ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230360

Declarative Encoding of Fairness in Logic Tensor Networks

Greta Greco[a, b;*], **Federico Alberici**[a], **Matteo Palmonari**[a] **and Andrea Cosentini**[b]

[a] Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy
[b]Data Science and Artificial Intelligence, Intesa Sanpaolo S.p.A., Turin, Italy

Abstract.

Algorithms are vulnerable to biases that might render their decisions unfair toward particular groups of individuals. Fairness comes with a range of facets that strongly depend on the application domain and that need to be enforced accordingly. However, most mitigation models embed fairness constraints as fundamental component of the loss function thus requiring code-level adjustments to adapt to specific contexts and domains. Rather than relying on a procedural approach, our model leverages declarative structured knowledge to encode fairness requirements in the form of logic rules capturing unambiguous and precise natural language statements. We propose a neuro-symbolic integration approach based on Logic Tensor Networks that combines data-driven network-based learning with high-level logical knowledge, allowing to perform classification tasks while reducing discrimination. Experimental evidence shows that performance is as good as state-of-the-art (SOTA) thus providing a flexible framework to account for non-discrimination often at a modest cost in terms of accuracy.

1 Introduction

Ensuring just algorithmic decisions is directly tied to the more ambitious intent of protecting the rights of individuals and avoiding harm. As society becomes more attuned to issues of discrimination and justice, there is a major focus on ensuring that AI systems are designed and deployed in ways that are fair and equitable. Governments and regulatory bodies around the world are increasingly requiring organizations to guarantee that their AI systems do not discriminate; businesses not paying attention to such constraints could face legal and reputational consequences.

This burst of interest has fueled the research in the field: literature has proposed a number of bias mitigation strategies that may apply according to different fairness definitions [32, 37, 39, 5] - and the corresponding metrics [6]. Nonetheless, most mitigation approaches expect fairness constraints to be embedded within the loss function, hindering the feasibility to make modifications and adapt to the specific domain's purpose. In turn, the option to express natural language fairness constraints into logical predicates allows for a wider leeway while addressing the task. The ability to embed declarative logical statements into neural models is one of the key features of neuro-symbolic integration (NSI) and introduces the chance to account for symbolic rules, in contrast to a mere procedural data-driven process.

* Corresponding Author. Email: g.greco43@campus.unimib.it.

This work explores the potential of employing a NSI approach to fairness leveraging on Logic Tensor Networks in the context of binary classification tasks based on numerical features (thus excluding Large Language Models). We focus on group fairness, and specifically, we aim to meet the requirement imposed by statistical parity, which expects an equal predicted positive rate among groups identified by sensitive attributes. Our contributions rely on a first-order logic axiomatization of fairness constraint, along with a solid theoretical discussion about the choice of optimal fuzzy logic operators for connectives and quantifiers. Experimental evidence shows that our approach reaches results that often outperform SOTA models while providing the additional advantage of greater flexibility and ease of constraint declaration. The novelty of our approach, in fact, lies in that the user can specify fairness constraint in a unique logic predicate, whose impact on the model as a whole, can be incrementally controlled by a corresponding weight. While implementing LTN in the fairness domain, we have observed a strong impact of the logical connectives and quantifier interpretation on the model efficacy and inference task. On this specific aspect, we propose a discussion that indeed provides a contribution in the field of neural-symbolic applications.

To the best of our knowledge, literature proposes a single work in the direction of NSI for fairness [36] that primarily insists on iterative querying to inspect biases through Shapely values and proposes interactive continual learning by adding knowledge through LTN. Conversely, this paper precisely focuses on fairness enforcement in binary classification tasks through first-order logic clauses instilled through LTN. Code and models trained in the experiments will be released to ensure replication.

The remainder of the paper is organised as follows. In Section2, we examine the literature on algorithmic fairness providing an overview of the key concepts and techniques proposed so far. We explore the reasons behind the idea of approaching fairness through an NSI approach along with a review of recent developments in the field. In Section3 we introduce our approach after introducing key features of LTN. In Section 4, we present our results and compare them with similar approaches in the literature, and we end the paper with conclusions and future work.

2 Background and related work

2.1 Fairness

The complex and dynamic nature of the topic has been reflected in a consistent response from the scientific community in the attempt to

fill the gap between ethical and regulatory demand and the increasing adoption of AI at scale. The most harmful impacts of AI on individuals arise when opportunities are unjustly denied or resources are unfairly distributed as a consequence of algorithmic decisions [3, 7]. This work focuses on feature-based binary classification tasks, whose relevance has fostered extended research [29] since in this circumstance fairness can be actually quantified by a number of different metrics [6]. The proliferation of fairness definitions reflects the increasing need of ensuring that AI models at scale avoid replicating or amplifying biases learned from training data, potentially impacting individuals even in the sphere of fundamental rights[31].

In fact, if one defines the scenario that leads to an unfavourable outcome and identifies the socio-demographic features that might give cause for discrimination, bias can be measured and eventually mitigated. The concept of *fair classification* can be considered under an *individual* or a *group* connotation[10, 38].

Under an *individual* interpretation, fairness is informed by the principle that similar individuals should be treated similarly[10]. In this perspective, irrelevant differences between people's features should not lead to significant differences in the model outcome [12]. To bring an example, in a financial lending application, this principle requires that the model predicts similar outcomes for male and female applicants characterised by similar features, except for gender.

