# LearnAD: Learning Interpretable Rules for Brain Networks in Alzheimer's Disease Classification

**Thomas Andrews**
Department of Computing
Imperial College London
ta4218@imperial.ac.uk

**Alessandra Russo**
Department of Computing
Imperial College London
a.russo@imperial.ac.uk

**Sara Ahmadi-Abhari**
School of Public Health
Imperial College London
s.ahmadi-abhari@imperial.ac.uk

**Mark Law**
ILASP LTD
mark@ilasp.com

## Abstract

We introduce LearnAD, a neuro-symbolic method for predicting Alzheimer's disease from brain magnetic resonance imaging data, learning fully interpretable rules. LearnAD applies statistical models, Decision Trees, Random Forests, or GNNs to identify relevant brain connections, and then employs FastLAS to learn global rules. Our best instance outperforms Decision Trees, matches Support Vector Machine accuracy, and performs only slightly below Random Forests and GNNs trained on all features, all while remaining fully interpretable. Ablation studies show that our neuro-symbolic approach improves interpretability with comparable performance to pure statistical models. LearnAD demonstrates how symbolic learning can deepen our understanding of GNN behaviour in clinical neuroscience.

## 1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disease characterised by the accumulation of extracellular amyloid-$\beta$ plaques with subsequent intracellular tau neurofibrillary tangles [1]. However, these misfolded proteins appear in the brain prior to presentation of clinical and cognitive symptoms, such as cognitive decline and dementia [2, 3]. Therefore, an improved understanding of the underlying mechanisms of the pathology's progression and spread is essential for the early diagnosis of AD and for identifying therapeutic targets. One avenue is to understand associations between structural brain networks and disease progression and manifestation. In this work, we utilise the Alzheimer's Disease Neuroimaging Study (ADNI) [4] (https://adni.loni.usc.edu/) to learn rules that distinguish between individuals who are cognitively normal (CN) and those who have AD.

Statistical machine learning methods have been used to predict Alzheimer's disease (AD) from structural and functional brain magnetic resonance imaging (MRI) [5, 6]. Early work relied on handcrafted features with classifiers such as Support Vector Machines (SVMs), Decision Trees (DTs), and Random Forests (RFs) [7, 8, 9]. DTs are generally more interpretable than SVMs and RFs, making them preferable when explanations of learned knowledge are required [10]. RFs aggregate predictions across multiple DTs, improving generalisation and typically outperforming DTs, albeit with reduced transparency [10].

Recently, with the availability of large datasets such as ADNI, deep convolutional neural networks (CNNs) have been shown to learn discriminative patterns from minimally preprocessed MRI data [11].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

Despite strong performance, particularly in cross-sectional AD, these models remain black boxes. Their prediction processes are non-transparent, reducing clinician trust and hindering verification that they learn true disease markers rather than confounds. Graph neural networks (GNNs) have been increasingly studied in the healthcare domain [12, 13]. They have emerged as a promising paradigm for modelling AD, offering a natural way to capture the brain's network organisation. Typically, nodes represent regions of interest (ROIs), while edges encode structural or functional connectivity derived from MRIs. Message-passing enables the use of local neighbourhoods for information flow, thus making predictions heavily influenced by the graph's natural local structure when learning global representations [14]. In a clinical context, GNNs learn structural patterns that are predictive of AD. However, when interpreting learned models, it is unclear what contribution specific node features, edges, and subgraphs make in the inference process [15].

Addressing interpretability when learning predictive models for AD remains an open challenge. Symbolic machine learning offers an alternative learning paradigm with formal guarantees of learning from structured data and domain knowledge, with sound and verifiable decision-making [16, 17]. Learned predictive models, represented in logical form, are inherently transparent and human-interpretable. However, the unstructured nature of MRI data and the complexity of predicting AD challenge their applicability and scalability, requiring the need for new methods that can enable symbolic learning from unstructured MRI data in a controlled search space, without reducing their predictive performance. Automatic engineering of relevant structured features from unstructured MRI data can aid in this endeavour.

In this work, we propose LearnAD, a new neuro-symbolic approach that combines statistical machine learning with symbolic learning to learn global interpretable rules for predicting AD from structural MRI of the brain. Firstly, a statistical machine learning component is trained on labelled features extracted from structural MRI data. The features most relevant in the classification of AD patients are selected from this trained model. In the second step, we use a symbolic machine learning system, called FastLAS [18], known for learning interpretable rule-based models from noisy (unstructured) data [19]. Symbolic examples are automatically generated from the labelled patient data (AD versus CN) using the selected patient instance values of the relevant features selected in the first step, but represented as structured contextual information. The selected features are also used to constrain the search space for the symbolic learner. The learned predictive model, generated by FastLAS, generalises across the different local instance-level connections of the brain and determines the most relevant affected brain connections responsible for differentiating AD from CN, and a semi-parametric bound over these regions.

We consider three different instances of our approach, using respectively different statistical machine learning models, DT, RF, and GNN, for learning the most relevant local subgraphs of brain regions. This is in order to evaluate which of these statistical machine learning models is most effective for feature selection when combined with the symbolic learner. We use cross-validation to evaluate the three different instances and show that the accuracy of the predictive model that uses features extracted from the DT outperforms the other two instances. We use a DT, RF, and GNN trained on the full set of features as baseline models to evaluate if the generalisation power of the symbolic learner would achieve similar accuracy with limited features.

Our results show that our best performing instance outperforms the DT, is as accurate as an SVM, and slightly underperforms against the RF and GNN when trained over the full set of features. In all cases, our approach is fully interpretable. Finally, we perform an ablation study to evaluate the benefit of a combined neuro-symbolic approach versus its respective pure statistical machine learning counterpart. In this case, we consider only the DT and RF counterpart models, as the GNN would not be trainable on the sparse networks. Our ablation study shows that our combined neuro-symbolic approach outperforms the DT, and slightly underperforms with respect to an RF trained on these same selected features, but the interpretability of our symbolically learned predictive model is much higher than the RF-trained model.

The paper is structured as follows. In Section 2, we introduce the techniques that our approach builds upon. Section 3 offers a description of the preprocessing of our ADNI dataset that we use in our experiments, followed by Section 4, which presents our neuro-symbolic approach. The results of our experiments are discussed in Section 5. Section 6 summarises related work in the area of AD prediction, and finally Section 7 concludes the paper, suggesting future research directions.

## 2 Background

In this section, we cover two main topics necessary to understand our proposed approach. We briefly describe the statistical machine learning methods that we use to instantiate our approach, and we present the Learning from Answer Sets framework [20] and its current state-of-the-art (SOTA) system, FastLAS [17].

