# Entropy-Based Logic Explanations of Neural Networks

**Pietro Barbiero**[1], **Gabriele Ciravegna** [2,3,4], **Francesco Giannini** [3], **Pietro Lió** [1], **Marco Gori** [3,4], **Stefano Melacci** [3],

[1] University of Cambridge (UK)
[2] Università di Firenze (Italy)
[3] Università di Siena (Italy)
[4] Université Côte d'Azur (France)

{pb737, pl213}@cam.ac.uk, gabriele.ciravegna@unifi.it, {francesco.giannini, marco.gori}@unisi.it, mela@diism.unisi.it

**Abstract**

Explainable artificial intelligence has rapidly emerged since lawmakers have started requiring interpretable models for safety-critical domains. Concept-based neural networks have arisen as explainable-by-design methods as they leverage human-understandable symbols (i.e. concepts) to predict class memberships. However, most of these approaches focus on the identification of the most relevant concepts but do not provide concise, formal explanations of how such concepts are leveraged by the classifier to make predictions. In this paper, we propose a novel end-to-end differentiable approach enabling the extraction of logic explanations from neural networks using the formalism of First-Order Logic. The method relies on an entropy-based criterion which automatically identifies the most relevant concepts. We consider four different case studies to demonstrate that: (i) this entropy-based criterion enables the distillation of concise logic explanations in safety-critical domains from clinical data to computer vision; (ii) the proposed approach outperforms state-of-the-art white-box models in terms of classification accuracy and matches black box performances.

## 1 Introduction

The lack of transparency in the decision process of some machine learning models, such as neural networks, limits their application in many safety-critical domains (EUGDPR 2017; Goddard 2017). For this reason, explainable artificial intelligence (XAI) research has focused either on *explaining* black box decisions (Zilke, Loza Mencía, and Janssen 2016; Ying et al. 2019; Ciravegna et al. 2020a; Arrieta et al. 2020) or on developing machine learning models *interpretable by design* (Schmidt and Lipson 2009; Letham et al. 2015; Cranmer et al. 2019; Molnar 2020). However, while interpretable models engender trust in their predictions (Doshi-Velez and Kim 2017, 2018; Ahmad, Eckert, and Teredesai 2018; Rudin et al. 2021), black box models, such as neural networks, are the ones that provide state-of-the-art task performances (Battaglia et al. 2018; Devlin et al. 2018; Dosovitskiy et al. 2020; Xie et al. 2020). Research to address this imbalance is needed for the deployment of cutting-edge technologies.

Most techniques *explaining* black boxes focus on finding or ranking the most relevant features used by the model to make predictions (Simonyan, Vedaldi, and Zisserman 2013; Zeiler and Fergus 2014; Ribeiro, Singh, and Guestrin 2016b; Lundberg and Lee 2017; Selvaraju et al. 2017). Such feature-scoring methods are very efficient and widely used, but they cannot explain how neural networks compose such features to make predictions (Kindermans et al. 2019; Kim et al. 2018; Alvarez-Melis and Jaakkola 2018). In addition, a key issue of most *explaining* methods is that explanations are given in terms of input features (e.g. pixel intensities) that do not correspond to high-level categories that humans can easily understand (Kim et al. 2018; Su, Vargas, and Sakurai 2019). To overcome this issue, *concept-based* approaches have become increasingly popular as they provide explanations in terms of human-understandable categories (i.e. the *concepts*) rather than raw features (Kim et al. 2018; Ghorbani et al. 2019; Koh et al. 2020; Chen, Bei, and Rudin 2020). However, fewer approaches are able to explain how such concepts are leveraged by the classifier and even fewer provide concise explanations whose validity can be assessed quantitatively (Ribeiro, Singh, and Guestrin 2016b; Guidotti et al. 2018; Das and Rad 2020).

**Contributions.** In this paper, we first propose an entropy-based layer (Sec. 3.1) that enables the implementation of *concept-based* neural networks, providing First-Order Logic explanations (Fig. 1). The proposed approach is not just a post-hoc method, but an *explainable by design* approach as it embeds additional constraints both in the architecture and in the learning process, to allow the emergence of simple logic explanations. This point of view is in contrast with post-hoc methods, which generally do not impose any constraint on classifiers: After the training is completed, the post-hoc method kicks in. Second, we describe how to interpret the predictions of the proposed neural model to distill logic explanations for individual observations and for a whole target class (Sec. 3.3). We demonstrate how the proposed approach provides high-quality explanations according to six *quantitative* metrics while matching black-box and outperforming state-of-the-art white-box models (Sec. 4) in terms of classification accuracy on four case studies (Sec. 5). Finally, we share an implementation of the entropy layer, with extensive documentation and all the experiments in the public repository: https://github.com/pietrobarbiero/entropy-lens.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

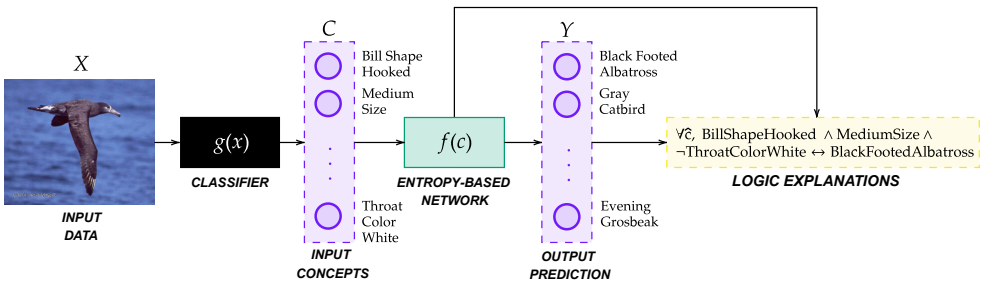


Figure 1: The proposed pipeline on one example from the CUB dataset. The neural network $f\colon C \mapsto Y$ maps concepts onto target classes and provide concise logic explanations (yellow – arguments of predicates are dropped for simplicity) of its own decision process. When the input data is non-interpretable (as pixels intensities), a classifier $g\colon X \mapsto C$ maps inputs to concepts.