Alternatively, the *group* interpretation of fairness aims to equalise statistical quantities over groups of individuals identified by protected attributes, such as gender or ethnicity[38]. In practice, this can be formalised in terms of properties of the joint distribution of the sensitive attribute, the classifier, the target variable, and, in some cases, other features. For instance, one could require an equal predicted positive rate among groups, in all those cases where the model outcome is supposed to be independent of the sensitive attribute, reflecting the notion of an even distribution of resources. Resuming to our running example, this principle can require an equal acceptance rate of credit requests among male and female applicants. In other circumstances, especially when the target variable is supposed to be unbiased, fairness can be determined in terms of the difference in error rates between groups. In this case, the rate of negated opportunities arising from predicting a negative response to worthy individuals shall be equally high among groups identified by gender.

None of the individual and group approaches comes with no shortcomings: if the first formalization suffers its heavy reliance on the definition of a distance metric to measure similarity between individuals, the latter introduces the possibility of undesirable outcomes [38] when unqualified individuals are assigned a positive outcome for the sole reason of belonging to an unprivileged group [2]. In addition, the profusion of existing definitions could be sometimes misleading and makes it theoretically impossible to satisfy even a few of them at once except in highly constrained special cases [19].

The existing means of mitigating algorithmic bias can be framed within three methods: pre-processing, in-processing, and post-processing, according to the step of the AI life cycle they operate on. Pre-processing is meant to reduce or eliminate bias in the dataset [40] by *relabelling* the target variable Y to satisfy a certain fairness measure [16, 25], by *reweighting* [4, 20] or *resampling* [17] representative but unbiased instances or by *learning an intermediate representation* that satisfies fairness constraints while preserving helpful information [38, 11]. In some cases, the latter approach is often ascribed to a form of implicit in-processing since several approaches do involve learning, e.g. the use of variational autoencoder [24] or adversarial learning [26]. All things considered, pre-processing techniques often lack the ability to achieve a user-defined trade-off between fairness and accuracy [35] and are not well suited to circumstances where the problem is caused by the algorithm.

Post-processing strategies have the advantage of being model-agnostic and do not require access to the training procedure [23]. Loosely speaking, they function by changing the predicted labels on a subset of samples, appropriately selected to meet fairness constraints. Proposed methods differ in the rationales behind the choice of the instances that undergo a switch in the predicted labels: some approaches construct randomised decision rules [14, 33], others operate on the uncertainty boundary or the disagreement region of ensemble models [17] or imposing separate thresholds for different groups [8, 30].

If the mitigation is designed to be enforced at training time, it comes down to learning unbiased models on biased training data [16]. Algorithms are then designed to maximise accuracy while minimising discrimination constraints. A prejudice remover regulariser has been proposed by [18], which enforces a classifier's independence from sensitive information, to be integrated into the loss function of a logit model. A fair neural network classifier (FNNC) was introduced by [32] and incorporates fairness constraints into the loss in the form of Lagrangian multipliers.

Leveraging on a novel measure of decision boundary (un)fairness, [37] implemented two complementary approaches that maximise Disparate Impact and accuracy, respectively. Taking into account tree-based classifiers, [5] presents FFTree, a method to find fair splits designed to work with different criteria and metrics.

Although characterised by different connotations, all the abovementioned approaches rely on the idea of hard-coding the fairness notion as a building block of the loss function. Nonetheless, the need to ensure fair outcomes in a diversity of application domains requires a meticulous choice among the broad availability of definitions, or even the urge to devise a bespoke constraint. This makes it difficult to conceive and implement models able to respond to the most diverse requirements unless it is possible to leverage declarative knowledge, rather than procedural. When analysing results in Section 4.1, our approach will be experimentally compared against feature-based (binary) classification models that account for group fairness at training time (in-processing) [32, 6, 26, 37], or that involve learning while retrieving a fair representation of the dataset [26, 38]. Specifically, we will take into account the major and most promising approaches that optimise Statistical Parity Difference or Disparate Impact, which will be further discussed in Section 4. Finally, one approach has been proposed that models fair classification using NSI as we do propose in this paper; we discuss the relationship with this previous work in Section 2.2.

2.2 Neural-symbolic Integration

The success of network-based machine learning is attributed to its ability to learn from complex and high-dimensional data by automatically extracting relevant features [22]. However, the flexibility to learn from examples becomes a disadvantage when data is limited or when (user-defined) explicit knowledge needs to be incorporated. Symbolic AI can handle these limitations by leveraging logical reasoning and explicit representation of knowledge and generalises through logical rules without requiring large amounts of data, yet loosing efficiency when learning from noisy information. As a result, NSI, which combines elements of symbolic reasoning with neural network-based machine learning, is introduced to support each other and overcome their respective limitations. Embedding techniques can enable the representation of relational knowledge in a dis-

tributed neural network, allowing for reasoning to take place through matrix computations over distance functions. This can potentially provide a bridge between distributed and localist representations for reasoning[13], allowing for the integration of learning and reasoning in a principled way.

The coupling of such two different views has resulted in the emergence of multiple approaches that can be broadly partitioned into three main categories[1]:

- frameworks exploiting neural architectures for logical reasoning
- methods that provide logical specifications of neural network architectures
- architectures designed to integrate inductive learning and deductive reasoning into a unique and differentiable network

For the intent of this paper, we concentrate on the latter class that, in turn, includes a number of implementations that significantly differs from each other in terms of the semantics of the logical language, richness of expressivity and the mechanism of integration.

Clauses enforcement, in fact, can be achieved by tweaking the prediction of the network through additional layers, as presented in Deep Logic Model (DLM) [28] that relies on fuzzy logic but only considers propositional connectives. Knowledge Enhanced Neural Network (KENN) [9] adopts a similar approach and introduces a function that modifies the predictions of a base NN exerting weighted constraints. Yet, it does not support full first-order logic and it is limited to universally quantified clauses.