### 2.1 Statistical Machine Learning

In this paper, we focus on three machine learning approaches, DT, RF, and GNN, and their use in binary classification tasks. Classification and Regression Tree (CART) is among the most common methods for learning binary classification decision trees [21]. Given a labelled dataset $D = \{(X, y)\}$, the CART algorithms learn a binary tree model that covers the labelled data. Internal nodes are decision points on features in $X$. Each branch of the learned tree is a decision rule, learned from the data features to maximise the coverage of the given labelled data.

CART is a recursive binary algorithm; at each node it chooses a feature and a split point that best separates the data by reducing the Gini Impurity: the measure of how mixed (or how homogeneous) the classes are at that node. Specifically, it measures the probability that a randomly chosen sample from that node would be incorrectly classified if it is labelled according to the class distribution in that node. Formally, the Gini Impurity is defined as follows:

$$I_G = 1 - \sum_{j=1}^{C} {p_j}^2. \tag{1}$$

where $C$ is the number of classes, and $p_j$ is the proportion of the samples at the node belonging to class $j$. The algorithm aims to partition the feature space into regions that are as "pure" as possible (i.e., have a minimal value of Gini Impurity), so that each region ideally contains samples of only one class. RF are extensions of DT, employing a number of DTs, trained on a bootstrapped subsample of features in $X$, utilising averaging to improve the predictive accuracy of the model [22].

Graph Neural Networks (GNNs) are a class of deep learning models designed to work directly on graph-structured data, where information is represented as nodes (entities) and edges (relationships). One of the most widely adopted GNNs is Graph Convolutional Networks (GCN) [23]. The objective of GCNs is to learn node embeddings and graph representations to perform downstream classification. A GCN is a binary undirected graph of $m$ nodes with $n$ features per node. It is represented as an adjacency matrix, $A \in \{0, 1\}^{m \times m}$, with a feature matrix $X \in \mathbb{R}^{m \times n}$. For a weighted graph, with real-numbered edge weights, the adjacency matrix is extended to $A \in \mathbb{R}^{m \times m}$. During training, a single GCN layer updates node embeddings as

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2} \hat{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}) \tag{2}$$

where $\hat{A} = A + I$, $\tilde{D}^{-1/2}$ is the diagonal degree matrix, $W^{(l)}$ is a learnable weight matrix, and $\sigma$ is the activation. This transformation aggregates information at each node from its local neighbourhood, producing updated node embeddings layer by layer. After a readout step that pools node embeddings into a global graph representation, a multi-layer perceptron produces the final classification.

To extract edges relevant to the prediction, we implement *GNNExplainer* [24]. *GNNExplainer* is chosen due to its simplicity, speed, and its model-agnostic design applicable to any GNN architecture. Given a target classification, in a graph, $G$, *GNNExplainer* aims to find a subgraph $G_s \in G$, and a reduced subset of features $X_s = \{x_i | v_i \in G_s\}$ that are important in a prediction, $y$. Importance is defined with respect to mutual information (MI), such that the *GNNExplainer* aims to maximise the function:

$$\max_{G_s}(Y, (G_s, X_s)) = H(Y) - H(Y | G = G_s, X = X_s). \tag{3}$$

We refer the reader to the original paper for optimisation details. For now, we summarise the final result; a soft mask is learned over the adjacency matrix for subgraphs $G_s$, $A_s \in [0, 1]^{n \times n}$.

### 2.2 Learning from Answer Sets

Learning from Answer Sets (LAS) [18] is a symbolic machine learning [25] paradigm for learning Answer Set Programs (ASP). First, we briefly introduce ASP programs.

Answer Set Programming is a symbolic formalism for representing knowledge and performing inference. Within the scope of this paper, ASP programs are constituted of *normal* rules and *constraints*. The former are of the form $h$ :- $b_1$,...,$b_m$, not $c_1$,...,not $c_m$, where $h$ is the *head* and $b_1$,...,$b_m$, not $c_1$,...,not $c_m$ is collectively the *body* of the rule, not represents negation as failure. A normal rule reads as "$h$ is true if all $b_i$ are true and none of the $c_j$ are proved to be true". Constraints take the form $:-\ b_1, \cdots, b_n$ and have the effect of ruling out solutions where the body is satisfied. The semantics of ASP programs is based on the notion of Herbrand interpretations. The Herbrand interpretation of a given program $P$ is a set of ground atoms constructed from the relations and objects that appear in $P$. Given an ASP program $P$ and a Herbrand interpretation $I$, the reduct program $P^I$ can be constructed using the following three steps: 1. Eliminate all rules containing negated atoms that appear in $I$; 2. Strip all negated atoms from the body of the remaining rules; 3. Replace constraint heads with $\bot$ (meaning falsity). The interpretation $I$ is an answer set (i.e. model) of the program $P$, if it is the minimal model of the reduct program $P^I$. A partial interpretation is expressed as the pair $e_{pi} = \langle e^{inc}, e^{exc} \rangle$, where $e^{inc}$ and $e^{exc}$ contain ground atoms for inclusions and exclusions respectively. An interpretation $I$ *extends* a partial interpretation $e_{pi}$ when $e^{inc} \subseteq I$ and $e^{exc} \cap I = \emptyset$.

A Learning from Answer Sets task is a tuple $\langle B, S_M, E \rangle$, where $B$ is an ASP called *background knowledge*, $S_M$ is a set of rules known as a *hypothesis space*, and $E$ is a set of *examples*. The hypothesis space is specified by means of a set of *mode declarations*, defining which predicates can appear in the head or body of a rule. In this work, we use the SOTA LAS system FastLAS [17]. Examples in FastLAS are referred to as *weighted context-dependent partial interpretation* (WCDPI). They are of the form $\langle e_{id}, e_{pen}, \langle e^{inc}, e^{exc} \rangle, e_{ctx} \rangle$, where $e_{id}$ is an identifier for the example, $e_{pen}$ is a penalty for not covering the example, $\langle e^{inc}, e^{exc} \rangle$ is a partial interpretation and $e_{ctx}$ is an ASP known as the *context* for the example used to provide example-specific information, such as example features. The solution of a given LAS task is called a *hypothesis* and it is an ASP program $H \subseteq S_M$. $H$ is said to *cover* an example $e$ if there exists an answer set of $B \cup e_{\text{ctx}} \cup H$ that contains every atom in $e^{inc}$ and no atom in $e^{exc}$. The score of a hypothesis $H$ is the sum of the length of $H$ (in terms of the number of atoms that appear in $H$) and the penalties of uncovered examples. An optimal solution is a hypothesis $H \subset S_M$ with minimal score.

