y_2=1.

5 EMPIRICAL EVALUATION

Due to the page limit, this study concerns only feature-based explanations. The datasets used in our experiments are publicly available and can be found at Kaggle or at UCI. Numerical and categorical attributes are binarized. The textual datasets used are pre-preprocessed and binarized. In order to enumerate our symbolic explanations for binary classifiers, we rely on two SAT-based oracles: the enumeration of counterfactuals is done using the *enumcs* tool(Grégoire et al., 2018) and the sufficient reasons are enumerated using the PySAT (Ignatiev et al., 2018) tool. The time limit for the enumeration of symbolic explanations was set to 300 seconds.

Table 4: Properties of the different data-sets used.

Dataset	#instances	#classes	#features	data type
Augmented MNIST	70000	13	784	Images
Yelp Review Analysis	10806	5	671	Textual
IMDB Movie Genre Prediction	65500	24	30	Textual
Patient Characteristics Survey (NYS 2015)	105099	5	63	Textual/ Numeric

5.1 Results

In order to generate entire-outcome explanations, each base classifier of the a binary relevance (BR) model is approximated using a random forest and then encoded into a CNF. Table 5 lists the average size and time of the encoding step computed over surrogate models. We can see that the average accuracy of the surrogate random forest classifiers is high meaning that the surrogate models can achieve high faithfulness levels wrt. the MLC. Regarding the size of the generated CNFs expressed as the number of variables (#Vars) and number of clauses (#Clauses), one can see that it is tractable and it is easily handled by the SAT-solver (in Step 2).

Table 6 shows the results of enumerating both sufficient reasons and counterfactuals. Using local surrogate models over multiple values of the radius, the symbolic explanations of each base classifier are enumerated, and then the average is computed and given in Table 6 and Table 7. The average time necessary to enumerate all the explanations for a given instance, this latter varies between 2 and 20 seconds. The same finding holds for the number of explanations where one can see that on average this number increases proportionally to the size of the feature set. We also notice that the number of *SR* explanations is of the same order as the number of *CF* ones. Interestingly enough, one can notice that the time required to find one sufficient reason (resp. counterfactual) explanation is very negligible, meaning that the proposed approach is feasible in practice.

6 CONCLUDING REMARKS

This paper proposed a declarative and model-agnostic multi-label classification explanation method. We defined several symbolic explanation types and showed how we can enumerate them using the existing SAT-based oracles. We introduced the concept of the label-based explanations in order to take advantage of the structural relationships between labels in order to reduce the number of generated explanations and improve their presentation to the user. It is worth noticing that the contributions of this work are not simple extensions from the multi-class framework to the multi-label one since there are, for example, concepts specific to the multi-label case such as label-based and fine-grained explanations.

ACKNOWLEDGMENT

This work was supported by the Région Hauts-de-France.

Table 6: Enumeration of entire-outcome counterfactual explanations.

Dataset	radius	min #CFs	avg #CFs	max #CFs	enumtime One CF (s)	min enumtime (s)	avg enumtime (s)	max enumtime (s)
YELP Review Analysis	60	1891	2025	6858	$\leq 10^{-3}$	$\leq 10^{-3}$	2.29	13.46
	180	2601	3203	9693	$\leq 10^{-3}$	0.009	4.5	29.97
Augmented MNIST	150	96	4971	9347	$\leq 10^{-3}$	0.02	15.61	33.27
	250	1158	5027	11323	$\leq 10^{-3}$	1.77	15.9	45.36
IMDB Movie Genre Pred	30	5	14	22	≈ 0	0.13	2.78	7.47
Patient Characteristics (NYS15)	63	134	1052	2399	$\leq 10^{-4}$	0.15	2.83	9.37

Table 7: Enumeration of entire-outcome sufficient reasons explanations.

Dataset	radius	min #SRs	avg #SRs	max #SRs	enumtime One SR (s)	min enumtime (s)	avg enumtime (s)	max enumtime (s)
YELP Review Analysis	60	13116	23167	38620	0.028	10.94	19.37	31.95
Augmented MNIST	150	11292	11956	12621	0.053	12.26	13.06	13.85
IMDB Movie Genre Pred	30	3	41.83	161	0.004	0.003	0.02	0.07

REFERENCES

Boumazouza, R., Cheikh-Alili, F., Mazure, B., and Tabia, K. (2021). Asteryx: A model-agnostic sat-based approach for symbolic and score-based explanations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 120–129.

Chen, S. (2021). Interpretation of multi-label classification models using shapley values. *CoRR*, abs/2104.10505.

Ciravegna, G., Giannini, F., Gori, M., Maggini, M., and Melacci, S. (2020). Human-driven fol explanations of deep learning. In *IJCAI*, pages 2234–2240.

Grégoire, É., Izza, Y., and Lagniez, J.-M. (2018). Boosting mcses enumeration. In *IJCAI*, pages 1309–1315.

Ignatiev, A., Morgado, A., and Marques-Silva, J. (2018). PySAT: A Python toolkit for prototyping with SAT oracles. In *SAT*, pages 428–437.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019a). Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1511–1519.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019b). On relating explanations and adversarial examples. In *Advances in Neural Information Processing Systems*, volume 32.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Narodytska, N., Kasiviswanathan, S., Ryzhyk, L., Sagiv, M., and Walsh, T. (2018). Verifying properties of binarized deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Panigutti, C., Guidotti, R., Monreale, A., and Pedreschi, D. (2019). Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence*, pages 97–110. Springer.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial intelligence*, 32(1):57–95.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD 2016*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Shih, A., Choi, A., and Darwiche, A. (2018). A symbolic approach to explaining bayesian network classifiers. In *IJCAI-18*, pages 5103–5111. International Joint Conferences on Artificial Intelligence Organization.

Shih, A., Choi, A., and Darwiche, A. (2019). Compiling bayesian network classifiers into decision graphs. In *Proceedings of the AAAI-19*, volume 33, pages 7966–7974.

Singla, K. and Biswas, S. (2021). Machine learning explanability method for the multi-label classification model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 337–340. IEEE.