An alternative technique consists in applying logical reasoning to the predictions of a neural network, as proposed in Deepproblog [27], which combines neural networks with expressive probabilistic-logical modeling and reasoning by associating facts and rules with probability values.

A third class of methodologies expects the knowledge to be infused as a part of the model optimization process. [15] encapsulates logic rules into the network parameters performing an iterative training based on the satisfiability of a joint loss. At each step, a student network learns from training data while a teacher network provides feedback based on the encoding of logical rules. Among its limitations, the model only supports universally quantified formulas. The potential of fully expressive first-order logic knowledge has been unlocked by Logic Tensor Networks (LTN) [34, 1] and its generalization LYRICS [28]. Here, low-level perception and high-level reasoning are tightly integrated and influence each other.

For this work, we choose to employ LTN over other approaches due to the following reasons:

- **Expressiveness**: LTN is based on differentiable first-order logic with fuzzy semantics, supporting different interpretations of logical connectives and quantifiers, which enables the use of the full expressiveness of FOL and several modeling choices.
- **Undirected graphical models**: LTNs, along with LYRICS, use undirected graphical models, which view logic as a constraint on a predictive model rather than focusing on causal relationships. This aligns better with the notion of fairness being viewed as a constraint on the model.
- **Clause weighting**: LTNs allow for the specification of fixed or learnable weights on clauses, which can be highly beneficial when dealing with conflicting tasks. This capability is particularly useful in balancing competing objectives and in our domain, to retain control over the desirable fairness level.
- **Versatility**: LTNs allow to mix predicates whose interpretation is learned from the data and predicates whose interpretation is fixed into individual formulas, which we found relevant to model fairness.

To this extent, if natural language facts or constraints, can be formalised into logical statements, then fairness can be instilled through a neural-symbolic approach and, interestingly enough, little research has been conducted in this direction.

As mentioned above, a first approach to neural-symbolic integration to fairness was proposed by [36] that, differently from our objectives, primarily focuses on the continuous interaction between the model and the human in the loop. Using explanatory features provided by Shapeley values, the authors recursively inspect bias in model outcomes and address it accordingly by injecting background knowledge. This work is based on a previous version of LTN that is no longer available and that differs from the current version[36] in a number of relevant theoretical and code-level aspects, and does not investigate the role of different axioms and interpretations of logical operators. By contrast, with our work, we propose an original axiomatization of fairness that, to the best of our knowledge, is unseen in literature, and provide a meticulous investigation of the role played by different mathematical interpretations of universal quantifiers and implication operators. In our experiments, we find that in settings similar to fair classification, non-default interpretations have a great impact on the effectiveness of the approach; therefore, we believe that our findings are interesting for the NSI, also beyond its application to fairness.

3 Modeling Fair Classification in LTN

Within this work, we concentrate on binary classification models where one of the possible outcomes represents an unfavourable decision towards groups of individuals identified by socio-demographic attributes. LTN [34] relies on Real Logic, a fully differentiable first-order knowledge representation system based on the manipulation of real-valued vectors, combined with data-driven machine learning. LTN can represent and compute tasks of deep learning - including classification - while taking into account logic-based constraints. We first summarize the main features of LTN and Real Logic and introduce the problem settings. Then we present the axioms, we discuss the interpretation of logical connectives and the universal quantifier, and provide a summary of the proposed approach.

3.1 Preliminaries and Problem Setting

Real logic constitutes a major component of LTN and its description can be summarized as follows (we emphasize the most distinguishing features in italics and refer to [1] for details).

- **Syntax**: Real Logic is defined on a first-order language $\mathcal{L}$ with a signature that contains a set $\mathcal{C}$ of constant symbols, a set $\mathcal{F}$ of function symbols, a set $\mathcal{P}$ of relation symbols (predicates), and a set $\mathcal{X}$ of variable symbols. Terms are constants and variables (objects), *sequences of terms*, and function symbols applied to terms. *Objects, functions, and predicates are typed*: a function **D** assigns types to the elements of $\mathcal{L}$ to the corresponding domain symbol $\mathcal{D}$. Formulas are defined as in FOL, provided that typing constraints are preserved: if t_1,t_2, and t are terms and P is a predicate, then $t_1 = t_2$ and $P(t)$ are atomic formulas; complex formulas can be defined inductively as usual with logical connectives.
- **Semantics** is inspired by standard abstract semantics of FOL and based on a *grounding function* $\mathcal{G}$, which provides the interpretation of the domain symbols in $\mathcal{D}$ and the non-logical symbols in

$\mathcal{L}$. $\mathcal{G}$ associates a tensor of real numbers to any term of $\mathcal{L}$, and a real number in the interval $[0, 1]$ to any formula ϕ. Intuitively, $G(t)$ are the numeric features of the objects denoted by t, and $G(\phi)$ represents the system's degree of confidence in the truth of ϕ; the higher the value, the higher the confidence. There are a few aspects of LTN semantics that are relevant to our work. The tensor operation that grounds a predicate P can be implemented with an arbitrary neural network; the semantics of logical connectives is based on the semantics of first-order fuzzy logic, for which different interpretations have been proposed (e.g., different t-norms and t-conorms for conjunction and disjunction), thus making LTN highly dependent on the selection of specific semantic interpretations; if domain knowledge, in our case fairness, highly depends on universally quantified implications, the interpretation of fuzzy implication have a deep impact on modeling (e.g., Gödel vs Łukasiewicz); different parametric interpretations of the universal quantifiers are possible: the truth value of a formula ϕ such as $\forall x A(x)$ (where $A(x)$ represents an arbitrary formula with a free variable x) estimates the truth value of ϕ based on some aggregation of the truth values estimated for instantiated formulas $A(c)$ found in the training data.

