From MLPs to Logic: An End-to-End Neuro-Symbolic Compilation Pipeline for Explainable Clinical Classification

Mashael Alsowail[1,*], Hsi-Ming Ho[1] and Ian Mackie[2]

[1]University of Sussex, Brighton, UK
[2]London South Bank University, London, UK

Abstract

Neural networks can be used as powerful predictors, but they are opaque. This is a serious drawback when they are used in areas such as medical diagnosis, where decisions must be auditable. In this paper, we present a pipeline that turns a trained neural network into a small set of human-readable logical rules and then rebuilds those compiled rules back into a neural-symbolic network with fixed rules for interpretable prediction. Starting from a standard feed-forward classifier, we extract rule-based specifications using an existing symbolic knowledge extraction tool, and recompile them into a differentiable logic model in which logical operations such as conjunction, disjunction, and numerical inequalities are realised as smooth layers. Rulesets extracted in this way are typically large and contain redundancies. We therefore introduce a selection procedure that combines systematic rule removal with a contribution measure based on Shapley values adapted from cooperative game theory, allowing us to identify the rules that genuinely drive predictions. Across two clinical classification tasks, the pipeline produces compact, interpretable approximations of the original neural models: four rules at 92% accuracy on the Wisconsin Breast Cancer dataset, and eight rules at 73% on the real-world Diabetes Health Indicators dataset. Broader empirical validation across further domains is left to future work.

Keywords

Neuro-Symbolic AI, Explainable AI, Rule Extraction, Structural Interpretability, Logic Programming, Shapley Values

1. Introduction

Artificial intelligence and machine learning have become increasingly influential in domains where decisions have significant practical, ethical, and social consequences. Neural networks, in particular, demonstrate strong predictive capacity in a wide range of classification and decision-support tasks [1]. Their capacity to model complex non-linear relationships often allows them to outperform simpler traditional models in terms of raw predictive accuracy. However, this performance advantage is frequently accompanied by a serious limitation: *the internal reasoning process of such models is entirely opaque to human users* [2, 3]. This opacity poses significant challenges to safety, trust, and validation. As these models are deployed in safety-critical environments, such as clinical diagnosis, empirical testing alone is no longer sufficient [4]. Clinicians, patients, and regulators often require not only accurate outputs, but also *understandable explanations* of how and why those outputs are produced. Formal logical frameworks offer a rigorous path to deeply understand, interpret, and structurally verify the behaviour of these models. Symbolic models express decision processes in an explicit and human-readable form, supporting transparency, inspection, and the incorporation of domain constraints.

This paper directly addresses the intersection of logic and neural network analysis, a rapidly evolving paradigm known as *neuro-symbolic AI* [5]. We investigate whether logic can provide structural guarantees of safety-critical properties, and whether logical formalisms can accurately represent continuous neural architectures with better interpretability. Our work demonstrates that by extracting exact logical structures from connectionist models and compiling them back into a differentiable framework, we can

LOGICNN 2026: Workshop on Logical Methods for Neural Network Analysis, July 25, 2026, Lisbon, Portugal
*Corresponding author.
ma2342@sussex.ac.uk (M. Alsowail); hh435@sussex.ac.uk (H. Ho); ian.mackie@lsbu.ac.uk (I. Mackie)
0000-0003-0387-4857 (H. Ho); 0000-0001-8954-7173 (I. Mackie)

build neural networks that achieve comparable accuracy while providing architecturally enforced and human-understandable transparency. Although motivated by clinical decision support, the proposed framework is not limited to healthcare. It is a general neuro-symbolic XAI pipeline for structured classification. Healthcare is used here as a high-stakes evaluation setting because diagnostic tasks require transparent and auditable reasoning; therefore, the framework is evaluated on two clinical tabular case studies: Wisconsin Breast Cancer and Diabetes Health Indicators.

To achieve this, we present an end-to-end neuro-symbolic pipeline. Our core technical contributions address four fundamental challenges in logical neural analysis:

- **Expressivity and Translation:** We demonstrate the exact translation of a trained Multilayer Perceptron (MLP) into declarative logical specifications using the PSyKE platform [6], followed by the recompilation of these rules into a fully differentiable neural network through NeuralLog [7].
- **Logical Interpretation:** We introduce a specialised data-parsing extension, enabling continuous numerical data to be directly expressed as logical facts in NeuralLog, bypassing the sparse, one-hot encodings traditionally required by neuro-symbolic compilers.
- **Structural Guarantees:** We present an extended computational graph architecture, which implements logical inequalities as strict, differentiable TensorFlow operations, ensuring that the network's output is bounded by human-readable rules.
- **Scalability and Tractability:** We introduce a comprehensive, three-tiered *rule selection framework* that mitigates the combinatorial explosion of logical rule evaluation, ensuring tractability.

2. Related Work

The application of formal logic to neural network analysis has expanded rapidly in recent years, yet significant challenges remain regarding continuous expressivity and tractability.

2.1. Expressivity Constraints in Probabilistic Logic

Frameworks such as DeepProbLog [8] extend traditional Prolog by allowing neural networks to evaluate the likelihood of probabilistic facts. DeepProbLog achieves this integration cleanly by associating the output of neural networks to the probability of ground atoms being true. However, DeepProbLog relies on *exact* probabilistic inference via knowledge compilation to Sentential Decision Diagrams (SDDs). When scaling to complex datasets with continuous variables, this exact inference faces severe performance bottlenecks.