FastLAS's scalability is rooted in the computation of an opt-sufficient-subset of the given hypothesis space that guarantees the existence of an optimal solution. It also allows for meta-bias over the search space and domain-specific scoring functions as mechanisms for controlling the size of the search space and enabling scalability. The reader is referred to [17] for further details on FastLAS.

## 3 Dataset and Processing

### 3.1 ADNI

Data were extracted as follows: we retrieved all available structural connectomes from ADNI2 and ADNI3 (phases two and three of ADNI). In keeping with case-control studies, each AD patient was matched to a CN patient by age and sex. The resulting dataset is balanced, with 152 CN patients and 152 AD patients (304 total) who were matched by age (CN: 76.7; AD: 76.4) and sex (F: 120; M: 184). Patients had images recorded at multiple time points; we extracted their latest available image. Patients labelled CN maintain a healthy brain throughout the duration of the study. Meanwhile, patients who are labelled AD have, at the time of imaging, been diagnosed with AD.

### 3.2 Connectome Extraction

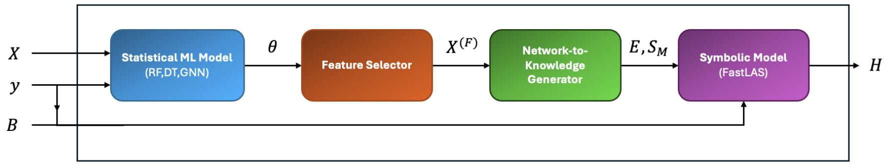
Structural connectomes were extracted with Clinica, a software platform for clinical neuroscience research [26]. Each patient had a T1W MRI with an associated DWI MRI. T1W MRIs were processed with the `t1-freesurfer` pipeline. This pipeline is a wrapper for different functionalities of the FreeSurfer software [27]. The processing includes segmentation of subcortical structures, extraction of cortical surfaces, cortical thickness estimation, spatial normalisation onto the FreeSurfer surface template (FsAverage), and parcellation of cortical regions [26].

The DWI MRIs are first processed with the `dwi-preprocessing-using-t1` pipeline in Clinica. The pipeline corrects for motion, eddy currents, magnetic susceptibility, and bias field distortions [26]. The pipeline utilises FSL [28], ANTs [29], and MRtrix3 [30]. Finally, the connectomes are extracted with the `dwi-connectome` pipeline in Clinica. The connectomes are weighted graphs that encode the structural connections between brain regions defined with a standardised brain atlas. In this work, we use the Desikan–Killiany atlas and obtain 84 brain regions of interest (ROI), consisting of 34 cortical ROIs and 8 subcortical ROIs [31]. This pipeline utilises MRtrix3 and FreeSurfer.

# 4 Methodology

## 4.1 Overview of the approach

A diagram of LearnAD is shown in Figure 1. LearnAD takes as input a labelled dataset of preprocessed structural MRI data and relevant domain-specific background knowledge $B$.



X
y
B
Statistical ML Model (RF,DT,GNN)
θ
Feature Selector
X^(F)
Network-to-Knowledge Generator
E, S_M
Symbolic Model (FastLAS)
H

Figure 1: Architecture of our neuro-symbolic approach. The model consists of 4 components: Statistical ML Model, Feature Selector, Network-to-Knowledge Generator, and Symbolic Model. The dataset $(X, y)$ is provided to the model, along with the background knowledge, $B$. $\theta$ defines the learned parameters of the statistical models, $X^{(F)}$ are the learned relevant features. $S_M$ is the hypothesis space and $E$ are the examples. The learned hypothesis, $H$, is the output.

The first step of our learning approach is to train a statistical machine learning model over the dataset $\{(X, y)\} \subseteq \mathcal{X} \times \mathcal{Y}$, and select from the trained model relevant features. $\mathcal{X}$ defines the space of structural brain networks, whose representation depends on the specific statistical machine learning method used by the approach. $\mathcal{Y}$ describes the space of classification labels, $\mathcal{Y} = \{CN, AD\}$. The Feature Selector identifies in the trained model the most important brain connections in predicting AD.

The selected features are subsequently input to the Network-to-Knowledge Generator (NKG) to define the learning task for the symbolic learner, i.e., construct the set of examples $E$ and the hypothesis space $S_M$. Finally, the FastLAS symbolic learner is used to solve the learning task and generate an optimal solution $H$ in terms of an ASP program that maximises the coverage of the set $E$ of examples. This learning step also takes into account the domain-specific background knowledge $B$.

## 4.2 Statistical Classifier

The first component of our learning task is the specification of our statistical classifier. Each data point is a symmetric weighted anatomical graph, or connectome. Modern neuroimaging techniques are unable to completely remove spurious brain connections, resulting in noisy connectomes [32]. We mask each connectome to remove these connections.

For the DT and RF, we learn a mapping $f : \mathcal{X} \mapsto \mathcal{Y}$, that maps feature vectors $X \in \mathcal{X}$ to a label $y \in \{CN, AD\}$. Each feature vector represents the flattened lower triangular matrix of the masked connectome, with each element defining an ROI-to-ROI connection.

The GCN is characterised by $f(X, \mathcal{E}, A)$, where $X$, $\mathcal{E}$, and $A$ are the feature matrix, edge-feature matrix, and adjacency matrix. In the formulation of our GCN, $X$ is an identity matrix, a positional encoding for the ROIs. As all connectomes are masked, $A$ is the same for all data points. Although the use of edge features alone may seem counterintuitive given that all graphs share the same $A$ and $X$, it is precisely these values that encode the anatomical variation relevant to the task driven by connectivity patterns within each node's local neighbourhood. We learn the mapping $f : (X, \mathcal{E}, A) \mapsto \hat{y}$ where $\hat{y} \in \mathcal{Y}$ approximates $y$.

### 4.3 Feature Selector

To improve the scalability of our symbolic reasoning model, we reduce the context space to $k$ ROI-to-ROI connections most relevant to the learning task. For the instances of our approach that use DT and RF as a statistical ML model, the top $k_{global}$ features (ROI-to-ROI connectivity strength) are extracted by an importance measure: the total reduction in the Gini Impurity atrributable to each feature, summed over all nodes (and averaged over all DTs for RFs). These features are "global" (at the model level) and therefore refer to the full dataset.

