




# FFNSL: Feed-Forward Neural-Symbolic Learner

Daniel Cunnington<sup>1,2</sup>  · Mark Law<sup>2,3</sup> · Jorge Lobo<sup>4</sup> · Alessandra Russo<sup>2</sup>

Received: 25 January 2022 / Revised: 2 September 2022 / Accepted: 7 November 2022 /  
Published online: 23 January 2023  
© The Author(s) 2023

## Abstract

Logic-based machine learning aims to learn general, interpretable knowledge in a data-efficient manner. However, labelled data must be specified in a structured logical form. To address this limitation, we propose a neural-symbolic learning framework, called *Feed-Forward Neural-Symbolic Learner (FFNSL)*, that integrates a logic-based machine learning system capable of learning from noisy examples, with neural networks, in order to learn interpretable knowledge from labelled unstructured data. We demonstrate the generality of FFNSL on four neural-symbolic classification problems, where different pre-trained neural network models and logic-based machine learning systems are integrated to learn interpretable knowledge from sequences of images. We evaluate the robustness of our framework by using images subject to distributional shifts, for which the pre-trained neural networks may predict incorrectly and with high confidence. We analyse the impact that these shifts have on the accuracy of the learned knowledge and run-time performance, comparing FFNSL to tree-based and pure neural approaches. Our experimental results show that FFNSL outperforms the baselines by learning more accurate and interpretable knowledge with fewer examples.

**Keywords** Neural-symbolic learning · Inductive logic programming · Logic-based machine learning · Distributional shift

## 1 Introduction

Logic-based machine learning (Muggleton, 1991; Law et al., 2019) learns interpretable knowledge expressed in the form of a logic program, called a *hypothesis*, that explains

---

Editors: Nikos Katzouris, Alexander Artikis, Luc De Raedt, Artur d’Avila Garcez, Ute Schmid, and Jay Pujara.

---

✉ Daniel Cunnington  
dancunnington@uk.ibm.com

<sup>1</sup> IBM Research Europe, Winchester, UK

<sup>2</sup> Imperial College London, London, UK

<sup>3</sup> ILASP Limited, Grantham, UK

<sup>4</sup> ICREA-Universitat Pompeu Fabra, Barcelona, Spain

labelled examples in the context of (optional) background knowledge. Recent logic-based machine learning systems have demonstrated the ability to learn highly complex and noise-tolerant hypotheses in a data efficient manner [e.g., Learning from Answer Sets (LAS) (Law et al., 2019)]. However, they require labelled examples to be specified in a structured logical form, which limits their applicability to many real-world problems. On the other hand, differentiable learning systems, such as (deep) neural networks, are able to learn directly from unstructured data, but they require large amounts of training data and their learned models are difficult to interpret (Gilpin et al., 2018).

Within neural-symbolic artificial intelligence, many approaches aim to integrate neural and symbolic systems with the goal of preserving the benefits of both paradigms (Besold et al., 2017; Garcez & Lamb, 2020). Most neural-symbolic integrations assume the existence of pre-defined knowledge expressed symbolically, or logically, and focus on training a neural network to extract symbolic features from raw unstructured data (Manhaeve et al., 2018; Yang et al., 2020; Serafini & d’Avila Garcez, 2016; Cohen, 2016; Riegel et al., 2020). In this paper, we introduce *Feed-Forward Neural-Symbolic Learner (FFNSL)*, a neural-symbolic learning framework that assumes the opposite. Given a pre-trained neural network, FFNSL uses a logic-based machine learning system robust to noise to learn a logic-based hypothesis whose symbolic features are constructed from neural network predictions. The motivation is to enable logic-based machine learning systems to utilise *pre-trained* neural networks<sup>1</sup> to learn symbolic features from unstructured data, and use these features to learn interpretable knowledge needed to solve a downstream classification task. FFNSL preserves the benefits of both paradigms, increasing the scope of the tasks logic-based machine learning systems can be applied to. The challenge in performing such an integration, is that neural networks are vulnerable to *distributional shifts*, where unstructured data belonging to a distribution different from that used for training often leads to incorrect predictions (Ovadia et al., 2019; Sensoy et al., 2018; Amodei et al., 2016). By using a logic-based machine learning system that is robust to noise, such as a LAS system, FFNSL is capable of learning robust logic-based hypotheses from examples generated from labelled unstructured data, which may contain incorrect or noisy features as a result of incorrect neural network predictions.

The novel aspect of our FFNSL framework is the *Data-to-Knowledge (D2K) generator* that bridges the neural and symbolic learning components. The D2K generator automatically constructs a symbolic representation of the features predicted from the unstructured data, and weights such knowledge with a level of truthfulness that reflects the confidence score of the neural network predictions. The symbolic features can then be used by the symbolic learning component to automatically generate weighted examples from which to learn general and interpretable knowledge needed to solve the given downstream task.

FFNSL is general enough to support the integration of any neural component capable of making discrete predictions from unstructured data (binary or multi-class classification), with any logic-based machine learning system capable of learning from noisy examples. In this paper, we present four instances of our framework, where the LAS systems, *ILASP* (Law, 2018) and *FastLAS* (Law et al., 2020), are used as the symbolic learning component, and different neural network architectures are used as the neural component. The LAS systems have been shown to learn optimal hypotheses from noisy examples (Law et al., 2018), and to be suitable for different forms of symbolic learning tasks. In these systems, a noisy example includes a *weight*, which defines the penalty paid by a hypothesis for not covering that example. FFNSL interprets this weight as a level of certainty of the example, and computes it using the confidence score of the related neural network predictions. In this way, the LAS systems

<sup>1</sup> Examples include <https://modelzoo.co/> and those listed here: <https://github.com/collections/ai-model-zoos>.

become biased towards learning a hypothesis that has minimal penalty, i.e., a hypothesis that covers examples generated from high confidence neural network predictions (examples with high weights). For each proposed instance of our FFNSL framework, we investigate: (1) whether FFNSL can learn an accurate and interpretable hypothesis from incorrect feature predictions of the neural component, (2) how robust the learned hypothesis is in the presence of distributional shifts applied to an increasing percentage of the unstructured data, (3) the impact of using an uncertainty-aware neural network component that provides more robust confidence estimates when distributional shifts are applied to the unstructured data, and (4) how FFNSL performs in comparison to other hybrid systems where the same pre-trained neural networks, used for predicting features from the unstructured data, are integrated with a random forest and deep neural networks trained to learn the knowledge required to solve the downstream task.

To evaluate our FFNSL framework, we use four neural-symbolic classification tasks, one for each proposed instance.<sup>2</sup> Firstly, the *Follow Suit Winner* task is a card game where 4 players each play a card and the goal is to predict the winning player. In order to solve the task, the neural network predicts the rank and suit of the playing card images and the rules of the game are learned as symbolic knowledge, where the winner is the player that plays the highest ranked card with the same suit as player 1. The second task is *Sudoku Grid Validity* classification, which consists of observing a sequence of images of handwritten MNIST digits, corresponding to the digits in a Sudoku grid, and predicting if the grid is valid or not. The neural network classifies each digit and the symbolic knowledge required to be learned is the definition of valid (or invalid) Sudoku grids. The final tasks are *Crop Yield Prediction* and *Indoor Scene Classification*, which demonstrate the applicability of FFNSL to real-world problems and datasets. The Crop Yield Prediction task requires predicting the quality of crop yield from an image containing potentially diseased crops, where the neural network predicts the crop's species and disease status, and the learned symbolic knowledge predicts the quality of yield. In the Indoor Scene Classification task, the neural network is pre-trained to predict the scene class from an image, and the learned symbolic knowledge maps scene classes to high-level super-classes. In the first task, the neural network is pre-trained on images of playing cards from a standard deck, but our FFNSL framework is applied on card images subject to distributional shifts, where a percentage of standard card images are replaced with images from alternative card decks. In the second task, the neural network is pre-trained on the standard MNIST dataset, and our FFNSL system is applied on an out-of-distribution MNIST dataset generated by rotating MNIST digits 90° clockwise. In the Crop Yield Prediction task, we pre-train the neural network on the Plant Village dataset (Hughes & Salathé, 2015), and apply distributional shifts using a hue filter. Finally, in the Indoor Scene Classification task, we adopt a neural network model pre-trained on the MIT Indoor Scene dataset (Quattoni & Torralba, 2009), and apply distributional shift using blur, hue, and rotation filters.

Our evaluation demonstrates that FFNSL outperforms the baselines on all four tasks. The hypotheses learned from unstructured data, subject to distributional shifts, are more interpretable and more accurate than those learned by the random forest and deep neural networks even when these baselines are trained with significantly more data. We have also evaluated the robustness of the FFNSL instances when applied to a test set that is also subject to distributional shifts. The results show that FFNSL outperforms the baselines, trained with the same amount of unstructured data, when up to ~80% of the test set is subject to distributional shifts.

<sup>2</sup> See <https://github.com/DanCunnington/FFNSL> for code and data.

The paper is structured as follows. Section 2 provides necessary background material on the LAS framework, alongside further discussion of the drawbacks of the standard neural network Softmax layer for providing robust confidence estimates, and details of the uncertainty-aware neural networks used in this paper. Section 3 presents our general FFNSL framework followed by four instances discussed in detail in Sect. 4. We introduce our evaluation methodology in Sect. 5 and present the results of each FFNSL instance on the Follow Suit Winner and Sudoku Grid Validity tasks in Sects. 6 and 7 respectively, followed by the Crop Yield Prediction and Indoor Scene Classification tasks in Sect. 8. Related work is discussed in Sects. 9 and 10 concludes the paper.

## 2 Background

This section provides an overview of the LAS framework and the neural network approaches used in FFNSL. We discuss the difference between confidence estimates of uncertainty-aware neural networks versus that of the standard Softmax layer, when applying these trained networks to out-of-distribution data. This is particularly relevant to our FFNSL framework, as FFNSL relies upon neural network predictions and their confidence scores to learn interpretable knowledge for solving a downstream task.

### 2.1 Learning from answer sets

LAS (Law et al., 2019) is a logic-based machine learning approach that extends the field of Logic Programming (ILP) (Muggleton, 1991) with systems ILASP (Law, 2018) and FastLAS (Law et al., 2020). ILASP and FastLAS are capable of learning interpretable knowledge, expressed in the language of Answer Set Programming (ASP) (Gelfond & Kahl, 2014), from noisy labelled examples in an effective and scalable manner. Typically, an ASP program includes four types of rules: normal rules, choice rules, and hard and weak constraints. In this paper, we consider ASP programs composed of *normal rules* only.<sup>3</sup> A normal rule is of the form  $h : -b_1, \dots, b_n, \text{not } c_1, \dots, \text{not } c_m$ , where  $h, b_1, \dots, b_n, c_1, \dots, c_m$  are atoms, “not” is negation as failure,  $h$  is the *head* of the rule and  $b_1, \dots, b_n, \text{not } c_1, \dots, \text{not } c_m$  is the *body* of the rule. The Herbrand Base of an ASP program  $P$ , denoted  $HB_P$ , is the set of ground (variable free) atoms that can be formed from predicates and constants in  $P$ . Subsets of  $HB_P$  are called *interpretations* of  $P$ . The semantics of an ASP program  $P$  is defined in terms of *answer sets*, a subset, denoted as  $AS(P)$ , of all interpretations of  $P$  that satisfy every rule in  $P$ . Given an answer set  $A$ , a ground normal rule is satisfied if the head is satisfied by  $A$  whenever all positive atoms and none of the negated atoms of the body are in  $A$ , that is when the body is satisfied. A *partial interpretation*,  $e_{pi}$ , is a pair of sets of ground atoms  $\langle e_{pi}^{inc}, e_{pi}^{exc} \rangle$ , called the *inclusion* and *exclusion* sets respectively. An interpretation  $I$  extends  $e_{pi}$  iff  $e_{pi}^{inc} \subseteq I$  and  $e_{pi}^{exc} \cap I = \emptyset$ .

In the LAS framework, labelled examples are specified as *Context-Dependent Partial Interpretations (CDPIs)*. A CDPI example  $e$  is a pair  $\langle e_{pi}, e_{ctx} \rangle$ , where  $e_{pi}$  is a partial interpretation and  $e_{ctx}$  is an ASP program called the context of  $e$ . An ASP program  $P$  is said to *accept*  $e$  if there is at least one answer set  $A$  of  $P \cup e_{ctx}$  that extends  $e_{pi}$ . Essentially, a CDPI states that a learned program  $P$ , together with  $e_{ctx}$ , should bravely entail<sup>4</sup> all inclusion atoms

<sup>3</sup> The reader is referred to Gelfond and Kahl (2014) for a full overview of ASP.

<sup>4</sup> A program  $P$  *bravely entails* an atom  $a$  if there is at least one answer set of  $P$  that contains  $a$ .

and none of the exclusion atoms of  $e$ . When a CDPI example is noisy, that is, the truthfulness of its context and/or partial interpretation is not guaranteed, it has a *weight* or *penalty* assigned to it, in the form of a positive integer. A Weighted Context-Dependant Partial Interpretation (WCDPI) is therefore a CDPI weighted with a penalty. It is formally defined as a tuple  $e = \langle e_{id}, e_{pen}, e_{pi}, e_{ctx} \rangle$  where  $e_{id}$  is a unique identifier of  $e$ ,  $e_{pen}$  is the penalty of  $e$ , and  $e_{pi}$  and  $e_{ctx}$  represent a CDPI. A LAS system that is noise-tolerant learns an ASP program  $H$ , called a *hypothesis*, from WCDPI examples. If a hypothesis  $H$  does not accept a WCDPI example, we say that it *pays the penalty* of that example. Informally, penalties are used to calculate the cost associated with a hypothesis for not covering examples. The cost function of a hypothesis  $H$  is the sum over the penalties of all of the examples that are not *covered* by  $H$ , augmented with the length of the hypothesis. A LAS learning task with noisy examples, consists of an ASP program denoting background knowledge, a hypothesis space defined by a language bias,<sup>5</sup> expressing the set of rules that can be used to construct a solution of the task, and a set of WCDPI examples. The goal of such a task is to find a hypothesis  $H$  in the hypothesis space that minimises a cost function with respect to a given set of noisy examples. This is formally defined below, adapted from Law (2018).

**Definition 1** An  $ILP_{LAS}^{noise}$  task  $T$  is a tuple  $T = \langle B, S_M, E \rangle$ , where  $B$  is an ASP program,  $S_M$  is a hypothesis space, and  $E$  is a set of WCDPIs. Given a hypothesis  $H \subseteq S_M$ ,

1.  $UNCOV(H, T)$  is the set consisting of all examples  $e \in E$  such that  $B \cup H$  does not accept  $e$ .
2. The penalty of  $H$ , denoted as  $PEN(H, T)$ , is the sum  $\sum_{e \in UNCOV(H, T)} e_{pen}$ .
3. The score of  $H$ , denoted as  $S(H, T)$ , is calculated as  $|H| + PEN(H, T)$ .
4.  $H$  is an *optimal inductive solution* of  $T$  if and only if  $\nexists H' \subseteq S_M$  such that  $S(H', T) < S(H, T)$ .

ILASP and FastLAS are two state-of-the-art systems capable of solving an  $ILP_{LAS}^{noise}$  task. The optimisation function used by both systems aims at learning a hypothesis  $H$  that jointly minimises the total penalty paid for the uncovered examples and its length. In practice, this creates a bias towards shorter, and therefore more general solutions that cover examples with a high penalty value.

## 2.2 Uncertainty-aware neural networks

Our FFNSL framework relies on pre-trained neural networks to extract symbolic features from unstructured data. The neural network prediction and its confidence score may therefore affect the accuracy of a learned hypothesis. In this paper, we consider two different types of neural networks as FFNSL neural components: a standard Convolutional Neural Network (CNN) that uses a Softmax layer, and an uncertainty-aware CNN that provides more robust confidence estimates when given data outside the training distribution.

Uncertainty can be formulated as either *aleatoric* or *epistemic* uncertainty (Hüllermeier & Waegeman, 2021; Pearce et al., 2021). In a machine learning classification task, aleatoric uncertainty can be thought of as the uncertainty along the class decision boundary, whereas epistemic uncertainty can be thought of as whether the sample falls into any of the classes at all. The confidence estimates output by a neural network Softmax layer in a classification task often only capture aleatoric uncertainty, as these outputs are based on a single probability distribution over a set of classes squashed into real values between 0 and 1. For example,

<sup>5</sup> For a detailed definition of a language bias see Law (2018).

given neural network output logits  $\mathbf{l}$  and  $k$  possible classes, the Softmax output  $\sigma(\mathbf{l})$  for class  $i$ , where  $1 \leq i \leq k$  is calculated as:

$$\sigma(\mathbf{l})_i = \frac{e^{l_i}}{\sum_{j=1}^k e^{l_j}}$$

where  $e = 2.71828\dots$  is the Euler number.<sup>6</sup> There are three challenges with this approach in terms of uncertainty quantification. Firstly, the exponent applied to neural network outputs inflates the confidence estimate. Secondly, as the Softmax output is a point-wise, multinomial distribution, it is only possible to compare the confidence of the predicted class among other classes, as opposed to estimating the predictive distribution variance (Sensoy et al., 2018). Finally, when Softmax is paired with the commonly used cross-entropy loss, the network is only trained to minimise prediction error, as opposed to expressing uncertainty robustly.

To address these challenges, many techniques have been proposed in the literature (Rasmussen, 2003; Mackay, 1995; Blundell et al., 2015; Abdar et al., 2021). In this paper we consider the *EDL-GEN* (Sensoy et al., 2020) approach, which is a neural network based on generative models of Evidential Deep Learning (EDL) systems (Sensoy et al., 2018) that have been shown to achieve state-of-the-art performance in handling epistemic uncertainty. An EDL (Sensoy et al., 2018) system replaces the Softmax layer in a neural network with a linear layer that represents the parameters of a Dirichlet distribution, a second-order distribution that inherently models the variance of a predictive distribution as opposed to the single point-wise output provided by Softmax. It then uses a new loss function that jointly minimises prediction error and the variance of the Dirichlet distribution, to reduce aleatoric uncertainty on the class decision boundary. EDL-GEN (Sensoy et al., 2020) extends this approach to also capture epistemic uncertainty by firstly treating the output of each class as a binary decision and secondly, using a variational auto-encoder to automatically generate out-of-distribution samples for training, in order to help the network discriminate between samples within and outside the training distribution.

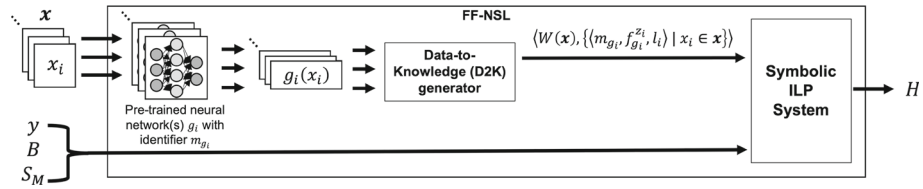
To better understand how the uncertainty estimation of neural network predictions impacts the overall accuracy of our FFNSL framework, we analyse in the evaluation Sects. 6 and 7, the predicted confidence scores generated by a standard CNN with a Softmax layer and an EDL-GEN neural network and evaluate how they affect the accuracy of FFNSL when increasing percentages of input training data are subject to distributional shifts.

### 3 FFNSL framework

In this section we present our general FFNSL framework. It consists of three components, a pre-trained neural network, a symbolic (logic-based) learning system and a D2K generator that bridges the neural and symbolic learning components. It takes as input a dataset  $D$  of labelled (sequences of) unstructured data, alongside a background knowledge  $B$  (if any) and a search space  $S_M$ . The output is a *hypothesis*  $H$  in the search space  $S_M$  ( $H \subseteq S_M$ ), that predicts the labels of (sequences of) unstructured data. An overview of the FFNSL architecture is presented in Fig. 1.

We now define each of the three components of our FFNSL architecture. Let us assume that the training dataset  $D$  is given by a finite set  $D = \{\langle \mathbf{x}_w, y_w \rangle \mid 1 \leq w \leq |D|\}$ . The downstream task is a classification task where the objective is to predict the target label  $y \in \mathcal{Y}$  given a sequence of unstructured data  $\mathbf{x} \in \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ . Note that  $\mathcal{X}_i$  could refer to different

<sup>6</sup> We have used bold  $e$  to avoid confusion with a WCDPI  $e$ .



**Fig. 1** FFNSL architecture and data flow generated for a single data point  $\langle x, y \rangle$ , where  $x$  is a sequence of images and  $y$  is a label for the sequence.  $B$  is the background knowledge,  $S_M$  is the hypothesis search space and  $H$  is the learned hypothesis. In practice, the architecture is applied on a set of data points from which the D2K generator produces a set of symbolic examples passed in as input to the symbolic learner

types of unstructured inputs and the sequence could also contain only a single input. The neural component of FFNSL contains up to  $n$  pre-trained neural network(s).<sup>7</sup> Each neural network  $g_i : \mathcal{X}_i \rightarrow [0, 1]^{k_i}$  returns a vector denoting relative assignment to  $k_i$  possible classes for an unstructured input  $x_i \in \mathbf{x}$ . Each possible class  $z_i \in \{1, \dots, k_i\}$  represents a set of symbolic feature and value pairs from a given set  $F_{g_i}$  of symbolic feature mappings associated with the neural network  $g_i$ . For example, in the Follow Suit Winner task,  $F_{g_i}$  contains all possible suit and rank values corresponding to the possible predictions of the neural network, when given an image of a playing card.