2.2. Logical Interpretation in Tensor Networks

Logic Tensor Networks (LTNs) [9] embed first-order logic into real-valued vector spaces, enabling the training of differentiable models that satisfy logical constraints. While LTNs efficiently enforce safety constraints during training using *continuous truth values* (Real Logic), their resulting knowledge representations become highly entangled. Because the logic is distributed across complex vector spaces, extracting discrete, sequential, and human-understandable rules post-training remains challenging.

2.3. Rule Extraction Methodologies

Rule extraction has a central role in neural network interpretability, focusing on transforming sub-symbolic knowledge into symbolic representations. These approaches are commonly classified into three categories [10]:

1. **Decompositional:** Methods that operate at the structural level, analysing internal activations and weights. While highly faithful, they are computationally demanding.
2. **Pedagogical:** Methods that treat the network as a black box, generating synthetic inputs to train a transparent surrogate model (e.g., CART or REAL decision trees). These have the benefit of being more scalable and architecture-agnostic.

3. **Eclectic:** Hybrid methods combining structural and behavioural analyses.

Unlike systems that loosely emulate logical inference, our framework relies on exact pedagogical translation. Through PSyKE, the CART surrogate approximates the predictive behavior of the trained MLP [6], while the extracted tree paths are converted into symbolic logic bounds and compiled into a structured NeuralLog architecture [7]. The hidden layers of our computational graph are defined explicitly by verifiable, human-readable predicates.

3. Expressivity and Translation

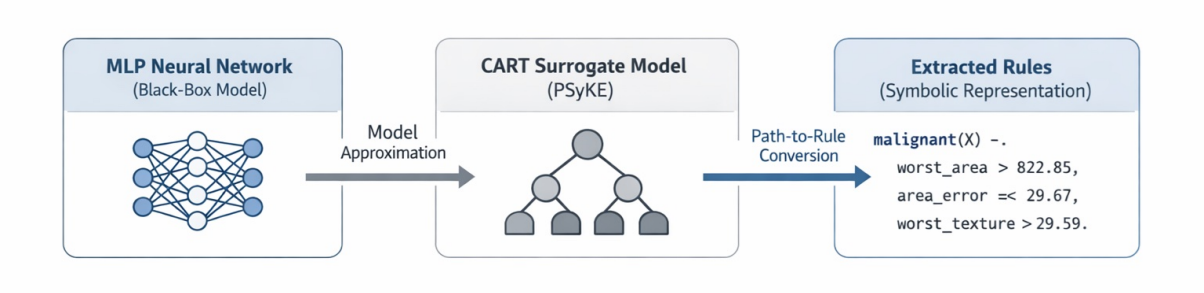
To accurately represent neural network behaviour using logical formalisms, we designed a multi-phase translation pipeline that extracts logic from subsymbolic weights and translates it into Prolog [11].

3.1. Translating Neural Architectures to Logic

We first train an MLP classifier on structured clinical tabular data, specifically the Wisconsin Breast Cancer Dataset [12], featuring 30 continuous numerical features and 569 cases. To ensure syntactic compatibility with the downstream logic compiler, the input space is systematically mapped to Prolog-compliant identifiers (e.g., transforming "mean radius" to `mean_radius`).

To map the subsymbolic network to logical rules, we use the PSyKE framework, specifically employing a CART-based surrogate extraction method. A surrogate decision tree mimics the MLP's non-linear boundaries, and the paths from the root to the leaf nodes are translated into conjunctive first-order rules, resulting in a total of 14 extracted symbolic rules for the Breast Cancer case study (Fig. 1). The same translation procedure is also applied to the Diabetes Health Indicators dataset [13, 14], producing 31 extracted symbolic rules.

Figure 1: Overview of the rule extraction process. A trained MLP neural network is approximated by a CART surrogate model, which is subsequently converted into formal symbolic Prolog specifications.



It is important to distinguish between two levels of correctness in this translation chain. The CART model is a surrogate approximation of the trained MLP, and therefore the extracted rules should not be interpreted as an exact semantic representation of the original neural network. Instead, the MLP-to-CART relationship is evaluated through surrogate fidelity, measured as the proportion of instances for which the CART surrogate and the MLP produce the same prediction:

$$\text{Fidelity}(f_{\text{MLP}}, f_{\text{CART}}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[f_{\text{MLP}}(x_i) = f_{\text{CART}}(x_i)\right]. \tag{1}$$

By contrast, the subsequent conversion from CART paths to logical rules is exact at the syntactic level, since it preserves the decision conditions and class labels represented by the surrogate tree. The NeuralLog recompilation step is therefore evaluated as rule-compilation correctness, while the relationship between the original MLP and the extracted rules is assessed through surrogate fidelity.