For the instance of our approach that uses a GCN classifier, the feature selector is more complex. It employs the *GNNExplainer* to extract edge-level explanations for each individual training graph $x \in \mathcal{X}$. Specifically, for each training graph, *GNNExplainer* identifies a set of salient edges $\mathcal{E}_s \subset \mathcal{E}$, where $|\mathcal{E}_s| = k$ (controlled by limiting the explanation size to $k_{instance}$ edges). As these explanations are at the data point level, we maintain a frequency count $C : \mathcal{E}_s \to \mathbb{N}$ for each edge over the full set of training instances. For each $\epsilon \in \mathcal{E}$, $C(\epsilon)$ gives the number of times the edge $\epsilon$ is returned by *GNNExplainer* among the top $k_{instance}$ relevant features, over the full training set. The selector returns the set $M$ (where $|M| = k_{total}$) of top edges with the highest count. We denote this set of globally relevant features as $\mathcal{E}_{\text{context}} \subseteq \mathcal{E}$. These features provide the contextual information needed when defining the set $E$ for the symbolic learner.

#### 4.3.1 Symbolic Learning

The output features $X^{(F)}$ generated by the Feature Selector need to be represented in the symbolic form accepted by the symbolic learner. In general, any symbolic machine learning system that can induce knowledge from noisy data could be used. In all three instances of our LearnAD approach, we use FastLAS because of its scalability, speed, and robustness to noisy examples. The NKG automatically instantiates the learning task for FastLAS. Recall that FastLAS requires, as a learning task, a tuple $\langle B, S_M, E \rangle$. The $S_M$ is automatically generated by the NKG as a set of rules whose head atom is the predicate $AD$ and the body conditions are all possible combinations of atoms of the form $connection(region(i), region(j), V_{strength})$, for every $(i, j)$ edge in the selected set $M$, and comparator operators ($\geq$, $>$, $<$, $\leq$) for the variable $V_{strength}$. The set $E$ is generated by creating for each datapoint $(X, AD)$ a WCDPI example $e$ defined as $\langle e_{id}, e_{pen}, \langle \{AD\}, \{CN\} \rangle, e_{ctx} \rangle$ where $e_{pen}$ is an integer greater than 0, and the context $e_{ctx}$ is defined as follows:

$$e_{ctx} = \{\texttt{connection}(\texttt{region}(i), \texttt{region}(j), \texttt{strength}(i, j)) \mid (i, j) \in \mathcal{E}_{\text{ctx}}\},$$

where $(i, j)$ indexes node pairs selected according to the masked node feature matrix. Here, the node feature matrix is treated as a binary mask over nodes, and $\mathcal{E}_{\text{ctx}} \subseteq V \times V$ denotes the set of node pairs such that both $i$ and $j$ are among the selected context nodes. The `strength` term refers to the edge weight between region $i$ and region $j$. Examples for CN patients are generated in a similar way, with the difference that their respective $e^{inc} = \{CN\}$ and $e^{exc} = \{AD\}$. As discussed, noisy data is common in neuroimaging data, and a finite noisy penalty is applied to account for this. A uniform penalty $e_{pen}$ is applied across all examples in $E$, as the distribution of classes is equal. A non-uniform penalty could have been implemented in examples with low prediction confidence. Similarly, a penalty could have downweighted brain regions less frequently observed during the explanation stage.

## 5 Experiments and Results

We use our LearnAD approach to learn interpretable rules that can differentiate between AD and CN. Patients with AD exhibit neurodegenerative changes in brain structure (e.g., reduced volume and cortical thinning). Our clinical hypothesis is that alterations in structural connectivity provide a discriminative signal for identifying AD manifestation. Our research question is whether symbolic machine learning, such as FastLAS, can learn interpretable, generalisable rules in this complex clinical setting, achieving comparable accuracy while improving interpretability relative to established statistical baselines.

We perform the task on structural connectomes derived from T1W and DWI MRI from ADNI. Each connectome comprises $84$ ROIs; self-connections (diagonals) are set to $0$. Because probabilistic tractography can introduce spurious edges, we apply proportional thresholding—retaining connections that occur in a sufficient proportion of subjects. We use a strict threshold that preserves $30\%$ of edges, yielding a network sparsity of $70\%$. We favour stringent thresholds, which have been shown to elicit stronger age associations [33].

We train statistical models on the full brain connectomes and evaluate them with repeated stratified cross-validation. We perform 10 repeats with distinct random seeds and report the mean training and validation scores across repeats. In each repeat, the seed is used to generate a new 90% subsample of the dataset and to derive the fold assignments. All subsampling and fold assignments are stratified by diagnosis, sex, and MRI machine manufacturer [34]. Consequently, each repeat operates on a distinct subsample and fold configuration, helping to minimise noise from manufacturer-induced artefacts. We evaluate Decision Tree (DT), Random Forest (RF), Graph Convolutional Network (GCN), and an SVM baseline (see Table 1).

For the Feature Selector, for each DT and RF we pass the top $k_{global}$ features to the Network-to-Knowledge Generator (NKG). For the GCN, we extract $k_{instance}$ features per training graph, then retain the top $k_{total}$ edges globally by frequency across all training graphs. The NKG then constructs the set $E$ of examples for FastLAS. Each example's context comprises facts encoding the two ROIs and the associated connectivity strength; strengths are rounded to four decimal places, then scaled by $1000$ (as the solver cannot handle real numbers) to improve scalability. Once an ASP program $H$ is learned, we use Clingo [35] at inference time to evaluate the accuracy of $H$. To further improve scalability, we partition AD examples into disjoint subsets; each AD subset is paired with the full CN set to form separate learning tasks. We set a base noisy penalty $e_{pen} = 1$ per example and rescale AD penalties by the CN-to-AD ratio in each task to balance positive and negative cases, yielding an effective uniform penalty across classes. The final hypothesis for each configuration is the union of rules learned across these tasks. For LearnAD(DT), $k_{global} = 3$, and the number of disjoint AD sets is 3. For LearnAD(RF), $k_{global} = 6$, and the number of disjoint AD sets is 4. For LearnAD(GCN), $k_{instance} = 10$, $k_{total} = 4$, and the number of disjoint AD sets is 3. Performance of DT$^*$, and RF$^*$, (where $^*$ denotes a model applied on the extracted feature set) and their corresponding symbolic models on the selected features is reported in Table 2.