## 2 Background

Classification is the problem of identifying a set of categories an observation belongs to. We indicate with $Y \subset \{0,1\}^r$ the space of binary encoded targets in a problem with $r$ categories. Concept-based classifiers $f$ are a family of machine learning models predicting class memberships from the activation scores of $k$ human-understandable categories, $f : C \mapsto Y$, where $C \subset [0,1]^k$ (see Fig. 1). Concept-based classifiers improve human understanding as their input and output spaces consists of interpretable symbols. When observations are represented in terms of non-interpretable input features belonging to $X \subset \mathbb{R}^d$ (such as pixels intensities), a "concept decoder" $g$ is used to map the input into a concept-based space, $g : X \mapsto C$ (see Fig. 1). Otherwise, they are simply rescaled from the unbounded space $\mathbb{R}^d$ into the unit interval $[0,1]^k$, such that input features can be treated as logic predicates.

In the recent literature, the most similar method related to the proposed approach is the $\psi$ network proposed by Ciravegna et al. (Ciravegna et al. 2020a,b), an end-to-end differentiable concept-based classifier *explaining its own decision process*. The $\psi$ network leverages the intermediate symbolic layer whose output belongs to $C$ to distill First-Order Logic formulas, representing the learned map from $C$ to $Y$. The model consists of a sequence of fully connected layers with sigmoid activations only. An $L1$-regularization and a strong pruning strategy is applied to each layer of weights in order to allow the computation of logic formulas representing the activation of each node. Such constraints, however, limit the learning capacity of the network and impair the classification accuracy, making standard white-box models, such as decision trees, more attractive.

## 3 Entropy-Based Logic Explanations of Neural Networks

The key contribution of this paper is a novel linear layer enabling entropy-based logic explanations of neural networks (see Fig. 2 and Fig. 3). The layer input belongs to the concept space $C$ and the outcomes of the layer computations are: (i) the embeddings $h^i$ (as any linear layer), (ii) a truth table $\mathcal{T}^i$ explaining how the network leveraged concepts to make predictions for the $i$-th target class. Each class of the problem requires an independent entropy-based layer, as emphasized

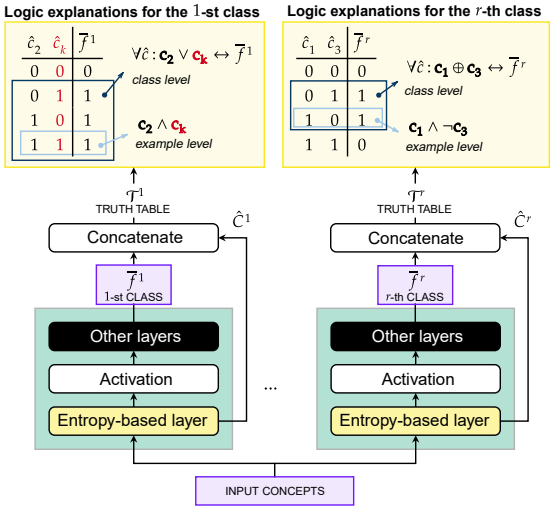


Figure 2: For each class $i$, the network leverages one "head" of the entropy-based linear layer (green) as first layer, and it provides: the class membership predictions $f^i$ and the truth table $\mathcal{T}^i$ (Eq. 6) to distill FOL explanations (yellow, top).

by the superscript $i$. For ease of reading and without loss of generality, all the following descriptions concern inference for a single observation (corresponding to the concept tuple $c \in C$) and a neural network $f^i$ predicting the class memberships for the $i$-th class of the problem. For multi-class problems, multiple "heads" of this layer are instantiated, with one "head" per target class (see Sec. 5), and the hidden layers of the class-specific networks could be eventually shared.

### 3.1 Entropy-Based Linear Layer

When humans compare a set of hypotheses outlining the same outcomes, they tend to have an implicit bias towards the simplest ones as outlined in philosophy (Soklakov 2002; Rathmanner and Hutter 2011), psychology (Miller 1956; Cowan 2001), and decision making (Simon 1956, 1957, 1979). The proposed entropy-based approach encodes this inductive bias in an end-to-end differentiable model. The purpose of the entropy-based linear layer is to encourage the neural model to pick a limited subset of input concepts, al-

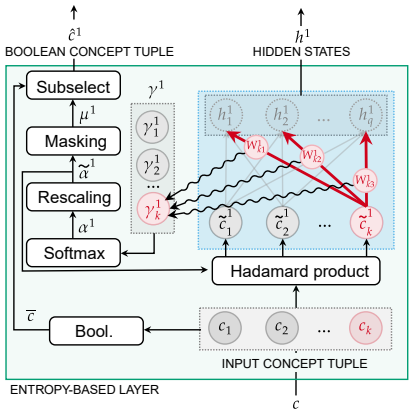


Figure 3: A detailed view on one "head" of the entropy-based linear layer for the 1-st class, emphasizing the role of the $k$-th input concept as example: (i) the scalar $\gamma_k^1$ (Eq. 1) is computed from the set of weights connecting the $k$-th input concept to the output neurons of the entropy-based layer; (ii) the relative importance of each concept is summarized by the categorical distribution $\alpha^1$ (Eq. 2); (iii) rescaled relevance scores $\tilde{\alpha}^1$ drop irrelevant input concepts out (Eq. 3); (iv) hidden states $h^1$ (Eq. 4) and Boolean-like concepts $\hat{c}^1$ (Eq. 5) are provided as outputs of the entropy-based layer.

lowing it to provide concise explanations of its predictions. The learnable parameters of the layer are the usual weight matrix $W$ and bias vector $b$. In the following, the forward pass is described by the operations going from Eq. 1 to Eq. 4 while the generation of the truth tables from which explanations are extracted is formalized by Eq. 5 and Eq. 6.