3.2. Logical Representation of Continuous Variables

The raw output from PSyKE requires strict syntactic normalisation to ensure compatibility with the logic compiler. We implemented a parsing module to standardise scientific notation and translate complex range conditions into distinct logical inequalities. Standard neuro-symbolic compilers process entities strictly via discrete dictionary mappings, heavily relying on binary truth values derived from sparse relation matrices. This paradigm struggles to natively represent continuous, safety-critical parameters. To resolve this expressivity gap, we introduce `LogicalFactDataset`. Instead of treating facts as static structural assertions that require arbitrary bucketing, our extension dynamically parses continuous atomic facts (e.g., `mean_radius(row_10, 16.02)`) and populates a tensor-ready feature dictionary:

$$\mathcal{T}_{features} = \begin{bmatrix} f_1^{(1)} & f_2^{(1)} & \cdots & f_n^{(1)} \\ f_1^{(2)} & f_2^{(2)} & \cdots & f_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(m)} & f_2^{(m)} & \cdots & f_n^{(m)} \end{bmatrix} \tag{2}$$

LogicalFactDataset is a custom NeuralLog dataset that reads numerical facts from Prolog-style input files and converts them into a tensor-compatible feature representation. Instead of converting continuous values into discrete symbols or one-hot encodings, it parses facts such as `mean_radius(row_10, 16.02)` and stores the value 16.02 under the corresponding row and feature identifier. During the NeuralLog evaluation, these stored numerical values are supplied directly to the custom inequality predicates, such as `greater` and `less_equal`. This enables dynamic evaluation of numerical thresholds inside the compiled logical network without discretising the original feature space.

This permits the direct injection of continuous numerical data into the logic network, preserving the exact mathematical interpretation of the parameters without lossy discretisation.

3.3. Structural Guarantees via Logical Operations

To ensure the network provides formal safety guarantees, the logical constraints must be hard-coded into the computational graph. We extended NeuralLog with a suite of custom TensorFlow literal functions. For example, a safety constraint evaluating if a feature exceeds a threshold (`greater`) is natively implemented as a differentiable logical operator:

```
@neural_log_literal_function("greater")
def tf_greater(inputs, threshold=0.0):
    result = tf.cast(
        tf.greater(
            tf.cast(inputs, tf.float32),
            tf.constant(threshold, dtype=tf.float32)
        ),
        tf.float32
    )
    return result
```

While the underlying tensor operations are standard, this binding mechanism is the critical bridge that allows discrete logical inequalities to be evaluated natively within a differentiable computational graph.

Each extracted rule is instantiated as a `LogicalRuleLayer`. During the forward pass, conditions within a single rule are aggregated using a probabilistic AND operation (element-wise multiplication). Conversely, multiple rules predicting the same target class are aggregated via a probabilistic OR operation, implemented natively in the computational graph using the inclusion-exclusion principle:

$$P(A \vee B) = P(A) + P(B) - P(A)P(B) \tag{3}$$

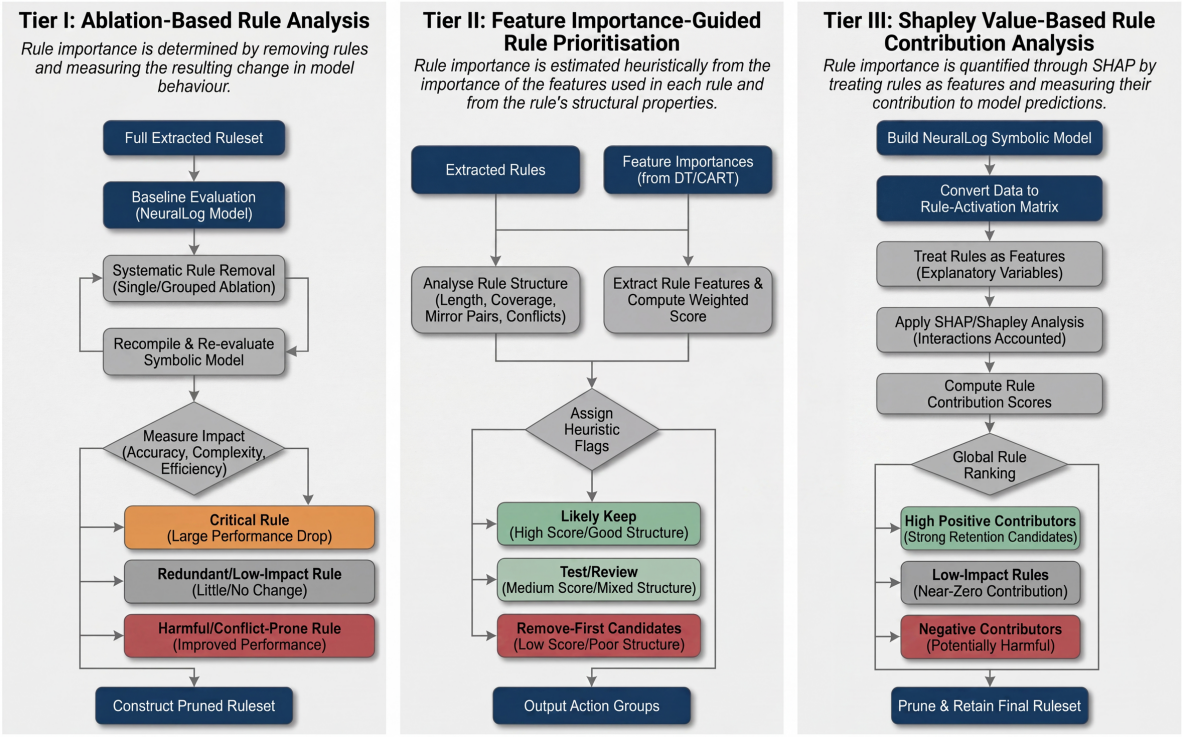
By replacing standard dense-layer aggregations with these probabilistic operations, the network topology is defined exclusively by the extracted logic. Consequently, it is mathematically impossible for the network to make a prediction outside of the explicit logical specification, providing a robust structural guarantee of behaviour.

