Distilling Neural Knowledge into Interpretable Belief Rule Bases

Xinlei Cao[1,2], Pengsen Liu[3*], Jun Liu[2], Bilal Ahmed Lodhi[2], Dongqiang Yang[1], Yunan Liu[1], Peng Su[4], Li Zou[1,5†], Peng Shi[6]

[1]School of Computer and Artificial Intelligence, Shandong Jianzhu University, Jinan, 250101, Shandong, China
[2]School of Computing, Ulster University, Belfast, BT15 1AP, Northern Ireland, UK
[3]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210093, Jiangsu, China
[4]School of Information Science and Engineering, Qilu Normal University, Jinan, 250200, Shandong, China
[5]Computational Intelligence Center, Shandong Jianzhu University, Jinan, 250101, Shandong, China
[6]chool of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, S.A. 5005, Australia
{cao-x5, j.liu, b.lodhi}@ulster.ac.uk, liups@lamda.nju.edu.cn, {ydq, liuyunan, zouli20}@sdjzu.edu.cn,
{cxl, supeng}@qlnu.edu.cn, peng.shi@adelaide.edu.au

Abstract

In recent years, deep learning has achieved remarkable progress in domains such as image recognition, natural language processing, and speech understanding. However, its inherent "black-box" nature restricts interpretability and undermines trust. As a representative symbolic reasoning method, the Belief Rule Base (BRB) offers strong interpretability and transparent inference for complex, uncertain decision-making. Nevertheless, traditional BRB models rely heavily on manually defined rules and parameters, which limits their scalability to large, data-driven tasks. To address this limitation, we propose a knowledge-distillation-based neuro-symbolic framework, termed Rule Distillation, in which a deep neural network acts as the teacher model to guide the training of a parameterized BRB student model. In this framework, rule weights, attribute weights, rule centers, and consequent belief distributions are treated as trainable parameters optimized via gradient descent. Simultaneously, the soft labels generated by the teacher model provide supervisory signals that enable the student model to capture complex class distributions effectively. Extensive experiments on 23 public datasets demonstrate that the proposed parameterized BRB not only inherits the predictive performance of its teacher model but also achieves faster convergence and stronger generalization, while maintaining interpretability. Overall, this study presents an effective pathway toward explainable artificial intelligence (XAI) by balancing predictive performance with model transparency.

Introduction

In recent years, Explainable Artificial Intelligence (XAI) has become a major research focus, drawing upon advanced AI models such as deep learning. Although these models achieve outstanding performance, their internal mechanisms and decision-making processes often remain opaque, limiting user understanding and trust (Hassija et al. 2024). The rapid proliferation of large-scale AI systems, such as ChatGPT (Cong-Lem et al. 2025), has further amplified concerns about the reliability of AI-generated outputs.

Among interpretable modeling approaches, the Belief Rule Base (BRB) has gained significant attention due to its strengths in uncertainty modeling and transparent reasoning. Originating from traditional IF-THEN rule-based systems (Sun, Ron. 1995), BRB integrates both qualitative knowledge and quantitative data while explicitly preserving uncertainty during inference. Its core reasoning mechanism-the Rule Inference Method using Evidential Reasoning (RIMER) (Yang et al. 2006). This combination enables robust modeling of data characterized by uncertainty, incompleteness, and nonlinearity.

Despite these advantages, parameter optimization in traditional BRB models typically relies on heuristic search methods such as genetic algorithms (Leyva et al. 2021) and particle swarm optimization (Qian et al. 2019). While these methods provide global search capabilities, they often exhibit low efficiency and limited precision when applied to large-scale data. A key limitation is the non-differentiability of the BRB inference process, which prevents end-to-end training with the Backpropagation (BP) algorithm (Werbos 1988).

Meanwhile, Knowledge Distillation (KD), first proposed by Hinton (Hinton G et al. 2015), has emerged as a powerful technique for model compression and knowledge transfer, bridging the gap between large-scale models and lightweight deployable ones. Its central idea is to use soft targets from a teacher model as supervisory signals, enabling the student model to approximate the teacher's learned distribution while maintaining a compact architecture. However, most existing studies have applied KD primarily to lightweight neural networks, mobile models, or decision trees, with limited exploration of its integration into explainable rule-based systems such as BRB.