Table 1: Statistical ML ($X$)

| Model | ACC (%) |
|---|---|
| SVM | 66.35 ± 1.42 |
| RF | 69.90 ± 1.61 |
| DT | 58.58 ± 2.66 |
| GCN | 68.94 ± 2.00 |

Table 2: Statistical ML & FastLAS ($X^{(F)}$)

| Model | ACC (%) |
|---|---|
| RF$^*$ | 70.70 ± 1.73 |
| LearnAD(RF) | 65.65 ± 2.55 |
| DT$^*$ | 63.86 ± 3.49 |
| LearnAD(DT) | 65.72 ± 2.38 |
| GCN$^*$ | — |
| LearnAD(GCN) | 62.97 ± 2.62 |

DTs are prone to overfitting; accordingly, our Feature Selector regularises the model via explicit feature selection, improving DT accuracy from the full network to DT$^*$ by $5.28\%$. RFs, as ensembles, are less susceptible to overfitting, so the gain for RF$^*$ is negligible. We use DT$^*$ and RF$^*$ as baselines for comparison with the symbolic models. LearnAD(DT) achieves accuracy comparable to DT$^*$, but DT$^*$ produces decision rules with bodies of up to $8$ conditions as determined by the maximum depth of the tree. In contrast, FastLAS favours compressed optimal hypotheses, yielding much shorter rules in $H$ that are more human-interpretable. RF$^*$ performs $5.05\%$ better than its symbolic counterpart; however, we consider this a reasonable trade-off for improved interpretability. We do not provide a GCN$^*$ baseline on the selected features because the resulting graphs are too sparse. LearnAD(GCN) is $5.97\%$ lower in accuracy than the GCN trained on the full network, suggesting that most of the GCN's predictive signal can be captured by simple rules. The residual gap may reflect noise introduced by *GNNExplainer* and the Feature Selector or, alternatively, the intrinsic difficulty of distilling message passing and convolutional operations into concise logical constraints. On interpretability, defined by the total number of atoms, our best-performing model has $23.48 \pm 1.6$. By comparison, the DT has $250.91 \pm 38.2$ atoms and the RF has $797.00 \pm 5.35$ atoms.

We display a representative set of rules learned by LearnAD(DT) during a learning task in Figure 2. Across all random seeds, rules involving the left temporal pole–left hippocampus connection appear consistently, and rules involving the right precuneus–right superior parietal connection occur in approximately $80\%$ of seeds. This stability across trials suggests that these connections are reliably associated with AD status in our cohort. The semi-parametric rules learned by FastLAS encode subject-level biomarkers over connection strengths, providing relative thresholds. The observed heterogeneity of AD yields multiple threshold boundaries across subjects, potentially reflecting distinct patterns of degradation or different disease subtypes. Notably, the left hippocampus, implicated in the formation of new memories, exhibits significant atrophy in AD [36], and reduced connectivity between the right precuneus and right superior parietal cortex is consistent with memory impairment observed in old-age AD patients [37].

## 6 Related Work

Machine learning has been applied to AD with a range of neuroimaging modalities and clinical data. Data modalities include structural MRI [5], functional MRI [38], PET data [39], genomic data [40] and proteomic data [41]. Classification tasks vary between the prediction of stages of dementia [42], scores in cognitive tests (MMSE) [43], and amyloid/tau status [40]. Deep learning methods implemented in these tasks include Support Vector Machines, logistic regression, k-means, CNN, reinforcement learning, transformers and GNNs; this is not an exhaustive list [44, 34, 45].

The structural connectome defines anatomical connections in the brain through white matter projections [46]. Accelerated white matter degeneration is observed in individuals with AD [47]. We can understand the progression of clinical-pathological correlations through changes in connectivity and function of distributed neural tracts; therefore, we posit that specific white matter connections should serve as biomarkers for the underlying pathology leading to cognitive decline [48]. In this work, we utilise structural connectomes, extracted from structural MRI and DWI via tract analysis. Machine learning applied to structural connectomes has moved into graph-based methods, due to the natural network structure of the brain [49]. Existing research has focused on developing novel graph-based architectures that learn new representations of the brain data with novel structural and positional encodings [42, 50, 34]. The performance of these models varies, depending on study, dataset size, and distribution. With limited work on developing benchmark datasets stemming from necessary privacy frameworks and missing biomarkers, interpretability is essential for progression in our understanding of the pathological trajectory of neurodegenerative diseases such as AD.

Interpretability of these graph-based methods has been limited to out-of-the-box explainers such as gradient-based saliency methods [6]. The integration of symbolic architectures into deep graph learning is less studied. Architectures that utilise symbolic methods in the explanation of GNNs include GLGExplainer [51], GraphTrail [52], and Monotonic GNNs [53]. Neuro-symbolic AI aims of marrying the reasoning capabilities of symbolic AI with the powerful learning capabilities of connectionist modern deep learning [54]. Advancements in neuro-symbolic learning include the injection of neural inferences to facilitate symbolic inference [19], the improvement in neural training through integration of symbolic knowledge and reasoning, and the mutual interaction of separate neural and symbolic components [55]. Prior work with ILP in the analysis of brain networks is nonexistent as far as the authors are aware. Therefore, this work represents the first step in learning interpretable and logically consistent rules for connection strengths implicated in the development of AD.

## 7 Conclusions

This paper presents a novel neuro-symbolic approach based on a SOTA symbolic machine learning system for investigating biomarkers present in AD-affected brain networks. We have compared three statistical machine learning components, DT, RF, and GCN, as neuro-components for learning relevant search spaces. Evaluation of the symbolic models highlights the ability of symbolic machine learning to learn interpretable rules that are comparable to statistical models. We have been able to generate a hypothesis for subgraphs extracted from a GNN. This work provides a baseline for which we can extend the work with time series data of structural MRI, as well as the inclusion of more imaging modalities, such as functional MRI and PET data.

## Acknowledgments and Disclosure of Funding

T. Andrews is supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1].

## References

[1] H Braak and E Braak. Neuropathological stageing of alzheimer-related changes. *Acta Neuropathologica*, 82:239–259, 1991.

[2] Rik Ossenkoppele, Rik van der Kant, and Oskar Hansson. Tau biomarkers in alzheimer's disease: towards implementation in clinical practice and trials. *The Lancet Neurology*, 21:726–734, 2022.

[3] Fan Yang, Samadrita Roy Chowdhury, Heidi I L Jacobs, Jorge Sepulcre, Van J Wedeen, Keith A Johnson, and Joyita Dutta. Longitudinal predictive modeling of tau progression along the structural connectome. *NeuroImage*, 237:118126, 2021.