4. Systematic Rule Selection

A fundamental challenge in logical neural network analysis is the combinatorial explosion of logical paths. Extracted rulesets frequently contain redundant, overlapping, or conflicting logic that bloats the computational graph, rendering inference intractable and explanations incomprehensible. To ensure tractability, we execute a rigorous, three-tiered *rule selection framework* (Fig. 2) prior to network compilation. We evaluate rule criticality based on three dimensions:

1. **Predictive Performance:** The primary indicator of a rule's impact on the decision boundary.
2. **Explainability:** Measured by the symbolic complexity, rule length, and total retained rules.
3. **Computational Efficiency:** The inference time and memory cost of the compiled DAG.

Figure 2: Overview of the proposed three-tiered rule selection framework. Strategy I employs systematic ablation to measure direct behavioural impact; Strategy II uses structural heuristics and CART feature importance for prioritisation; and Strategy III quantifies global rule contributions using Shapley values.



The circular arrow in the ablation tier denotes iterative testing rather than circular reasoning: each candidate rule subset is removed, recompiled, evaluated, and then the procedure continues with the next candidate subset until a compact and stable rule subset is obtained.

4.1. Tier I: Ablation-Based Rule Analysis

The first strategy is an empirical selection by removal and testing. The overall workflow of this process is detailed in Fig. 3. Starting from the full ruleset $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$, rules are systematically removed. The reduced rulesets are recompiled by NeuralLog and re-evaluated. For the Breast Cancer case study, the baseline accuracy, defined as the performance of the full extracted ruleset compiled and evaluated

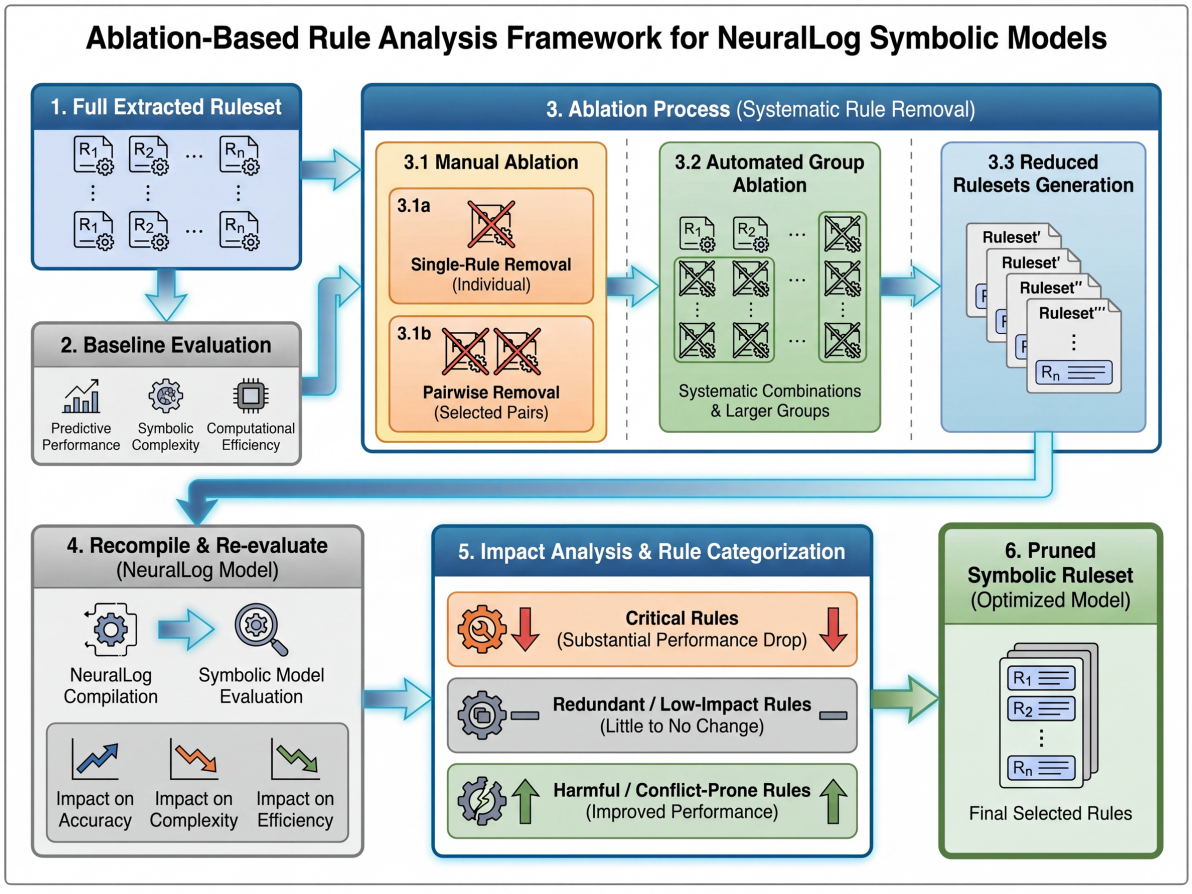
within NeuralLog, is 0.89, while the original MLP achieves an accuracy of 0.94, serving as the reference upper bound for the final comparison. The same ablation procedure is subsequently applied to the Diabetes Health Indicators dataset, where the full compiled ruleset achieves a baseline accuracy of 0.72, while the original MLP achieves 0.77.