This study addresses this gap by introducing a Rule Distillation framework that integrates BRB with deep learning. Specifically, we parameterize the BRB so that rule weights, belief degrees, and membership function parameters can be optimized via gradient descent, using the soft predictions of a neural network teacher as supervision. In this framework, the neural network acts as the teacher, and the BRB serves as the student. To enhance differentiability, Gaussian ker-

*These authors contributed equally.
†Corresponding author.

nel functions are incorporated into the rule-matching degree calculation, thereby avoiding the non-differentiability problem of Euclidean distance at zero. Moreover, rule centers, rule weights, attribute weights, membership bandwidths, and consequent belief distributions are all formulated as trainable parameters. The overall reasoning process remains within the Evidential Reasoning (ER) framework, ensuring both interpretability and transparency of inference.

The main contributions of this study are summarized as follows:

- Neural network-BRB integration via Rule Distillation: We propose a novel distillation-based framework in which the BRB student model learns from the soft predictions of a neural network teacher. This design allows the student to achieve both high predictive accuracy and interpretability, offering a new paradigm for the automated optimization of rule bases.
- Comprehensive parameterization and differentiable optimization of BRB: For the first time, rule centers, rule weights, membership bandwidths, and consequent belief distributions are jointly modeled as trainable parameters. This formulation enables end-to-end gradient-based optimization within the Evidential Reasoning (ER) framework, overcoming the limitations of heuristic methods.
- Label-guided clustering for rule center initialization: We introduce a label-guided clustering strategy for initializing rule centers, which are then treated as trainable parameters. This approach ensures a meaningful rule distribution, reduces the number of rules, and alleviates the combinatorial explosion problem. Consequently, the training process becomes more efficient while inference remains transparent and interpretable.

Related Work

Knowledge Distillation

Knowledge Distillation (KD) was first explored by Caruana (Buciluǎ et al. 2006) for model compression, transferring "knowledge" from an ensemble to a single network. Hinton et al. (2015) formalised the framework by softening teacher outputs with a temperature parameter, enabling a compact student to capture the teacher's distributed knowledge. KD has since been widely used, particularly to deploy models in mobile and realtime settings. For instance, FitNet (Romero et al. 2014) guides the student with intermediate teacher features, and Born-Again Networks (Furlanello et al. 2018) repeatedly distil knowledge to improve a single model-reducing model size while maintaining accuracy in vision and speech tasks.

Beyond efficiency, KD has been used to improve interpretability by transferring knowledge from black-box models to more transparent structures. Lu and Lee (2025) distilled deep models into decision trees (KDDT), producing structurally stable and interpretable predictors.

Multilayer perceptrons (MLPs) also play an important role in KD. Compared with deep convolutional or graph models, MLPs are simple, easy to implement, and effective for structured or high-dimensional feature inputs. Many frameworks adopt MLPs as teachers or students. For example, in GLNN, Li et al. (2024) use a graph neural network as teacher and an MLP as student; owing to its simplicity, the MLP achieves competitive results using raw node features and integrates readily with other components for large-scale deployment.

BRB Theoretical Basis

Yang et al. (2006) proposed belief rule-based (BRB) inference grounded in Dempster-Shafer (DS) theory. BRB often suffers from combinatorial explosion as antecedent attributes and evaluation grades grow. To mitigate this, prior work has pursued hierarchical structures (e.g., HFS-BRB with XGBoost-based feature selection (Chang et al. 2013); multilayer tree structures, MTS-BRB (Yang et al. 2023)), structural learning for rule reduction (GT, MDS, Isomap, PCA (Chang et al. 2013)), ensemble and bagging strategies (You et al. 2021), and vector-based BRB that ranks attributes by contribution and transforms them into attribute vectors (Zhang et al. 2021).

Despite many BRB variants (Badhon et al. 2025; Gutiérrez-Urzúa, et al. 2025; Zhang et al.2025), combinatorial growth remains a challenge. Liu et al. (2013) introduced the Extended Belief Rule Base (EBRB), which represents antecedents with belief degrees and learns rules in a data-driven way, alleviating explosion and improving the handling of uncertainty. EBRB, however, introduces issues such as rule boundlessness and inconsistency. To address these, researchers have proposed dynamic rule activation for incomplete or inconsistent data (Calzada et al. 2014) and adaptive activation factor adjustment (Ren et al. 2021); pruning and consolidation using DEA with Wang-Mendel synthesis (Yang et al. 2017), density-based clustering (Zhang et al. 2020), and greedy selection over candidate rules (Gao et al. 2021; Bi et al. 2022); and acceleration via KNN graphs built with HNSW to limit activated rules (Fu et al. 2024). Yang et al. (2022) further proposed cumulative BRB (CBRB), which fuses nearest-neighbour information to activate consistent cumulative rules, balancing interpretability, efficiency, and accuracy.