- **Learning** is mainly governed by grounding that plays a key role in the task of inductive inference. After fixing some choices, e.g., interpretation of connectives, interpretation of quantifiers, and boundaries of domain grounding, the parameters that underpin the representations of language elements can be learned in such a way as to maximize the satisfiability of a set of axioms, which include factual propositions available in a training set as well as generalized propositions encoding general constraints. The learning process eventually searches the optimal set of parameters from a hypothesis space that maximises the satisfiability of a theory $\mathcal{T} = \langle \mathcal{K}, \mathcal{G}_\theta \rangle$ namely the tuple composed by the set of closed predicates $\mathcal{K}$ defined on a set of symbols, and the parametric grounding for symbols and logical operators $\mathcal{G}_\theta$.

In modeling fair classification, we leverage an *explicit and fixed grounding of constants*, which model instances with known features that must be classified, and concentrate on the key learning task of finding the optimal grounding of logical predicates, which encode the separation of instances in groups based on their label. Observe that, since the choice of operators that interpret connectives and quantifiers deeply impacts the grounding of the formulas, this choice also impacts significantly on the satisfaction of predicates, and, eventually, on the performance of a classifier that depends on predicates' grounding. Therefore, another focus of our paper is on finding the choices that better fit modeling fair classification.

3.2 Classification with Known Instance Features

Framing a classification task in LTN requires the definition of the knowledge base along with encoding the dataset features and classification labels using Real Logic. To begin with, let `Instances` denote the **domain** of the examples in the dataset. If the training examples are described by n features, then the grounding of the domain can be expressed as

$$\mathcal{G}(Instances) \subseteq \mathbb{R}^n \tag{1}$$

Therefore, each example k of domain `Instances` is a **constant** symbol whose grounding $\mathcal{G}(k)$ is represented by a tensor in $\mathcal{G}(\mathbf{D}(k)) = \mathbb{R}^n$. We can then introduce a **variable** x that represents a finite sequence of m individuals, each described by n features:

$$\mathcal{G}(x) = \mathbb{R}^{m \times n} \tag{2}$$

To retrieve information about the target variable, we introduce the function *Label* that maps each instance k to its corresponding label $y \in 0, 1$. This helps in identifying two variables x_+ and x_- indicating respectively the sequence of positive and negative training examples. The grounding of the variable indicating positive instances can be expressed as follows:

$$\mathcal{G}(x_+) = \langle d \in \mathcal{G}(x) \mid Label(d) = 1 \rangle \tag{3}$$

Finally, we introduce a set of **predicates** that operate on the domain. The classification task is accomplished by $C(x)$, a trainable classifier with grounding

$$\mathcal{G}(C \mid \theta) \colon x \rightarrow \texttt{sigmoid}(\texttt{MLP}_\theta(x)) \tag{4}$$

where `MLP` is a Multilayer Perceptron with learning parameters θ and single output neuron that returns values between 0 and 1, interpreted as truth values. The learning process is meant to optimise the parameters of the predicate functions $C(x)$ while satisfying the following constraints:

$$\begin{array}{c} \forall x_+ C(x_+) \\ \forall x_- \neg C(x_-) \end{array} \tag{5}$$

The grounding of instances remains stable and is not subject to training. This possibility is encompassed by the framework we use and could be explored in the future; for the intent of this work however, we focus on a pure in-processing technique without learning a (fair) representation of instances.

3.3 Fairness Axiom and Interpretation of Connectives

After expressing the classification task, we finally introduce symbolic knowledge to encode fairness constraints. As mentioned above, we focus on group fairness and specifically, we begin accounting for statistical parity. In addition to reflecting the notion of equality and even distribution of resources, it has convenient technical properties, and it is frequently discussed in legislative contexts related to disparate impact[11] thus providing a common ground for comparison to other bias mitigation strategies. This principle is based on the assumption of independence and requires that the probability of a positive prediction, given a sensitive attribute, should be equal across all groups. Formally, this can be expressed as:

$$\mathbb{P}\{\hat{Y} = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1, A = b\} \tag{6}$$

where $A = a, b$ corresponds to different groups identified by protected attributes whose symbolic representation needs to be encapsulated within the domain. To this end, we introduce a pair of first-order logic (FOL) predicates identifying whether an instance belongs to the privileged - or unprivileged - subgroup: $Priv(x)$ and $Unpriv(x)$.

Precisely, $Priv(x)$ and $Unpriv(x)$ are non-trainable predicates that map each example x to a truth value based on whether the example belongs to the privileged or non-privileged group, respectively.

At this point, the knowledge base holds all the essential elements required to render the probabilistic formulation of statistical parity into a first-order logic axiom. Precisely, it takes the form of an equivalence between two implications:

$$\forall x(Priv(x) \rightarrow C(x)) \longleftrightarrow \forall x(Unpriv(x) \rightarrow C(x)) \tag{7}$$

The axiomatization captures an equivalence between the probability of a privileged group (denoted as $Priv(x)$) being assigned a particular target label and the probability of an unprivileged group (represented by $Unpriv(x)$) being assigned the same target label, by providing a satisfiability interpretation for it in Real Logic. This formalization aims to ensure fairness by asserting that the predicted target $C(x)$ is applied equally to both privileged and unprivileged groups. For instance, in the context of recidivism prediction, if $Recid(x)$ represents the likelihood of recidivism and $White(x)$ and $Black(x)$ denote two racial groups, the axiomatization asserts that the probability of assigning a recidivism label should be equivalent between these groups:

$$\forall x(White(x) \rightarrow Recid(x)) \longleftrightarrow \forall x(Black(x) \rightarrow Recid(x))$$

By establishing this equivalence, the axiom states unbiased recidivism label assignment, regardless of an individual's racial background. It is worth noting that such a simple, intuitive, and easy-to-interpret axiom is also the one that performs better, among a few variants we tested (omitted for space constraints).