For each rule $r_i \in \mathcal{R}$, a reduced ruleset is formed as $\mathcal{R}^{(-i)} = \mathcal{R} \setminus \{r_i\}$. The marginal impact of removing rule r_i is quantified by the difference between the baseline accuracy and the accuracy of the reduced compiled model:

$$\Delta_i = Acc(\mathcal{R}) - Acc(\mathcal{R}^{(-i)}) \tag{4}$$

A larger positive Δ_i indicates a critical rule. A near-zero value suggests redundancy. A negative Δ_i implies that removing the rule improves performance, identifying the rule as harmful or conflict-prone. To capture interaction effects, this analysis was subsequently extended to pairwise and triplet grouped ablations, generating a total of 469 distinct compilation experiments.

Figure 3: Detailed workflow of Strategy I. The process systematically removes individual, paired or group rules from the baseline NeuralLog model, recompiles the computational graph, and categorises rules into critical, redundant, or harmful subsets based on their impact on predictive performance.



4.2. Tier II: Feature Importance-Guided Prioritisation

Exhaustive ablation scales poorly as the ruleset grows, due to the inherent combinatorial explosion. To optimise the search, we employ a fast heuristic: Because extracted rules are built from input features, the inherent importance of those features in the underlying decision tree provides a proxy for rule relevance. For a given rule r, containing the feature set $\mathcal{F}(r)$, we compute a weighted importance score $W(r)$ by summing the Gini importance values (Imp) of its constituent features:

$$W(r) = \sum_{f \in \mathcal{F}(r)} Imp(f) \tag{5}$$

To account for rule complexity, we normalise this score by the rule length $L(r)$ (the number of literals in the rule body), yielding $W(r)/L(r)$. We further integrate structural rule analysis to identify 'mirror pairs', in this paper, a mirror pair refers to two rules whose bodies contain identical or highly overlapping feature conditions, but whose heads predict opposite classes. In the strictest case, the two rules use the same splitting features and threshold structure, possibly with complementary inequalities, while one concludes `benign(X)` and the other concludes `malignant(X)`. Such pairs are important because they may indicate boundary-sensitive or conflict-prone regions of the extracted ruleset, and therefore require further empirical validation through ablation.

4.3. Tier III: Shapley Value-Based Contribution Analysis

To obtain a mathematically rigorous estimate of rule importance, we ground our analysis in cooperative game theory via Shapley values, employing the SHAP (SHapley Additive exPlanations) framework [15] to compute these contributions. Crucially, rather than applying SHAP to the raw input features, we treat the logical rules themselves as the explanatory variables. We convert the dataset into a binary rule-activation matrix $\mathbb{Z} \in \{0,1\}^{N \times |\mathcal{R}|}$, where N is the number of instances and each column corresponds to an extracted rule. Each entry indicates whether rule r_i fires for instance n. The Shapley value ϕ_i for rule r_i measures its average marginal contribution across all possible coalitions of rules S:

$$\phi_i = \sum_{S \subseteq \mathcal{R} \setminus \{r_i\}} \frac{|S|!(|\mathcal{R}| - |S| - 1)!}{|\mathcal{R}|!} [v(S \cup \{r_i\}) - v(S)] \tag{6}$$

Where $v(S)$ represents the predictive output of the NeuralLog model utilising only the subset of rules S. Rules with near-zero SHAP values across all instances are considered redundant and are pruned.

5. Evaluation

5.1. Experimental Setup

The same protocol was applied to both clinical tabular case studies. Each dataset was split into training and test subsets using a stratified split to preserve the class distribution. An MLP was trained on the training subset as the black-box reference, and PSyKE's CART-based extraction produced a rule-based surrogate of the trained MLP, whose rules were normalised into NeuralLog-compatible clauses.

The full compiled ruleset, evaluated in NeuralLog, served as the symbolic baseline. Rules were then selected using three complementary strategies: ablation, feature-importance and structural analysis, and Shapley analysis. The final rulesets were recompiled by NeuralLog and evaluated on the same held-out test set, so that the original MLP, the full extracted-rule model, and the selected neuro-symbolic model were compared under a consistent setting.

We report accuracy, surrogate fidelity, rule complexity, inference time, and peak memory, with bias and fairness as additional indicators. *Accuracy* is the proportion of correctly classified test instances and *surrogate fidelity* the proportion on which the CART surrogate agrees with the MLP. *Rule complexity* combines the number of retained rules with clause-length statistics (average and maximum literals per rule, and a combined score). *Inference time* and *peak memory* capture computational efficiency, where lower values indicate a more efficient model. *Bias* and *fairness* are defined as

$$\text{Bias} = 1 - \frac{H(\hat{y})}{\log K}, \qquad H(\hat{y}) = -\sum_{k=1}^{K} p_k \log(p_k + \epsilon), \tag{7}$$

where p_k is the predicted-class proportion and K the number of classes, and

$$\text{Fairness} = 1 - \left(\max_c R_c - \min_c R_c\right), \tag{8}$$

where R_c is the recall for class c; for binary classification this reduces to $\text{Fairness} = 1 - |R_0 - R_1|$. Lower bias and higher fairness (closer to 1) indicate more balanced predictions across classes.