The relevance of each input concept can be summarized in a first approximation by a measure that depends on the values of the weights connecting such concept to the upper network. In the case of network $f^i$ (i.e. predicting the $i$-th class) and of the $j$-th input concept, we indicate with $W_j^i$ the vector of weights departing from the $j$-th input (see Fig. 3), and we introduce

$$\gamma_j^i = ||W_j^i||_1 \ . \tag{1}$$

The higher $\gamma_j^i$, the higher the relevance of the concept $j$ for the network $f^i$. In the limit case ($\gamma_j^i \to 0$) the model $f^i$ drops the $j$-th concept out. To select only few relevant concepts for each target class, concepts are set up to compete against each other. To this aim, the relative importance of each concept to the $i$-th class is summarized in the categorical distribution $\alpha^i$, composed of coefficients $\alpha_j^i \in [0,1]$ (with $\sum_j \alpha_j^i = 1$), modeled by the softmax function:

$$\alpha_j^i = \frac{e^{\gamma_j^i/\tau}}{\sum_{l=1}^k e^{\gamma_l^i/\tau}} \tag{2}$$

where $\tau \in \mathbb{R}^+$ is a user-defined temperature parameter to tune the softmax function. For a given set of $\gamma_j^i$, when using high temperature values ($\tau \to \infty$) all concepts have nearly the same relevance. For low temperatures values ($\tau \to 0$), the probability of the most relevant concept tends to $\alpha_j^i \approx 1$, while it becomes $\alpha_k^i \approx 0,\ k \neq j$, for all other concepts. For further details on the impact of $\tau$ on the model predictions and explanations (see Appendix). As the probability distribution $\alpha^i$ highlights the most relevant concepts, this information is directly fed back to the input, weighting concepts by the estimated importance. To avoid numerical cancellation due to values in $\alpha^i$ close to zero, especially when the input dimensionality is large, we replace $\alpha^i$ with its normalized instance $\tilde{\alpha}^i$, still in $[0,1]^k$, and each input sample $c \in C$ is modulated by this estimated importance,

$$\tilde{c}^i = c \odot \tilde{\alpha}^i \qquad \text{with} \qquad \tilde{\alpha}_j^i = \frac{\alpha_j^i}{\max_u \alpha_u^i}, \tag{3}$$

where $\odot$ denotes the Hadamard (element-wise) product. The highest value in $\tilde{\alpha}^i$ is always 1 (i.e. $\max_j \tilde{\alpha}_j^i = 1$) and it corresponds to the most relevant concept. The embeddings $h^i$ are computed as in any linear layer by means of the affine transformation:

$$h^i = W^i \tilde{c}^i + b^i. \tag{4}$$

Whenever $\tilde{\alpha}_j^i \to 0$, the input $\tilde{c}_j^i \to 0$. This means that the corresponding concept tends to be dropped out and the network $f^i$ will learn to predict the $i$-th class without relying on the $j$-th concept.

In order to get logic explanations, the proposed linear layer generates the truth table $\mathcal{T}^i$ formally representing the behaviour of the neural network in terms of Boolean-like representations of the input concepts. In detail, we indicate with $\bar{c}$ the Boolean interpretation of the input tuple $c \in C$, while $\mu^i \in \{0,1\}^k$ is the binary mask associated to $\tilde{\alpha}^i$. To encode the inductive human bias towards simple explanations (Miller 1956; Cowan 2001; Ma, Husain, and Bays 2014), the mask $\mu^i$ is used to generate the binary concept tuple $\hat{c}^i$, dropping the least relevant concepts out of $c$,

$$\hat{c}^i = \xi(\bar{c}, \mu^i) \quad \text{with} \quad \mu^i = \mathbb{I}_{\tilde{\alpha}^i \geq \epsilon} \quad \text{and} \quad \bar{c} = \mathbb{I}_{c \geq \epsilon}, \tag{5}$$

where $\mathbb{I}_{z \geq \epsilon}$ denotes the indicator function that is 1 for all the components of vector $z$ being $\geq \epsilon$ and 0 otherwise (considering the unbiased case, we set $\epsilon = 0.5$). The function $\xi$ returns the vector with the components of $\bar{c}$ that correspond to 1's in $\mu^i$ (i.e. it sub-selects the data in $\bar{c}$). As a results, $\hat{c}^i$ belongs to a space $\hat{C}^i$ of $m_i$ Boolean features, with $m_i < k$ due to the effects of the subselection procedure.

The truth table $\mathcal{T}^i$ is a particular way of representing the behaviour of network $f^i$ based on the outcomes of processing multiple input samples collected in a generic dataset $\mathcal{C}$. As the truth table involves Boolean data, we denote with $\hat{\mathcal{C}}^i$ the set with the Boolean-like representations of the samples in $\mathcal{C}$ computed by $\xi$, Eq. 5. We also introduce $\bar{f}^i(c)$ as the Boolean-like representation of the network output, $\bar{f}^i(c) = \mathbb{I}_{f^i(c) \geq \epsilon}$. The truth table $\mathcal{T}^i$ is obtained by stacking data of $\hat{\mathcal{C}}^i$ into a 2D matrix $\hat{\mathbf{C}}^i$ (row-wise), and concatenating the result with the column vector $\bar{\mathbf{f}}^i$ whose elements are $\bar{f}^i(c)$, $c \in \mathcal{C}$, that we summarize as

$$\mathcal{T}^i = \left( \hat{\mathbf{C}}^i \,\middle\|\, \bar{\mathbf{f}}^i \right). \tag{6}$$

To be precise, any $\mathcal{T}^i$ is more like an empirical truth table than a classic one corresponding to an $n$-ary boolean function, indeed $\mathcal{T}^i$ can have repeated rows and missing Boolean

tuple entries. However, $\mathcal{T}^i$ can be used to generate logic explanations in the same way, as we will explain in Sec. 3.3.