The second component of FFNSL is the D2K generator that outputs a symbolic representation of the sequence of neural network predictions, together with an aggregated confidence value. Specifically, for a given sequence of unstructured data  $\mathbf{x} = \langle x_1, \dots, x_n \rangle$ , the D2K generator takes each neural network output  $g_i(x_i)$ , and computes the corresponding prediction  $z_i$ . Each  $z_i$ , for  $1 \leq i \leq n$ , is obtained by using the standard “arg max” function, i.e., the class with the maximum confidence score:

$$z_i = \arg \max_{j \in \{1, \dots, k_i\}} (g_i(x_i)[j])$$

The D2K generator then uses the set  $F_{g_i}$ , associated with  $g_i$  and generates the set  $f_{g_i}^{z_i} \subseteq F_{g_i}$  of symbolic feature and value pairs corresponding to the prediction  $z_i$ . As an example, in the Follow Suit Winner task,  $z_i$  is an identifier for one of 52 playing cards, and  $f_{g_i}^{z_i}$  contains two feature and value pairs, one for the suit, and one for the rank of the card  $z_i$ . The D2K generator also generates a set  $l_i$  of pairs containing additional symbolic meta-data, associated with each input  $x_i$ . Again, each pair in  $l_i$  contains a name and a value. In the Follow Suit Winner task,  $l_i$  contains one pair indicating which player played the card  $z_i$ . The generated set of tuples  $\{(m_{g_i}, f_{g_i}^{z_i}, l_i) \mid x_i \in \mathbf{x}\}$ , where  $m_{g_i}$  is a unique identifier for the neural network  $g_i$ , defines the symbolic features extracted from a sequence of unstructured data  $\mathbf{x}$ , based on the neural network predictions. Finally, the D2K generator computes an aggregated confidence value  $W(\mathbf{x})$  for the generated symbolic features, representing the combined confidence scores of the neural network predictions:

$$W(\mathbf{x}) = \min(\{g_i(x_i)[z_i] \mid x_i \in \mathbf{x}\}) \tag{1}$$

$W(\mathbf{x})$  is a generalisation of the binary Gödel  $t$ -norm used in fuzzy logic to encode fuzzy conjunctions (Metcalf et al., 2008). So, given a sequence of unstructured inputs  $\mathbf{x} = \langle x_1, \dots, x_n \rangle$  and the predicted vector  $\langle g_1(x_1)[z_1], \dots, g_n(x_n)[z_n] \rangle$  from the neural network, the output of

<sup>7</sup> If the input sequence contains the same type of data, only one neural network is required.



the D2K generator is formally defined as:

$$D2K(\mathbf{x}) = \langle W(\mathbf{x}), \{ \langle m_{g_i}, f_{g_i}^{z_i}, l_i \rangle \mid x_i \in \mathbf{x} \} \rangle \quad (2)$$

A pseudo-code implementation of the D2K generator is presented in Algorithm 1. Note that some aspects are task specific, such as the set of feature value pairs  $f_{g_i}^{z_i}$ , and meta-data  $l_i$ . These are left general in Algorithm 1, and specified in more detail for each task in Sect. 4.

---

#### Algorithm 1 D2K generator

---

```

Input:  $\mathbf{x}$  // A sequence of images
 $SF = \emptyset$ 
 $CS = \emptyset$ 
for  $x_i \in \mathbf{x}$  do
  // Obtain the neural network prediction for each input  $x_i$ 
   $z_i = \arg \max_{j \in \{1, \dots, k_i\}} (g_i(x_i)[j])$ 
  // Accumulate symbolic facts associated with the neural network
  // prediction  $z_i$ , using the set of feature value pairs  $f_{g_i}^{z_i} \subseteq F_g$ , and
  // meta-data  $l_i$ 
   $SF = SF \cup \{ \langle m_{g_i}, f_{g_i}^{z_i}, l_i \rangle \}$ 
  // Accumulate the neural network confidence of prediction  $z_i$ 
   $CS = CS \cup \{ g_i(x_i)[z_i] \}$ 
end for
// Calculate the aggregated weight penalty
 $W = \min(CS)$ 
// Return the D2K output
return  $\langle W, SF \rangle$ 

```

---

The third component of our FFNSL framework is a symbolic logic-based machine learning system. For each labelled unstructured data  $\langle \mathbf{x}, y \rangle \in D$ , the symbolic learning system takes as input  $D2K(\mathbf{x})$  and the label  $y$ , and generates a *weighted symbolic labelled example* denoted as the tuple  $\langle W'(\mathbf{x}), e_{\langle \mathbf{x}, y \rangle} \rangle$  where  $W'(\mathbf{x})$  is a penalty for the example, calculated from the aggregated confidence score  $W(\mathbf{x})$ , and  $e_{\langle \mathbf{x}, y \rangle}$  is a labelled example. The syntactic form of  $e_{\langle \mathbf{x}, y \rangle}$  and the calculation of the penalty  $W'(\mathbf{x})$  depends on the specific symbolic learning system used in the instantiation of the framework. In Sect. 4 we present two specific instances of FFNSL where the symbolic learning systems are LAS systems and we show how weighted symbolic labelled examples are defined as WCDPI examples. We denote with  $E$  the set of weighted symbolic labelled examples defined by the symbolic learning system for all  $\langle \mathbf{x}, y \rangle \in D$ . A symbolic learning task  $T = \langle B, S_M, E \rangle$  is then generated where  $B$  and  $S_M$  are respectively the background knowledge and a search space given as input to FFNSL. The symbolic learner then computes an optimal solution  $H$  for this task  $T$  as the output of FFNSL.

Formally, an FFNSL learning task is a tuple  $T = \langle B, S_M, D \rangle$  where  $D$  is a set of labelled unstructured data,  $B$  is a set of optional background knowledge and  $S_M$  is a search space of possible solutions for  $T$ . A *hypothesis*  $H \subseteq S_M$  is an inductive solution of  $T$  if and only if  $H$  is an *optimal inductive solution* of the symbolic learning task  $\langle B, S_M, E \rangle$ , where  $E$  is the set of weighted symbolic labelled examples automatically generated relative to the given set  $D$ . In the next section, we present four specific instances of our FFNSL framework and give specific examples of the components described here.



## 4 FFNSL with LAS systems

The generality of our FFNSL framework allows it to be instantiated differently, using alternative neural and/or symbolic learning components, depending on the nature of the classification task in hand. We have considered four different classification tasks, called *Follow Suit Winner*, *Sudoku Grid Validity*, *Crop Yield Prediction* and *Indoor Scene Classification* respectively. The first requires the learning of concepts that are not directly observed in the labels, but linked to the label through the background knowledge, whereas the other tasks require the learning of concepts that define the classification label. Because of the different types of symbolic learning, we consider instantiations of our FFNSL framework with different LAS systems. In what follows we introduce these tasks, their datasets, define the respective FFNSL learning tasks and describe in more detail the FFNSL instances we have implemented to solve these tasks. Firstly, let us define the weighted symbolic labelled examples within a LAS system, based on the output from the D2K generator. Essentially, the predicted symbolic features and meta-data define the *context* of a LAS example, represented as a conjunction of facts, and the aggregated confidence score  $W(\mathbf{x})$  is used to calculate the associated *weight penalty*:

$$W'(\mathbf{x}) = \lfloor 100 \times W(\mathbf{x}) \rfloor + 1 \quad (3)$$

which converts  $W(\mathbf{x})$  to an integer  $W'(\mathbf{x}) > 0$  as required by the LAS systems. Given the output generated by D2K, a LAS system constructs a weighted symbolic labelled example  $\langle W'(\mathbf{x}), e_{(x,y)} \rangle$  as a WCDPI of the form  $\langle e_{id}, e_{pen}(\mathbf{x}), e_{pi}(y), e_{ctx}(\mathbf{x}) \rangle$ , where  $e_{id}$  is a unique identifier,  $e_{pen}(\mathbf{x}) = W'(\mathbf{x})$ ,  $e_{pi}(y)$  is the partial interpretation  $\{\{y\}, \mathcal{Y} \setminus \{y\}\}$ , defined in terms of the label  $y$  and its domain  $\mathcal{Y}$ , and the context  $e_{ctx}(\mathbf{x})$  is a conjunction of facts created from the predicted symbolic features and meta-data. The components  $e_{pi}(y)$  and  $e_{ctx}(\mathbf{x})$  together constitute the labelled example  $e_{(x,y)}$ .

Given a set  $E'$  of WCDPIs, a background knowledge  $B$ , and a search space  $S_M$ , a hypothesis  $H \subseteq S_M$  is learned such that  $H$  is an optimal inductive solution of the task  $T_{LAS}^{noise} = \langle B, S_M, E' \rangle$ . Let us now present the tasks used in our evaluation, alongside examples of each instantiated FFNSL component.

### 4.1 Follow suit winner

This is a classification task where 4 players each play 1 card and the goal is to predict the winning player. The symbolic knowledge required to solve the task defines the rules of the game, that is *the winner is the player that plays the highest ranked card with the same suit as player 1*. Each  $\langle \mathbf{x}, y \rangle \in D$  is composed of a sequence  $\mathbf{x}$  of 4 card images corresponding to the cards played by players 1, ..., 4, and a label  $y \in \{1, 2, 3, 4\}$  denoting the player who wins the 4 card trick.

Let us assume  $\mathbf{x} = [\img alt="10 of hearts" data-bbox="315 758 335 773"], [\img alt="jack of hearts" data-bbox="338 758 358 773"], [\img alt="4 of clubs" data-bbox="361 758 381 773"], [\img alt="8 of spades" data-bbox="384 758 404 773]]$  which contains images of the cards *10 of hearts*, *jack of hearts*, *4 of clubs* and *8 of spades* played by player 1, 2, 3 and 4 respectively. For this trick, the ground truth label is  $y = 2$  indicating that player 2 is the winner since player 2 has played the highest ranked card with the same suit as player 1. Since the unstructured inputs in the sequence  $\mathbf{x}$  are of the same type (i.e., card images), FFNSL can simply use a single neural network  $g$  pre-trained to predict the features of a card image, that is the rank and suit of each card. Therefore,  $g$  has two associated symbolic features *rank* and *suit* each with values  $\{2, \dots, 10, \text{jack, queen, king, ace}\}$  and  $\{\text{hearts, clubs, spades, diamonds}\}$  respectively. For each input  $x_i$ , there are 52 possible predictions, one for each combination of rank and suit, i.e.,  $g: \mathcal{X} \rightarrow [0, 1]^{52}$ , where  $\mathcal{X}$  is the set of possible card images.  $g$  has an associated feature

value mapping  $F_g$  which gives for each card prediction  $z_i \in \{1, \dots, 52\}$ , a unique set of two pairs, each containing a feature and value, i.e.,  $f_g^{z_i} = \{\langle rank, v_{rank} \rangle, \langle suit, v_{suit} \rangle\}$ , where  $v_{rank}$  is one of the 13 rank values and  $v_{suit}$  is one of the 4 suit values. Furthermore, each input  $x_i$  also has associated symbolic meta-data  $l_i = \{\langle player, v_{player} \rangle\}$  where  $v_{player} \in \{1, 2, 3, 4\}$  indicates the player that has played card  $x_i$ .

We instantiate our FFNSL framework as follows. Given a sequence  $\mathbf{x}$  of 4 card images, the neural component of FFNSL generates 4 vectors  $g(x_i)$ , where  $1 \leq i \leq 4$ . The D2K component generates for each  $x_i$ , the card prediction  $z_i$  and its corresponding symbolic features and meta-data, thus computing the tuple  $D2K(x_i) = \langle card, f_g^{z_i}, l_i \rangle$ , where  $card$  is the identifier for the network  $g$  (i.e.,  $m_g = card$ ).

**Example 1** Consider the sequence  $\mathbf{x} = [\img alt="10 of hearts" data-bbox="245 271 265 286"], [\img alt="Jack of hearts" data-bbox="271 271 291 286"], [\img alt="4 of clubs" data-bbox="297 271 317 286"], [\img alt="8 of spades" data-bbox="323 271 343 286]]$  and  $y = 2$ . Let us assume that the neural network  $g$  computes the outputs  $g(x_1), \dots, g(x_4)$  from which the D2K generator generates the correct card predictions  $z_1 = 10, z_2 = 11, z_3 = 17$ , and  $z_4 = 34$ . Let us also assume the neural network confidence scores for these predictions are:

$$g(x_1)[z_1] = 0.95; g(x_2)[z_2] = 0.92; g(x_3)[z_3] = 0.80; g(x_4)[z_4] = 0.94;$$

$D2K(\mathbf{x})$  is given by the following tuple:

$$D2K(\mathbf{x}) = \langle 0.80, \{ \langle card, \{ \langle rank, 10 \rangle, \langle suit, hearts \rangle \} \rangle, \{ \langle player, 1 \rangle \} \rangle, \\ \langle card, \{ \langle rank, jack \rangle, \langle suit, hearts \rangle \} \rangle, \{ \langle player, 2 \rangle \} \rangle, \\ \langle card, \{ \langle rank, 4 \rangle, \langle suit, clubs \rangle \} \rangle, \{ \langle player, 3 \rangle \} \rangle, \\ \langle card, \{ \langle rank, 8 \rangle, \langle suit, spades \rangle \} \rangle, \{ \langle player, 4 \rangle \} \} \rangle.$$

In this task, FFNSL uses the symbolic learner ILASP. The concept to be learned is not directly expressed as a label, but is related to it. The label is a single winning player for a trick, but the learned concept requires reasoning over the conditions of the suit and rank values of the other players' cards. We encode as background knowledge, possible suit and rank values, the four players, as well as the definition of a higher rank predicate. ILASP is particularly suited for solving such learning tasks, known as non-observational predicate learning. The full background knowledge  $B$  and language bias used to construct the search space  $S_M$  for this classification task are given in Appendix F. To generate its learning task, ILASP has to generate its set  $E'$  of WCDPI examples based on the output of the D2K component. For example, the WCDPI generated from the D2K output and the corresponding label in Example 1 is:

$$\langle e_{id}, 81, \{ \{2\}, \{1, 3, 4\} \}, e_{ctx} \rangle$$

where  $e_{id}$  is a unique identifier and  $e_{ctx}$  is the set of facts  $\{card(1, 10, hearts), card(2, jack, hearts), card(3, 4, clubs), card(4, 8, spades)\}$ .

### 4.2 Sudoku grid validity

Our second classification task is *Sudoku Grid Validity*. This consists of observing a sequence of images of handwritten MNIST digits, corresponding to the digits in a Sudoku grid, and predicting if the grid is valid or not.<sup>8</sup> The learned symbolic knowledge required to solve this task is the definition of a valid Sudoku grid. In this task, each  $\langle \mathbf{x}, y \rangle \in D$  contains a sequence

<sup>8</sup> We assume a Sudoku grid has been pre-processed to return images of digits in different cells and we do not process blank cells.

of digit images  $\mathbf{x}$  with a label  $y \in \{0, 1\}$  for valid and invalid respectively. The length of the sequence depends on the size of the grid. We consider  $4 \times 4$  and  $9 \times 9$  Sudoku grids as two separate tasks, with respective datasets  $D_{4 \times 4}$  and  $D_{9 \times 9}$  where the maximum length of the sequence in input is given by  $n = 16$  and  $n = 81$  respectively. As the images are all MNIST digits, FFNSL uses two neural networks  $g_{4 \times 4}$  and  $g_{9 \times 9}$ , depending on the grid size, pre-trained to predict the feature *digit* of a single image  $x_i$  in  $\mathbf{x}$ . So  $g_{k \times k} : \mathcal{X} \rightarrow [0, 1]^k$ , where  $\mathcal{X} = \text{MNIST}$ . In the case of  $D_{4 \times 4}$ ,  $n = 16$  and  $k = 4$  whereas in the case of  $D_{9 \times 9}$ ,  $n = 81$  and  $k = 9$ . The neural network  $g_{k \times k}$  has associated a feature value mapping  $F_{g_{k \times k}}$  which gives for each digit prediction  $z_i \in \{1, \dots, k\}$  a unique set of pairs  $f_{g_{k \times k}}^{z_i} = \{\langle \text{value}, v \rangle\}$ , where  $v$  is one of the  $k$  digits that can appear in a Sudoku grid of size  $k \times k$ . The meta-data related to each  $x_i$  is a set of two feature value pairs denoting the row and column that the image  $x_i$  has in the Sudoku grid, i.e.,  $l_i = \{\langle \text{row}, v_{\text{row}} \rangle, \langle \text{col}, v_{\text{col}} \rangle\}$ , where  $v_{\text{row}}, v_{\text{col}} \in \{1, \dots, k\}$ .

The instantiated FFNSL framework for this classification task is defined as follows. Given a sequence,  $\mathbf{x}$ , of MNIST digit images, for each  $x_i \in \mathbf{x}$ , the pre-trained neural network  $g_{k \times k}$  computes the vector  $g(x_i)$ . The D2K component generates for each  $x_i$  the tuple  $D2K(x_i) = \langle \text{digit}, f_{g_{k \times k}}^{z_i}, l_i \rangle$  where *digit* is the network identifier,  $f_{g_{k \times k}}^{z_i}$  is the set of symbolic feature values associated with the prediction  $z_i$ , and  $l_i$  is the set of symbolic meta-data feature value pairs associated with  $x_i$ .

**Example 2** Consider the task of predicting the validity of a  $4 \times 4$  Sudoku grid. Let  $\mathbf{x} = [\mathbf{2}, \mathbf{4}, \mathbf{1}, \mathbf{3}, \mathbf{4}]$ , with label  $y = 1$ , and associated symbolic meta-data:

$$\begin{aligned} l_1 &= \{\langle \text{row}, 1 \rangle, \langle \text{col}, 1 \rangle\} \\ l_2 &= \{\langle \text{row}, 1 \rangle, \langle \text{col}, 3 \rangle\} \\ l_3 &= \{\langle \text{row}, 1 \rangle, \langle \text{col}, 4 \rangle\} \\ l_4 &= \{\langle \text{row}, 3 \rangle, \langle \text{col}, 2 \rangle\} \\ l_5 &= \{\langle \text{row}, 4 \rangle, \langle \text{col}, 3 \rangle\} \end{aligned}$$

Let us assume the neural network  $g = g_{4 \times 4}$  and  $g$  computes the outputs  $g(x_1), \dots, g(x_5)$  from which the D2K generator generates the correct digit predictions  $z_1 = 2, z_2 = 4, z_3 = 1, z_4 = 3$ , and  $z_5 = 4$ . Let us also assume the neural network confidence scores for these predictions are:  $g(x_1)[z_1] = 0.88, g(x_2)[z_2] = 0.93, g(x_3)[z_3] = 0.87, g(x_4)[z_4] = 0.97$ , and  $g(x_5)[z_5] = 0.99$ . The aggregated confidence score  $W(\mathbf{x}) = 0.87$ .  $D2K(\mathbf{x})$  is given by the following tuple:

$$\begin{aligned} D2K(\mathbf{x}) = \langle 0.87, \{ & \langle \text{digit}, \{\langle \text{value}, 2 \rangle\}, \{\langle \text{row}, 1 \rangle, \langle \text{col}, 1 \rangle\}, \\ & \langle \text{digit}, \{\langle \text{value}, 4 \rangle\}, \{\langle \text{row}, 1 \rangle, \langle \text{col}, 3 \rangle\}, \\ & \langle \text{digit}, \{\langle \text{value}, 1 \rangle\}, \{\langle \text{row}, 1 \rangle, \langle \text{col}, 4 \rangle\}, \\ & \langle \text{digit}, \{\langle \text{value}, 3 \rangle\}, \{\langle \text{row}, 3 \rangle, \langle \text{col}, 2 \rangle\}, \\ & \langle \text{digit}, \{\langle \text{value}, 4 \rangle\}, \{\langle \text{row}, 4 \rangle, \langle \text{col}, 3 \rangle\} \} \rangle. \end{aligned}$$

In this task FFNSL uses the FastLAS symbolic learner because the task is to learn the definition of the classification label, and FastLAS has been shown, for these types of learning tasks, to be more scalable than ILASP (Law et al., 2020). For both  $4 \times 4$  and  $9 \times 9$  Sudoku grids, the knowledge of the grid is encoded as part of the background knowledge  $B$ , given in Appendix F together with the language bias used to construct the search space  $S_M$ . For each  $(\mathbf{x}, y)$ , FastLAS takes as input  $D2K(\mathbf{x})$  and generates a WCDPI example. For instance, the WCDPI generated for the D2K output and the corresponding label in Example 2 is:

$$\langle e_{\text{id}}, 88, \{\{1\}, \{0\}\}, e_{\text{ctx}} \rangle$$

where  $e_{\text{id}}$  is a unique identifier and  $e_{\text{ctx}}$  is given by the set of facts  $\{\text{digit}(1, 1, 2), \text{digit}(1, 3, 4), \text{digit}(1, 4, 1), \text{digit}(3, 2, 3), \text{digit}(4, 3, 4)\}$ .

### 4.3 Crop yield prediction

To demonstrate the application of FFNSL to a real-world problem and dataset, consider the *Crop Yield Prediction* task. The goal is to classify the quality of yield, given an image and the location of a particular crop. The symbolic knowledge required to solve the task defines the quality of yield according to the crop's location, species, and any disease that may be present. Each  $(x, y) \in D$  is composed of a sequence  $x$  containing a single image, and a label  $y \in \{0, 1, 2\}$  denoting the quality of yield as *poor*, *moderate*, and *strong* respectively.

Let us assume  $x = \langle \text{img} \rangle$  which contains an image of a peach crop with the bacterial spot disease. Given symbolic meta-data denoting the location of this crop, let us assume the label  $y = 0$ , indicating poor yield. In this task, we use one neural network  $g$  to predict the features of a crop image, which are the crop species and disease. In total, there are 38 possible combinations of crop species and diseases, and  $g$  is trained to classify each combination. To assist with neural network training, the image dataset also contains a background class with unrelated images.<sup>9</sup> Therefore,  $g : \mathcal{X} \rightarrow [0, 1]^{39}$ , where  $\mathcal{X}$  is the set of possible crop and background images.  $g$  has an associated feature value mapping  $F_g$ , which specifies for the crop prediction  $z_i \in \{1, \dots, 38\}$ , a unique set of feature and value pairs  $f_g^{z_i} = \{\langle \text{species}, v_{\text{species}} \rangle, \langle \text{disease}, v_{\text{disease}} \rangle\}$ , where  $v_{\text{species}}$  and  $v_{\text{disease}}$  are the crop species and disease values respectively. Also, each input  $x_i$  has associated symbolic meta-data  $l_i = \{\langle \text{location}, v_{\text{location}} \rangle\}$  where  $v_{\text{location}} \in \{1, \dots, 19\}$  is the location of the crop.<sup>10</sup>

We instantiate our FFNSL framework as follows. Given a sequence  $x$  containing a single crop image, the neural component generates a single vector  $g(x_i)$ . The D2K component generates the prediction  $z_i$  and its corresponding symbolic features and meta-data, thus computing the tuple  $D2K(x_i) = \langle \text{crop}, f_g^{z_i}, l_i \rangle$ , where *crop* is the identifier for the network  $g$  (i.e.,  $m_g = \text{crop}$ ).