[4] Clifford R Jack Jr., Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M Dale, Joel P Felmlee, Jeffrey L Gunter, Derek L G Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S DeCarli, Gunnar Krueger, Heidi A Ward, Gregory J Metzger, Katherine T Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P Debbins, Adam S Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W Weiner. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27:685–691, 2008.

[5] Gurur Gamgam, Alkan Kabakcioglu, Demet Yüksel Dal, and Burak Acar. Disentangled Attention Graph Neural Network for Alzheimer's Disease Diagnosis . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15010. Springer Nature Switzerland, October 2024.

[6] Zhepeng Wang, Runxue Bao, Yawen Wu, Guodong Liu, Lei Yang, Liang Zhan, Feng Zheng, Weiwen Jiang, and Yanfu Zhang. Self-guided Knowledge-injected Graph Neural Network for Alzheimer's Diseases . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15002. Springer Nature Switzerland, October 2024.

[7] E Gerardin, M Chupin, R Cuingnet, B Dubois, S Lehéricy, L Garnero, and O Colliot. Svm classification of patients with alzheimer's disease and mild cognitive impairment using hippocampal shape features. *NeuroImage*, 47:S57, 2009. Organization for Human Brain Mapping 2009 Annual Meeting.

[8] Dana AL-Dlaeen and Abdallah Alashqur. Using decision tree classification to assist in the prediction of alzheimer's disease. In *2014 6th International Conference on Computer Science and Information Technology (CSIT)*, pages 122–126, 2014.

[9] Alessia Sarica, Antonio Cerasa, and Aldo Quattrone. Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, Volume 9 - 2017, 2017.

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[11] Youhao, Zhou Feng, Liu Yong, Tu Liyun Zhou Yanjie, and Li. Learning with domain-knowledge for generalizable prediction of alzheimer's disease from multi-site structural mri. In Anant, Mousavi Parvin, Salcudean Septimiu, Duncan James, Syeda-Mahmood Tanveer, Taylor Russell Greenspan Hayit, and Madabhushi, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 452–461. Springer Nature Switzerland, 2023.

[12] Thomas N., Bloem Peter, van den Berg Rianne, Titov Ivan, Welling Max Schlichtkrull Michael, and Kipf. Modeling relational data with graph convolutional networks. In Roberto, Vidal Maria-Esther, Hitzler Pascal, Troncy Raphaël, Hollink Laura, Tordai Anna, Alam Mehwish Gangemi Aldo, and Navigli, editors, *The Semantic Web*, pages 593–607. Springer International Publishing, 2018.

[13] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. JMLR.org, 2017.

[14] Miru Tang, Baiqing Li, and Hongming Chen. Application of message passing neural networks for molecular property prediction. *Current Opinion in Structural Biology*, 81:102616, 2023.

[15] Nima Dehmamy, AlUCHert-Laszlo Barabasi, and Rose Yu. Understanding the representation power of graph neural networks in learning graph topology. In H Wallach, H Larochelle, A Beygelzimer, F d Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[16] Mark Law. Conflict-driven inductive logic programming. *Theory Pract. Log. Program.*, 23(2):387–414, 2023.

[17] Mark Law, Alessandra Russo, Elisa Bertino, Krysia Broda, and Jorge Lobo. Fastlas: Scalable inductive logic programming incorporating domain-specific optimisation criteria. 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*, pages 2877–2885. AAAI Press, 2020.

[18] Mark Law, Alessandra Russo, Elisa Bertino, Krysia Broda, and Jorge Lobo. Fastlas: Scalable inductive logic programming incorporating domain-specific optimisation criteria. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:2877–2885, 4 2020.

[19] Daniel Cunnington, Mark Law, Jorge Lobo, and Alessandra Russo. Ffnsl: Feed-forward neural-symbolic learner. *Machine Learning*, 112:515–569, 2023.

[20] Mark Law, Alessandra Russo, and Krysia Broda. The complexity and generality of learning answer set programs. *Artif. Intell.*, 259:110–146, 2018.

[21] L Breiman, J Friedman, R Olshen, and C J Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1 edition, 1984.

[22] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[23] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[24] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks, 2019.

[25] Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8:295–318, 1991.

[26] Alexandre Routier, Ninon Burgos, Mauricio Díaz, Michael Bacci, Simona Bottani, Omar El-Rifai, Sabrina Fontanella, Pietro Gori, Jérémy Guillon, Alexis Guyot, Ravi Hassanaly, Thomas Jacquemont, Pascal Lu, Arnaud Marcoux, Tristan Moreau, Jorge Samper-González, Marc Teichmann, Elina Thibeau-Sutre, Ghislain Vaillant, Junhao Wen, Adam Wild, Marie-Odile Habert, Stanley Durrleman, and Olivier Colliot. Clinica: An open-source software platform for reproducible clinical neuroscience studies. *Frontiers in Neuroinformatics*, Volume 15 - 2021, 2021.

[27] Bruce Fischl. Freesurfer. *NeuroImage*, 62:774–781, 2012. 20 YEARS OF fMRI.

[28] Mark Jenkinson, Christian F Beckmann, Timothy E J Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *NeuroImage*, 62:782–790, 2012. 20 YEARS OF fMRI.

[29] Brian B Avants, Nicholas J Tustison, Michael Stauffer, Gang Song, Baohua Wu, and James C Gee. The insight toolkit image registration framework. *Frontiers in Neuroinformatics*, Volume 8 - 2014, 2014.

[30] J-Donald Tournier, Robert Smith, David Raffelt, Rami Tabbara, Thijs Dhollander, Maximilian Pietsch, Daan Christiaens, Ben Jeurissen, Chun-Hung Yeh, and Alan Connelly. Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*, 202:116137, 2019.

[31] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, Marilyn S Albert, and Ronald J Killiany. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31:968–980, 2006.

[32] James A Roberts, Alistair Perry, Gloria Roberts, Philip B Mitchell, and Michael Breakspear. Consistency-based thresholding of the human connectome. *NeuroImage*, 145:118–129, 2017.

[33] Colin R Buchanan, Mark E Bastin, Stuart J Ritchie, David C Liewald, James W Madole, Elliot M Tucker-Drob, Ian J Deary, and Simon R Cox. The effect of network thresholding and weighting on structural brain networks in the uk biobank. *NeuroImage*, 211:116443, 2020.

[34] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25586–25599. Curran Associates, Inc., 2022.

[35] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Multi-shot ASP solving with clingo. *CoRR*, abs/1705.09811, 2017.