Integration of BRB with Other Models

Recent work increasingly combines BRB with neural networks to marry data-driven learning with symbolic, interpretable reasoning. Neural networks can predict or adapt BRB parameters (e.g., attribute weights, rule weights, belief distributions), reducing reliance on expert specification and improving performance. Some approaches use deep features as BRB inputs and optimise BRB parameters with heuristic or semiautomatic procedures, often requiring stage-wise training (Zhou et al. 2023). Others use neural networks purely as feature extractors, passing outputs to BRB for transparent reasoning, with promising results in visual sentiment analysis (Zisad et al. 2021) and air pollution forecasting (Kabir et al. 2020). However, because the BRB component is typically non-differentiable, these pipelines are trained separately and cannot exploit end-to-end optimisation; moreover, as the number of rules increases, combinatorial complexity can persist (You et al. 2022).

Overall, integrating BRB with neural and fuzzy methods and distilling neural knowledge into symbolic rules provides a foundation for developing rule-distillation models that retain interpretability without sacrificing accuracy.

Rule Distillation

In this study, we construct a rule distillation model comprising a neural-network teacher and a Belief Rule Base (BRB) student (Figure 1). The framework implements knowledge distillation: a deep neural network (MLP) serves as the teacher, and an evidential-reasoning BRB serves as the student. The teacher produces soft labels that supervise training of the student, enabling the BRB to mimic the teacher's outputs while retaining interpretability.

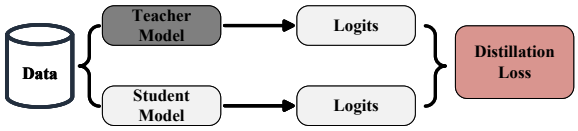

Figure 1: Rules Distillation Framework.

Teacher Model

Within the rule distillation framework, the teacher provides high-quality soft labels that act as supervisory signals for the BRB student. Because our focus is the interpretable BRB rather than the representational power of a very deep network, we adopt a Multi-Layer Perceptron (MLP) as the teacher to balance capacity and controllability. The architecture is shown in Figure 2.

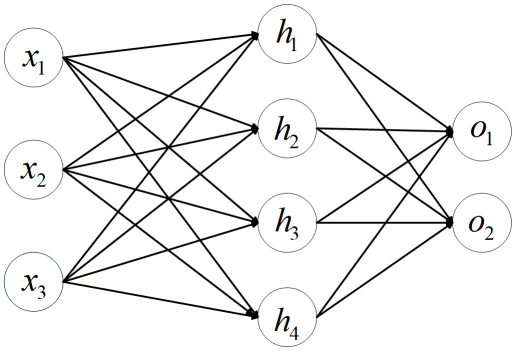

Figure 2: MLP Framework.

Input layer: The input dimension equals the number of premise attributes T in the samples: $x \in \mathbb{R}_T$, where, x denotes the input feature vector.

Hidden layer: The hidden layer contains H neurons and uses ReLU to enhance nonlinearity: $h = \mathrm{ReLU}(0, W_1 x + b_1)$, where, $W_1 \in \mathbb{R}_{H*T}$ denotes the weight matrix and $b_1 \in \mathbb{R}_H$ the bias vector.

Output layer: Hidden features are mapped to a C-classes probability distribution via a linear transform and softmax: $O = W_2 h + b_2, p_t = \mathrm{Softmax}(O)$, where, $W_2 \in \mathbb{R}_{C*T}$ represents the weight matrix, $b_2 \in \mathbb{R}_C$ the bias vector, and p_t the predicted class distribution produced by the teacher model for a given input sample.

Student Model

This section introduces the Backpropagation-based BRB (**BP-BRB**). First, rule centers are initialized using a label-guided Fuzzy C-Means (FCM) strategy (Fu et al. 2021). Second, inference proceeds via the Evidential Reasoning (ER) method. Finally, we show that the end-to-end training of BRB parameters with backpropagation is feasible.