**Example 3** Consider the sequence  $x = \langle \text{img} \rangle$  and  $y = 0$ . Let us assume the neural network  $g$  computes the output  $g(x_1)$  from which the D2K generator generates the correct crop prediction  $z_1 = 17$ . Let us also assume the neural network predicts with confidence  $g(x_1)[z_1] = 0.98$ , and this crop is in location 5.  $D2K(x)$  is given by the following tuple:

$$D2K(x) = \langle 0.98, \{ \langle \text{crop}, \{ \langle \text{species}, \text{peach} \rangle, \langle \text{disease}, \text{bacterial\_spot} \rangle \} \rangle, \langle \text{location}, 5 \rangle \} \rangle.$$

In this task FFNSL uses the FastLAS symbolic learner which is shown to be more scalable than ILASP. The background knowledge  $B$  contains a rule that ensures a classification is performed, i.e., given a crop, disease, and a location, the learned hypothesis should output only one class of crop yield. This rule, alongside the language bias used to construct the search space  $S_M$  is given in Appendix F. For each  $(x, y)$ , FastLAS takes as input  $D2K(x)$  and generates a WCDPI example. For instance, the WCDPI generated from the D2K output and the corresponding label given in Example 3 is:

$$\langle e_{\text{id}}, 99, \{ \{0\}, \{1, 2\} \}, e_{\text{ctx}} \rangle$$

where  $e_{\text{id}}$  is a unique identifier and  $e_{\text{ctx}}$  is given by the set of facts  $\{\text{species}(\text{peach}), \text{disease}(\text{bacterial\_spot}), \text{location}(5)\}$ .

<sup>9</sup> In FFNSL, if the neural network predicts the background class, no WCDPI example is generated by the D2K component.

<sup>10</sup> The dataset consists of two unique crops in each location, hence 19 possible location values.



**Fig. 2** Example bookstore image from the MIT Indoor Scenes dataset. Quattoni and Torralba (2009)

#### 4.4 Indoor scene classification

Our final instantiation of FFNSL is with the *Indoor Scene Classification* task, where both neural and symbolic components are trained with real data. The goal is to learn symbolic knowledge that maps indoor scene classes (e.g., bedroom, bathroom, kitchen) into higher level super-classes (e.g., home), given images of indoor scenes. Each  $\langle x, y \rangle \in D$  is composed of a sequence  $x$  containing a single indoor scene image, and a label  $y \in \{0, \dots, 4\}$  denoting the super-class as *store*, *home*, *public space*, *leisure*, and *working place* respectively.

Let us assume  $x = \{\text{image}\}$  which contains an image of a bookstore (also shown in Fig. 2). The label for this example is  $y = 0$  (i.e., *store*). We use one neural network  $g$  to predict the scene class. In total, there are 67 different classes of various indoor scenes, and therefore  $g : \mathcal{X} \rightarrow [0, 1]^{67}$ , where  $\mathcal{X}$  is the set of possible images in the MIT Indoor Scene dataset.  $g$  has an associated feature value mapping  $F_g$ , which for the scene prediction  $z_i \in \{1, \dots, 67\}$ , gives a pair that denotes the symbolic scene name  $v_{scene}$ , i.e.,  $f_g^{z_i} = \{\{scene, v_{scene}\}\}$ . In this task there is no symbolic meta-data associated with each input  $x \in \mathcal{X}$ . Given a sequence  $x$  containing a single scene image, the neural component generates a single vector  $g(x_i)$ . The D2K component generates the prediction  $z_i$  and its corresponding symbolic feature, thus computing the tuple  $D2K(x_i) = \{\text{image}, f_g^{z_i}, \{\}\}$ , where *image* is the identifier for the network  $g$  (i.e.,  $m_g = \text{image}$ ).

**Example 4** Consider the sequence  $x = \{\text{image}\}$  and  $y = 0$ . Let us assume the neural network  $g$  computes the output  $g(x_1)$  from which the D2K generator generates the correct scene prediction  $z_1 = 8$ . Let us also assume the neural network predicts with confidence  $g(x_1)[z_1] = 0.96$ .  $D2K(x)$  gives as output the following tuple:

$$D2K(x) = \{0.96, \{\{\text{image}, \{\{scene, bookstore\}\}, \{\}\}\}\}$$

The FastLAS symbolic learner is also used in this task. No background knowledge is required, and the language bias is given in Appendix F. For each  $\langle x, y \rangle$ , FastLAS takes as input  $D2K(x)$  and generates a WCDPI example. For instance, the WCDPI generated for the D2K output in Example 4 is:

$$\langle e_{id}, 96, \{\{0\}, \{1, 2, 3, 4\}\}, e_{ctx} \rangle$$

where  $e_{id}$  is a unique identifier and  $e_{ctx}$  is given by the set  $\{\text{scene}(\text{bookstore})\}$ .

## 5 Evaluation methodology

In this section we describe the methodology used to evaluate the FFNSL framework. In the first two tasks, the focus is on learning complex first-order knowledge involving negation as failure and predicate invention, which are essential aspects of common-sense learning and reasoning. In the second two tasks, we demonstrate FFNSLs applicability to real-world problems and datasets. For each of the four classification tasks, we divide the evaluation into two types. Firstly, we evaluate the symbolic *learning* capability of FFNSL, where the goal is to learn interpretable knowledge from symbolic features extracted from pre-trained neural network predictions. Secondly, we evaluate the *inference* capability of FFNSL, where the pre-trained neural networks together with the learned knowledge are used to make a downstream classification of *unseen* unstructured data. We refer to the first type of evaluation as the *learned hypothesis evaluation* and the second type as the *FFNSL framework evaluation*, since this targets both neural and symbolic components. Let us now describe each evaluation type in more detail.

### 5.1 Learned hypothesis evaluation

We evaluate the learned hypothesis in terms of accuracy, interpretability and learning time. To measure accuracy, we use a symbolic test set containing ground truth symbolic features. This ensures that the evaluation only targets the accuracy of the learned hypothesis. For each example in the test set, the symbolic features are used with the learned hypothesis to make a prediction of the downstream label. This prediction is compared to the ground truth label in the test set and accuracy is computed using the standard measure. Since FFNSL learns knowledge from a pre-trained neural network, we consider the hypotheses that have been learned at each (increasing) percentage of distributional shift and evaluate the accuracy of the knowledge that FFNSL learns in the presence of incorrect neural network predictions. Note that the symbolic test set remains unchanged and is not affected by the distributional shifts, as we want to evaluate in this case just the accuracy of the learned hypotheses.

To perform a deeper analysis of the accuracy of the learned hypotheses, we take into consideration the following measures. Firstly, the accuracy and confidence score distribution of the pre-trained neural network(s) in classifying unstructured data in the training set  $D$ . Since the neural networks were pre-trained on a dataset different from  $D$ , this measure enables us to understand the reliability of the pre-trained neural network predictions over new unseen input data (For more dataset details, see Appendix C.). Secondly, we measure the percentage of WCDPI examples generated by the LAS system, that contains features in the context which are incorrect with respect to the label in the inclusion set. This enables us to understand the relationship between incorrect neural network predictions and the accuracy of the learned hypotheses, as well as analyse how many correct WCDPI examples are needed to learn hypotheses with a certain level of accuracy. Thirdly, we calculate the weight penalty ratio  $r$  over the generated WCDPI examples, defined as

$$r = \frac{\sum_{e \in E'_{\text{correct}}} e_{\text{pen}}}{\sum_{e \in E'} e_{\text{pen}}}$$

where  $E'_{\text{correct}}$  is the set of correctly generated WCDPI examples, (i.e., WCDPI examples with features in the context that are consistent with the label in the inclusion set) and  $E'$  is the complete set of generated WCDPI examples. This enables us to measure the bias given to the LAS system by the weights of the WCDPI examples, which are based on the neural network



confidence scores. Ideally, FFNSL should allocate a higher proportion of the total weight penalty to WCDPI examples that contain correct neural network predictions. We compare the accuracy of the knowledge learned from these WCDPI examples with that of knowledge learned from corresponding WCDPI examples where we fix the penalty to be constant for all examples, as a baseline. To measure interpretability, we count the total number of atoms in a learned hypothesis: a hypothesis with a lower number of atoms is considered to be more interpretable (Lakkaraju et al., 2016). Finally, we measure the wall-clock time taken to learn a hypothesis at each percentage of distributional shift.

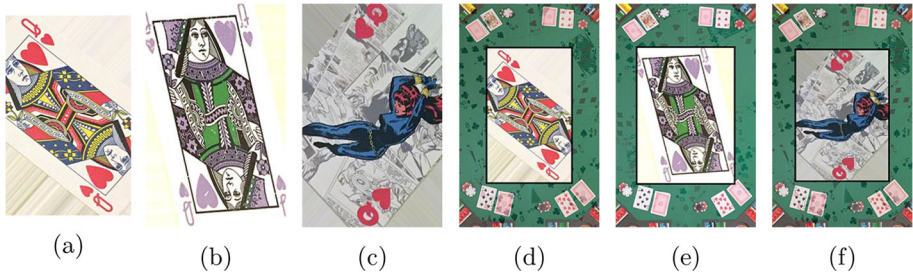
## 5.2 FFNSL framework evaluation

When a hypothesis has been learned, the entire FFNSL framework can be evaluated using a test set containing unseen labelled unstructured data. In this case, the neural network component of FFNSL classifies each element of a sequence of unstructured data. The symbolic features predicted from the neural network classification are added to the background knowledge alongside the learned hypothesis. The symbolic component of FFNSL is used to compute the downstream prediction. This is compared to the ground-truth label associated with the sequence of unstructured data, and the accuracy is computed with the standard measure. To assist the evaluation, and to provide insight into where mistakes are being made, we evaluate the neural network accuracy in predicting the symbolic features from the unstructured data in the test set with respect to ground truth information. This enables us to identify whether any downstream classification error is due to neural network feature prediction, the learned hypothesis, or both. We also evaluate FFNSL under distributional shifts. We inject into the test data the same percentages of distributional shifts used during the learning of hypotheses, and evaluate the accuracy of FFNSL. This evaluates the performance of FFNSL in realistic scenarios where distributional shifts occur during learning and inference.

## 5.3 Experimental setting

In the next four sections, we present the results of the Follow Suit Winner, Sudoku Grid Validity, Crop Yield Prediction, and Indoor Scene Classification tasks, using the evaluation methodology outlined in this section. In the first three tasks we pre-train a Softmax CNN and an EDL-GEN neural network and when used in combination with a symbolic learning system in FFNSL, we refer to these as *FFNSL Softmax* and *FFNSL EDL-GEN* respectively. For the Indoor Scene Classification task we adopt a pre-trained network, called *Semantic Aware Scene Recognition (SASR)*, tailored to the task of scene classification. We use a Random Forest (RF) and neural network as baseline rule learning approaches in all tasks, and they both use the same pre-trained Softmax neural network for feature extraction as used in FFNSL Softmax, and are trained to learn the knowledge needed to predict the downstream label, given Softmax neural network predictions. In the Indoor Scene Classification task the SASR network is used. The RF is chosen as a powerful decision tree approach, known for being a lightweight model that is quick to train and exhibits a certain level of interpretability. In the Follow Suit Winner and Crop Yield Prediction tasks, the neural network is a Fully Connected Network (FCN), chosen to evaluate a deeper architecture, and in the Sudoku Grid Validity task, the neural network is a Convolutional Neural Network-Long Short Term Memory (CNN-LSTM) designed for sequence classification problems where the CNN component can learn spatial dependencies in the Sudoku grid. Full details of the baseline architectures are given in Appendix D. To measure interpretability of the RF baseline, we used the first tree in the forest and extract a





**Fig. 3** Example playing card images

rule from each branch (from root to leaf) of this tree. For the neural network baselines, we fitted a surrogate decision tree model (Molnar, 2019) to approximate black-box predictions, and applied the same rule extraction methods as that used for the RF. Let us now present our results.

## 6 Follow suit winner

In this section we present the results of the Follow Suit Winner task. We start with Softmax and EDL-GEN neural networks pre-trained on standard playing card images and apply *minor* and *major* distributional shifts by substituting standard playing card images with images from alternative decks. Example images are shown in Fig. 3 for the queen of hearts card taken from the *Standard* Fig. 3a, *Batman Joker* Fig. 3b, *Captain America* Fig. 3c, *Adversarial Standard* Fig. 3d, *Adversarial Batman Joker* Fig. 3e and *Adversarial Captain America* Fig. 3f decks.

The Batman Joker and Captain America decks represent *minor* distributional shifts; the adversarial decks represent instead *major* distributional shifts where card images from each of the Standard, Batman Joker and Captain America decks are placed against a background containing additional card images from the Standard deck. These adversarial decks are designed to trick the neural networks into predicting incorrectly as images from the Standard deck are from the same distribution as the card images used during neural network pre-training. In order to understand the challenge faced by the LAS system when learning from neural network feature predictions in the presence of distributional shifts, Fig. 4 presents the accuracy and confidence score distribution of pre-trained neural networks when evaluated on different playing card decks than the one used for pre-training.

In Fig. 4, each row shows the type of neural network, the playing card deck used for evaluation, the predictive accuracy, and the confidence score distribution. As one would expect, the accuracy was very high when classifying playing card images from the standard deck, as this was the deck used for pre-training. For the Softmax neural network the confidence score was also very high in this case, whereas EDL-GEN had more distributed confidence scores. When evaluating the pre-trained networks on decks different from the one used in training, the Softmax neural network still reported high confidence despite its overall low accuracy, whereas the EDL-GEN network reported comparable low accuracy but with much lower confidence. For example, evaluating the networks over the *Captain America* deck (see 3rd and 9th rows), 96% of Softmax predictions were made with confidence in the interval  $[0.95, 1]$ , despite an accuracy of 0.0697, whereas only 10% of EDL-GEN predictions were made within this same confidence interval. As for the overall accuracy, EDL-GEN performed slightly better than Softmax over decks representing minor distributional shifts,

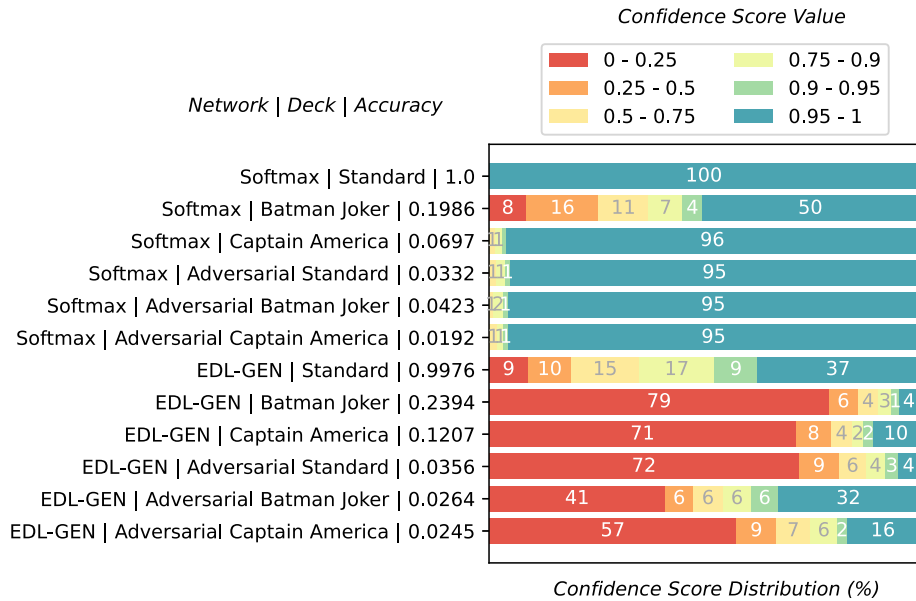


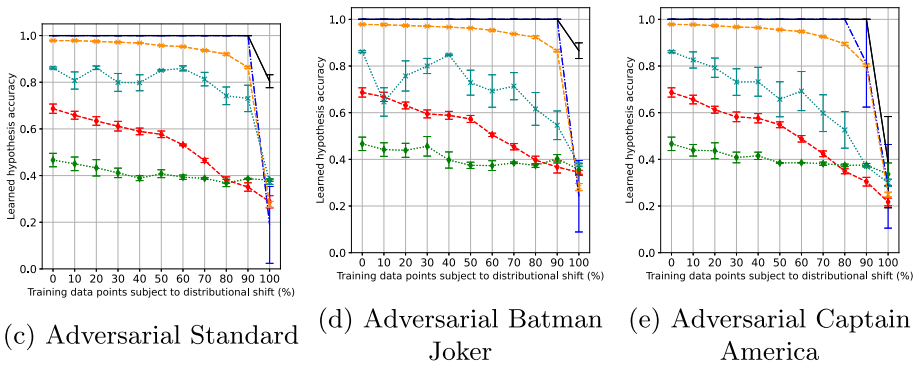
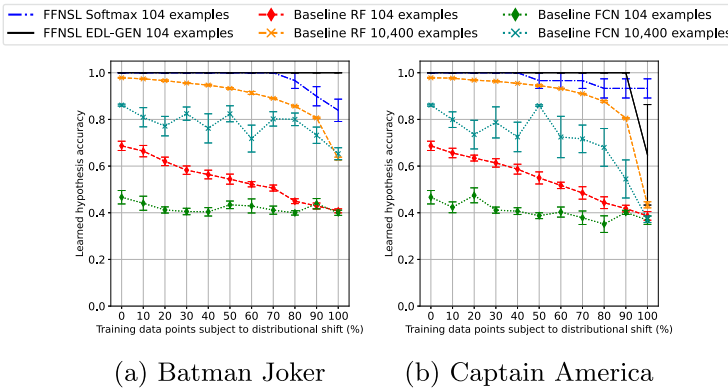
Fig. 4 Neural network performance under distributional shifts

whereas both networks performed in a similar way when applied to decks representing major distributional shifts. This highlights the challenge for our FFNSL framework in learning knowledge when presented with out-of-distribution data, as neural network predictions are likely to be incorrect, and may potentially be made with high confidence.

### 6.1 Learned hypothesis evaluation

Figure 5 presents the accuracy of the learned hypotheses when an increasing percentage of labelled unstructured data were subject to distributional shifts, applied with cards from the alternative decks. The reported accuracy is the mean accuracy over 5 repeats and the error bars indicate standard error.

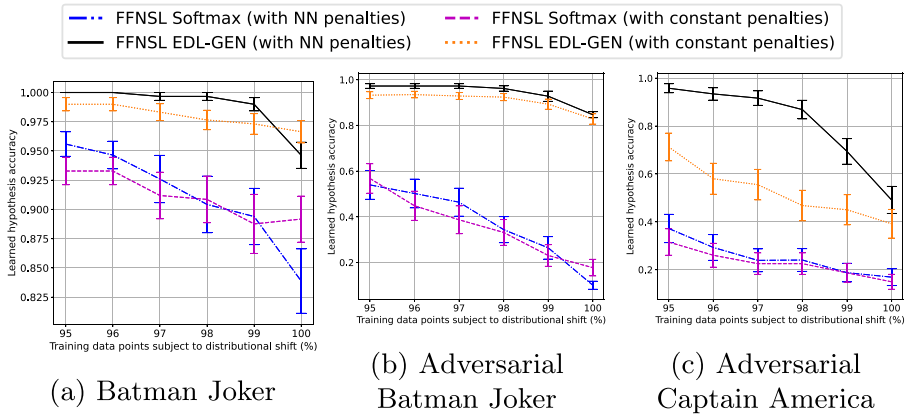
FFNSL outperformed the baselines and learned far superior hypotheses when up to 90% of labelled unstructured data were subject to distributional shifts. This was the case for both instances of FFNSL. The baselines required 100× the number of examples in order to perform close to FFNSL, and despite the significant increase in the amount of data used by the baselines, FFNSL still learned more accurate hypotheses. Figures 5a and b refer to the injection of minor distributional shifts. In these two cases, when the percentage of distributional shift was very high (above 90%), the accuracy of the FFNSL learned hypotheses decreased, but still remained between ~ 70–100%, whereas the accuracy of the baselines trained with the same amount of data reduced to ~ 40%. Figure 5c–e refer to the injection of major distribution shifts. FFNSL Softmax had similar performance in Fig. 5c and d but a much lower accuracy than that shown with minor distributional shifts when 90% or more of the unstructured data were subject to distributional shifts. The FFNSL EDL-GEN maintained instead a higher accuracy in these cases. We now perform a more in-depth analysis to explore the reasons for dropping accuracy in the presence of high percentages of distributional shifts. Given the two groups of similar behaviours we consider only two representative cases: *Batman*



**Fig. 5** Accuracy of learned hypotheses with increasing percentages of data subject to distributional shifts, Follow Suit Winner task. 5 repeats

*Joker*, as minor distributional shift, and *Adversarial Batman Joker* as major distributional shift. A full set of analysis results, with respect to all the other card decks, is given in Appendix A.

In particular, we explore whether FFNSL EDL-GEN provides a performance benefit over FFNSL Softmax, and if so, what are the contributing factors. Specifically we analyse the accuracy performance in relation to either or both (i) better neural network predictive accuracy, when classifying out-of-distribution data, and (ii) more informative weight penalties of the generated WCDPI examples, calculated from the neural network confidence scores. For this analysis we focus on high percentages of distributional shifts, 95–100%, as this was when FFNSL instances deteriorated in their learned hypothesis accuracy. We run 50 experimental repeats to generate statistically significant results. In order to isolate the effect of the example weight penalties, we also run two additional baseline FFNSL instances where the weight penalties of the generated WCDPI examples are all constant and equal to 10. The results are shown in Fig. 6. We have also included the performances with respect to distributional shifts given by the Adversarial Captain America deck (Fig. 6c), since Fig. 5e shows that in this case the accuracy of both FFNSL instances decreased to around 40% when nearly 100% of the data were subject to distributional shifts. Full analysis of this deck is presented in Appendix A.



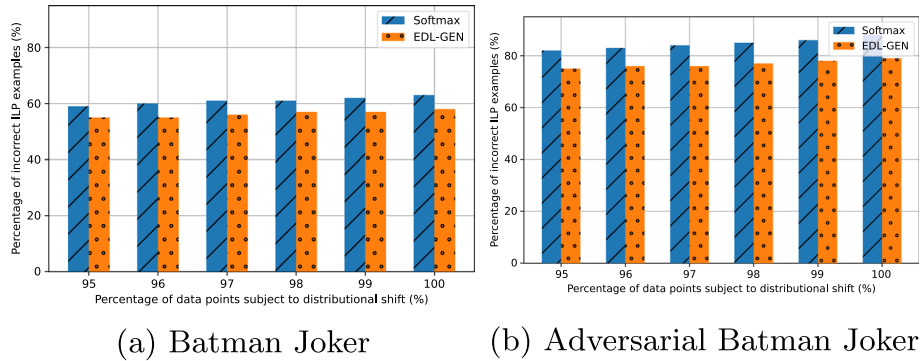
**Fig. 6** FFNSL Softmax vs. FFNSL EDL-GEN. Accuracy of learned hypotheses with 95–100% distributional shifts using 50 repeats. Follow Suit Winner task

For both Batman Joker and Adversarial Batman Joker, FFNSL EDL-GEN outperformed FFNSL Softmax. Note, however, the difference in y-axis scale between Fig. 6a and b and the difference in FFNSL performance. This was due to the fact that both Softmax and EDL-GEN neural networks predicted more accurately on the Batman Joker deck than the Adversarial Batman Joker deck, as presented in Fig. 4. The improved performances of FFNSL EDL-GEN versus FFNSL Softmax did not seem to depend on the more informative weights of WCDPI examples with weights calculated from neural network confidence scores, versus constant weights, since the accuracy of FFNSL instances (denoted “...with NN penalties”) was similar to that of the respective baselines with constant weight penalties. However, Fig. 6c shows that when the distributional shift was more severe,<sup>11</sup> the decrease in accuracy of FFNSL EDL-GEN was less drastic than that of its corresponding baseline with constant penalty, whereas there was no difference in the case of FFNSL Softmax. Even though the overall accuracy of the framework was lower than that reported for less drastic forms of distributional shifts, the more informative weight penalties of WCDPI examples, calculated from the EDL-GEN neural network confidence scores, provided a clear benefit compared to using constant weights, in particular when the percentage of distributional shift was very high.