[36] Yori Pusparani, Chih-Yang Lin, Yih-Kuen Jan, Fu-Yu Lin, Ben-Yi Liau, John Sahaya Rani Alex, Jeetashree Aparajeeta, Wen-Hung Chao, and Chi-Wen Lung. Hippocampal volume asymmetry in alzheimer disease: A systematic review and meta-analysis. *Medicine*, 104, 2025.

[37] Pukovisa Prawiroharjo, Ken ichiro Yamashita, Koji Yamashita, Osamu Togao, Akio Hiwatashi, Ryo Yamasaki, and Jun ichi Kira. Disconnection of the right superior parietal lobule from the precuneus is associated with memory impairment in oldest-old alzheimer's disease patients. *Heliyon*, 6, 7 2020. doi: 10.1016/j.heliyon.2020.e04516.

[38] Xinmei Qiu, Fan Wang, Yongheng Sun, Chunfeng Lian, and Jianhua Ma. Towards Graph Neural Networks with Domain-Generalizable Explainability for fMRI-Based Brain Disorder Diagnosis . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15002. Springer Nature Switzerland, October 2024.

[39] Giovanna Castellano, Andrea Esposito, Eufemia Lella, Graziano Montanaro, and Gennaro Vessio. Automated detection of alzheimer's disease: a multi-modal approach with 3d mri and amyloid pet. *Scientific Reports*, 14:5210, 2024.

[40] Varuna H Jasodanand, Sahana S Kowshik, Shreyas Puducheri, Michael F Romano, Lingyi Xu, Rhoda Au, and Vijaya B Kolachalama. Ai-driven fusion of multimodal data for alzheimer's disease biomarker assessment. *Nature Communications*, 16:7407, 2025.

[41] Alexa Pichet Binette, Chris Gaiteri, Malin Wennström, Atul Kumar, Ines Hristovska, Nicola Spotorno, Gemma Salvadó, Olof Strandberg, Hansruedi Mathys, Li-Huei Tsai, Sebastian Palmqvist, Niklas Mattsson-Carlgren, Shorena Janelidze, Erik Stomrud, Jacob W Vogel, and Oskar Hansson. Proteomic changes in alzheimer's disease associated with progressive a plaque and tau tangle pathologies. *Nature Neuroscience*, 27:1880–1891, 2024.

[42] Hongchao Jiang and Chunyan Miao. Anatomy-Aware Gating Network for Explainable Alzheimer's Disease Diagnosis . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.

[43] Pieter J van der Veere, Jeroen Hoogland, Leonie N C Visser, Argonde C Van Harten, Hanneke F Rhodius-Meester, Sietske A M Sikkes, Vikram Venkatraghavan, Frederik Barkhof, Charlotte E Teunissen, Elsmarieke van de Giessen, for the Alzheimer's Disease Neuroimaging Initiative (ADNI), Johannes Berkhof, and Wiesje M Van Der Flier. Predicting cognitive decline in amyloid-positive patients with mild cognitive impairment or mild dementia. *Neurology*, 103:e209605, 2024.

[44] Michael Cabanillas-Carbonell and Joselyn Zapata-Paulini. Evaluation of machine learning models for the prediction of alzheimer's: In search of the best performance. *Brain, Behavior, & Immunity - Health*, 44:100957, 2025.

[45] Krishnakant Saboo, Anirudh Choudhary, Yurui Cao, Gregory Worrell, David Jones, and Ravishankar Iyer. Reinforcement learning based disease progression model for alzheimer's disease. In M Ranzato, A Beygelzimer, Y Dauphin, P S Liang, and J Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20903–20915. Curran Associates, Inc., 2021.

[46] Parinaz Babaeeghazvini, Laura M Rueda-Delgado, Jolien Gooijers, Stephan P Swinnen, and Andreas Daffertshofer. Brain structural and functional connectivity: A review of combined works of diffusion magnetic resonance imaging and electro-encephalography. *Frontiers in Human Neuroscience*, 15, 2021.

[47] Jeffrey S Phillips, Nagesh Adluru, Moo K Chung, Hamsanandini Radhakrishnan, Christopher A Olm, Philip A Cook, James C Gee, Katheryn A Q Cousins, Sanaz Arezoumandan, David A Wolk, Corey T McMillan, Murray Grossman, and David J Irwin. Greater white matter degeneration and lower structural connectivity in non-amnestic vs. amnestic alzheimer's disease. *Frontiers in Neuroscience*, Volume 18 - 2024, 2024.

[48] Paul M Matthews, Nicola Filippini, and Gwenaëlle Douaud. Brain structural and functional connectivity and the progression of neuropathology in alzheimer's disease. *Journal of Alzheimer's disease*, 33 Suppl 1:S163–S172, 2013. ObjectType-Article-2.

[49] Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. Braingb: A benchmark for brain network analysis with graph neural networks. *IEEE Transactions on Medical Imaging*, 42(2):493–506, February 2023.

[50] Liang Yang, Yuwei Liu, Jiaming Zhuo, Di Jin, Chuan Wang, Zhen Wang, and Xiaochun Cao. Do we really need message passing in brain network modeling? In *Forty-second International Conference on Machine Learning*, 2025.

[51] Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Liò, and Andrea Passerini. Global explainability of gnns via logic combination of learned concepts, 2023.

[52] Burouj Armgaan, Manthan Dalmia, Sourav Medya, and Sayan Ranu. Graphtrail: Translating GNN predictions into human-interpretable logical rules. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[53] David Jaime Tena Cucala, Bernardo Cuenca Grau, Egor V. Kostylev, and Boris Motik. Explainable GNN-based models over knowledge graphs. In *International Conference on Learning Representations*, 2022.

[54] Johanna Ott, Arthur Ledaguenel, Céline Hudelot, and Mattis Hartwig. How to Think About Benchmarking Neurosymbolic AI? In *CEUR Workshop Proceedings*, Sienne, Italy, July 2023.