By generating rules via clustering and optimizing within the interpretable BRB framework using gradient descent, the proposed method addresses the long-standing limitation that BRB models could not be trained end-to-end. Figure 3 provides an overview; details follow.

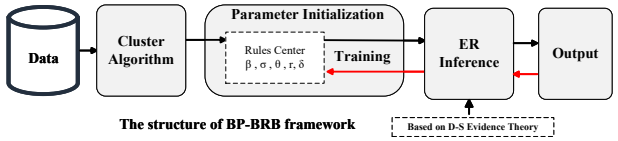

Figure 3: BP-BRB Framework.

Constructing Rule Centers: Constructing the rule base is crucial in rule-based systems. We generate rule centers using FCM with label guidance. This initialization yields a more reasonable rule distribution, facilitates parameterization, and accelerates convergence.

FCM is a soft clustering algorithm that generalizes K-Means by introducing a fuzziness factor F, which controls the degree of cluster overlap. Unlike K-Means, FCM allows each sample to belong to multiple clusters with varying membership degrees, making it particularly suitable for problems with ambiguous boundaries or multi-class tendencies. By modeling the fuzzy relationship between samples and clusters through a membership function, FCM offers stronger representational capacity and is especially effective in uncertainty modeling and fuzzy rule-based systems.

The FCM clustering algorithm is:

$$J = \sum_{c=0}^{C} \sum_{k=1}^{K_c} \sum_{i=1}^{N} u_{ki}^{F} \|x_i - R_k^c\|^2, \sum_{k=1}^{K_c} u_{ki} = 1 \qquad (1)$$

where, J denotes the cost function of FCM; K_c is the number of rule centers; N is the total number of samples; u_{ki} represents the membership degree of sample $\mathbf{x}_i$ to cluster center R_k; F is the fuzziness coefficient (typically set to 2) controlling the degree of fuzziness; and C is the number of classes.

Figure 4 visualizes FCM clustering on the IRIS dataset after projecting the original four-dimensional features to two dimensions for display. Different colors denote different classes; small circles denote samples; large circles are cluster centers. With 150 samples, 12 centers are obtained and well positioned within primary class regions. Class 0

and Class 2 centers show compact aggregation, whereas Class 1 is more dispersed and exhibits overlap near boundaries—consistent with known class interactions. These results support the rationality of label-guided fuzzy clustering for initializing rule centers in Rule Distillation.

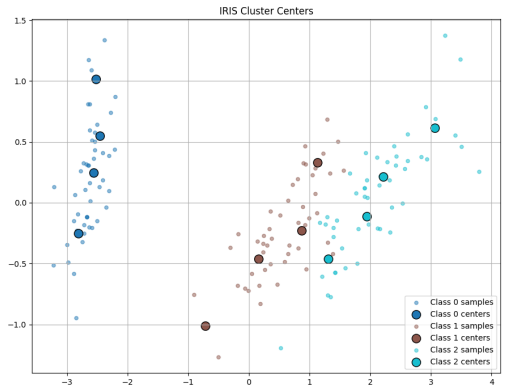

Figure 4: IRIS cluster center.

In a parameterized BRB, a rule center represents a typical premise-attribute pattern for a subset of inputs. A rule can be expressed as:

$$R_k^c : \text{IF}(r_1^c, \ldots, r_T^c)\text{THEN}(\beta_1^c, \ldots, \beta_N^c), \sum_{n=1}^{N} \beta_n^c \leq 1 \quad (2)$$

with rule weight θ_k and attribute weight $\{\delta_1, \ \ldots, \delta_T\}$ and $c \in C$. where, R_k denote the kth rule center, and C the total number of classes. The term r_T^c refers to the Tth attribute of the kth rule center for class c. The vector β_N^c represents the belief degree distribution of the kth rule center in class c; higher values indicate greater reliability of the corresponding conclusion or decision. A rule is considered complete if $\sum_{n=1}^{N} \beta_N^p = 1$; otherwise, it is incomplete. The scalar θ_k denotes the rule weight of the kth rule center, while δ_T represents the weight of the Tth attribute. Both are trainable parameters.