It still remains open the question as to why FFNSL EDL-GEN performed better than FFNSL Softmax in Fig. 6a and b. For the percentages of distributional shifts between 95–100%, the pre-trained neural networks both reported low average accuracy. So a natural question to ask is whether EDL-GEN led to more consistent symbolic feature predictions than Softmax. This is important to investigate because LAS systems are capable of learning accurate hypotheses from few “good” examples. So, we investigated the percentage of incorrect WCDPI examples generated when 95–100% of the unstructured data were subject to distributional shifts. These were examples whose contextual symbolic features were inconsistent with the ground-truth label due to incorrect neural network predictions.

Figure 7 shows, first of all, that the percentage of generated incorrect WCDPI examples was lower than the corresponding percentage of data subject to distributional shifts. This indicated that some correct WCDPI examples could be generated even when the neural

<sup>11</sup> Recall that the accuracy of Softmax and EDL-GEN neural networks was the lowest for the Adversarial Captain America test set, as shown in Fig. 4.



**Fig. 7** The effect of distributional shifts on percentage of incorrect WCDPI examples generated. Follow Suit Winner task

networks made incorrect predictions. Incorrect predictions made over the 4 cards played could in combination lead to predicted symbolic features for the trick whose winning player would match the ground truth label. Secondly, more correct WCDPI examples were generated when the distributional shift was given by the Batman Joker deck, compared to that of the Adversarial Batman Joker deck. This was because, as indicated in Fig. 4, the neural network accuracy for the former was better than that for the latter. Furthermore, EDL-GEN led to a lower number of incorrect WCDPI examples compared to that of Softmax in both forms of distributional shifts, and this difference was bigger in the case of the Adversarial Batman Joker deck. Given the relatively small number of WCDPI examples used by the LAS system (104), this difference contributed to the larger gap in accuracy between FFNSL Softmax and FFNSL EDL-GEN in Fig. 6b than Fig. 6a.

Now, how did the weight penalty, generated from the neural network confidence score, effect the accuracy of learned hypotheses? Clearly, EDL-GEN provided improved confidence scores than the Softmax neural network which in-turn, improved the accuracy of FFNSL. This explains why the accuracy of each FFNSL approach was higher in Fig. 6a than that shown in Fig. 6b. However, Fig. 6 shows that for FFNSL Softmax, using WCDPI example weight penalties calculated from Softmax neural network confidence scores appears to have no benefit compared to using WCDPI examples with constant weight penalties. However, this was different in the case of FFNSL EDL-GEN. To investigate this further, we calculated the weight penalty ratio for the WCDPI examples generated from both Softmax and EDL-GEN neural network confidence scores. The analysis is shown in Fig. 8 for each deck and 95–100% distributional shifts.

Figure 8 shows that the weight penalty ratio calculated from EDL-GEN confidence scores provided a clear benefit than that calculated from the Softmax neural network confidence scores, which was instead very similar to the weight penalty ratio given by constant penalties. At 100% distributional shifts, the benefits of calculating WCDPI example weight penalties with the neural network confidence scores reduced, as there were very few correct examples. This explains why the gap between the accuracy of the FFNSL EDL-GEN with neural network penalties and that of FFNSL EDL-GEN with constant penalties, in both decks, reduced as distributional shifts increase towards 100% (see Fig. 6). In summary, improved accuracy of the neural network predictions led to a higher percentage (even if small) of correct WCDPI examples and improved neural network confidence scores led to an improved penalty ratio of correct WCDPI examples. Together they provided an improved bias for the LAS system

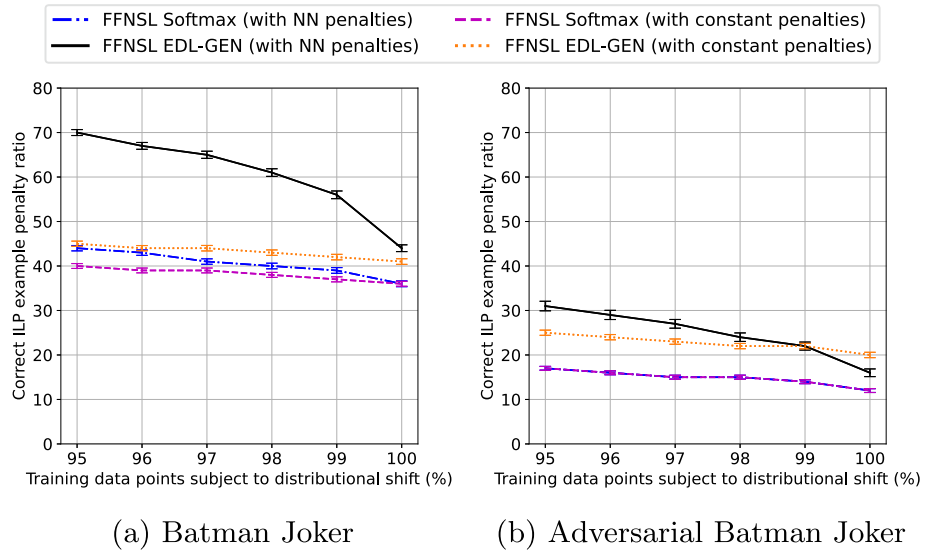


Fig. 8 WCDPI example weight penalty ratio. Follow Suit Winner task

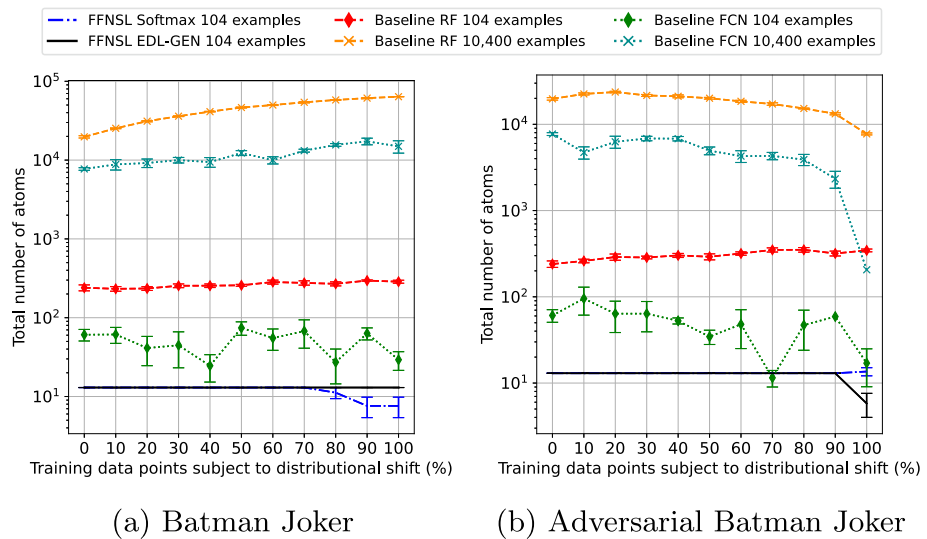


Fig. 9 Interpretability of the learned hypotheses, Follow Suit Winner task

which even if it was reduced to learn from a small percentage of correct examples, these had improved penalty weight to guide the search for optimal solutions.

Let us now investigate the interpretability of the hypotheses learned using our FFNSL framework compared to that of the baseline approaches. Figure 9 shows the results, where interpretability was measured in terms of the number of atoms that formed the learned hypothesis.

FFNSL learned significantly more interpretable knowledge than the baseline approaches (note the logarithmic scale on the y-axis). In the case of the minor form of distributional

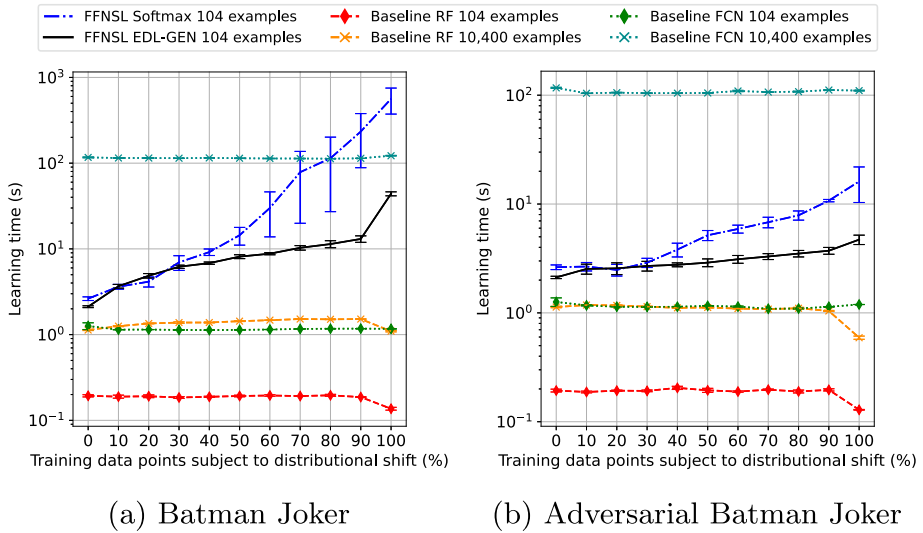


Fig. 10 Learning time. Follow Suit Winner task

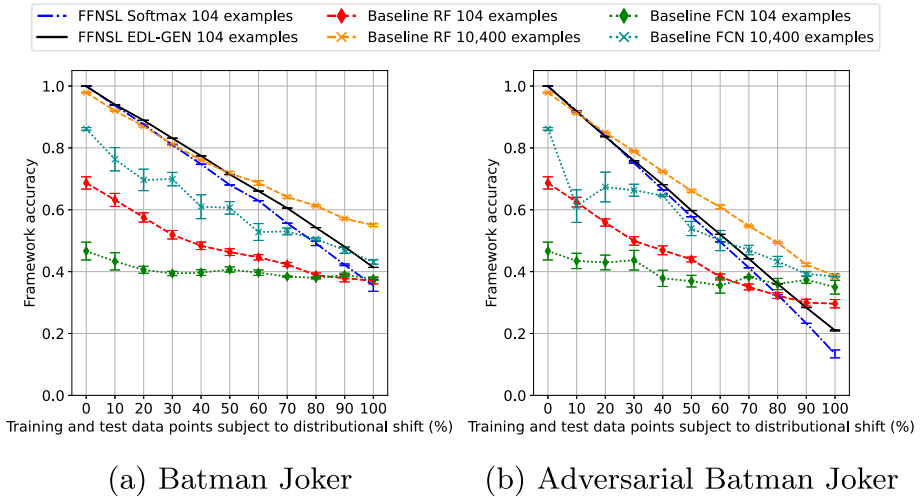
shift (see Fig. 9a), the interpretability of the baseline models trained with  $100\times$  the amount of data decreased as distributional shift increased. These models reached high accuracy by training over a much larger dataset (see Fig. 5a), but they did so at the cost of much lower interpretability. This was because they learned a more complex mapping between input and output, instead of learning general rules, as was the case for our FFNSL approach. The FCN trained with the same amount of data as FFNSL had similar interpretability to that of FFNSL, because the model learned to largely predict the same class and the surrogate decision tree was very small. This was reflected in the poor performance of the FCN shown in Fig. 5d for the Adversarial Batman Joker deck. Examples of interpretable knowledge learned by our FFNSL approaches are presented in Appendix A.

Finally, to investigate the scalability of FFNSL, we have also computed the time required to learn an interpretable hypothesis. The results are shown in Fig. 10.

Both FFNSL approaches learned with an order of magnitude of time similar to that of the FCN trained with the same number of examples when no distributional shifts were applied to the data. As distributional shifts increased, FFNSL took longer because the ILASP system required more iterations to prove optimality with respect to minimising the total penalty on the examples. However, the learning time of FFNSL EDL-GEN did not increase as quickly, when compared to FFNSL Softmax. This was because the WCDPI example weight penalties were much more informative (see Fig. 8) and the ILASP learning system required fewer iterations overall to prove optimality.

In conclusion, our analysis shows that FFNSL outperformed the baseline approaches in terms of accuracy and interpretability, even when the baselines were trained with  $100\times$  the amount of data. FFNSL EDL-GEN outperformed FFNSL Softmax, in the accuracy of the learned hypotheses, as EDL-GEN neural network predictions were more accurate, and this influenced the downstream performance of the FFNSL framework more than the neural network confidence scores. When major distributional shifts were applied, the EDL-GEN uncertainty-aware neural network led to significantly more informative WCDPI example weight penalties compared to the Softmax neural network, although this benefit diminished





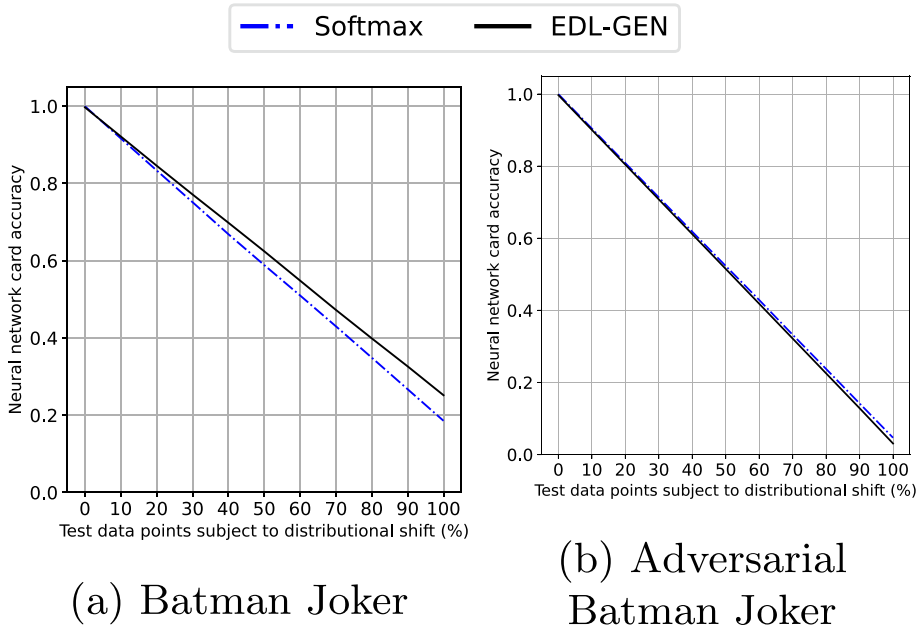
**Fig. 11** Accuracy of the FFNSL framework when training and test data were subject to distributional shifts. Follow Suit Winner

as the percentage of input data subject to distributional shifts approached 100%. Finally, we have shown that more informative WCDPI example weight penalties resulted in faster hypothesis learning times, when the iterative ILASP system was used.

### 6.2 FFNSL framework evaluation

Figure 11 presents the accuracy of the entire FFNSL framework when evaluated over a test data subject to the same types of distributional shifts used during the learning of interpretable knowledge. The mean accuracy is reported and the error bars denote standard error over 5 repeats.

FFNSL outperformed the baselines trained with the same amount of data at each percentage of distributional shift on the Batman Joker deck, and until ~80% distributional shift on the Adversarial Batman Joker deck. The baselines required 100× the amount of data in order to match or outperform FFNSL. On the Adversarial Batman Joker deck, the performance was lower for all approaches when the percentage of distributional shift was high, due to the neural networks predicting with lower accuracy (see Fig. 4). The baselines trained with the same amount of data outperformed FFNSL for > 80% distributional shifts. This was because they largely predicted player 1 and was sufficient to reach approximately 40% of accuracy on the test set: the Follow Suit Winner task is biased towards player 1 because winning depends on playing the highest ranked card with the same suit as player 1. In the test set, 38.6% of the data was indeed labelled with player 1 as the winner, which roughly corresponds to the performance of the baselines trained with the same number of examples at 100% shifts in Fig. 11b. FFNSL EDL-GEN outperformed FFNSL Softmax because of two reasons. Firstly, the rules learned by FFNSL EDL-GEN, in the presence of high percentage of distributional shifts, were more accurate (see Figs. 5a and d) because of the lower number of incorrect WCDPI examples when the EDL-GEN neural network was used (Fig. 7). In addition, the EDL-GEN neural network provided more informative bias to the LAS system through better WCDPI example weight penalties (Fig. 8). Finally, the decrease in perfor-



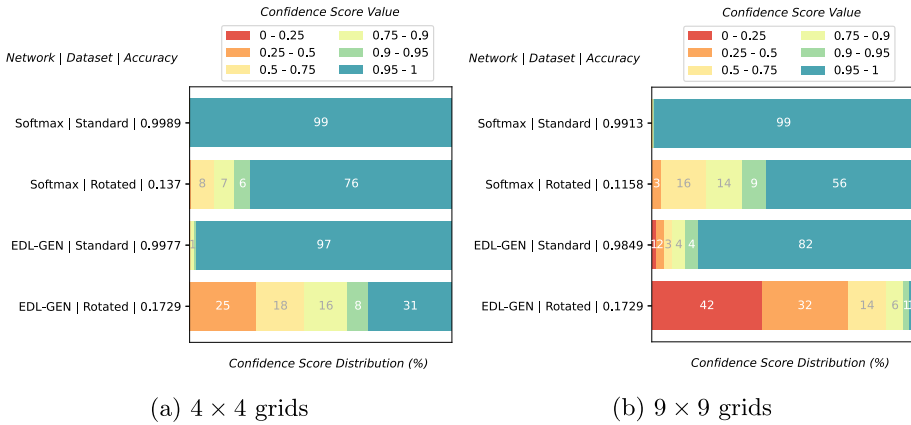
**Fig. 12** Neural network card accuracy when test data points were subject to distributional shifts. Follow Suit Winner task

mance of the FFNSL approaches over unseen data subject to distributional shift seemed to be linear in the percentage of applied distributional shift. This was primarily due to the accuracy of the neural network feature predictions. Figure 12 shows that indeed the accuracy of neural network predictions over unseen card images decreased linearly with the increase of the percentage of distributional shifts.

As shown in Fig. 12, the EDL-GEN neural network was more accurate than Softmax in predicting unseen playing cards in the case of minor distributional shift given by the Batman Joker deck. However, the accuracy of the neural networks was the same in the case of major distributional shift given by the Adversarial Batman Joker deck. This was why in Fig. 11, FFNSL EDL-GEN's showed better performance on the Batman Joker deck. For the Adversarial Batman Joker deck, FFNSL EDL-GEN's better performance than FFNSL Softmax was primarily due to more accurate hypotheses.

## 7 Sudoku grid validity

Having presented in detail the performance of our FFNSL approaches on the Follow Suit Winner task, we now explore whether the approach can generalise to other tasks. We have applied our approach to a different classification task, the Sudoku Grid Validity task and we present the results in this section. We consider two cases: a  $4 \times 4$  Sudoku grid size, for which the sequence  $\mathbf{x}$  of unstructured data is much longer than that used for the Follow Suit Winner task. Therefore, each generated WCDPI example contains more contextual features that are likely to be predicted incorrectly, as a result of distributional shifts applied to input images. We then evaluate the scalability of the FFNSL framework even further by considering  $9 \times 9$



**Fig. 13** Neural network performance under distributional shifts, Sudoku Grid Validity task

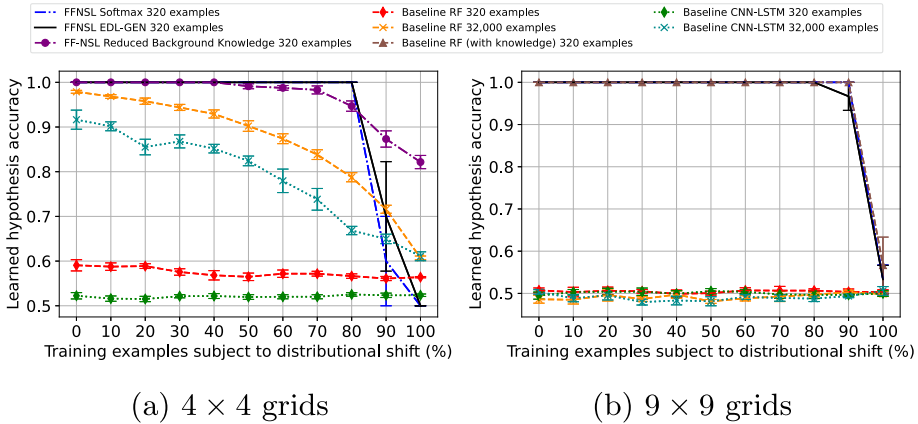
Sudoku grid sizes. For the Sudoku grid validity tasks, the FFNSL instance makes use of the FastLAS system, which has been shown to scale to handle large hypothesis spaces (Law et al., 2020).

We first pre-train both Softmax and EDL-GEN neural networks on standard images from the MNIST training set. In all experiments, we used MNIST digits 1–4 and 1–9 for the respective Sudoku grid size tasks. Figure 13 shows the accuracy and confidence score distribution of the pre-trained neural networks for the 4 × 4 and 9 × 9 grid tasks, on two test sets: a *standard* MNIST test set, and a test set where the MNIST digits have been *rotated* 90° clockwise, representing a distributional shift. The test sets also contain MNIST digits 1-4 or 1-9, depending on the Sudoku grid size.

The results are similar to the Follow Suit Winner task. The Softmax neural network predicted with high confidence also over data subject to distributional shift, despite its low test set accuracy. The EDL-GEN neural network predicted more accurately than Softmax on data subject to distributional shift, but Softmax was slightly more accurate on the standard test sets.