### 3.2 Loss Function

The entropy of the probability distribution $\alpha^i$ (Eq. 2),

$$\mathcal{H}(\alpha^i) = -\sum_{j=1}^{k} \alpha_j^i \log \alpha_j^i \tag{7}$$

is minimized when a single $\alpha_j^i$ is one, thus representing the extreme case in which only one concept matters, while it is maximum when all concepts are equally important. When $\mathcal{H}$ is jointly minimized with the usual loss function for supervised learning $L(f, y)$ (being $y$ the target labels–we used the cross-entropy in our experiments), it allows the model to find a trade off between fitting quality and a parsimonious activation of the concepts, allowing each network $f^i$ to predict $i$-th class memberships using few relevant concepts only. Overall, the loss function to train the network $f$ is defined as,

$$\mathcal{L}(f, y, \alpha_1, \ldots, \alpha_r) = L(f, y) + \lambda \sum_{i=1}^{r} \mathcal{H}(\alpha^i), \tag{8}$$

where $\lambda > 0$ is the hyperparameter used to balance the relative importance of low-entropy solutions in the loss function. Higher values of $\lambda$ lead to sparser configuration of $\alpha$, constraining the network to focus on a smaller set of concepts for each classification task (and vice versa), thus encoding the inductive human bias towards simple explanations (Miller 1956; Cowan 2001; Ma, Husain, and Bays 2014). For further details on the impact of $\lambda$ on the model predictions and explanations (see Appendix). It may be pointed out that a similar regularization effect could be achieved by simply minimizing the $L_1$ norm over $\gamma^i$. However, as we observed in the Appendix, the $L_1$ loss does not sufficiently penalize the concept scores for those features which are uncorrelated with the predicted category. The Entropy loss, instead, correctly shrink to zero concept scores associated to uncorrelated features while the other remains close to one.

### 3.3 First-Order Logic Explanations

Any Boolean function can be converted into a logic formula in Disjunctive Normal Form (DNF) by means of its truth-table (Mendelson 2009). Converting a truth table into a DNF formula provides an effective mechanism to extract logic rules of increasing complexity from individual observations to a whole class of samples. The following rule extraction mechanism is applied to any empirical truth table $\mathcal{T}^i$ for each task $i$.

**FOL extraction.** Each row of the truth table $\mathcal{T}^i$ can be partitioned into two parts that are a tuple of binary concept activations, $\hat{q} \in \hat{C}^i$, and the outcome of $\bar{f}^i(\hat{q}) \in \{0, 1\}$. An *example-level* logic formula, consisting in a single minterm, can be trivially extracted from each row for which $\bar{f}^i(\hat{q}) = 1$, by simply connecting with the logic AND ($\wedge$) the true concepts and negated instances of the false ones. The logic formula becomes human understandable whenever concepts appearing in such a formula are replaced with human-interpretable strings that represent their name (similar consideration holds for $\bar{f}^i$, in what follows). For example, the following logic formula $\varphi_t^i$,

$$\varphi_t^i = \mathbf{c}_1 \wedge \neg \mathbf{c}_2 \wedge \ldots \wedge \mathbf{c}_{m_i}, \tag{9}$$

is the formula extracted from the $t$-th row of the table where, in the considered example, only the second concept is false, being $\mathbf{c}_z$ the name of the $z$-th concept. Example-level formulas can be aggregated with the logic OR ($\vee$) to provide a *class-level* formula,

$$\bigvee_{t \in S_i} \varphi_t^i, \tag{10}$$

being $S_i$ the set of rows indices of $\mathcal{T}^i$ for which $\bar{f}^i(\hat{q}) = 1$, i.e. it is the support of $\bar{f}^i$. We define with $\phi^i(\hat{c})$ the function that holds true whenever Eq. 10, evaluated on a given Boolean tuple $\hat{c}$, is true. Due to the aforementioned definition of support, we get the following class-level First-Order Logic (FOL) explanation for all the concept tuples,

$$\forall \hat{c} \in \hat{C}^i : \ \phi^i(\hat{c}) \leftrightarrow \bar{f}^i(\hat{c}). \tag{11}$$

We note that in case of non-concept-like input features, we may still derive the FOL formula through the "concept decoder" function $g$ (see Sec. 2),

$$\forall x \in X : \ \phi^i\left(\xi(\overline{g(x)}, \mu^i)\right) \leftrightarrow \bar{f}^i\left(\xi(\overline{g(x)}, \mu^i)\right) \tag{12}$$

An example of the above scheme for both example and class-level explanations is depicted on top-right of Fig. 2.

**Remarks.** The aggregation of many example-level explanations may increase the length and the complexity of the FOL formula being extracted for a whole class. However, existing techniques as the Quine–McCluskey algorithm can be used to get compact and simplified equivalent FOL expressions (McColl 1878; Quine 1952; McCluskey 1956). For instance, the explanation (*person* $\wedge$ *nose*) $\vee$ ($\neg$*person* $\wedge$ *nose*) can be formally simplified in *nose*. Moreover, the Boolean interpretation of concept tuples may generate colliding representations for different samples. For instance, the Boolean representation of the two samples $\{(0.1, 0.7), (0.2, 0.9)\}$ is the tuple $\bar{c} = (0, 1)$ for both of them. This means that their example-level explanations match as well. However, a concept can be eventually split into multiple finer grain concepts to avoid collisions. Finally, we mention that the number of samples for which any example-level formula holds (i.e. the support of the formula) is used as a measure of the explanation importance. In practice, example-level formulas are ranked by support and iteratively aggregated to extract class-level explanations, until the aggregation improves the accuracy of the explanation over a validation set.

## 4 Related Work

In order to provide explanations for a given black-box model, most methods focus on identifying or scoring the most relevant input features (Simonyan, Vedaldi, and Zisserman 2013; Zeiler and Fergus 2014; Ribeiro, Singh, and Guestrin 2016b,a; Lundberg and Lee 2017; Selvaraju et al.