5.2. Results of Ablation Analysis

Table 1 details the results of the single-rule ablation analysis for the Wisconsin Breast Cancer case study. The empirical data establishes a clear hierarchy of rule importance. Rule 3 and Rule 13 are absolutely critical; removing Rule 3 causes the model's accuracy to collapse to **0.36** ($\Delta = +0.52$). Conversely, removing Rules 5, 7, and 14 actually improves accuracy. Through grouped ablation, we found that simultaneously removing the harmful triplet {5, 7, 14} yielded an optimised 11-rule network with an accuracy of **0.92**, effectively closing the gap with the black-box MLP while significantly reducing the size of the computational graph.

The same ablation procedure was then applied to the Diabetes Health Indicators case study to test whether rule-level behavioural analysis remains useful beyond the Breast Cancer dataset. As shown in Table 2, the single-rule ablation results again reveal a clear hierarchy of rule contribution. Rule 2 is identified as the most critical rule, because its removal causes the largest reduction in accuracy, decreasing performance from approximately 0.72 to 0.6778 ((Δ Acc = +0.0417)). Rules 1 and 7 also make important contributions, since removing them reduces the accuracy to 0.6954 and 0.7083, respectively. The grouped ablation analysis further shows that some rules can be removed jointly without damaging performance. In particular, removing the rule group {28,12,6} produced the best reduced diabetes model, improving the accuracy to approximately **0.73** and yielding an optimised 28-rule network. Overall, the two case studies show that ablation analysis provides the most direct behavioural evidence of rule importance, because each candidate rule or rule group is evaluated by its actual effect on the recompiled neuro-symbolic model. At the same time, the diabetes case shows that rule importance may be more distributed in larger or more complex tabular datasets, making grouped ablation useful for identifying redundant or conflict-prone rule combinations.

Table 1
Summary of single-rule ablation results for the Breast Cancer ruleset. Accuracy is reported after removing each rule, relative to the full-ruleset baseline accuracy of approximately 0.89.

Rank	Rule ID(s)	Category	Acc. dropped	Δ_{Acc}	Interpretation
1	3	Critical	0.362	+0.527	Most important benign rule; removing it causes the largest accuracy collapse.
2	13	Critical	0.655	+0.235	Strong malignant rule; removal substantially reduces performance.
3	11	Helpful	0.877	+0.012	Small but positive contribution to the full ruleset.
4	1, 2, 4, 6, 8, 9, 10, 12	Redundant	0.888	+0.001	Minimal or no individual impact; removing these rules does not materially change performance.
5	14	Harmful	0.894	-0.004	Removal slightly improves accuracy.
6	5	Harmful	0.900	-0.010	Removal improves accuracy, suggesting conflict or redundancy.
7	7	Harmful	0.906	-0.016	Most harmful individual rule; removal gives the largest improvement.

5.3. Results of Heuristic and Structural Analysis

As detailed in Table 3, the feature-importance heuristic provides an efficient initial screening of rules, correctly identifying Rule 14 for removal due to its reliance on a zero-importance feature and highlighting Rules 1, 2, and 5 as critical. However, ablation analysis reveals additional contributions not captured by the heuristic alone; therefore, ablation analysis is applied to the test rules requiring validation. This further evaluation shows that Rules 3 and 13 also contribute significantly to improving predictive

Table 2
Summary of single-rule ablation results for the Diabetes Health Indicators ruleset. Accuracy is reported after removing each rule, relative to the full-ruleset baseline accuracy of approximately 0.72.

Rank	Rule ID(s)	Category	Acc. dropped	Δ_{Acc}	Interpretation
1	2	Critical	0.677	+0.041	Largest performance drop; most critical rule.
2	1	Critical	0.695	+0.024	Strong positive contribution.
3	7	Critical	0.708	+0.011	Important rule; accuracy drops when removed.
4	11	Helpful	0.709	+0.010	Moderate positive contribution.
5	29	Helpful	0.709	+0.010	Moderate positive contribution.
6	25	Helpful	0.714	+0.004	Small positive contribution.
7	22	Helpful	0.715	+0.003	Small positive contribution.
8	15	Helpful	0.717	+0.001	Small positive contribution.
9	3, 4, 5, 8, 9, 10, 13, 16, 18, 19, 20, 21, 23, 24, 26, 27	Redundant	0.7185–0.7204	≈ 0	Minimal individual effect.
10	30	Harmful	0.721	-0.001	Removal slightly improves accuracy.
11	17	Harmful	0.721	-0.001	Removal slightly improves accuracy.
12	0	Harmful	0.722	-0.002	Removal improves accuracy.
13	14	Harmful	0.722	-0.002	Removal improves accuracy.
14	6	Harmful	0.722	-0.002	Removal improves accuracy.
15	12	Harmful	0.724	-0.004	Clear improvement when removed.
16	28	Harmful	0.725	-0.005	Largest improvement when removed.

performance. The final selected ruleset $\{1, 2, 3, 5, 13\}$, therefore, reflects a combined heuristic and empirical selection method, producing a highly compact symbolic model with substantially reduced computational graph complexity and a final accuracy of 0.90.