### 7.1 Learned hypothesis evaluation

Figure 14 presents the accuracy of the hypotheses learned from unstructured data with increasing percentages of distributional shift, given by rotating MNIST digit images. We plot the mean accuracy over 5 repeats and the error bars denote standard error. In both Sudoku Grid Validity tasks the FFNSL approaches have as input a background knowledge that encodes the concept of a Sudoku grid (see Appendix F for details). For the 4 × 4 task, we created an additional, more challenging task with a reduced background knowledge where facts about column, row and block were not given but implicitly inferred from a more general notion of division and cell coordinates (given as meta-data). For the 9 × 9 task, we also created an additional training task for the RF, (which was the best performing baseline), where pre-trained neural network predictions were post-processed into 3 Boolean features: whether digits were in the same row, column or block, which was given as input to the RF. This type of input effectively encoded the Sudoku grid knowledge into the RF learning task, and constituted even more information than what was provided to our FFNSL approaches. We demonstrate



**Fig. 14** Accuracy, over 5 repeats, of the learned hypotheses with increasing percentages of data subject to distributional shifts, Sudoku Grid Validity task

that FFNSL performed similarly to this baseline with additional background information. Full FastLAS task listings are given in Appendix F.

FFNSL approaches outperformed the baselines in both  $4 \times 4$  and  $9 \times 9$  tasks by learning far more accurate hypotheses. In the  $4 \times 4$  task, the baselines required  $100\times$  the amount of data to reach an accuracy closer to that of FFNSL, whereas in the  $9 \times 9$  task, the baselines failed completely. In Fig. 14a the purple line is FFNSL with the explicit background knowledge about the Sudoku grid removed. In this case, the FFNSL approach used the EDL-GEN neural network, and it outperformed the baselines. It also outperformed the other two FFNSL approaches, which used explicit background knowledge about the Sudoku grid, when 90% and 100% distributional shifts were applied to the data. This was because with less explicit facts about the grid, the symbolic learner FastLAS was less constrained and alternative hypotheses could be learned which better accommodated the (incorrect) predictions of the neural networks. With explicit facts about the Sudoku grid, the hypothesis space contained rules that performed either very well or very poorly. In Fig. 14b, the brown line shows the accuracy of the RF with the 3 Boolean input features, post-processed from the pre-trained neural network predictions, indicating if digits were in the same row, column or block. FFNSL approaches performed similarly to this baseline that used extra input knowledge.

We investigate our results further to understand whether using the pre-trained EDL-GEN neural network provides a benefit over Softmax in the presence of high percentages of distributional shifts, also in this domain. We focused on 80–96% distributional shifts for the  $4 \times 4$  Sudoku Grid Validity task and 95–99% distributional shifts for the  $9 \times 9$  task, as this was where the performance of FFNSL deteriorated. Similarly to the Follow Suit Winner task, we run 50 experimental repeats and run two baseline FFNSL approaches with constant weight penalties. Figure 15 shows our further experimental results.

Firstly, in both cases of  $4 \times 4$  and  $9 \times 9$  grids, the FFNSL Softmax and FFNSL EDL-GEN that used WCDPI example weight penalties calculated from neural network confidence scores, outperformed the corresponding FFNSL with constant weight penalties. To investigate this further, we explore the WCDPI example weight penalty ratio for the  $4 \times 4$  and  $9 \times 9$  tasks.

Figure 16 shows that both FFNSL Softmax and FFNSL EDL-GEN with neural network penalties had a larger weight penalty ratio than the corresponding FFNSL with constant

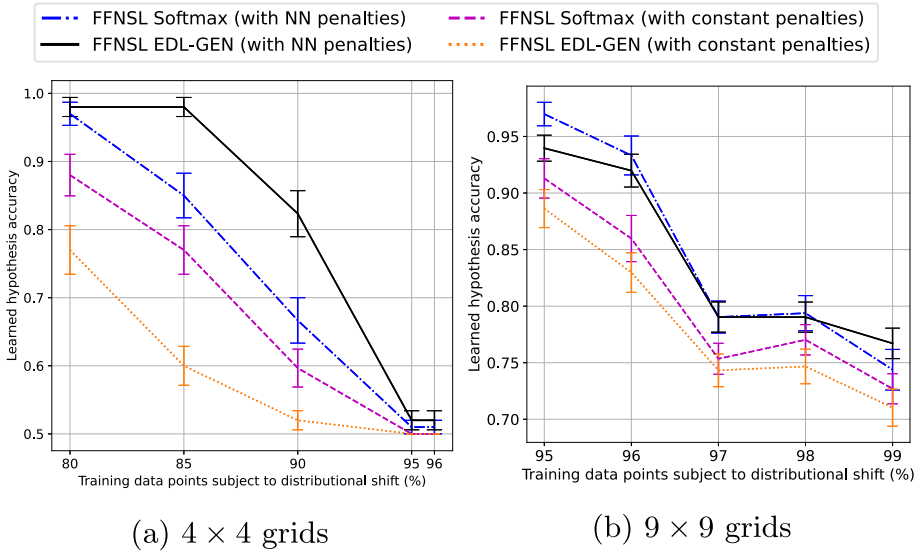


Fig. 15 FFNSL Softmax versus FFNSL EDL-GEN. Average accuracy of learned hypotheses over 50 repeats. Sudoku Grid Validity task

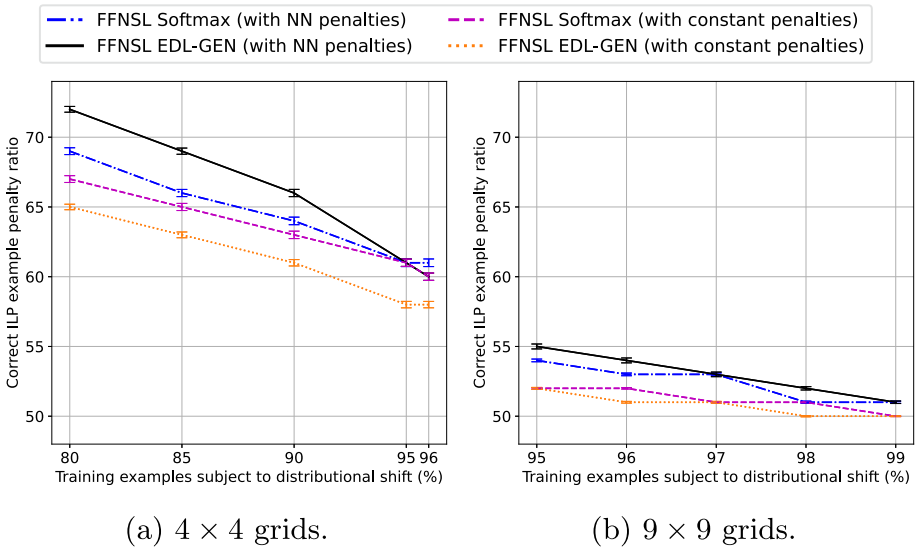
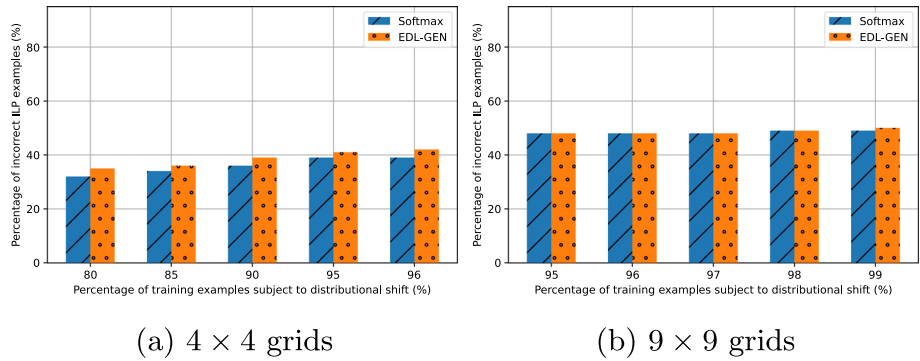


Fig. 16 ILP example weight penalty ratio, Sudoku Grid Validity task

penalties. For FFNSL EDL-GEN this was expected, and for FFNSL Softmax, this is explained by the fact that, as shown in Fig. 13, the Softmax neural network had a more varied confidence score distribution over data subject to distributional shift (rotated digits). The difference between the WCDPI example weight penalty ratio with neural network penalties and constant penalties explains the performance gain of FFNSL with neural network penalties versus FFNSL with constant penalties in Fig. 15. Similarly, the difference between the example



**Fig. 17** The effect of applying distributional shifts on the percentage of incorrect ILP examples, Sudoku Grid Validity task

weight penalty ratio of FFNSL EDL-GEN and that of FFNSL Softmax, with neural network penalties, also explains why FFNSL EDL-GEN outperformed FFNSL Softmax in Fig. 15a for  $4 \times 4$  grids. For the  $9 \times 9$  task the difference between the example weight penalty ratio of FFNSL EDL-GEN and that of FFNSL Softmax, with neural network penalties, is very small and this explains why FFNSL EDL-GEN and FFNSL Softmax show a similar performance in Fig. 15b.

Now, in Fig. 15a, FFNSL Softmax with constant weight penalties outperformed FFNSL EDL-GEN with constant weight penalties, whereas in Fig. 15b these two approaches performed similarly. To investigate this further, we consider the percentage of incorrect ILP examples when distributional shifts were applied. The results are shown in Fig. 17.

For the  $4 \times 4$  task, despite the pre-trained EDL-GEN neural network predicting on average more accurately (see Fig. 13), it led to a higher percentage of incorrect WCDPI examples than the pre-trained Softmax neural network, as shown in Fig. 17a. This explains the lower performance in Fig. 15a of FFNSL EDL-GEN with constant penalties. Using WCDPI example weight penalties calculated with EDL-GEN neural network confidence scores was able to rectify this and bias the LAS system to focus on learning a hypothesis from WCDPI examples containing correct neural network predictions. For the  $9 \times 9$  task both pre-trained neural networks led to a similar percentage of incorrect WCDPI examples (see Fig. 17b), which explains the similar performance of FFNSL EDL-GEN and FFNSL Softmax with constant penalties (shown in Fig. 15b).

Let us now investigate the interpretability of FFNSL compared to the baseline approaches. The results are shown in Fig. 18.

Again, FFNSL learns significantly more interpretable hypotheses than the baseline approaches. Example learned hypotheses are presented in Appendix A. As for the learning time, results are shown in Fig. 19.

In both  $4 \times 4$  and  $9 \times 9$  tasks, the learning time for FFNSL did not increase as the percentage of input data subject to distributional shifts increased. This was because the FastLAS learning system used by FFNSL learned a hypothesis by solving an optimisation problem with respect to all generated WCDPI examples. This was not the case for the Follow Suit Winner task where the ILASP system learned an optimal hypothesis iteratively over the examples. It is interesting to note that in Fig. 19a, FFNSL's learning time had the same order of magnitude as that of the CNN-LSTM trained with  $100\times$  the amount of data, which had lower accuracy up to 90% of distributional shifts.

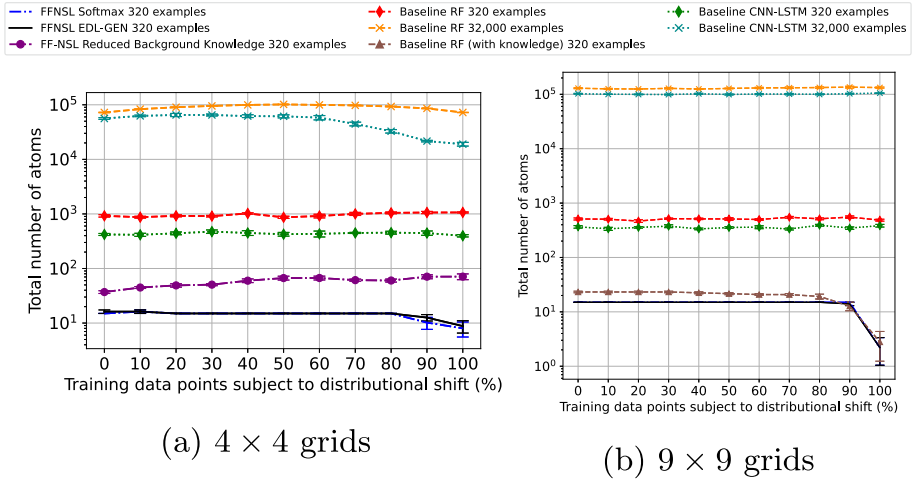


Fig. 18 Interpretability of the learned hypotheses, Sudoku Grid Validity task

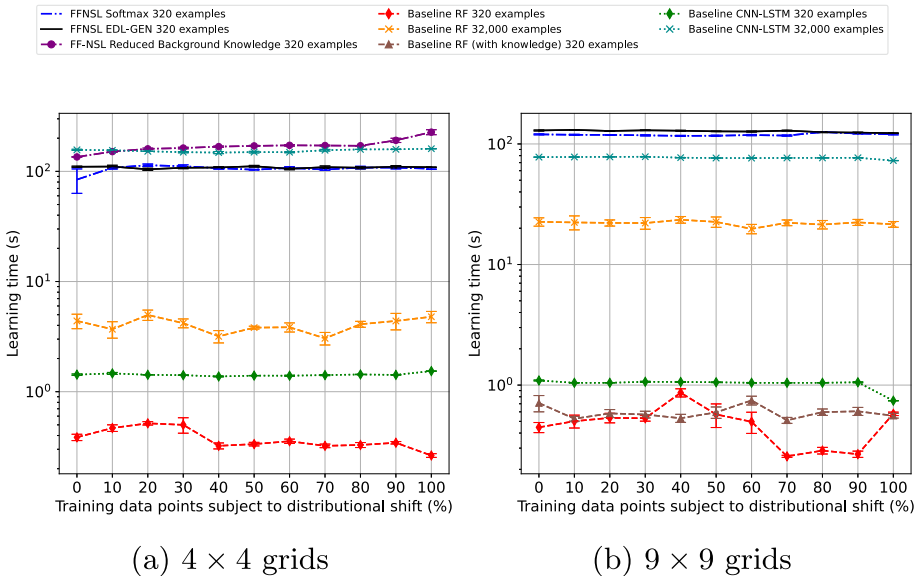
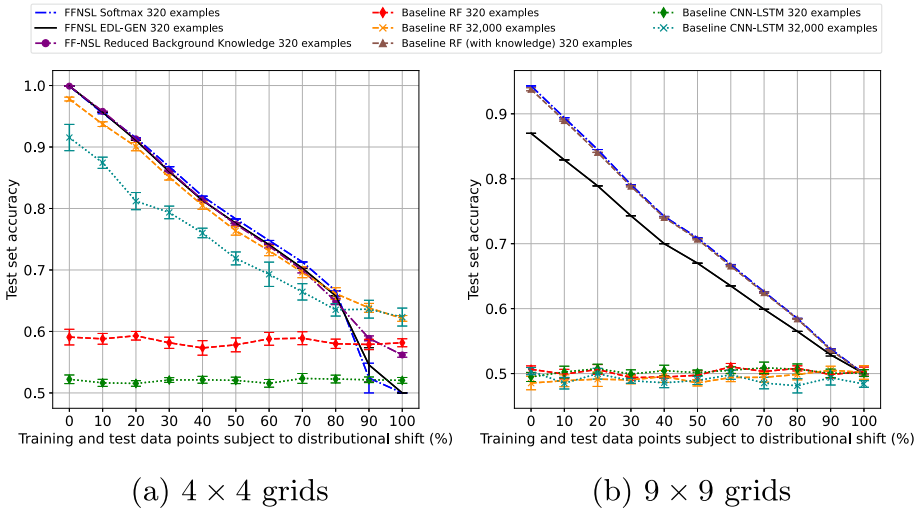


Fig. 19 Hypothesis learning time, Sudoku Grid Validity task

In conclusion, we have shown that for the learned hypotheses evaluation, FFNSL outperformed the baseline approaches in terms of accuracy and interpretability, even when the baselines were trained with  $100\times$  the amount of data. Furthermore, in this task, WCDPI example weight penalties had a larger impact on the performance of FFNSL. We have also shown that FFNSL can scale to learning hypotheses where many more unstructured data points  $x_i$  are observed per labelled input  $(x, y)$ , and in these cases FFNSL learns in a timely manner.





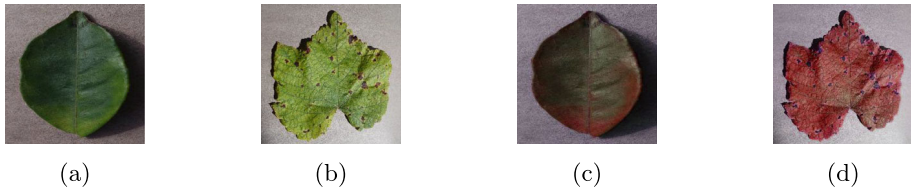
**Fig. 20** Accuracy of the FFNSL framework when training and test data were subject to distributional shifts. Sudoku Grid Validity task

### 7.2 FFNSL framework evaluation

The final evaluation is the accuracy of the overall FFNSL framework when it is applied to a test set of unseen data also subject to distributional shifts. Figure 20 shows the mean accuracy over 5 repeats and the error bars denote standard error.

In the  $4 \times 4$  task, FFNSL outperformed the baselines until 80% of the test data were subject to distributional shifts, even when the baselines were trained with  $100\times$  the amount of data. In the  $9 \times 9$  task, FFNSL outperformed all baselines, with the exception of the RF with additional background knowledge, which performed similarly to FFNSL Softmax. It is indeed interesting to analyse why for the  $9 \times 9$  task FFNSL Softmax outperformed FFNSL EDL-GEN especially when low percentages of test data were subject to distributional shift.

Aside from the accuracy of the learned hypotheses, there were two contributing factors to the test set accuracy shown in Fig. 20b. Firstly, the ability to correctly predict test examples when input data was subject to distributional shifts, and secondly, the ability to correctly predict test examples when no distributional shifts were applied. In the  $9 \times 9$  task, there were many more digit images on the grid for the neural network to predict. For test examples that were not subject to distributional shift, just one single incorrect neural network prediction may have led to a miss-classified example. At 0% shifts in Fig. 20b, FFNSL Softmax outperforms FFNSL EDL-GEN. Now, the Softmax neural network accuracy over unseen and *non-rotated* MNIST digits was 0.9927, whereas that of EDL-GEN neural network was 0.9861. This explains the drop in performance for FFNSL EDL-GEN at 0% shifts. As distributional shifts were applied to the test set for percentages ranging between 10–80%, both FFNSL Softmax and FFNSL EDL-GEN failed to classify most examples subject to distributional shifts, but FFNSL EDL-GEN also failed to classify more examples that were not subject to distributional shifts, when compared to FFNSL Softmax. At distributional shifts  $> 80\%$ , the accuracy of the learned rules also became a factor and the performance of both approaches deteriorated towards 50% accuracy.



**Fig. 21** Example crop images from the Plant Village dataset with and without distributional shift

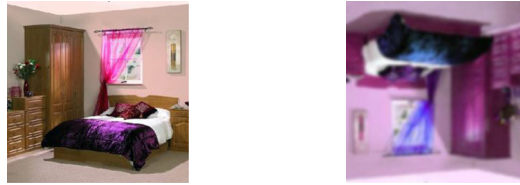
## 8 Real-world datasets

In order to demonstrate FFNSLs applicability to real-world problems and datasets, in this section we present evaluations of two additional tasks: (1) *Crop Yield Prediction*, and (2) *Indoor Scene Classification*, introduced in Sects. 4.3 and 4.4 respectively. Let us now summarise each task.

**Crop yield prediction** The goal is to classify the quality of yield given an image of a particular crop, and symbolic information denoting the crop’s location. Softmax and EDL-GEN neural networks were trained to output species and disease information for each crop image, and the symbolic learner learned knowledge that identifies which predicted crop features correspond to different qualities of yield. We used the Plant Village dataset containing images of healthy and diseased crops (Hughes & Salathé, 2015), and generated a synthetic symbolic dataset for yield prediction. A distributional shift was applied to crop images using a hue filter, after the neural networks were pre-trained on the unaltered images. Example images are shown in Fig. 21 for a grape crop. Fig. 21a and b show *standard* images of a healthy and black measles image respectively, and Fig. 21c and d are the same as Fig. 21a and b respectively, with the distributional shift applied.

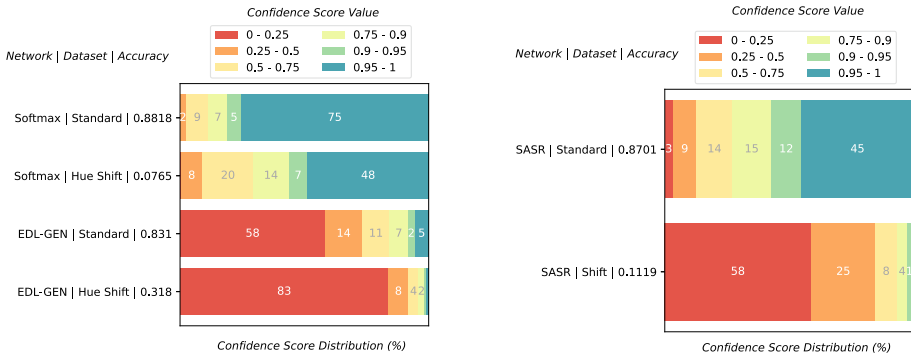
**Indoor scene classification** The goal is to learn knowledge that maps scene level classifications (e.g., bedroom, bathroom, living room) into higher-level super-classes that correspond to a collection of scenes (e.g., home). In this task, we used a state-of-the art neural network called *Semantic Aware Scene Recognition (SASR)* (López-Cifuentes et al., 2020). SASR is a dual-branch CNN that is trained to output scene level classifications, utilising semantic segmentation information, and raw image RGB pixel data on each CNN branch respectively. The symbolic learner then learned the super-class of each scene. Both neural and symbolic datasets are real and were constructed from the MIT Indoor Scenes dataset (Quattoni & Torralba, 2009). To apply a distributional shift, we transformed each image using a Gaussian blur, hue shift, and 180° rotation, after the neural network was trained on unaltered images. An example image for a *bedroom* scene is shown in Fig. 22. In order to obtain results in a timely manner, in this task we implemented a timeout for FastLAS to return the most optimal hypothesis found after 10 min. Also, in contrast to the other tasks, all models were trained with the same dataset size, as the baselines performed strongly when no distributional shift was applied. Finally, only one experimental repeat was performed as the image train/test split was already defined in the dataset (Quattoni & Torralba, 2009).