2017). Feature scores are usually computed sample by sample (i.e. providing *local explanations*) analyzing the activation patterns in the hidden layers of neural networks (Simonyan, Vedaldi, and Zisserman 2013; Zeiler and Fergus 2014; Selvaraju et al. 2017) or by following a model-agnostic approach (Ribeiro, Singh, and Guestrin 2016a; Lundberg and Lee 2017). To enhance human understanding of feature scoring methods, concept-based approaches have been effectively employed for identifying common activations patterns in the last nodes of neural networks corresponding to human categories (Kim et al. 2018; Kazhdan et al. 2020) or constraining the network to learn such concepts (Chen, Bei, and Rudin 2020; Koh et al. 2020). Either way, feature-scoring methods are not able to explain *how* neural networks compose features to make predictions (Kindermans et al. 2019; Kim et al. 2018; Alvarez-Melis and Jaakkola 2018) and only a few of these approaches have been efficiently extended to provide explanations for a whole class (i.e. providing *global explanations*) (Simonyan, Vedaldi, and Zisserman 2013; Ribeiro, Singh, and Guestrin 2016a). By contrast, a variety of rule-based approaches have been proposed to provide concept-based explanations. Logic rules are used to explain how black boxes predict class memberships for indivudal samples (Guidotti et al. 2018; Ribeiro, Singh, and Guestrin 2018), or for a whole class (Sato and Tsukimoto 2001; Zilke, Loza Mencía, and Janssen 2016; Ciravegna et al. 2020a,b). Distilling explanations from an existing model, however, is not the only way to achieve explainability. Historically, standard machine-learning such as Logistic Regression (McKelvey and Zavoina 1975), Generalized Additive Models (Hastie and Tibshirani 1987; Lou, Caruana, and Gehrke 2012; Caruana et al. 2015) Decision Trees (Breiman et al. 1984; Quinlan 1986, 2014) and Decision Lists (Rivest 1987; Letham et al. 2015; Angelino et al. 2018) were devised to be intrinsically interpretable. However, most of them struggle in solving complex classification problems. Logistic Regression, for instance, in its vanilla definition, can only recognize linear patterns, e.g. it cannot to solve the XOR problem (Minsky and Papert 2017). Further, only Decision Trees and Decision Lists provide explanations in the from of logic rules. Considering decision trees, each path may be seen as a human comprehensible decision rule when the height of the tree is reasonably contained. Another family of concept-based XAI methods is represented by rule-mining algorithms which became popular at the end of the last century (Holte 1993; Cohen 1995). Recent research has led to powerful rule-mining approaches as Bayesian Rule Lists (BRL) (Letham et al. 2015), where a set of rules is "pre-mined" using the frequent-pattern tree mining algorithm (Han, Pei, and Yin 2000) and then the best rule set is identified with Bayesian statistics. In this paper, the proposed approach is compared with methods providing logic-based, global explanations. In particular, we selected one representative approach from different families of methods: Decision Trees (white-box, https://scikit-learn.org/stable/modules/tree), BRL (rule mining, https://github.com/tmadl/sklearn-expertsys) and $\psi$ Networks (explainable neural models, https://github.com/pietrobarbiero/logic_explainer_networks).

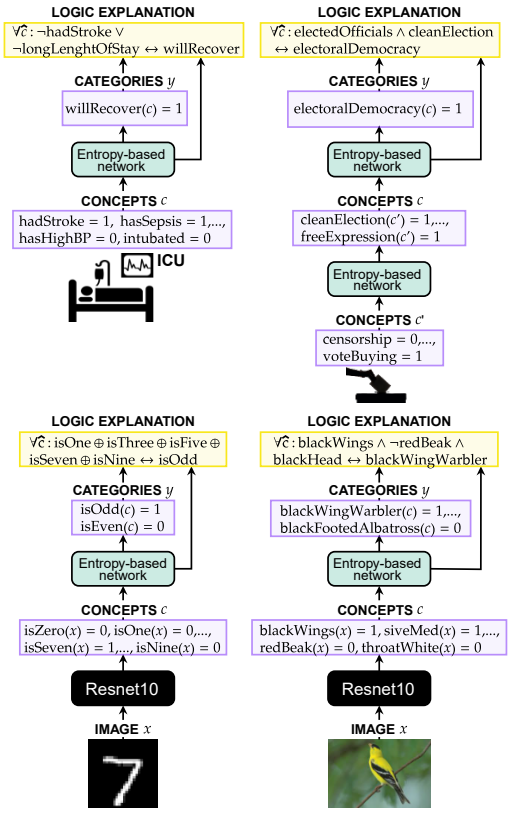


Figure 4: The four case studies show how the proposed Entropy-based networks (green) provide concise logic explanations (yellow) of their own decision process in different real-world contexts. When input features are non-interpretable, as pixel intensities, a "concept decoder" (ResNet10) maps images into concepts. Entropy-based networks then map concepts into target classes.

## 5 Experiments

The quality of the explanations and the classification performance of the proposed approach are quantitatively assessed and compared to state-of-the-art white-box models. A visual sketch of each classification problem (described in detail in Sec. 5.1) and a selection of the logic formulas found by the proposed approach is reported in Fig. 4. Six quantitative metrics are defined and used to compare the proposed approach with state-of-the-art methods. Sec. 5.2 summarizes the main findings. Further details concerning the experiments are reported in the Appendix.

### 5.1 Classification Tasks and Datasets

Four classification problems ranging from computer vision to medicine are considered. Computer vision datasets (e.g. CUB) are annotated with low-level concepts (e.g. bird attributes) used to train concept bottleneck pipelines (Koh et al. 2020). In the other datasets, the input data is rescaled into a categorical space ($\mathbb{R}^k \rightarrow C$) suitable for concept-based networks. Please notice that this preprocessing step is performed for all white-box models considered in the ex-

| | Entropy net | Tree | BRL | $\psi$ net | Neural Network | Random Forest |
|---|---|---|---|---|---|---|
| **MIMIC-II** | $\mathbf{79.05^{\star} \pm 1.35}$ | $77.53 \pm 1.45$ | $76.40 \pm 1.22$ | $77.19 \pm 1.64$ | $77.81 \pm 2.45$ | $\mathbf{78.88 \pm 2.25}$ |
| **V-Dem** | $\mathbf{94.51 \pm 0.48}$ | $85.61 \pm 0.57$ | $91.23 \pm 0.75$ | $89.77 \pm 2.07$ | $\mathbf{94.53^{\star} \pm 1.17}$ | $93.08 \pm 0.44$ |
| **MNIST** | $\mathbf{99.81 \pm 0.02}$ | $99.75 \pm 0.01$ | $99.80 \pm 0.02$ | $99.79 \pm 0.03$ | $99.72 \pm 0.03$ | $\mathbf{99.96^{\star} \pm 0.01}$ |
| **CUB** | $\mathbf{92.95 \pm 0.20}$ | $81.62 \pm 1.17$ | $90.79 \pm 0.34$ | $91.92 \pm 0.27$ | $\mathbf{93.10^{\star} \pm 0.51}$ | $91.88 \pm 0.36$ |