For the Diabetes Health Indicators case study, the same feature-importance and structural analysis was applied to the extracted ruleset. As shown in Table 4, the heuristic screening stage identifies eleven rules as strong candidates for retention: Rules 1,2,5,7,10,11,12,13,18,19, and 20. These rules are marked as *Keep* because they contain core features, relatively compact structures, or high feature-importance scores. In contrast, Rules 6,14,15,16,17,23,24,25,26,27,28, and 31 are marked as *Prune*, as they are deeper, broader, mirrored, or structurally less useful under the heuristic criteria. The remaining rules are marked as *Validate* because their mirror or overlapping structure requires empirical testing before a final decision.

As in the Breast Cancer case study, the heuristic ranking was not used as the sole basis for the final selection. The rules retained by the screening stage were further examined by targeted ablation validation, which confirmed the same 11-rule subset $\{1, 2, 5, 7, 10, 11, 12, 13, 18, 19, 20\}$. The resulting diabetes model achieved an accuracy of 0.66. Overall, these results show that feature-importance and structural analysis is useful as a fast screening mechanism for larger rulesets, but empirical validation remains necessary because structural importance does not always correspond directly to behavioural importance after NeuralLog recompilation.

5.4. Results of Shapley Contribution Analysis

Applying SHAP to the rule-activation matrix provided a rigorous quantitative distillation of the network. Global rule importance was computed by summing the mean |SHAP| values across all classes, where each class payoff is modelled via a logical OR aggregation over that class's rule activations. For the Wisconsin Breast Cancer case study, the ranking revealed that for both Benign and Malignant classes, Rule 3 completely dominated predictive behaviour, with a mean $|SHAP|$ value of ≈ 0.458, while the vast majority of remaining rules exhibited near-zero contributions across all instances. This sharp drop-off provides mathematical justification for pruning the neural DAG down to the top 4 most influential rules $\{3, 10, 11, 13\}$. The performance of this distilled neuro-symbolic model is presented in Table 5.

For the Diabetes Health Indicators case study, the same SHAP-based rule contribution analysis was applied to the rule-activation matrix. The results show that rule importance is more distributed than in the Breast Cancer case, although a small number of rules still dominate the predictive behaviour of

Table 3
Initial rule categorisation for the Breast Cancer dataset based on feature-importance and structural properties. Rules are grouped only when they form mirror pairs or share similar key features and importance scores. The action column summarises the preliminary screening decision: Keep, Validate, or Prune.

Rule ID(s)	Class	Key Features	L	W	W/L	Heuristic Flags	Action
1	Mal.	area_error, worst_area, worst_texture	3	0.892	0.297	core	Keep
5	Ben.	area_error, symmetry_error, worst_area	3	0.842	0.281	core	Keep
2	Mal.	fractal_dimension_error, worst_area, worst_concave_points	4	0.878	0.220	core	Keep
8, 9	Both	worst_area, worst_fractal_dimension, worst_texture	3	0.835	0.278	core; mirror	Validate
3, 4	Both	worst_area, worst_concave_points, worst_radius	4	0.867	0.217	core; mirror	Validate
6, 7	Both	fractal_dimension_error, symmetry_error, worst_area	5	0.899	0.180	core; mirror	Validate
10	Mal.	area_error, mean_perimeter, worst_texture	3	0.128	0.043	low importance	Validate
11, 12	Both	area_error, mean_perimeter, worst_compactness	4	0.128	0.032	mirror	Validate
13	Mal.	worst_texture	1	0.070	0.070	broad	Validate
14	Ben.	mean_fractal_dimension	1	0.005	0.005	broad; zero importance	Prune

Note. L: number of literals; W: total feature-importance weight; W/L: importance per condition.

Table 4
Initial rule categorisation for the Diabetes Health Indicators dataset based on feature-importance and structural properties. Rules are grouped only when they form mirror pairs or share similar key features and importance scores. The action column summarises the preliminary screening decision: Keep, Validate, or Prune.

Rule ID(s)	Class	Key Features	L	W	W/L	Heuristic Flags	Action
3, 4	Both	bmi, genhlth, highbp	3	0.658	0.219	core; mirror	Validate
2	Diab.	education, genhlth, highbp	3	0.614	0.205	core	Keep
10	Healthy	education, genhlth, highbp, highchol	4	0.667	0.167	core	Keep
20	Healthy	age, genhlth, highbp, income	4	0.667	0.167	core	Keep
1	Diab.	bmi, education, genhlth, highbp, highchol	5	0.762	0.152	core	Keep
5, 11	Diab.	age, education, genhlth, highbp, highchol	5	0.735	0.147	core; mirror	Keep
6, 16, 17, 23, 24	Both	age, bmi, education, genhlth, highbp, highchol	6	0.830	0.138	deep leaf; mirror	Prune
8, 9, 21, 22	Both	education, genhlth, highbp, highchol, hvyalcoholconsump	5	0.684	0.137	core; mirror	Validate
14, 15	Both	age, bmi, education, genhlth, highbp, highchol, menthlth	7	0.857	0.122	deep leaf; mirror	Prune
7, 12, 13	Healthy	age, highbp, highchol	5	0.510	0.102	core	Keep
18, 19	Diab.	age, bmi, highbp, highchol	6	0.605	0.101	core; deep leaf	Keep
27, 28	Both	age, bmi, education, highbp, highchol	7	0.655	0.094	deep leaf; mirror	Prune
25, 26	Both	age, education, highbp, highchol, sex	7	0.575	0.082	deep leaf; mirror	Prune
29, 30	Both	age, bmi, income	3	0.199	0.066	mirror	Validate
31	Diab.	education	1	0.050	0.050	broad	Prune

Note. L: number of literals; W: total feature-importance weight; W/L: importance per condition.

each class. For Class 0, corresponding to the healthy class, Rule 3 has the strongest contribution, with a mean absolute SHAP value of 0.30. a blood pressure, a body mass index below the selected threshold, and better general health. Its dominance indicates that healthy predictions are largely driven by a single highly influential rule, with a smaller number of secondary rules.