Figure 23 presents the neural network performance on both tasks, in terms of both accuracy and confidence score distribution when classifying unseen images with and without distributional shift. For the Crop Yield Prediction task in Fig. 23a, the Softmax neural network achieved 88.18% accuracy for the standard images, and performed poorly when classifying hue shift images. In both datasets, predictions were made with very high confidence. The



(a) Standard image      (b) Dist. shift applied

**Fig. 22** Example *bedroom* image from the MIT Indoor scene dataset with (b) and without (a) distributional shift applied



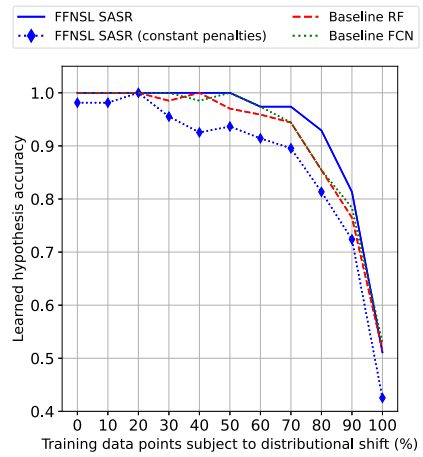
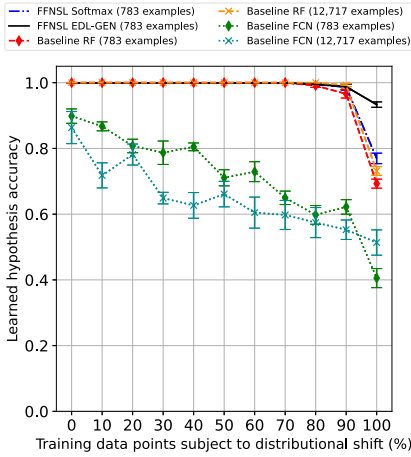
(a) Crop images      (b) Indoor scene images

**Fig. 23** Neural network performance under distributional shifts

EDL-GEN neural network achieved 83.1% accuracy on the standard images, and 31.8% accuracy on the hue shift images which is much higher than Softmax. Crucially, the EDL-GEN neural network predicted with much lower confidence than Softmax on both datasets which better reflects the predictive accuracy. However, the confidence for standard images was somewhat lower than expected, as 58% of predictions were made with less than 25% confidence, despite 83% accuracy. For the Indoor Scene Classification task in Fig. 23b, the distributional shift reduced the network accuracy from 87.01 to 11.19%, although the confidence scores from the SASR network appropriately reflected the reduced accuracy when distributional shift was applied. Although the SASR network does not have an uncertainty-aware architecture like the EDL-GEN networks used in the other tasks, SASR was able to predict with low confidence under our distributional shift. We suspect this was due to the shifted samples falling between the decision boundary of the 67 scene classes, rather than being completely out-of-distribution, enabling the network to better reflect its uncertainty amongst the possible classes. We now present our evaluation of the learned hypotheses in each task.

### 8.1 Learned hypothesis evaluation

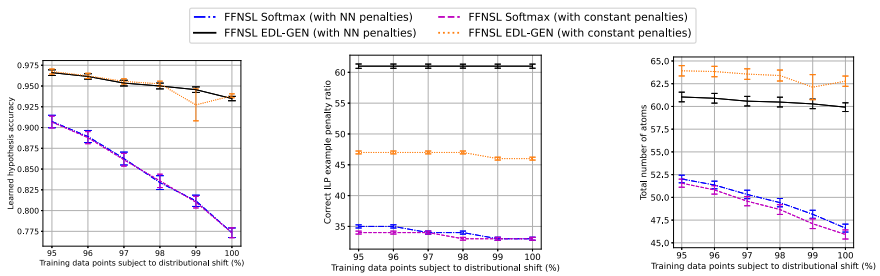
Figure 24 shows the accuracy of the learned hypotheses in each task, when an increasing percentage of labelled unstructured data were subject to distributional shift. In the Crop Yield Prediction task (Fig. 24a), the reported accuracy is the mean accuracy over 5 repeats, and



(a) Crop Yield Prediction

(b) Indoor Scene Classification

Fig. 24 Learned hypothesis accuracy



(a) Hypothesis accuracy (b) Weight penalty ratio (c) Hypothesis length

Fig. 25 The effect of setting ILP example weight penalties based on neural network confidence scores, compared to constant penalties, Crop Yield Prediction task. 95–100% shifts, 50 repeats

the error bars indicate standard error. In this task, both instances of FFNSL learned accurate hypotheses until 90% of the data were subject to distributional shift, and outperformed all the baseline approaches, even when the baselines were trained with significantly more data. In the Indoor Scene Classification task (Fig. 24b), FFNSL also outperformed the baseline approaches, learning the correct hypothesis up to 50% of the data subject to distributional shift. Setting the weight penalties for the examples based on the neural network confidence scores led to more accurate hypotheses, compared to using constant penalties. This is because using the neural network-based weight penalties enabled FastLAS to find a better optimal solution within the 10 min timeout, as the more informative weight penalties gave a clearer optimisation signal for the final solving stage. With constant weight penalties, the optimisation took significantly longer as the distributional shift increased (see Fig. 27b).

Exploring deeper the effect of using example weight penalties set by neural network confidence scores, compared to using constant penalties, we ran 50 experimental repeats between 95–100% shifts. Figure 25 shows the accuracy, weight penalty ratio, and hypothesis length comparison for the Crop Yield Prediction task.

The EDL-GEN instances of FFNSL clearly outperformed the Softmax FFNSL instances (see Fig. 25a). This is because the EDL-GEN neural network predicted with greater accuracy than Softmax when distributional shifts were applied (see Fig. 23a). However, setting the weight penalties of the examples for the symbolic learner based on neural network confidence scores made very little difference in the FFNSL EDL-GEN instances. Therefore, we investigated the weight penalty ratio (Fig. 25b). As expected, both FFNSL Softmax instances had a similar weight penalty ratio, due to the Softmax neural network predicting with high confidence when distributional shift was applied (see Fig. 23a). The EDL-GEN instances do however show a difference, and the neural network-based weight penalties did provide a more informative signal. The question therefore, is why did this not translate into an improvement in learned hypothesis accuracy? It turns out that the benefit was realised in the length of the learned hypothesis (Fig. 25c), as FFNSL EDL-GEN with neural network weight penalties learned a shorter hypothesis than when constant penalties were used. Comparing Fig. 25c to a, you can see that at 99% shifts, when the accuracy of FFNSL EDL-GEN with constant penalties decreased, the length of the learned hypotheses also decreased, whilst FFNSL EDL-GEN with neural network weight penalties achieved a higher accuracy with a shorter hypothesis. With constant penalties, to account for the level of noise, the symbolic learner had to learn more rules that map additional values of location type, plant species and disease to crop yield, in order to maintain the same level of accuracy as when neural network-based penalties were used.

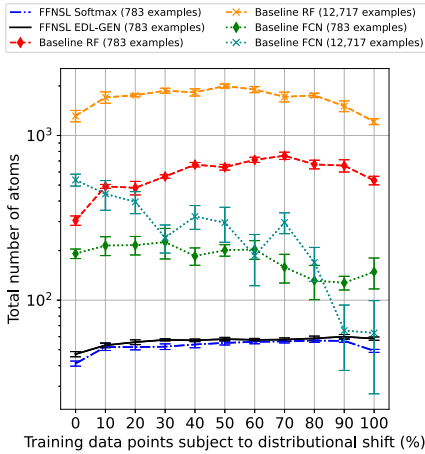
Finally, Figs. 26 and 27 present the interpretability and learning time results for both tasks. FFNSL learned significantly more interpretable hypotheses than the baseline approaches in both tasks. In terms of learning time, FFNSL learned a hypothesis faster than the baselines trained with more examples in the Crop Yield Prediction task, and was slower than the baselines in the Indoor Scene Classification task. Figure 27b clearly shows the computational benefit of setting ILP example weight penalties based on neural network confidence scores, as a hypothesis was learned significantly faster than when constant penalties were used. The near constant learning times at 80–100% shifts for FFNSL SASR with neural network-based penalties, and 30–100% shifts with constant penalties, was due to the 10 min timeout imposed on each FastLAS learning task.

## 8.2 FFNSL framework evaluation

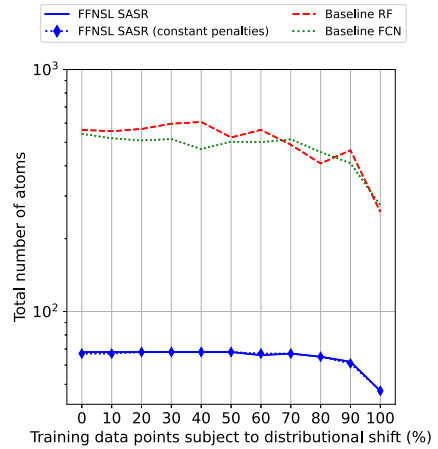
Figure 28 presents the accuracy of the entire FFNSL framework when evaluated over test data also subject to the same percentage of distributional shift as used during learning.

In the Crop Yield Prediction task (Fig. 28a), FFNSL EDL-GEN outperformed all other methods, and FFNSL Softmax outperformed all other methods trained with the same amount of data. The next best approach, the random forest, required significantly more data to match the performance of FFNSL Softmax. The superior performance of FFNSL EDL-GEN compared to FFNSL Softmax was due to the EDL-GEN neural network predicting more accurately for images subject to distributional shift (see Fig. 23a). In the Indoor Scene Classification task, FFNSL performed similarly to the best baseline approach, and all approaches degraded gracefully as the percentage of data points subject to distributional shift increased.

To conclude, this evaluation of FFNSL to real-world datasets shows that the framework can support a wide range of neural modules, and the D2K component is flexible enough to support the interface between different neural and symbolic modules. When taking into account the Follow Suit Winner and Sudoku Grid Validity results, we have also shown that FFNSL can learn complex, first-order symbolic knowledge, using essential aspects of

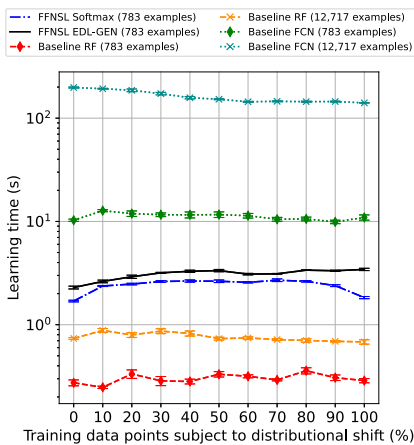


(a) Crop Yield Prediction

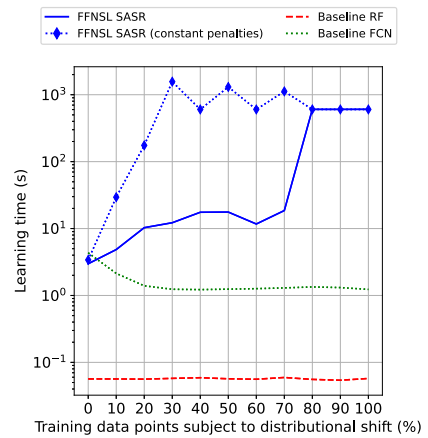


(b) Indoor Scene Classification

Fig. 26 Interpretability of the learned hypotheses



(a) Crop Yield Prediction



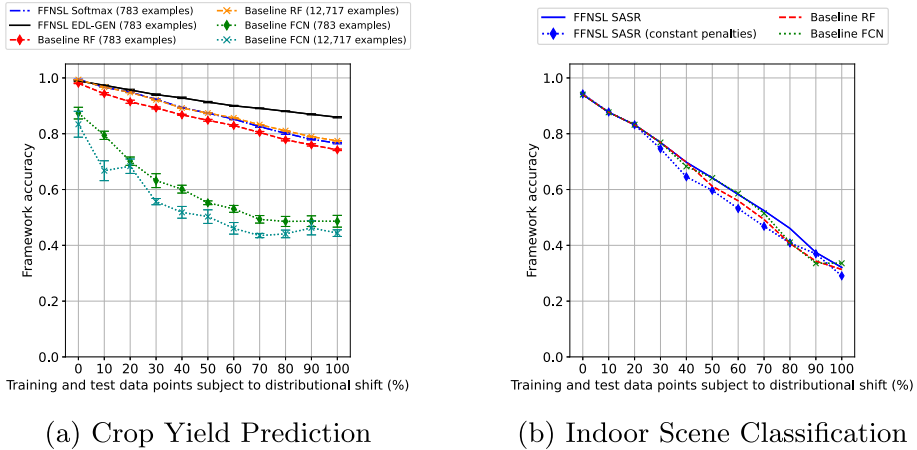
(b) Indoor Scene Classification

Fig. 27 Hypothesis learning time

common-sense learning and reasoning such as negation as failure and predicate invention. In the next section, we discuss related work before concluding the paper.

### 9 Related work

Our proposed FFNSL approach is a specific form of a neural-symbolic learning and reasoning system that, differently from other neural-symbolic methods, uses pre-trained neural networks and logic-based machine learning systems to learn interpretable, logic-based knowledge from unstructured data that can be used to solve a given task. Most of the recently proposed neural-



**Fig. 28** Accuracy of the FFNSL framework when training and test data were subject to distributional shift

symbolic approaches focus on ways in which a given fixed knowledge can be used to improve the training of a neural network (Serafini & d’Avila Garcez, 2016; Donadello et al., 2017; Riegel et al., 2020; Manhaeve et al., 2018; Yang et al., 2020; Tsamoura et al., 2021). These approaches leverage the notions of Real Logic (Serafini & d’Avila Garcez, 2016; Donadello et al., 2017) or t-norm functions (Flaminio & Marchioni, 2006) to enable the injection of logical reasoning in data-driven relational machine learning. This is the case, for instance, of the Logic Tensor Network approaches proposed in Serafini and d’Avila Garcez (2016); Donadello et al. (2017). Our FFNSL approach also uses a similar notion of t-norms but not to embed logic into the differentiable setting, rather to “combine” neural network predictive approximations with logic-based learning optimisation so enabling the composition of these two different machine learning paradigms.

Neural-symbolic approaches that preserve the composition of neural and symbolic inference include DeepProbLog (Manhaeve et al., 2018), NeurASP (Yang et al., 2020) and NeuroLog (Tsamoura et al., 2021). They compose deep learning architectures with symbolic reasoning in order to use existing background knowledge, expressed as logic programs, to train deep learning models. DeepProbLog (Manhaeve et al., 2018) uses ProbLog (De Raedt et al., 2007) to interpret network outputs as probabilistic atoms, and symbolic knowledge compiled into an arithmetic circuit, to train the network. NeurASP extends ASP with neural predicates, expressed as choice rules, to symbolically capture possible network outputs. The probability of each model of the ASP program is computed based on the network predictions, which is in turn used to optimise a semantic loss function for training the network (Xu et al., 2018). NeuroLog also trains the neural network using a semantic loss function, although uses abduction to prune the space of possible pseudo-label revisions for the latent concepts, instead of considering all possibilities as in NeurASP. Although compositional in their architectural solution, and novel in their end-to-end approach for differentiable training of the neural networks, these methods require the logic-based knowledge to be manually engineered. Our FFNSL approach, on the other hand, enables the learning of logic-based knowledge from unstructured data exploiting pre-trained neural models. The semantics of the underlying logic-based learning algorithm in FFNSL is the Answer Set semantics, as it is the case for the symbolic component of the NeurASP system, but with the advantage in



FFNSL that knowledge expressed in ASP programs is learned instead of being fully encoded as input.

Contrary to the end-to-end feature of DeepProbLog and NeurASP neural-symbolic systems, our FFNSL adopts a pipeline approach. It is therefore somewhat related to the Concept Bottleneck Model architecture proposed in Koh et al. (2020), which advocates the idea of training first a model to predict “primary” concepts and then using these concepts to train a downstream model for predicting the labels. These models are however differentiable and even though they can be trained in an end-to-end fashion to improve the overall accuracy (Koh et al., 2020), the trained downstream model is not interpretable. Their interpretability is limited to extracting correlations between the primary concepts and the final label. In our FFNSL approach, the use of LAS logic-based machine learning systems allow the learning of knowledge that is fully interpretable and that is more robust to distributional shifts and noise in the data. In fact, the CNN-LSTM and FCN baselines used in our tasks could be considered as independent concept bottleneck models, and FFNSL outperformed both of these models in our evaluation.

The compositional aspect of our framework could, in principle, make it amenable to instantiations where the symbolic component is a probabilistic rule learning system. Different probabilistic rule learning and statistical relational learning systems have been proposed, such as ProbFOIL (De Raedt et al., 2015), SLIPCOVER (Bellodi & Riguzzi, 2013), Markov Logic Networks (Richardson & Domingos, 2006) and Credal-FOIL (Tuckey et al., 2020). They adopt a probabilistic notion of uncertainty which is different from the notion of WCDPI example weight penalties used in our FFNSL approach. Such systems would, however, make FFNSL not applicable to tasks where non-observational predicate learning with negation as failure is required, like our Follow Suit Winner task, and limit its scalability. This is because it still remains to be shown whether current probabilistic rule learning systems are scalable to a large number of probabilistic facts and large hypothesis search spaces.

Related approaches that support the learning of interpretable knowledge from (unstructured) data in a neural-symbolic manner include  $\delta$ ILP (Evans & Grefenstette, 2018), and NeuralLP (Yang et al., 2017). They make use of rule templates and differentiable reasoning to approximate the inference process and learn instances of the rule templates that cover given labelled examples or to answer given queries. Such approaches, preserve the symbolic, logic-based representation of the knowledge, but replace the logic-based inference process with a purely differentiable one. Our FFNSL approach uses instead a pure symbolic inference process to learn interpretable knowledge, leveraging on state-of-the-art logic-based machine learning systems such as ILASP and FastLAS. The composition of these systems with differentiable feature extraction from unstructured data enables FFNSL to learn knowledge that is more expressive than the definite clausal form supported by  $\delta$ ILP and NeuralLP, broadening the applicability of FFNSL to real-world problems where non-monotonicity and preference learning are required. Results in Law et al. (2018) have already demonstrated that, in the case of structured data, the ILASP system used by our FFNSL framework outperforms  $\delta$ ILP when learning interpretable knowledge from noisy examples.

Neural-symbolic systems such as Neural-Theorem Prover (Rocktäschel & Riedel, 2017) and its extensions, adopt instead a counterpart approach whereby knowledge is expressed as dense vector embedding representations that are learned in a differentiable manner by using a symbolically inspired backward chaining algorithm and (soft) unification. In these systems, the knowledge is represented in a high-dimensional differentiable space and the inference is symbolically inspired. More recently, a fully differentiable rule induction approach based on Logical Neural Networks has been proposed (Sen et al., 2021) that uses differentiable operators from fuzzy and real logic to learn rules from structured data within a very controlled

search space expressed using templates. Although some of these systems have recently shown to be somewhat scalable over large knowledge bases (Minervini et al., 2020, ?), they are all limited in the expressivity of the knowledge that they can learn and they are not guaranteed to learn (mathematically provable) optimal solutions. These are two main properties that our FFNSL framework instead benefits from, making our approach particularly suited for safe and trusted AI applications where data are unstructured, complex, and interpretable knowledge is required to solve complex tasks.

Recent approaches train a neural network to extract primary concepts from raw data, whilst learning interpretable symbolic knowledge in an end-to-end fashion (Dai et al., 2019; Dai & Muggleton, 2021). These methods don't require labels for the primary concepts, and train a neural network from scratch whilst simultaneously learning knowledge. The Abductive Learning framework (ABL) (Dai et al., 2019) learns ground operation facts that complete a symbolic knowledge base, to map neural network outputs to downstream labels. This knowledge is then used to abduce revised pseudo-labels to improve the training of the neural network. Crucially, Dai et al. (2019) cannot perform program induction, and assumes monotonicity of the background knowledge, as ground operation facts are abduced and accumulated during an iterative sampling process over the training data. In contrast, our approach learns first-order rule-based programs, which contain universally quantified variables, and are therefore applicable to a range of input sizes greater than the sizes used for training. We can also handle non-monotonicity, thus enabling the learning of more complex knowledge. The *MetaAbd* approach (Dai & Muggleton, 2021) extends (Dai et al., 2019) to perform rule induction using the Metagol symbolic learner (Muggleton et al., 2015). The key drawback of *MetaAbd* is that Metagol can only learn symbolic knowledge expressed as definite logic programs without function symbols, which can compute only polynomial functions (Dantsin et al., 2001). *MetaAbd* cannot learn more expressive knowledge involving defaults, exceptions, constraints and choice, which are essential aspects of common-sense learning and reasoning. In FFNSL, we learn first-order complex knowledge expressed in the language of ASP, which is more general than symbolic learning of definite clauses (Law, 2018; Law et al., 2020, 2018), and can solve computationally harder problems (Karp, 1972). Also, due to the high level of difficulty of such an end-to-end neuro-symbolic task, *MetaAbd* has only been applied to very simple classification problems. Our architecture is motivated by a completely different requirement, that of using already trained and therefore possibly much more complex neural components for extracting features from challenging raw data.

## 10 Conclusion

This paper introduces a neural-symbolic learning framework, *FFNSL*, that learns interpretable knowledge from unstructured data that is robust to distributional shifts. Three main instantiations of this framework have been presented, which use the ILASP and FastLAS logic-based machine learning systems, according to the type of symbolic learning task required. In each instantiation, pre-trained neural networks have been used for extracting symbolic features from the unstructured data. The novel component of FFNSL is the D2K generator, which generates symbolic features, weighted by neural network confidence scores, that together with a label, form the input to the logic-based machine learning system which then learns interpretable knowledge needed to solve the given downstream task.

Our evaluation on four neural-symbolic classification tasks, Follow Suit Winner, Sudoku Grid Validity, Crop Yield Prediction and Indoor Scene Classification, demonstrates that

FFNSL is robust to distributional shifts in the input data, outperforming random forest and deep neural network baselines. FFNSL learns more accurate and interpretable knowledge than the baselines even when the latter are trained with significantly more data. The application of FFNSL learned knowledge to unseen data also subject to similar proportions of distributional shifts shows that FFNSL is again capable of outperforming the baseline approaches trained with the same amount of data up to ~80% of data subject to distributional shifts. A detailed analysis of the performance in accuracy of our FFNSL framework shows that using an uncertainty-aware neural network provides an improved bias to the logic-based machine learning system compared to Softmax neural networks, with a greater proportion of the total weight penalty allocated to WCDPI examples containing correct contextual information extracted from the unstructured data.

**Acknowledgements** This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

**Author Contributions** DC defined the FFNSL method, performed the experimental evaluation and wrote the initial version of the paper. ML provided support with running the ILASP and FastLAS systems and helped define the correct encoding for each task. ML also suggested the Follow Suit Winner task and provided feedback on the final paper. AR and JL both equally contributed to the papers positioning, the generalised FFNSL method and gave suggestions for the experimental approach. AR and JL also contributed to the writing of the paper.