Given that the bias mitigation strategy we propose in this paper strongly relies on Axiom 7, it is noteworthy to delve into the details concerning its implementation and pose a theoretical background behind our choice of the fuzzy implication operand. Broadly speaking, implications are employed in two well-known rules of inference: *modus ponens* and *modus tollens*. Considering the implication $\forall x\phi(x) \rightarrow \psi(x)$, *Modus ponens* states that if $\phi(x)$ is known to be true, then $\psi(x)$ is also true. *Modus tollens* instead, poses its accent on the consequent and when $\neg\psi(x)$ is known to be true, then $\neg\phi(x)$ is true as well, this is because if the antecedent were true, the consequent should have been true as well.

When the classification predicate predicts a scenario with a false implication, there are multiple ways to rectify it. Consider the following implication from the left-hand side of Axiom 7:

$$\forall x(Priv(x) \rightarrow C(x))$$

This formula implies that all privileged examples are positive examples. Four categories emerge: positive privileged examples (PPE), negative non-privileged examples (NNPE), positive non-privileged examples (PNPE), and negative privileged examples (NPE). Assuming an NPE is observed, which is inconsistent with the background knowledge, there are two options:

- *Modus Ponens* trusts the observation of a privileged example and believes it is a positive example (PPE). The truth of the consequent is believed if the antecedent is true.
- *Modus Tollens* trusts the observation of a negative example and believes it is a non-privileged example (NNPE). The antecedent is believed to be false if the consequent is false.

Given that the predicate $Priv(x)$ has a fixed grounding, opting for a *modus ponens* reasoning is a preferable alternative. In a differentiable fuzzy logic setting, this means that when the antecedent is high, the consequent is increased. The *modus tollens* reasoning would be ineffective since it operates by decreasing the antecedent, which cannot be changed due to its inherent fixed nature.

If in FOL the implication has a well-defined semantic, we discussed how its interpretation can vary in fuzzy logic. There are two primary classes of implications generated from the fuzzy logic operators for negation, conjunction, and disjunction:

- *Strong implications* are dfined using a fuzzy negation and fuzzy disjunction as $x \rightarrow y = \neg x \vee y$.
- *Residuated implications* are defined using a fuzzy conjunction and can be understood as a generalization of *modus ponens*, where the consequent is at least as true as the (fuzzy) conjunction of the antecedent and the implication.

Consequently, the latter represents the most suitable alternative, and we opt for the Gödel implication, defined as follows:

$$I(x,y) = \begin{cases} 1, & \text{if } x \leq y \\ y, & \text{otherwise} \end{cases} \tag{8}$$

This implication makes strong discrete choices and increases at most one of its outputs. The Gödel implication increases the consequent whenever is smaller than the antecedent, which, in turn, is never changed [21]. Moreover, as the derivative of the negated antecedent is always 0, it can never choose the *modus tollens* correction, as intended.

A last choice concerns the aggregation function used in universal quantifiers, for which we choose the parametric A_{pME}[1]. With this choice, universal quantifiers are represented by the generalised mean w.r.t the error that measures to what extent each value deviates from the ground truth:

$$A_{pME}(x_1,\ldots,x_n) = 1 - \left(\frac{1}{n}\sum_{i=1}^{n}(1-x_i)^p\right)^{\frac{1}{p}} \tag{9}$$

retrieving the arithmetic mean for $p = 1$. On the other hand, given the same input and increasing p, the value for the quantifier starts diminishing as A_{pME} converges to the min operator. Consequently, the parameter p provides flexibility in formulating more or less strict formulas, thereby accommodating or limiting the impact of outliers.

Our approach is summarized in Figure 1. The only trainable predicate is C, the one that models the classification outcomes. Its parameters are trained to jointly optimize the satisfiability of the facts in the training set (via Axiom 5) and the fairness constraint specified by Axiom 7, whose compositional interpretation is shown in the figure.

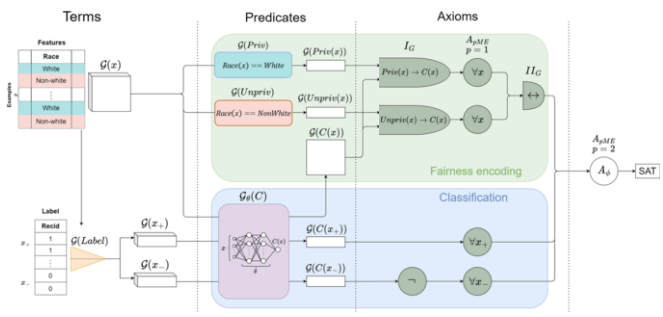

Figure 1. Conceptual representation of the elements of Real Logic and their role within the proposed approach. Classification task and Fairness constraints are declared separately through their respective axioms.