Table 1: Classification accuracy (%). Left group, the compared white-box models. Right group, two black box models. We indicate in bold the best model in each group, with a star the best model overall.

periments for a fair comparison. Further descriptions of each dataset and links to all sources are reported in Appendix.

**Will we recover from ICU? (MIMIC-II).** The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II, (Saeed et al. 2011; Goldberger et al. 2000)) is a public-access intensive care unit (ICU) database consisting of 32,536 subjects (with 40,426 ICU admissions) admitted to different ICUs. The task consists in identifying recovering or dying patients after ICU admission. An end-to-end classifier $f : C \rightarrow Y$ carries out the classification task.

**What kind of democracy are we living in? (V-Dem).** Varieties of Democracy (V-Dem, (Pemstein et al. 2018; Coppedge et al. 2021)) dataset contains a collection of indicators of latent regime characteristics over 202 countries from 1789 to 2020. The database include $k_1 = 483$ low-level indicators $k_2 = 82$ mid-level indices. The task consists in identifying electoral democracies from non-electoral ones. We indicate with $C_1$, $C_2$ the spaces associated to the activations of the two levels of concepts. Classifiers $f_1$ and $f_2$ are trained to learn the map $C_1 \rightarrow C_2 \rightarrow Y$. Explanations are given for classifier $f_2$ in terms of concepts $c_2 \in C_2$.

**What does parity mean? (MNIST Even/Odd).** The Modified National Institute of Standards and Technology database (MNIST, (LeCun 1998)) contains a large collection of images representing handwritten digits. The task we consider here is slightly different from the common digit-classification. Assuming $Y \subset \{0, 1\}^2$, we are interested in determining if a digit is either odd or even, and explaining the assignment to one of these classes in terms of the digit labels (concepts in $C$). The mapping $X \rightarrow C$ is provided by a ResNet10 classifier $g$ (He et al. 2016) trained from scratch. while the classifier $f$ learn both the final mapping and the explanation as a function $C \rightarrow Y$.

**What kind of bird is that? (CUB).** The Caltech-UCSD Birds-200-2011 dataset (CUB, (Wah et al. 2011)) is a fine-grained classification dataset. It includes 11,788 images representing $r = 200$ ($Y = \{0, 1\}^{200}$) different bird species. 312 binary attributes (concepts in $C$) describe visual characteristics (color, pattern, shape) of particular parts (beak, wings, tail, etc.) for each bird image. The mapping $X \rightarrow C$ is performed with a ResNet10 model $g$ trained from scratch while the classifier $f$ learns the final function $C \rightarrow Y$.

**Quantitative metrics.** Measuring the classification quality is of crucial importance for models that are going to be applied in real-world environments. On the other hand, assessing the quality of the explanations is required for a safedeployment. In contrast with other kind of explanations, logic-based formulas can be evaluated quantitatively. Given a classification problem, first a set of rules are extracted for each target category from each considered model. Each explanation is then tested on an unseen set of test samples. The results for each metric are reported in terms of mean and standard error, computed over a 5-fold cross validation (Krzywinski and Altman 2013). For each experiment and for each model model ($f : C \rightarrow Y$ mapping concepts to target categories) six quantitative metrics are measured. (i) The MODEL ACCURACY measures how well the explainer identifies the target classes on unseen data (see Table 1). (ii) The EXPLANATION ACCURACY measures how well the extracted logic formulas identifies the target classes (Fig. 5). This metric is obtained as the average of the F1 scores computed for each class explanation. (iii) The COMPLEXITY OF AN EXPLANATION is computed by standardizing the explanations in DNF and then by counting the number of terms of the standardized formula (Fig. 5): the longer the formula, the harder the interpretation for a human being. (iv) The FIDELITY OF AN EXPLANATION measures how well the extracted explanation matches the predictions obtained using the explainer (Table 2). (v) The RULE EXTRACTION TIME measures the time required to obtain an explanation from scratch (see Fig. 6), computed as the sum of the time required to train the model and to extract the formula from a trained explainer. (vi) The CONSISTENCY OF AN EXPLANATION measures the average similarity of the extracted explanations over the 5-fold cross validation runs (see Table 3), computed by counting how many times the same concepts appear in a logic formula over different iterations.

## 5.2 Results and Discussion

Experiments show how entropy-based networks outperform state-of-the-art white box models such as BRL and decision trees and interpretable neural models such as $\psi$ networks on challenging classification tasks (Table 1). Moreover, the entropy-based regularization and the adoption of a concept-based neural network have minor affects on the classification accuracy of the explainer when compared to a standard black box neural network directly working on the input data, and a Random Forest model applied on the concepts.At the same time, the logic explanations provided by entropy-based networks are better than $\psi$ networks and almost as accurate as the rules found by decision trees and BRL, while being far more concise, as demonstrated in Fig. 5. More precisely, logic explanations generated by the proposed approach represent non-dominated solutions (Marler and Arora 2004) *quantitatively* measured in terms of complexity and classification error of the explanation. Furthermore, the time

| | Entropy net | $\psi$ net |
|---|---|---|
| **MIMIC-II** | $\mathbf{79.11 \pm 2.02}$ | $51.63 \pm 6.67$ |
| **V-Dem** | $\mathbf{90.90 \pm 1.23}$ | $69.67 \pm 10.43$ |
| **MNIST** | $\mathbf{99.63 \pm 0.00}$ | $65.68 \pm 5.05$ |
| **CUB** | $\mathbf{99.86 \pm 0.01}$ | $77.34 \pm 0.52$ |