**Funding** This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001.

**Availability of data and material** The Sudoku and Follow Suit Winner datasets introduced in this paper are available at the following GitHub repository: <https://github.com/DanCunnington/FFNSL>.

**Code availability** All the experimental code is also available at the GitHub repository.

## Declarations

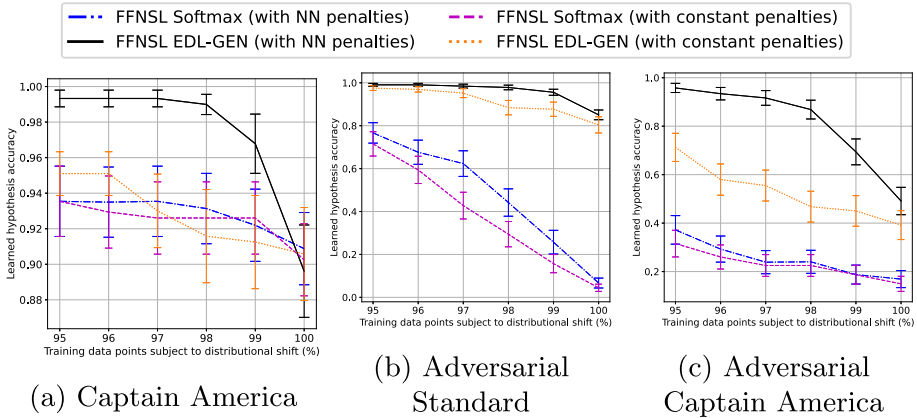
**Conflict of interest** Daniel Cunnington, Jorge Lobo and Alessandra Russo have no relevant financial or non-financial interests to disclose. Mark Law is the director of ILASP Limited, which owns the intellectual property of the ILASP system used in this paper.

**Ethics approval** Not applicable to this paper.

**Consent to participate** Not applicable to this paper as no humans were used to conduct the experimental evaluations.

**Consent for publication** Not applicable, all data, figures and tables are original and are generated synthetically, with the exception of the MNIST dataset LeCun et al. (1998).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.



**Fig. 29** FFNSL Softmax vs. FFNSL EDL-GEN. Accuracy of learned hypotheses, 95–100% distributional shifts, 50 repeats

### A Additional follow suit winner results

In this section we present additional Follow Suit Winner results and analysis when the Captain America, Adversarial Standard and Adversarial Captain America decks were used to apply distributional shifts to the input data points. These results supplement the results and analysis presented in Sect. 6 for the Batman Joker and Adversarial Batman Joker decks.

#### A.1 Learned hypothesis evaluation

Firstly, let us present the comparison of FFNSL EDL-GEN versus FFNSL Softmax, for 95–100% distributional shifts and 50 experimental repeats. The results are presented in Fig. 29, which extend the results presented in Fig. 6.

The results for the Captain America deck in Fig. 29a are very similar to the Batman Joker deck presented in Fig. 6a, with the exception of FFNSL EDL-GEN with constant penalties that performed similarly to the FFNSL Softmax approaches. The Adversarial Standard deck results in Fig. 29b are very similar to the Adversarial Batman Joker results in Fig. 6b, however, the Adversarial Captain America results in Fig. 29c are different to the other decks. The two FFNSL Softmax approaches had much lower accuracy at 95% shifts compared to the other decks, and there was a significant gap between FFNSL EDL-GEN with neural network penalties and FFNSL EDL-GEN with constant penalties. Let us now investigate each of these decks w.r.t. the percentage of incorrect ILP examples and the ILP example weight penalty ratios calculated from neural network confidence scores. The incorrect ILP example analysis is presented in Figure 30.

Firstly, with the Captain America deck in Fig. 30a, both Softmax and EDL-GEN had a very similar percentage of incorrect ILP examples, more similar than the other decks with the exception of Adversarial Captain America. This explains why FFNSL EDL-GEN with constant penalties performs similarly to FFNSL Softmax approaches in Fig. 29a. For the Adversarial Standard deck in Fig. 30b, the Softmax neural network resulted in a significantly higher percentage of incorrect ILP examples compared to EDL-GEN, which explains

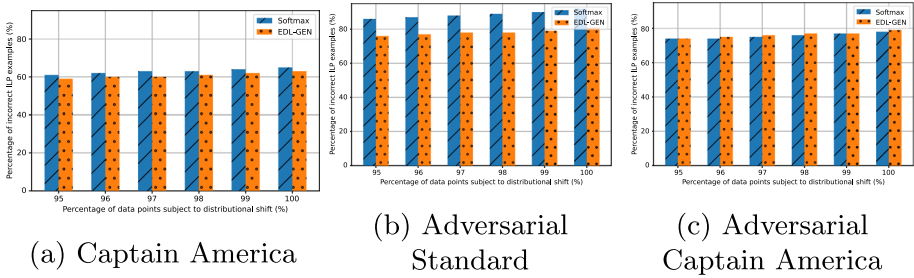


Fig. 30 The effect of applying distributional shifts on the percentage of incorrect ILP examples, Follow Suit Winner task

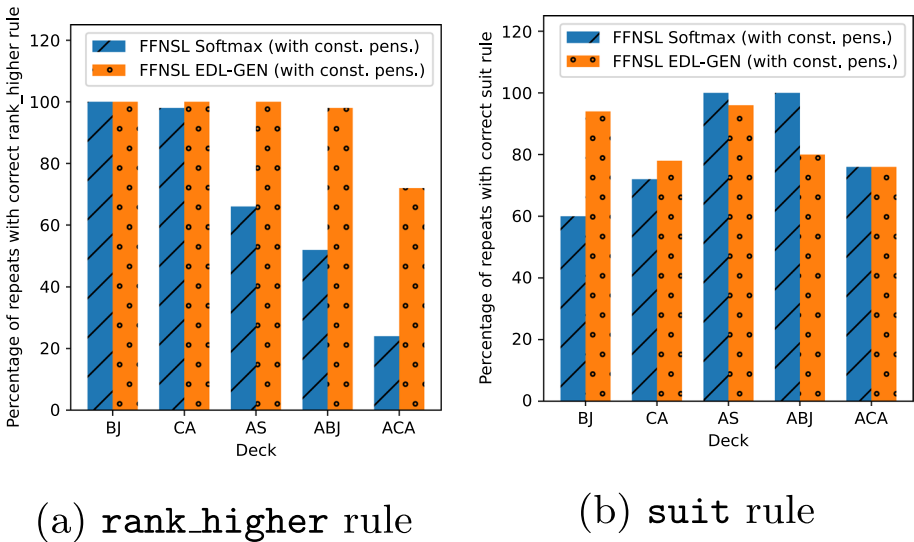
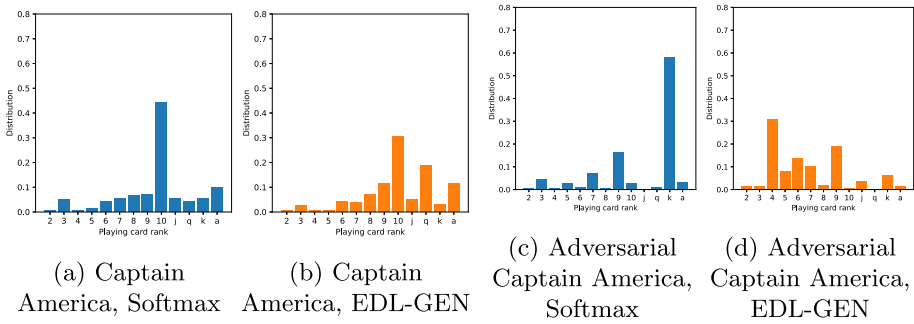


Fig. 31 The percentage of experimental repeats that learned the correct follow suit rules, when 95% of input data points were subject to distributional shifts

the large performance gap between FFNSL EDL-GEN and FFNSL Softmax approaches in Fig. 29b. For the Adversarial Captain America deck, using the EDL-GEN neural network resulted in a higher percentage of incorrect ILP examples in Fig. 30c, yet FFNSL EDL-GEN clearly outperformed FFNSL Softmax in Fig. 29c. To investigate this further, we look at the percentage of experimental repeats for both FFNSL approaches with constant penalties that learned the correct rank\_higher and suit rules (respectively; the winning player had a higher ranked card than other players, and the winning player also had the same suit as player 1), both of which are key to solving the task successfully. The analysis is presented in Fig. 31 when 95% of the input data points were subject to distributional shifts. The x-axis labels are the abbreviated names for each deck.

With the Adversarial Captain America deck (ACA in Fig. 31), only 24% of experimental repeats with FFNSL Softmax learned the correct rank\_higher rule, compared to 72% with FFNSL EDL-GEN (both with constant penalties). There isn't much difference between FFNSL Softmax and FFNSL EDL-GEN in learning the correct suit rule



**Fig. 32** Distribution of playing card rank predictions for Softmax and EDL-GEN neural networks when the Captain America and Adversarial Captain America decks were used to apply distributional shift for 95% of input data points

(Fig. 31b). So, as FFNSL Softmax fails to learn the correct rank\_higher rule, this explains why in Fig. 29c, the FFNSL Softmax approaches performed worse than FFNSL EDL-GEN approaches. The question now becomes, why does ILASP fail to learn the correct rank\_higher rule for 76% of the experimental repeats with FFNSL Softmax, despite FFNSL Softmax having a similar or lower percentage of incorrect ILP examples compared to FFNSL EDL-GEN? To answer this question, Fig. 32 shows the distribution of playing card rank predictions for both Softmax and EDL-GEN neural networks when the Captain America and Adversarial Captain America decks were used to apply distributional shift for 95% of input data points, as these two decks have a similar percentage of incorrect ILP examples between Softmax and EDL-GEN (see Fig. 30a and c).

The Softmax neural network predicted the same playing card rank more often than the EDL-GEN neural network. With the Captain America deck, 45% of playing cards were predicted with rank 10 (Fig. 32a), and with the Adversarial Captain America deck, nearly 60% of playing cards were predicted with rank King (Fig. 32c). The EDL-GEN neural network predicted with a more even distribution. Now, with the Softmax neural network and the Adversarial Captain America deck, ILASP didn't learn the correct rank\_higher rule very often. Investigating the neural network card predictions within the generated ILP examples when 95% of distributional shifts were applied, we calculate the percentage of examples where the ground-truth winner has a predicted card with a higher rank than the other players. For FFNSL Softmax, only 9% of the examples contained a higher ranked card for the ground-truth winning player, compared to 19% with FFNSL EDL-GEN. Looking at the Captain America deck, 35% of the examples for FFNSL Softmax contained a higher ranked card for the ground-truth winning player, compared to 37% for FFNSL EDL-GEN.

This explains why, in the Adversarial Captain America deck FFNSL Softmax struggled to learn the rank\_higher rule and therefore, why there was a drop in performance in Fig. 29c for FFNSL Softmax. As the Softmax neural network failed to predict playing card ranks correctly, the ILP examples didn't contain a higher ranked card for the ground-truth winner, and therefore 76% of the experimental repeats failed to learn the correct rank\_higher rule. Comparing with the Captain America deck in Fig. 29a, the FFNSL Softmax approaches performed much better, because there was a higher number of generated ILP examples that contained higher ranked card predictions for the ground-truth winning player.

Finally, in Fig. 33c we now investigate the ILP example weight penalty ratio to explain why there was a significant gap between the two FFNSL EDL-GEN approaches in Fig. 29c for the Adversarial Captain America deck.

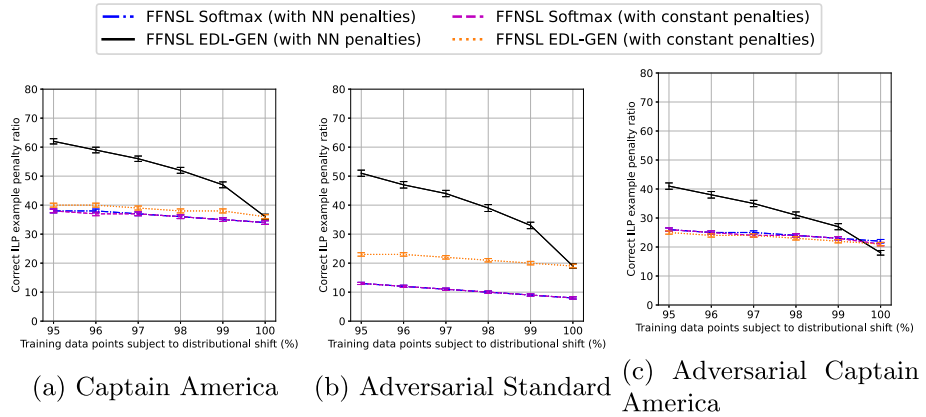


Fig. 33 ILP example weight penalty ratio, 95–100% shifts

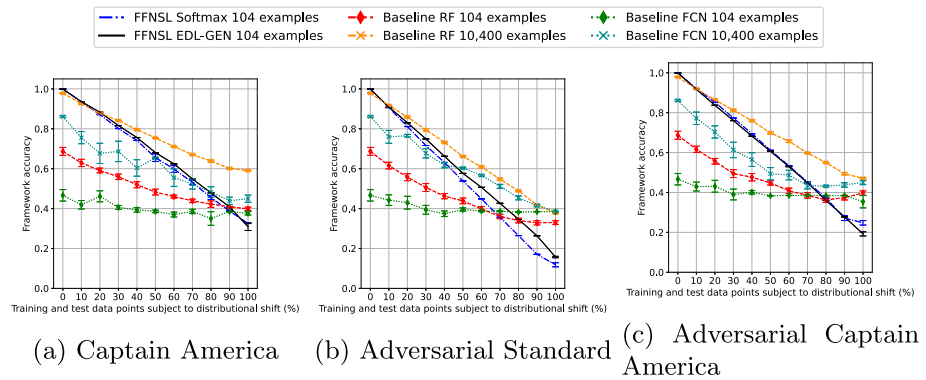


Fig. 34 Accuracy of the FFNSL framework when training and test data points were subject to distributional shifts

The weight penalty ratio of FFNSL EDL-GEN with penalties calculated from neural network confidence scores outperformed FFNSL EDL-GEN with constant penalties (Fig. 33c). As the Adversarial Captain America deck was more challenging for ILASP in terms of the predictions from the neural networks, the ILP example weight penalties had more impact on the accuracy of the learned hypotheses, as ILASP was able to focus on covering the ILP examples that contained the correct neural network predictions.

### A.2 FFNSL framework evaluation

Figure 34 presents the accuracy of the entire FFNSL framework when both training and test data points were subject to distributional shifts. The results for the Captain America deck in Fig. 34a are very similar to the results for the Batman Joker deck presented in Fig. 11a, and the results for the Adversarial Standard and Adversarial Captain America decks in Fig. 34b and c are very similar to the results presented in Fig. 11b for the Adversarial Batman Joker deck.



## B Learned hypotheses

In this section we present a sample of the hypotheses learned by FFNSL when distributional shifts were applied to input data points.

### B.1 Follow suit winner

When no distributional shifts were applied, i.e., at points 0% on the x-axes in Fig. 5, the following hypothesis was learned by FFNSL:

```
winner(X) :- not p1(X), player(X).
p1(V1) :- V2 != V3; suit(1,V2); suit(V1,V3); player(V1); suit(V2); suit(V3).
p1(V1) :- rank_higher(V2,V1); suit(1,V3); suit(V2,V3); player(V1); player(V2);
           suit(V3).
```

The first rule states that player  $X$  is a winner if *neither* of the bottom two rules hold. The second rule holds if the suit of player  $X$  is different to the suit of player 1, and the final rule holds if there is another player with a higher ranked card with the same suit as player 1. The `player`, `suit` and `rank_higher` predicates were defined in the background knowledge (for details, see Appendix F). As an example, in the case of 100% distributional shifts using the Batman Joker deck in Fig. 5a, the following rules were learned on one experimental repeat:

```
winner(X) :- not p1(X), player(X).
p1(V1) :- rank_higher(V2,V1); player(V1); player(V2).
```

In this hypothesis, a player is the winner if they have played the highest ranked card. In this case, the rule denoting the suit having to match the suit of player 1 was missed.

### B.2 Sudoku grid validity

For the  $4 \times 4$  and  $9 \times 9$  grid Sudoku tasks, when no distributional shifts were applied (i.e., at points 0% on the x-axes in Fig. 14), the following hypothesis was learned by FFNSL, which states that a Sudoku grid is invalid if there are two of the same digits in a block, column or row:

```
invalid :- neq(V2,V1), digit(V1,V3), block(V2,V0), block(V1,V0), digit(V2,V3).
invalid :- neq(V1,V0), digit(V0,V2), digit(V1,V2), row(V0,V3), row(V1,V3).
invalid :- neq(V1,V0), digit(V0,V3), digit(V1,V3), col(V0,V2), col(V1,V2).
```

The `neq`, `digit`, `block`, `row` and `col` predicates were defined in the background knowledge (for details, see Appendix F). For the  $4 \times 4$  grid task, when 100% of the training data points were subject to distributional shifts in Fig. 14a, the following hypothesis was learned:

```
invalid :- neq(V2,V1), neq(V3,V1), neq(V3,V2), block(V2,V0), block(V3,V0),
           block(V1,V0).
invalid :- not block(V1,V0), block(V2,V0), col(V2,V3), col(V1,V3).
invalid :- not block(V2,V0), block(V1,V0), row(V1,V3), row(V2,V3).
```

The first argument in the `block`, `row` and `col` predicates is a string representing cell coordinates and the second argument is an identifier (e.g., block 1, block 2, etc...). The first

rule states that a grid is invalid if there are three cells in the same block that have different coordinates. The second rule states that a grid is invalid if there are two cells within the same column that are in different blocks and the third rule states that a grid is invalid if there are two cells within the same row that are in different blocks. Therefore, this hypothesis always returned invalid at test-time.

### B.3 Crop yield prediction

When no distributional shift was applied, the following hypothesis was learned by FFNSL. Note that we trim the number of rules for compactness and the full listing is available in the experiment code:<sup>12</sup>

```
yield(0) :- disease(bacterial_spot), location(18).
yield(0) :- disease(black_rot), location(18).
yield(2) :- disease(late_blight), location(6).
yield(2) :- disease(leaf_scorch).
yield(2) :- location(16).
yield(2) :- disease(healthy), location(7).
yield(2) :- disease(healthy), location(17).
yield(2) :- location(11).
...
```

At 100% shifts, the following hypothesis was learned:

```
yield(0) :- location(19), disease(early_blight).
yield(1) :- location(7), disease(bacterial_spot).
yield(2) :- location(16).
yield(2) :- location(11).
yield(2) :- species(corn), disease(healthy).
yield(2) :- disease(powdery_mildew), location(7).
yield(0) :- location(19), species(potato).
```

Here you can see the yield quality has changed for the bacterial spot disease, and this also depends on a different location. This is due to incorrect neural network predictions for the bacterial spot disease.

### B.4 Indoor scene classification

When no distributional shift was applied, the following hypothesis was learned by FFNSL. Note that we also trim the number of rules for compactness and the full listing is available in the experiment code:

```
label(4) :- image(meeting_room).
label(2) :- image(inside_subway).
label(2) :- image(elevator).
label(3) :- image(bowling).
label(0) :- image(shoeshop).
label(4) :- image(classroom).
...
```

<sup>12</sup> <https://github.com/DanCunnington/FFNSL>.

At 100% shifts, the following hypothesis was learned:

```
label(4) :- image(meeting_room).
label(2) :- image(inside_subway).
label(2) :- image(elevator).
label(3) :- image(bowling).
label(0) :- image(shoeshop).
label(2) :- image(airport_inside).
...
```

Analysing these rules further, it appears that at 100% shifts, whilst some super-class rules are correct, others were not learned at all. For example, the `classroom` rule is missing. In total, there were 46 rules learned at 100% shifts, compared to the full set of 67 rules at 0% shifts. In Figure 26b you can see the number of rules in the learned hypothesis decreases at 100% shifts.

## C Dataset details

### *Follow suit winner*

The Follow Suit Winner dataset was generated by simulating multiple games, where each game began with a randomly shuffled deck of playing cards split between the four players. Each game consisted of 13 tricks and the card played by each player along with the winner of each trick was stored. The *small* training datasets contained 104 example tricks from 8 games and the *large* training datasets contained 10,400 example tricks from 800 games. A test set was created containing 1001 example tricks from 77 games. For the neural network, an image was taken of every playing card in a standard deck. The *ImageDataGenerator* class from the Keras image pre-processing library<sup>13</sup> was used to apply transformations to each playing card image, generating 750 variations of each image. We set the *rotation range* to 55, *brightness range* to 0.5-1.5, *shear range* to 15, *channel shift range* to 2.5, *zoom range* to 0.1 and enable *horizontal flip*. From a total of 39,000 images, we created a training set of 27,300 images and a test set of 11,700 images (70%/30% split), maintaining an equal representation of each playing card. Similarly to the Sudoku Grid Validity task, the test set was further split into two datasets (~70%/30%), maintaining an equal representation of each playing card, as follows. The first, denoted `CARDS_TEST_A` contains 8164 images and was used to create FFNSL training sets for learning a hypothesis. Playing cards in the Follow Suit Winner training sets were replaced with a random image of the corresponding playing card from `CARDS_TEST_A`. The second split, denoted `CARDS_TEST_B` contained 3536 images and was used to create a hold out test set such that FFNSL can be evaluated on unseen data once a hypothesis has been learned.

Distributional shifts were applied by replacing playing card images from the standard deck with playing card images from alternative decks in an increasing percentage of data points in the Follow Suit Winner training sets. We used playing card images from *Batman Joker* and *Captain America* decks and also created *adversarial* data points from each deck, placing the candidate playing card image on a background containing playing card images from the standard deck. We applied the same image transformations to the alternative decks such that standard playing card images can be directly swapped with a corresponding card image from an alternative deck. Figure 3 shows an example queen of hearts playing card

<sup>13</sup> <https://keras.io/api/preprocessing/image/>.

image from each deck: *Standard* (3a), *Batman Joker* (3b), *Captain America* (3c), *Adversarial Standard* (3d), *Adversarial Batman Joker* (3e) and *Adversarial Captain America* (3f).

### **Sudoku grid validity**

The Sudoku Grid Validity datasets were generated using valid  $4 \times 4$  and  $9 \times 9$  Sudoku starting configurations obtained from Hanssen's Sudoku puzzle generator.<sup>14</sup> Invalid starting configurations were obtained by taking a valid example (that didn't exist in the set of valid data points) and changing one digit at random in a row, column or block to match another digit in the same row, column or block. All sets of invalid data points contained an equal distribution of data points containing two of the same digit in a row, column or block. The *small* training datasets contained 320 data points, each consisting of 160 *valid* starting configurations and 160 *invalid* starting configurations. The *large* training datasets contained 32,000 data points, with 16,000 valid and 16,000 invalid data points. Finally, separate test sets were created for  $4 \times 4$  and  $9 \times 9$  grids, which contained 1000 data points: 500 valid and 500 invalid.