4 Experiments and Results

This section collects the results of our experiments and aims at evaluating to what extent our approach is able to optimise observational group fairness metrics derived from eq. 6 without excessively yielding on accuracy. In particular we account for Statistical Parity Difference and Disparate Impact

$$SPD = \mathbb{P}\{\hat{Y} = 1, A = a\} - \mathbb{P}\{\hat{Y} = 1, A = b\} \tag{10}$$

$$DI = \frac{\mathbb{P}\{\hat{Y} = 1, A = a\}}{\mathbb{P}\{\hat{Y} = 1, A = b\}} \tag{11}$$

for all demographic groups a, b. Optimum values for SPD are close to zero while DI implies equity for values close to one. Despite at first sight the two metrics may seem interchangeable, it is important to remark that one might be more suitable than the other depending on the specific context: in applications where the rate of positive labels is extremely low (e.g. fraud detection), minimising SPD is not an advisable choice since its value would be indeed negligible ever for very diverse values of $\mathbb{P}\{\hat{Y} = 1\}$ among groups.

Our mitigated model will be confronted with a baseline model implemented in LTN with ho fairness constraint to verify the eventual drop in accuracy while optimising non-discrimination. In addition, debiased results will be compared against SOTA in-processing approaches proposed in the literature and metrics for accuracy, SPD and DI will be checked against.

The MLP that implements the grounding of the classification predicate C is composed of a two-layered feed-forward network with (100, 50) hidden neurons, for all the datasets. Due to their statistical nature, observational group fairness measures only make sense when calculated across samples containing a number of instances from both the sensitive groups, we approximate group fairness metrics (SPD and DI) using batch training. All the results presented in this paper are averaged over a 5-fold cross-validation.

Our model is trained and evaluated on three benchmark datasets used in the fairness domain [29], available from the UCI ML-repository:

- *Adult* income dataset has 45,222 instances. The target variable indicates whether or not income exceeds $50K per year based on census data, with gender as the protected attribute.
- *German* credit risk dataset is composed by 1,000 entries and is meant to classify bad credits based on a set of attributes encompassing demographic and financial information, including gender, that is used as a protected attribute.
- *COMPAS* dataset includes 7,918 records collecting demographics and criminal history to predict someone's recidivism. Here, race is considered the protected attribute - restricted to white and black defendants.

The inherent bias corresponding to different metrics computed on the three datasets is reported in Table 1

Table 1. Inherent bias over the dataset under examination

	Dataset - Protected Attribute		
Metric	Adult - Gender	Compas - Race	German - Gender
SPD	0.200	0.095	0.067
DI	0.362	0.805	0.907

We train our model to enforce the fairness clause formalised in Axiom 7, along with the classification task expressed in Axiom 5. The weight associated with the classifier is kept fixed, while we vary the weight of the fairness constraint: in principle, our approach allows for a fine-grained control on the degree of fairness given that the fairness axiom can be associated with a weight determining its relevance within the overall computation of the satisfiability. Hence, we investigate how accuracy and bias metrics vary according to different weights. We establish our baseline to be a plain classifier based on LTN, with no additional constraints. Being our model optimised on the probabilistic formulation of statistical parity and not on the associated metrics, we are interested in evaluating fairness both in terms of statistical parity difference and disparate impact.

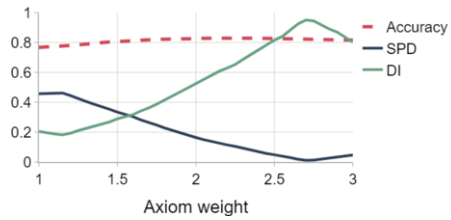

Figure 2. Mitigation results on Adult Income Dataset

In observing results, we wish to remark that enforcing fairness principles like statistical parity, which do not condition on the true target label, cannot reach perfect accuracy even with a perfect model since in that case, one would retrieve the inherent fairness level. Therefore, in these cases, there needs to be a trade-off between fairness and accuracy, which shall be taken into account while analysing the results from the mitigation. We perform a thorough analysis of the Adult dataset as it has a stronger inherited bias and represents the most interesting scenario to evaluate in terms of room for fairness improvements. For the remaining dataset, we report performance and fairness metrics for the optimal parameter configuration. According to what emerges from Figure 2, increasing fairness axiom weight does have an impact both on SPD and DI, simultaneously, that reach almost perfect equality for weights close to 2.7. Despite the inherent bias of the Adult dataset being rather severe and one would expect a significant drop in accuracy as a result of the mitigation, the plot shows that accuracy remains relatively stable. This reveals a highly efficient optimization process: to equalise statistical parity, a significant number of predictions shall be changed, increasing the rate of positive predictions among the unprivileged group. If this relevant adjustment merely impacts accuracy, the model is mainly changing predictions to individuals that were incorrectly classified in the first place and that most likely were assigned a predicted probability close to the decision boundary.

Despite the architectural optimization of the network and its parameters falling outside the scope of this paper, it is noteworthy to remark that this tuning has not resulted in significant improvements in model outcomes in terms of accuracy. Indeed, the satisfiability of logical clauses - and consequently the learning process - strongly depends on the choice of the operators approximating the connectives and quantifiers. Taking into account that some first-order fuzzy semantics are better suited for gradient-descent optimization, the best-performing implementation for conjunction uses the product t-norm T_P with its dual t-conorm S_P for disjunction, together with standard negation N_s.

We also evaluate the impact of parameter p used in the A_{pME} interpretation of the universal quantifier (see Equation9). In a learning setting, assigning an excessively high value to p may lead to a "single-passing" operator that overly focuses on outliers at each step. This can result in gradients overfitting one input at that step, which may adversely affect the training of other inputs. This can be experimentally observed in Figure 3 where we observe values of DI and accuracy for different values of p.