For Class 1 corresponding to the diabetic class, the most influential rule is 12, with a mean absolute SHAP value of 0.23. The second most important rule is Rule 7, with a mean absolute SHAP value of 0.118. These values indicate that diabetic predictions are based on a small set of influential rules rather than on a single dominant rule. Based on this ranking, the SHAP-selected diabetes ruleset is $\{2, 3, 7, 8, 12, 23, 30, 31\}$. This provides a compact rule-level explanation of the diabetes model while preserving the main predictive patterns identified by the full extracted ruleset. The performance of the distilled diabetes neuro-symbolic model is presented in Table 5.

Table 5
Final SHAP-selected neuro-symbolic model performance across the two case studies.

Metric	Breast Cancer	Diabetes
Selected Rules	4	8
Accuracy	0.92	0.73
Bias	0.031	0.001
Fairness	0.97	0.95
Complexity Score	18.00	27.50
Avg. Literals/Rule	6.00	7.50
Inference Time (s)	3	18
Peak Memory (MB)	3.67	9.05

6. Conclusion

We have presented an end-to-end neuro-symbolic framework that mitigates the inherent opacity of Multilayer Perceptrons (MLPs) by compiling them into transparent, logic-driven neural architectures. By extending the logic compiler to directly process continuous numerical data—executing inequalities as native, differentiable layers—we successfully overcome traditional expressivity bottlenecks. Furthermore, our three-tiered rule selection framework, culminating in a SHAP-based mathematical distillation, it strictly bounds the combinatorial explosion of logical paths, ensuring the compiled network remains highly tractable.

This paper has evaluated the proposed framework on two clinical case studies to provide an in-depth structural analysis of the distillation process. The Wisconsin Breast Cancer dataset was used as the primary case study; the final distilled model in this dataset achieved 92% predictive accuracy using only four explicit rules. To strengthen the empirical evaluation beyond a single diagnostic benchmark, the framework was also applied to the real-world Diabetes Health Indicators dataset. In this second case study, the pipeline produced a compact eight-rule neuro-symbolic model while retaining approximately 73% predictive accuracy. These results demonstrate that bridging formal logic and neural networks is a viable, highly robust path toward developing auditable, safety-critical AI systems equipped with exact structural guarantees of behaviour. Future work will detail the framework's generalisability across broader multi-class, and higher-dimensional domains.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3.1 Pro to aid with the grammar and readability of the manuscript. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. doi:10.1038/nature14539.

[2] Z. C. Lipton, The mythos of model interpretability, ACM Queue 16 (2018) 31–57.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys (CSUR) 51 (2018) 1–42. doi:10.1145/3236009.

[4] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:0.1038/s42256-019-0048-x.

[5] V. Belle, G. Marcus, The future is neuro-symbolic: Where has it been, and where is it going?, in:

Proceedings of the 40th Annual AAAI Conference on Artificial Intelligence, volume 40, AAAI Press, 2026, pp. 40954–40961. doi:10.1609/aaai.v40i48.42130.

[6] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, On the design of PSyKE: A platform for symbolic knowledge extraction, in: WOA 2021 - 22nd Workshop "From Objects to Agents", volume 29 of *CEUR Workshop Proceedings*, 2021, p. 48.

[7] V. Guimarães, et al., Neurallog: a neural logic language, arXiv preprint arXiv:2107.03225 (2021).

[8] R. Manhaeve, S. Dumančić, A. Kimmig, T. Demeester, L. De Raedt, Deepproblog: Neural probabilistic logic programming, in: Advances in Neural Information Processing Systems (NeurIPS), volume 31, 2018.

[9] L. Serafini, A. S. d. Garcez, Logic tensor networks: Deep learning and logical reasoning from data and knowledge, arXiv preprint arXiv:1606.04422 (2016).

[10] A. B. Tickle, R. Andrews, M. Golea, J. Diederich, The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks, IEEE Transactions on Neural Networks 9 (1998) 1057–1068.

[11] D. S. Warren, V. Dahl, T. Eiter, M. V. Hermenegildo, R. Kowalski, F. Rossi (Eds.), Prolog: The Next 50 Years, volume 13900 of *Lecture Notes in Artificial Intelligence*, Springer Nature, 2023.

[12] W. N. Street, W. H. Wolberg, O. L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, in: Biomedical Image Processing and Biomedical Visualization, volume 1905, SPIE, 1993, pp. 861–870.

[13] A. Teboul, Diabetes health indicators dataset, Kaggle dataset, 2021. URL: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset, accessed: 25 June 2026.

[14] Centers for Disease Control and Prevention, 2015 brfss survey data and documentation, Behavioral Risk Factor Surveillance System, 2015. URL: https://www.cdc.gov/brfss/annual_data/annual_2015.html, accessed: 25 June 2026.

[15] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, volume 30, 2017, p. 4768–4777.