For the neural network used in the  $4 \times 4$  grids, we used digit classes 1-4 from the standard MNIST dataset (LeCun et al., 1998) and created a training set of 24,674 data points and a test set of 4,160 data points. The MNIST test set was further split ( $\sim 70\%/30\%$ ), maintaining an equal representation of digits, into two datasets as follows. The first, denoted *MNIST\_TEST\_A* contained 2910 images and was used to create FFNSL training sets for learning a hypothesis. Digits in the Sudoku training sets were replaced with a random image of the corresponding digit from *MNIST\_TEST\_A*. The second split, denoted *MNIST\_TEST\_B* contained 1249 images and was used to create a hold out test set such that FFNSL could be evaluated on unseen data once a hypothesis was learned. Digits in the Sudoku test set were replaced with a random image of the corresponding image from *MNIST\_TEST\_B*.

For the neural network used in the  $9 \times 9$  grids, we used digit classes 1-9 from the standard MNIST dataset (LeCun et al., 1998) and created a training set of 54,078 data points and a test set of 9,021 data points. The MNIST test set was further split ( $\sim 70\%/30\%$ ), maintaining an equal representation of digits, into two datasets as follows. The first, denoted *MNIST\_TEST\_C* contained 6310 images and was used to create FFNSL training sets for learning a hypothesis. Digits in the Sudoku training sets were replaced with a random image of the corresponding digit from *MNIST\_TEST\_C*. The second split, denoted *MNIST\_TEST\_D* contained 2710 images and was used to create a hold out test set such that FFNSL could be evaluated on unseen data once a hypothesis was learned. Digits in the Sudoku test set were replaced with a random image of the corresponding image from *MNIST\_TEST\_D*. Note that data observed by FFNSL at learning time was completely unseen by the neural network and was therefore vulnerable to distributional shifts. Also, data observed by FFNSL at evaluation time was completely unseen by the neural network and also FFNSL itself during learning.

Distributional shifts were applied by rotating MNIST digit images  $90^\circ$  clockwise in an increasing percentage of data points in the Sudoku training sets. When we evaluated with unstructured test data, the same procedure applied to the Sudoku test set, i.e., when we evaluated a hypothesis learned from a training set with 20% of the data points containing rotated images, 20% of the test set data points also contained rotated images.

<sup>14</sup> <https://www.menneske.no/sudoku/2>.

## D Neural network and baseline details

### *Follow suit winner*

Firstly, for FFNSL Softmax, we trained a Softmax-based CNN with 4 2D convolutional layers and 2 fully connected layers for 20 epochs in PyTorch. The network accepts 3-channel RGB input with images of size 274x174 pixels and outputs a 52 dimensional Softmax vector to predict each playing card. Secondly, for FFNSL EDL-GEN, we trained an uncertainty-aware neural network based on evidential deep learning (Sensoy et al., 2020). We used the available architecture and implementation in TensorFlow,<sup>15</sup> and modified  $k$ , the number of outputs to 52 and the layer dimensions to accept 274x174 RGB card images. We also trained this neural network for 20 epochs.

The baseline random forest model was implemented with scikit-learn 0.23.2 and tuned on the first *small* dataset with 0 data points subject to distributional shift. The number of estimators was tuned across: {10, 20, 50, 100, 200}. The best performing parameter value of 100 estimators was chosen and used for all Follow Suit Winner experiments. The random seed was set to 0 to enable reproducibility.

The baseline FCN consists of 3 fully connected layers with the ReLU activation function applied to each layer. Dropout was also applied after the first and second layers. Finally, a Softmax layer squashed the final logits into 4 classes, representing each possible winner. The input consisted of one-hot encoded suit values and the rank value of the playing card for each player. Therefore, the input size to the first fully connected layer was 20. We implemented the architecture in PyTorch v1.7.0.

To tune the FCN, we sampled the number of output units in the first and second layers, i.e.,  $l_1 \in \{20, 32, 46, 52\}$  and  $l_2 \in \{52, 64, 74, 80\}$  respectively, along with the dropout probability in both dropout layers  $dr \in \{0.1, 0.2, 0.5\}$ . We sampled all possible parameter combinations and tuned on the first *small* dataset, with no data points subject to distributional shift, trained for 50 epochs. The best performing parameter values of  $l_1 = 20$ ,  $l_2 = 74$  and  $dr = 0.1$  were chosen. These parameters were then fixed for all models trained and following tuning, each model was trained for 50 epochs. Finally, the random seed was set to 0 to enable reproducibility.

### *Sudoku grid validity*

Within FFNSL, we trained two types of neural networks. Firstly, for FFNSL Softmax, we adopted the CNN architecture available in the MNIST PyTorch tutorial<sup>16</sup> and replaced the *LogSoftmax* layer with a *Softmax* layer and the *Negative Log Likelihood* loss function with *Cross-Entropy Loss*. This is to satisfy the neural network definition in Sect. 3 such that a confidence score  $c \in [0, 1]^k$  is returned for  $k$  possible feature values. For the two grid sizes,  $4 \times 4$  and  $9 \times 9$ , we train two separate networks. For  $4 \times 4$  grids, we set  $k = 4$  and train on digits 1-4 inclusive, whilst for  $9 \times 9$  grids we set  $k = 9$  and train on digits 1-9 inclusive. We adopted all existing hyper-parameter values and trained for 20 epochs.

Secondly, for FFNSL EDL-GEN, we trained two uncertainty-aware neural networks (Sensoy et al., 2020) using the available architecture and implementation in TensorFlow<sup>15</sup>, and set  $k$ , the number of outputs, to 4 and 9, for  $4 \times 4$  and  $9 \times 9$  grids respectively. We used existing hyper-parameter values and trained for 20 epochs.

The baseline random forest model was implemented with scikit-learn 0.23.2 and tuned on the first *small* dataset with no data points subject to distributional shift. The number of estimators was tuned across: {10, 20, 50, 100, 200}. The best performing parameter value of

<sup>15</sup> <https://murasensoy.github.io/gen.html>.

<sup>16</sup> <https://github.com/pytorch/examples/tree/master/mnist>.

100 estimators was chosen and used for all Sudoku Grid Validity experiments. The random seed was set to 0 to enable reproducibility.

The baseline CNN-LSTM consisted of an embedding layer, followed by a 1D convolutional layer with a kernel size of 3 and the ReLU activation function. Then, a 1D max pooling layer with pool size 2 was used, followed by a dropout layer, an LSTM layer and a second dropout layer. Finally, a dense fully connected layer with the sigmoid activation function was used to produce a binary classification of the input digit sequence. The input sequence length to the embedding layer was 16 for  $4 \times 4$  grids and 81 for  $9 \times 9$  grids, representing each cell on the Sudoku grid. We implemented the architecture in PyTorch v1.7.0.

To tune the CNN-LSTM, we sampled the learning rate  $lr \in \{0.1, 0.001, 0.0001\}$ , the embedding dimension of the embedding layer  $ed \in \{32, 96, 256\}$ , the number of output channels of the 1D convolution layer  $oc \in \{64, 96\}$ , the number of hidden features in the LSTM layer  $lh \in \{32, 96, 128\}$  and the dropout probability  $dr \in \{0.01, 0.05, 0.1\}$  in both dropout layers. We performed 10 samples and evaluated the model on the first *large* dataset with 0 data points subject to distributional shift, trained for 2 epochs. The best performing parameter values of  $lr = 0.0001$ ,  $ed = 96$ ,  $oc = 64$ ,  $lh = 96$  and  $dr = 0.01$  were chosen. These parameters were then fixed for all models trained and following tuning, each model was trained for 5 epochs. Finally, the random seed was set to 0 to enable reproducibility.

## E System details

All experiments in this paper (with the exception of the deep neural network baselines) were run on the same machine with the following specifications:

**Hardware:** QEMU KVM virtual machine standard PC (i440FX + PIIX 1996) with 10 nodes of 8-core AMD EPYC Zen 2 CPUs (80 cores total), 16GB RAM.

**Operating System:** Ubuntu 18.04.4 LTS.

**Software:** FastLAS 1.1 (FastLAS 3 for  $4 \times 4$  Sudoku Grid Validity with reduced background knowledge), ILASP 4, Python 3.7.3, PyTorch 1.7.0, TensorFlow 1.14.0, Keras 2.4.0, scikit-learn 0.23.2, numpy 1.19.1, problog 2.1.0.42. The neural network baselines were run on a machine with the following specifications:

**Hardware:** x86 compute node with 24 cores (CPU) and an NVIDIA Tesla K80 GPU, 512GB RAM.

**Operating System:** Red Hat Enterprise Linux 7.6.

**Software:** Same as above.

## F ILP task listings

### F.1 Follow suit winner

For the Follow Suit Winner task, we used the ILASP (Law, 2018) ILP system as ILASP supports *predicate invention* (Stahl, 1993). Predicate invention was required for this task to link the winning player to the suit and rank of other players cards. We encoded as background knowledge possible suit and rank values, the four players, as well as the definition of the `rank_higher` predicate. The set of body mode declarations included a `suit` predicate, which linked a player's card to a suit, alongside the `rank_higher` predicate. The set of head mode declarations included a player variable, specified to support predicate invention. The

hypothesis space for this task contained 96 possible rules (therefore  $2^{96}$  potential hypotheses, computed as the power set).

### ***Background Knowledge***

```
% Suits
suit(h).
suit(s).
suit(d).
suit(c).
```

```
% Ranks
rank(a).
rank(2).
rank(3).
rank(4).
rank(5).
rank(6).
rank(7).
rank(8).
rank(9).
rank(10).
rank(j).
rank(q).
rank(k).
```

```
% Rank Value
rank_value(2, 2).
rank_value(3, 3).
rank_value(4, 4).
rank_value(5, 5).
rank_value(6, 6).
rank_value(7, 7).
rank_value(8, 8).
rank_value(9, 9).
rank_value(10, 10).
rank_value(j, 11).
rank_value(q, 12).
rank_value(k, 13).
rank_value(a, 14).
```

```
% 4 Players
player(1..4).
```

```
% Definition of higher rank
rank_higher(P1, P2) :- card(P1, R1, _), card(P2, R2, _), rank_value(R1, V1),
rank_value(R2, V2), V1 > V2.
```

```
% Link player's card to suit
suit(P1, S) :- card(P1, _, S).
```



**Mode declarations**

```

P(X) :- Q(X), identity(P, Q).
P(X) :- player(X), not Q(X), inverse(P, Q).
#mode(2, inverse(target/1, invented/1)).
#mode(2, identity(target/1, invented/1)).
#predicate(target, winner/1).
#predicate(invented, p1/1).

#constant(player, 1).
#constant(player, 2).
#constant(player, 3).
#constant(player, 4).
#modeh(p1(var(player))).
#modeb(1, var(suit) != var(suit)).
#modeb(1, suit(var(player), var(suit)), (positive)).
#modeb(1, suit(const(player), var(suit)), (positive)).
#modeb(1, rank_higher(var(player), var(player)), (positive)).

```

**F.2 Sudoku grid validity**

There are two variations of ILP tasks presented in this paper, where knowledge of the Sudoku grid was specified, and where grid knowledge was removed and replaced with a division predicate, which enabled FastLAS to learn column, row and block identifiers, based on the cell coordinates given in the example contexts. Both of these variations are presented below, with an example for  $9 \times 9$  grids with the grid knowledge, and  $4 \times 4$  grids without the grid knowledge. For each variation, we present the background knowledge specified and the mode declarations used. The argument in quotes for each column, row and block fact is a unique identifier for each cell. The subset of the hypothesis space computed by FastLAS for both  $4 \times 4$  and  $9 \times 9$  grids contained 2350 possible rules (therefore  $2^{2350}$  potential hypotheses, computed as the power set).

**Encoding the Sudoku grid: background knowledge** For  $9 \times 9$  Sudoku grids:

```

col("1, 1", 1).
col("1, 2", 2).
col("1, 3", 3).
col("1, 4", 4).
col("1, 5", 5).
col("1, 6", 6).
col("1, 7", 7).
col("1, 8", 8).
col("1, 9", 9).
...

row("1, 1", 1).
row("1, 2", 1).
row("1, 3", 1).
row("1, 4", 1).

```

```

row("1, 5", 1).
row("1, 6", 1).
row("1, 7", 1).
row("1, 8", 1).
row("1, 9", 1).
...

block("1, 1", 1).
block("1, 2", 1).
block("1, 3", 1).
block("2, 1", 1).
block("2, 2", 1).
block("2, 3", 1).
block("3, 1", 1).
block("3, 2", 1).
block("3, 3", 1).
...

```

**Encoding the sudoku grid: mode declarations** For  $9 \times 9$  Sudoku grids:

```

#modeh( invalid ).
#modeb( digit( var( cell ), var( num ) ) ).
#modeb( row( var( cell ), var( row ) ) ).
#modeb( col( var( cell ), var( col ) ) ).
#modeb( block( var( cell ), var( block ) ) ).
#modeb( neq( var( cell ), var( cell ) ) ).
#maxv( 4 ).
num( 1..9 ).
row( 1..9 ).
col( 1..9 ).
block( 1..9 ).
cell( C ) :- digit( C, _ ).
neq( X, Y ) :- cell( X ), cell( Y ), X != Y.

```

**Without encoding the Sudoku grid: Background knowledge** For  $4 \times 4$  Sudoku grids:

```

div_same1( X, Y, C ) :- ( X - 1 ) / C = ( Y - 1 ) / C, idx1( X ), idx1( Y ), X < Y, quotient( C ).
div_same2( X, Y, C ) :- ( X - 1 ) / C = ( Y - 1 ) / C, idx2( X ), idx2( Y ), X < Y, quotient( C ).

quotient( 1..3 ).
idx1( 1..4 ).
idx2( 1..4 ).

```

**Without encoding the Sudoku grid: Mode Declarations** For  $4 \times 4$  Sudoku grids:

```

#modeh( invalid ).
#modeb( digit( var( idx1 ), var( idx2 ), var( num ) ) ).
#modeb( div_same1( var( idx1 ), var( idx1 ), const( quotient ) ) ).
#modeb( div_same2( var( idx2 ), var( idx2 ), const( quotient ) ) ).

#maxv( 5 ).
num( 1..4 ).

#bias( "penalty( 1, head )." ).
#bias( "penalty( 1, body( X ) ) :- in_body( X )." ).
#ground_without_replacement.

```

### F.3 Crop yield prediction

#### Background Knowledge and Mode Declarations

```

:- yield(X), yield(Y), X < Y.

yield_type(0).
yield_type(1).
yield_type(2).

#modeh(yield(const(yield_type))).
#modeb(1, location(const(location))).
#modeb(1, species(const(species))).
#modeb(1, disease(const(disease))).

```

### F.4 Indoor scene classification

#### Mode Declarations

```

label_type(0).
label_type(1).
label_type(2).
label_type(3).
label_type(4).

#modeh(label(const(label_type))).
#modeb(1,image(const(image))).

```

## References

- Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. <http://arxiv.org/abs/1606.06565>
- Bellodi, E., & Riguzzi, F. (2013). Structure learning of probabilistic logic programs by searching the clause space. *Theory and Practice of Logic Programming*, 15.
- Besold, T., Garcez, A., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kühnberger, K.-U., Lamb, L., Lowd, D., Lima, P., de Penning, L., Pinkas, G., Poon, H., & Zaverucha, G. (2017). Neural-symbolic learning and reasoning: A survey and interpretation. <http://arxiv.org/abs/1711.03902>
- Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning* (pp. 1613–1622).
- Cohen, W. W. (2016). Tensorlog: A differentiable deductive database. <http://arxiv.org/abs/1605.06523>
- Dai, W.-Z., & Muggleton, S. (2021). Abductive knowledge induction from raw data. In: Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 1845–1851). <https://doi.org/10.24963/ijcai.2021/254>.
- Dai, W.-Z., Xu, Q., Yu, Y., & Zhou, Z.-H. (2019). Bridging machine learning and logical reasoning by abductive learning. *Advances in Neural Information Processing Systems*, 32.
- Dantsin, E., Eiter, T., Gottlob, G., & Voronkov, A. (2001). Complexity and expressive power of logic programming. *ACM Computing Surveys (CSUR)*, 33(3), 374–425.
- De Raedt, L., Dries, A., Thon, I., Van den Broeck, G., & Verbeke, M. (2015). Inducing probabilistic relational rules from probabilistic examples. In *Proceedings of 24th international joint conference on artificial intelligence (IJCAI)* (Vol. 2015-January, pp. 1835–1842). IJCAI-INT JOINT CONF ARTIF INTELL, United States.
- De Raedt, L., Kimmig, A., Toivonen, H. (2007). Problog: A probabilistic prolog and its application in link discovery. In *IJCAI* (Vol. 7, pp. 2462–2467).

- Donadello, I., Serafini, L., & d'Avila Garcez, A. S. (2017). Logic tensor networks for semantic image interpretation. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 1596–1602). IJCAI, California, USA.
- Evans, R., & Grefenstette, E. (2018). Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61, 1–64.
- Flaminio, T., & Marchioni, E. (2006). T-norm based logics with an independent involutive negation. *Fuzzy Sets and Systems*, 157, 3125–3144.
- Garcez, A.d., & Lamb, L.C. (2020). Neurosymbolic AI: the 3rd wave. <http://arxiv.org/abs/2012.05876>
- Gelfond, M., & Kahl, Y. (2014). *Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach*. Cambridge: Cambridge University Press.
- Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89).
- Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. CoRR abs/1511.08060.
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3), 457–506.
- Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations* (pp. 85–103).
- Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *International conference on machine learning* (pp. 5338–5348).
- Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675–1684).
- Law, M. (2018). Inductive learning of answer set programs. PhD thesis, Imperial College London.
- Law, M., Russo, A., & Broda, K. (2019). Logic-based learning of answer set programs. In *Reasoning Web. Explainable Artificial Intelligence - 15th International Summer School 2019, Bolzano, Italy, September 20-24, 2019, Tutorial Lectures* (pp. 196–231).
- Law, M., Russo, A., Bertino, E., Broda, K., & Lobo, J. (2020). Fastlas: scalable inductive logic programming incorporating domain-specific optimisation criteria. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 2877–2885).
- Law, M., Russo, A., & Broda, K. (2018). Inductive learning of answer set programs from noisy examples. *Advances in Cognitive Systems*, 7, 57–76.
- Law, M., Russo, A., & Broda, K. (2018). The complexity and generality of learning answer set programs. *Artificial Intelligence*, 259, 110–146.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- López-Cifuentes, A., Escudero-Viñolo, M., & Bescós, J. (2020). Álvaro García-Martín: Semantic-aware scene recognition. *Pattern Recognition*, 102, 107256. <https://doi.org/10.1016/j.patcog.2020.107256>.
- Mackay, D. J. C. (1995). Probable networks and plausible predictions—A review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3), 469–505.
- Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. In *Advances in neural information processing systems* (pp. 3749–3759).
- Metcalfe, G., Olivetti, N., & Gabbay, D. M. (2008). *Proof Theory for Fuzzy Logics* (Vol. 36). Springer.
- Minervini, P., Bosnjak, M., Rocktäschel, T., Riedel, S., & Grefenstette, E. (2020). Differentiable reasoning on large knowledge bases and natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020* (pp. 5182–5190).
- Minervini, P., Riedel, S., Stenetorp, P., Grefenstette, E., & Rocktäschel, T. (2020). Learning reasoning strategies in end-to-end differentiable proving. In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13-18 July 2020, Virtual event* (pp. 6938–6949).
- Molnar, C. (2019). Interpretable Machine Learning, Online.
- Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, 8(4), 295–318.
- Muggleton, S. H., Lin, D., & Tamaddoni-Nezhad, A. (2015). Meta-interpretive learning of higher-order dyadic datalog: Predicate invention revisited. *Machine Learning*, 100(1), 49–73.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *33rd conference on neural information processing systems (NeurIPS)* (pp. 13969–13980).

- Pearce, T., Brintrup, A., & Zhu, J. (2021). Understanding Softmax confidence and uncertainty. <http://arxiv.org/abs/2106.04972>
- Quattoni, A., & Torralba, A. (2009). Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 413–420). IEEE.
- Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning* (pp. 63–71)
- Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62(1–2), 107–136.
- Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., Qian, H., Fagin, R., Barahona, F., Sharma, U., et al. (2020). Logical neural networks. <http://arxiv.org/abs/2006.13155>
- Rocktäschel, T., & Riedel, S. (2017). End-to-end differentiable proving. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4-9, 2017, Long Beach, CA, USA* (pp. 3788–3800).
- Sen, P., de Carvalho, B. W., Riegel, R., & Gray, A. (2021). Neuro-symbolic inductive logic programming with logical neural networks. <http://arxiv.org/abs/2112.03324>
- Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *Advances in neural information processing systems* (pp. 3179–3189).
- Sensoy, M., Kaplan, L., Cerutti, F., & Saleki, M. (2020). Uncertainty-aware deep classifiers using generative models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 5620–5627).
- Serafini, L., & d’Avila Garcez, A. S. (2016). Logic tensor networks: Deep learning and logical reasoning from data and knowledge. <http://arxiv.org/abs/1606.04422>
- Stahl, I. (1993). Predicate invention in ilp-an overview. In *European conference on machine learning* (pp. 311–322).
- Tsamoura, E., Hospedales, T., & Michael, L. (2021). Neural-symbolic integration: A compositional perspective. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 5051–5060).
- Tuckey, D., Broda, K., & Russo, A. (2020). Towards structure learning under the credal semantics. In: C. Dodaro, G. A. Elder, W. Faber, J. Fandinno, M. Gebser, M. Hecher, E. LeBlanc, M. Morak, & J. Zangari (Eds.), *International Conference on Logic Programming 2020 Workshop Proceedings Co-located with 36th International Conference on Logic Programming (ICLP 2020), Rende, Italy, September 18-19, 2020*. CEUR Workshop Proceedings, vol. 2678. CEUR-WS.org, Italy.
- Xu, J., Zhang, Z., Friedman, T., Liang, Y., & Broeck, G. (2018). A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning* (pp. 5502–5511). PMLR.
- Yang, Z., Ishay, A., & Lee, J. (2020). Neurasp: Embracing neural networks into answer set programming. In C. Bessiere (Ed.) *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 1755–1762).
- Yang, F., Yang, Z., & Cohen, W. W. (2017). Differentiable learning of logical rules for knowledge base reasoning. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4-9, 2017, Long Beach, CA, USA* (pp. 2319–2328).