[55] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

# A Technical Appendix and Supplementary Materials

## A.1 Hyperparameters and Training Details

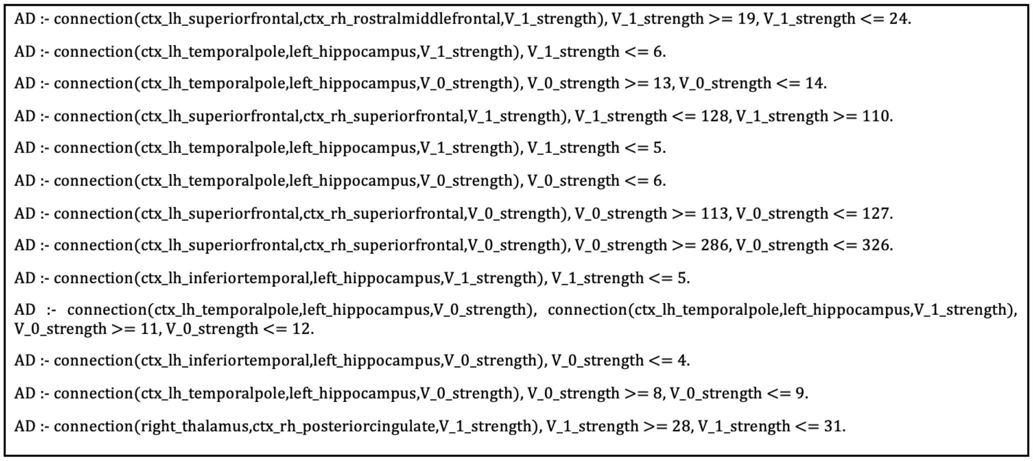
We evaluate Decision Tree (DT), Random Forest (RF), Graph Convolutional Network (GCN), and an SVM baseline (see Table 1). For scikit-learn baselines (SVM, DT, RF; v1.2.1), we use defaults unless noted. The DT uses `max_depth`=8, `criterion`='gini', and `min_samples_split`=2; `random_state` varies per repeat. The RF uses `n_estimators`=100, `max_depth`=8, and `max_features`='sqrt', with `random_state` varied per repeat. The SVM uses scikit-learn defaults unless otherwise specified.

GCN experiments run on an NVIDIA GeForce RTX 3080 GPU. We stack two GCN layers with hidden dimension 32, each followed by ReLU and $\ell_2$ normalisation; input node features use dropout 0.1. Graph-level embeddings are formed by concatenating global sum and max pooling and passed to a linear classifier. Training uses 100 epochs, batch size 8, and Adam optimizer with learning rate $5 \times 10^{-3}$.

We used the `torch_geometric.explain` implementation of *GNNExplainer* provided in PyTorch Geometric (v2.5.2). The explainer was trained for 90 epochs with a learning rate of 0.1. We employed an explanation type of "phenomenon" to characterise the patterns captured by the model, and specified an edge mask type of "object" to identify the most influential edges. The node mask type was disabled (None).

### A.2 LearnAD(DT) Rule Visualization

We show the representative rules learned by LearnAD(DT) in Figure 2.

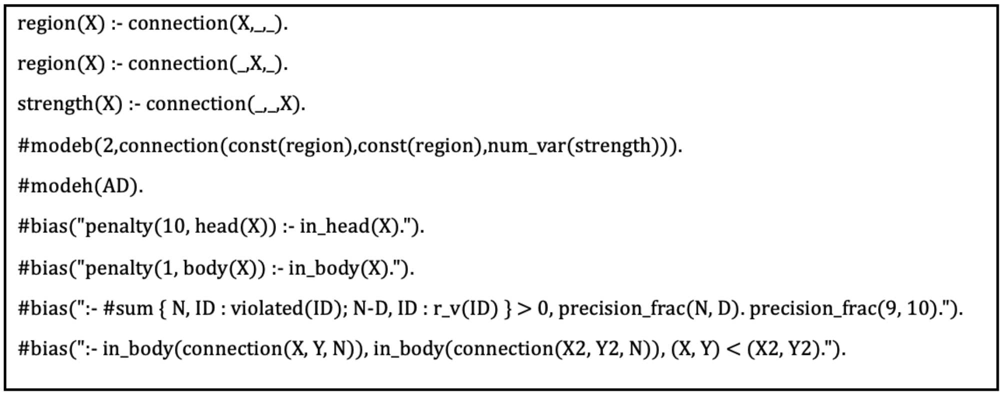


```
AD :- connection(ctx_lh_superiorfrontal,ctx_rh_rostralmiddlefrontal,V_1_strength), V_1_strength >= 19, V_1_strength <= 24.
AD :- connection(ctx_lh_temporalpole,left_hippocampus,V_1_strength), V_1_strength <= 6.
AD :- connection(ctx_lh_temporalpole,left_hippocampus,V_0_strength), V_0_strength >= 13, V_0_strength <= 14.
AD :- connection(ctx_lh_superiorfrontal,ctx_rh_superiorfrontal,V_1_strength), V_1_strength <= 128, V_1_strength >= 110.
AD :- connection(ctx_lh_temporalpole,left_hippocampus,V_1_strength), V_1_strength <= 5.
AD :- connection(ctx_lh_temporalpole,left_hippocampus,V_0_strength), V_0_strength <= 6.
AD :- connection(ctx_lh_superiorfrontal,ctx_rh_superiorfrontal,V_0_strength), V_0_strength >= 113, V_0_strength <= 127.
AD :- connection(ctx_lh_superiorfrontal,ctx_rh_superiorfrontal,V_0_strength), V_0_strength >= 286, V_0_strength <= 326.
AD :- connection(ctx_lh_inferiortemporal,left_hippocampus,V_1_strength), V_1_strength <= 5.
AD :- connection(ctx_lh_temporalpole,left_hippocampus,V_0_strength), connection(ctx_lh_temporalpole,left_hippocampus,V_1_strength),
V_0_strength >= 11, V_0_strength <= 12.
AD :- connection(ctx_lh_inferiortemporal,left_hippocampus,V_0_strength), V_0_strength <= 4.
AD :- connection(ctx_lh_temporalpole,left_hippocampus,V_0_strength), V_0_strength >= 8, V_0_strength <= 9.
AD :- connection(right_thalamus,ctx_rh_posteriorcingulate,V_1_strength), V_1_strength >= 28, V_1_strength <= 31.
```

Figure 2: Representative rules learned by LearnAD(DT).

### A.3 Inductive Bias and Example Construction

We illustrate the inductive bias used in FastLAS in Figure 3.



```
region(X) :- connection(X,_,_).
region(X) :- connection(_,X,_).
strength(X) :- connection(_,_,X).
#modeb(2,connection(const(region),const(region),num_var(strength))).
#modeh(AD).
#bias("penalty(10, head(X)) :- in_head(X).").
#bias("penalty(1, body(X)) :- in_body(X).").
#bias(":- #sum { N, ID : violated(ID); N-D, ID : r_v(ID) } > 0, precision_frac(N, D). precision_frac(9, 10).").
#bias(":- in_body(connection(X, Y, N)), in_body(connection(X2, Y2, N)), (X, Y) < (X2, Y2).").
```

Figure 3: Inductive bias used in a FastLAS task (examples redacted for privacy).