Lastly, another crucial aspect influencing the training procedure involves the choice of implication operator, as thoroughly discussed in Section 3: in contrast to what reported in other contexts and unlike

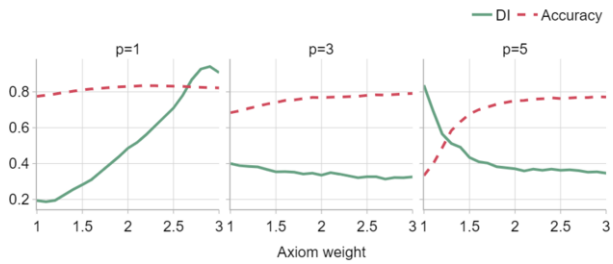

Figure 3. Different optimization curves for increasing value of parameter p of the universal quantifier, as a function of the fairness axiom weight. Optimal results are achieved for A_{pME} converging to arithmetic mean

authors' recommendation [1], we argue - and then experimentally observe, that Gödel represents the optimal choice concerning the implication operator.

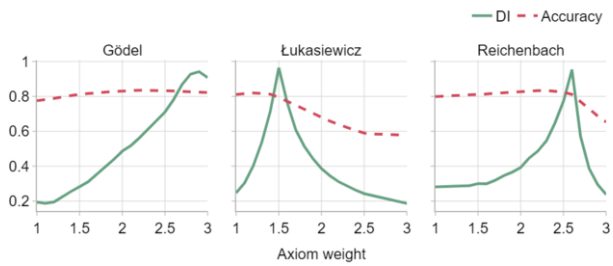

Figure 4. Despite all implementation of implication operator are capable of optimizing fairness, Gödel do not experience a noticeable accuracy descent in correspondence of the peak in DI.

Wrapping up, the parameters' space spans three dimensions: fairness axiom weight, implication operator and universal quantifier's exponent p. Table 2 reports accuracy and fairness metrics for all the dataset under consideration, for the optimal configuration identified by $p = 1$ and Gödel implication operator, while the optimal value for w varies according to the dataset.

Table 2. Experimental results

	Accuracy		DI		SPD	
	mitig	baseline	mitig	baseline	mitig	baseline
Adult	**0.823**	0.805	**0.949**	0.362	**0.012**	0.200
COMPAS	**0.643**	0.631	**0.957**	0.805	**0.020**	0.095
German	0.699	**0.675**	**0.966**	0.907	**0.021**	0.067

On each dataset, our approach is able to perform a mitigation that is close to complete debias, no matter the magnitude of the initial inherited bias. On mitigated predictions, the loss in accuracy is negligible, and in some cases performance increases. A possible explanation for this phenomenon is that the baseline model is keen on overfitting and the fairness enforcement act as a regulariser. It is necessary to take into account that neural-symbolic integration models offer way different advantages rather than a mere accuracy optimization.

4.1 Comparative results

In this section, we wish to compare our experiments with the results obtained by similar approaches. Instead of reproducing each model, we directly report the results from the original papers. It is noteworthy that prior studies differ from our approach in that they do not explicitly enforce SP and DI jointly, whereas they separately formulate different optimizations. Our model, in fact, reaches the best values for the two metrics for the same input parameters and configuration, as evident from Figure 2. Furthermore, literature has often focused on reporting accuracy for DI > 0.8, not taking into account that inherited bias in COMPAS and German dataset already satisfy this condition, which is likely to be replicated by a non-mitigated model. Instead, in Figure 5 we report results at a higher threshold to better capture the behavior of our model and show it can reach way higher fairness values, much closer to perfect equality.

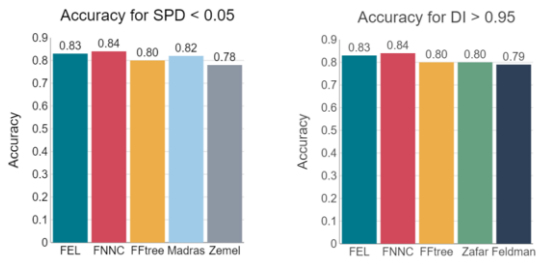

Figure 5. Comparative results, FEL refers to the proposed approach of Fairness Encoding in LTN

In comparing results, we choose the thresholds for SPD to be 0.05, higher than what reported in other works, in order to include all the approaches under consideration. This does not affect the soundness of the comparison since the chosen value is itself very close to the reachable minimum. Similarly, we deviate from the convention of evaluating accuracy for DI $\geq$ 0.8 since, as discussed above, we consider it a too-mild requirement.

As evinced by Figure 5, when evaluated on Adult dataset, our approach of Fairness Encoding in LTN (FEL) is outperformed by FNNC [32] in both metrics but is able to keep accuracy higher than any other model taken into consideration.

5 Discussion and Conclusion

We proposed and explored an approach that encodes fairness constraints in a binary classification setting, exploiting the declarative power of first-order logic and its fuzzy implementation in Logic Tensor Networks, a neural-symbolic integration framework. We instill the fairness principle based on independence and, to the best of our knowledge, present the first method that remains at a higher level of abstraction and optimises on a satisfiability constraint rather than on a numerical metric. In every setting and configuration, we concurrently reach the best values for demographic parity difference and disparate impact, often at a small cost in terms of accuracy. The adherence to non-discrimination constraint can be incrementally controlled by axiom weight, allowing to achieve the required trade-off between fairness and accuracy. We provide a theoretical grounding of the choices in terms of universal quantifier and implication operator interpretations, which are supported by experimental evidence and provide conclusive insights on the best choices to model fair classification with LTN. Experimentally, we contrast our results to similar models presented in literature, often outperforming SOTA approaches, despite using a simple formalization of fairness with interpretable semantics. While we focus on two well-known quantitative definitions of fairness, our model encompasses many others and one could consider using this framework and extending the axiomatisation to include equalised odds or predictive parity for instance.