Table 2: Out-of-distribution fidelity (%)

| | Entropy net | Tree | BRL | $\psi$ net |
|---|---|---|---|---|
| **MIMIC-II** | 28.75 | **40.49** | 30.48 | 27.62 |
| **V-Dem** | 46.25 | 72.00 | **73.33** | 38.00 |
| **MNIST** | **100.00** | 41.67 | **100.00** | 96.00 |
| **CUB** | 35.52 | 21.47 | **42.86** | 41.43 |

Table 3: Consistency (%)

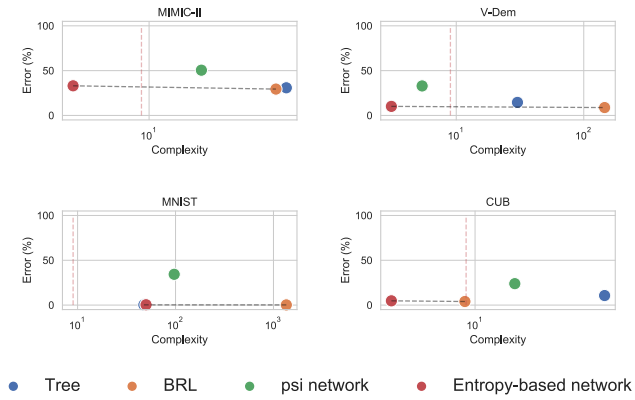


Figure 5: Non-dominated solutions (Marler and Arora 2004) (dotted black line) in terms of average explanation complexity and average explanation test error. The vertical dotted red line marks the maximum explanation complexity laypeople can handle (i.e. complexity $\approx$ 9, see (Miller 1956; Cowan 2001; Ma, Husain, and Bays 2014)). Notice how the explanations provided by the Entropy-based Network are always one of the non-dominated solution.

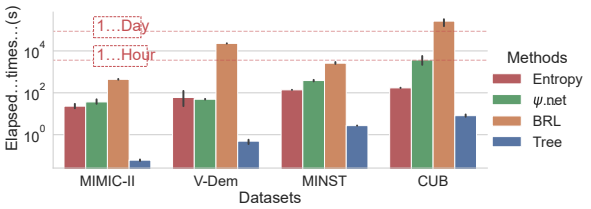


Figure 6: Time required to train models and to extract the explanations. Our model compares favorably with the competitors, with the exception of Decision Trees. BRL is by one to three order of magnitude slower than our approach.

required to train entropy-based networks is only slightly higher with respect to Decision Trees but is lower than $\psi$ Networks and BRL by one to three orders of magnitude (Fig. 6), making it feasible for explaining also complex tasks. The fidelity (Table 2) of the formulas extracted by the entropy-based network is always higher than 90% with the only exception of MIMIC. This means that almost any prediction made using the logic explanation matches the corresponding prediction made by the model, making the proposed approach very close to a white box model. These results empirically shows that our method represents a viable solution for a safe deployment of *explainable* cutting-edge models.

The reason why the proposed approach consistently outperform $\psi$ networks across all the key metrics (i.e. classification accuracy, explanation accuracy, and fidelity) can be explained observing how entropy-based networks are far less constrained than $\psi$ networks, both in the architecture (our approach does not apply weight pruning) and in the loss function (our approach applies a regularization on the distributions $\alpha^i$ and not on all weight matrices). Likewise, the main reason why the proposed approach provides a higher classification accuracy with respect to BRL and decision trees may lie in the smoothness of the decision functions of neural networks which tend to generalize better than rule-based methods, as already observed by Tavares et al. (Tavares et al. 2020). For each dataset, we report in the Appendix a few examples of logic explanations extracted by each method, as well as in Fig. 4. We mention that the proposed approach is the only matching the ground-truth explanation for the MNIST even/odd experiment, i.e. $\forall x, \text{isOdd}(x) \leftrightarrow \text{isOne}(x) \oplus \text{isThree}(x) \oplus \text{isFive}(x) \oplus \text{isSeven}(x) \oplus \text{isNine}(x)$ and $\forall x, \text{isEven}(x) \leftrightarrow \text{isZero}(x) \oplus \text{isTwo}(x) \oplus \text{isfour}(x) \oplus \text{isSix}(x) \oplus \text{isEight}(x)$, being $\oplus$ the exclusive OR. In terms of formula consistency, we observe how BRL is the most consistent rule extractor, closely followed by the proposed approach (Table 3).

## 6 Conclusions

This work contributes to a safer adoption and greater impact of deep learning by making neural models explainable-by-design, thanks to an entropy-based approach that yields FOL-based explanations. Moreover, as the proposed approach provides logic explanations for how a model arrives at a decision, it can be effectively used to reverse engineer algorithms, processes, to find vulnerabilities, or to improve system design powered by deep learning models. From a scientific perspective, formal knowledge distillation from state-of-the-art networks may enable scientific discoveries or falsification of existing theories. However, the extraction of a FOL explanation requires symbolic input and output spaces. In some contexts, such as computer vision, the use of concept-based approaches may require additional annotations and attribute labels to get a consistent symbolic layer of concepts. Recent works on automatic concept extraction may alleviate the related costs, leading to more cost-effective concept annotations (Ghorbani et al. 2019; Kazhdan et al. 2020).

## Acknowledgments

This work was partially supported by TAILOR and GODS-21 European Union's Horizon 2020 research and innovation programmes under GA No 952215 and 848077.

# References


Ahmad, M. A.; Eckert, C.; and Teredesai, A. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 559–560.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*.

Angelino, E.; Larus-Stone, N.; Alabi, D.; Seltzer, M.; and Rudin, C. 2018. Learning Certifiably Optimal Rule Lists for Categorical Data. arXiv:1704.01701.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Barbiero, P.; Ciravegna, G.; Georgiev, D.; and Giannini, F. 2021. PyTorch, Explain! A Python library for Logic Explained Networks. *arXiv preprint arXiv:2105.11697*.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. A. 1984. *Classification and regression trees*. CRC press.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1721–1730.