References

[1] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger, 'Logic tensor networks', *Artificial Intelligence*, **303**, 103649, (2 2022).

[2] Solon Barocas and Andrew D. Selbst, 'Big data's disparate impact', *SSRN Electronic Journal*, (2018).

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, 'Fairness in criminal justice risk assessments: The state of the art', *Sociological Methods and Research*, **50**, (2021).

[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy, 'Building classifiers with independency constraints', *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, 13–18, (2009).

[5] Alessandro Castelnovo, Andrea Cosentini, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica, 'Fftree: A flexible tree to handle multiple fairness criteria', *Information Processing & Management*, **59**, 103099, (2022).

[6] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini, 'A clarification of the nuances in the fairness metrics landscape', *Scientific Reports*, **12**, (2022).

[7] Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big Data*, **5**, (2017).

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, 'Algorithmic decision making and the cost of fairness', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **Part F129685**, (2017).

[9] Alessandro Daniele and Luciano Serafini, 'Knowledge enhanced neural networks', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **11670 LNAI**, (2019).

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, 'Fairness through awareness', *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, (2012).

[11] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, 'Certifying and removing disparate impact', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **2015-August**, (2015).

[12] Will Fleisher, 'What's fair about individual fairness?', *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, (2021).

[13] Artur d'Avila Garcez and Luis C Lamb, 'Neurosymbolic ai: The 3 rd wave', *Artificial Intelligence Review*, 1–20, (2023).

[14] Moritz Hardt, Eric Price, and Nathan Srebro, 'Equality of opportunity in supervised learning', *Advances in Neural Information Processing Systems*, 3323–3331, (2016).

[15] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric P. Xing, 'Harnessing deep neural networks with logic rules', *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, **4**, (2016).

[16] Faisal Kamiran and Toon Calders, 'Classifying without discriminating', *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*, (2009).

[17] Faisal Kamiran and Toon Calders, 'Data preprocessing techniques for classification without discrimination', *Knowledge and Information Systems*, **33**, (2012).

[18] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma, 'Fairness-aware learning through regularization approach', *Proceedings - IEEE International Conference on Data Mining, ICDM*, 643–650, (2011).

[19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, 'Inherent trade-offs in the fair determination of risk scores', *arXiv preprint arXiv:1609.05807*, (2016).

[20] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris, 'Adaptive sensitive reweighting to mitigate bias in fairness-aware classification', *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, (2018).

[21] Emile Van Krieken, Erman Acar, and Frank Van Harmelen, 'Analyzing differentiable fuzzy implications', *17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020*, **2**, (2020).

[22] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep learning', *Nature*, **521**, 436–444, (5 2015).

[23] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri, 'Bias mitigation post-processing for individual and group fairness', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, **2019-May**, (2019).

[24] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, 'The variational fair autoencoder', *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, (2016).

[25] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini, 'k-nn as an implementation of situation testing for discrimination discovery and prevention', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2011).

[26] David Madras, Elliot Creager, Toniann Pitassi, and Richards Zemel, 'Learning adversarially fair and transferable representations', *35th International Conference on Machine Learning, ICML 2018*, **8**, (2018).

[27] Robin Manhaeve, Angelika Kimmig, Sebastijan Dumančić, Thomas Demeester, and Luc De Raedt, 'Deepproblog: Neural probabilistic logic programming', *Advances in Neural Information Processing Systems*, **2018-December**, (2018).

[28] Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, and Marco Gori, 'Integrating learning and reasoning with deep logic models', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **11907 LNAI**, (2020).

[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, 'A survey on bias and fairness in machine learning', *ACM computing surveys (CSUR)*, **54**(6), 1–35, (2021).

[30] Aditya Krishna Menon and Robert C Williamson, 'The cost of fairness in binary classification', *Proceedings of Machine Learning Research*, **81**, 1–12, (2018).

[31] Cathy O'Neil, *Weapons of math destruction*, Penguin Books, 2017.

[32] Manisha Padala and Sujit Gujar, 'Fnnc: Achieving fairness through neural networks', *IJCAI International Joint Conference on Artificial Intelligence*, **2021-January**, (2020).

[33] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger, 'On fairness and calibration', *Advances in Neural Information Processing Systems*, **2017-December**, (2017).

[34] Luciano Serafini and Artur S. d'Avila Garcez, 'Learning and reasoning with logic tensor networks', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **10037 LNAI**, (2016).

[35] Ying Sun, Benjamin C.M. Fung, and Fariborz Haghighat, 'In-processing fairness improvement methods for regression data-driven building models: Achieving uniform energy prediction', *Energy and Buildings*, **277**, (12 2022).

[36] Benedikt Wagner and Artur d'Avila Garcez, 'Neural-symbolic integration for fairness in ai', *CEUR Workshop Proceedings*, **2846**, (2021).

[37] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi, 'Fairness constraints: A flexible approach for fair classification', *Journal of Machine Learning Research*, **20**, 1–42, (2019).

[38] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork, 'Learning fair representations', *30th International Conference on Machine Learning, ICML 2013*, (2013).

[39] Lu Zhang, Yongkai Wu, and Xintao Wu, 'Fairness-aware classification: Criterion, convexity, and bounds', *Association for the Advancement of Artificial Intelligence*, (2018).

[40] Zhe Zhang, Shenhang Wang, and Gong Meng, 'A review on preprocessing methods for fairness in machine learning', *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 1185–1191, (2022).