Chen, Z.; Bei, Y.; and Rudin, C. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12): 772–782.

Ciravegna, G.; Giannini, F.; Gori, M.; Maggini, M.; and Melacci, S. 2020a. Human-driven FOL explanations of deep learning. In *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}*, 2234–2240. International Joint Conferences on Artificial Intelligence Organization.

Ciravegna, G.; Giannini, F.; Melacci, S.; Maggini, M.; and Gori, M. 2020b. A Constraint-Based Approach to Learning and Explanation. In *AAAI*, 3658–3665.

Cohen, W. W. 1995. Fast effective rule induction. In *Machine learning proceedings 1995*, 115–123. Elsevier.

Coppedge, M.; Gerring, J.; Knutsen, C. H.; Lindberg, S. I.; Teorell, J.; Altman, D.; Bernhard, M.; Cornell, A.; Fish, M. S.; Gastaldi, L.; et al. 2021. V-Dem Codebook v11.

Cowan, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1): 87–114.

Cranmer, M. D.; Xu, R.; Battaglia, P.; and Ho, S. 2019. Learning symbolic physics with graph networks. *arXiv preprint arXiv:1909.05862*.

Das, A.; and Rad, P. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *ArXiv*, abs/2006.11371.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Doshi-Velez, F.; and Kim, B. 2018. Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and interpretable models in computer vision and machine learning*, 3–17. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

EUGDPR. 2017. GDPR. General data protection regulation. https://gdpr.eu/. Accessed: 2021-08-20.

Ghorbani, A.; Wexler, J.; Zou, J.; and Kim, B. 2019. Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129*.

Goddard, M. 2017. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6): 703–705.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; and Giannotti, F. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*.

Han, J.; Pei, J.; and Yin, Y. 2000. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2): 1–12.

Hastie, T.; and Tibshirani, R. 1987. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398): 371–386.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Holte, R. C. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1): 63–90.

Kazhdan, D.; Dimanov, B.; Jamnik, M.; Liò, P.; and Weller, A. 2020. Now You See Me (CME): Concept-based Model Extraction. *arXiv preprint arXiv:2010.13233*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2019. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 267–280. Springer.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, 5338–5348. PMLR.

Krzywinski, M.; and Altman, N. 2013. Error bars: the meaning of error bars is often misinterpreted, as is the statistical significance of their overlap. *Nature methods*, 10(10): 921–923.

LeCun, Y. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Letham, B.; Rudin, C.; McCormick, T. H.; Madigan, D.; et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3): 1350–1371.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lou, Y.; Caruana, R.; and Gehrke, J. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 150–158.

Lundberg, S.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Ma, W. J.; Husain, M.; and Bays, P. M. 2014. Changing concepts of working memory. *Nature neuroscience*, 17(3): 347.

Marler, R. T.; and Arora, J. S. 2004. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6): 369–395.

McCluskey, E. J. 1956. Minimization of Boolean functions. *The Bell System Technical Journal*, 35(6): 1417–1444.

McColl, H. 1878. The calculus of equivalent statements (third paper). *Proceedings of the London Mathematical Society*, 1(1): 16–28.

McKelvey, R. D.; and Zavoina, W. 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4(1): 103–120.

Mendelson, E. 2009. *Introduction to mathematical logic*. CRC press.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63: 81–97.

Minsky, M.; and Papert, S. A. 2017. *Perceptrons: An introduction to computational geometry*. MIT press.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825–2830.

Pemstein, D.; Marquardt, K. L.; Tzelgov, E.; Wang, Y.-t.; Krusell, J.; and Miri, F. 2018. The V-Dem measurement model: latent variable analysis for cross-national and cross-temporal expert-coded data. *V-Dem Working Paper*, 21.

Quine, W. V. 1952. The problem of simplifying truth functions. *The American mathematical monthly*, 59(8): 521–531.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning*, 1(1): 81–106.

Quinlan, J. R. 2014. *C4. 5: programs for machine learning*. Elsevier.

Rathmanner, S.; and Hutter, M. 2011. A philosophical treatise of universal induction. *Entropy*, 13(6): 1076–1136.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016a. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016b. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

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.

Rivest, R. L. 1987. Learning decision lists. *Machine learning*, 2(3): 229–246.

Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; and Zhong, C. 2021. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*.

Saeed, M.; Villarroel, M.; Reisner, A. T.; Clifford, G.; Lehman, L.-W.; Moody, G.; Heldt, T.; Kyaw, T. H.; Moody, B.; and Mark, R. G. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, 39(5): 952.

Sato, M.; and Tsukimoto, H. 2001. Rule extraction from neural networks via decision tree induction. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 3, 1870–1875. IEEE.

Schmidt, M.; and Lipson, H. 2009. Distilling free-form natural laws from experimental data. *science*, 324(5923): 81–85.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simon, H. A. 1956. Rational choice and the structure of the environment. *Psychological review*, 63(2): 129.

Simon, H. A. 1957. *Models of man; social and rational.* New York: John Wiley and Sons, Inc.

Simon, H. A. 1979. Rational decision making in business organizations. *The American economic review*, 69(4): 493–513.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Soklakov, A. N. 2002. Occam's razor as a formal basis for a physical theory. *Foundations of Physics Letters*, 15(2): 107–135.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5): 828–841.

Tavares, A. R.; Avelar, P.; Flach, J. M.; Nicolau, M.; Lamb, L. C.; and Vardi, M. 2020. Understanding boolean function learnability on deep neural networks. *arXiv preprint arXiv:2009.05908*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10687–10698.

Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32: 9240.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zilke, J. R.; Loza Mencía, E.; and Janssen, F. 2016. DeepRED – Rule Extraction from Deep Neural Networks. In Calders, T.; Ceci, M.; and Malerba, D., eds., *Discovery Science*, 457–473. Cham: Springer International Publishing. ISBN 978-3-319-46307-0".