In the parameterized BRB framework, these elements are represented as vectors in continuous space. The rule centers are initialized using the results of Fuzzy C-Means (FCM) clustering: $[r_1^c, \ldots, r_T^c, \beta_1^c, \ldots, \beta_N^c] \in \mathbb{R}_K, \ \beta_i \in [0, 1]$. Rule weight initialization: $\theta_k \in [0, 1], k \in K$. Gaussian Kernel Width: $\sigma_j = 1, \ \ j \in K$. Attribute weight initialization: $\delta_i = 1, \ i \in T$.

As an example, for the IRIS dataset illustrated in Figure 4, 12 cluster centers were generated. The corresponding trainable parameters are summarized in Table 1.

Parameter	Shape	Num
Rule_Centers	$[12, 4]$	48
Rule_Weight	$[12]$	12
sigma	$[12]$	12
Attribute_weight	$[4]$	4
Belief_Degrees	$[12, 3]$	36
total		112

Table 1: Number of Parameters

Parameterized Rule Base Forward Inference: Classical BRB models often compute rule-input similarity using the Euclidean norm, whose gradient is undefined at the origin. To obtain smooth, well-behaved gradients for Backpropagation, we replace the norm with a Gaussian kernel:

$$S_k(x) = \exp\left(-\frac{\sum_{i=1}^{K}\left(x_i^{\delta_i} - R_{k,i}^{\delta_i}\right)^2}{2\sigma_k^2}\right), R_k \in \mathbb{R}_K \quad (3)$$

where, R_k denotes the kth rule center, σ_k is the bandwidth parameter of the membership, δ_i represents the attribute weight. Both σ_k and δ_i are trainable parameters.

The activation weight of the rule center:

$$w_k = \frac{\theta_k S_k}{\sum_{j=1}^{L^*} \theta_j S_j} \quad (4)$$

where, L^* denotes the number of clustering rule center; θ represents the rule weight, S was given by (3).

After obtaining the similarity and activation weights. Yang et al. (2007) has been equivalently transformed into the Evidential Reasoning (ER) analytical algorithm. The ER method is employed to combine the rules and generate the inference output, which can be formulated as (5):

$$\begin{aligned}
\beta_s(x) &= \frac{\mu\left[\prod_{k=1}^{L^*}\left(w_k(x)\beta_{sk} + 1 - w_k(x)\sum_{s=1}^{N}\beta_{sk}\right) - \prod_{k=1}^{L^*}\left(1 - w_k(x)\sum_{s=1}^{N}\beta_{sk}\right)\right]}{1 - \mu\prod_{k=1}^{L^*}\left(1 - w_k(x)\sum_{s=1}^{N}\beta_{sk}\right)} \\
\mu &= \left[\sum_{s=1}^{N}\prod_{k=1}^{L^*}\left(w_k(x)\beta_{sk} + 1 - w_k(x)\sum_{s=1}^{N}\beta_{sk}\right) - (N-1)\prod_{k=1}^{L^*}\left(1 - w_k(x)\sum_{s=1}^{N}\beta_{sk}\right)\right]^{-1}
\end{aligned} \quad (5)$$

where, μ denotes the normalization factor, w_k represents the activation weight of the kth rule,was given by (4).

After obtaining the inference output, the interpretation differs depending on the task type. For regression problems, let $u(D_n)$ denote the confidence of result D_n in attribute D, satisfying $u(D_1) \leq u(D_2) \leq \ldots \leq u(D_n)$. The inference

output is described as:

$$f(x) = \sum_{n=1}^{N} u(D_n)\beta_n \tag{6}$$

For classification problems, D_n represents the nth class, and the inference output is described as (7):

$$f(x) = D_n, \quad n = \arg\left\{ \max_{i=1,\ldots,N} [\beta_i] \right\} \tag{7}$$

Training Loss and Objectives: The training objective consists of two components: (1) Supervised loss with hard labels: The student model is trained to fit the ground-truth labels using cross-entropy loss, which ensures that it does not deviate from the actual classification targets. (2) Distillation loss with soft labels: To encourage the student model to mimic the teacher's predictions, a Kullback-Leibler (KL) divergence loss is employed between the student's output distribution and the teacher's softened probability distribution. The softening is controlled by a temperature parameter T, which smooths the logits, making secondary classes more distinguishable. A scaling factor T^2 is applied to prevent the loss from becoming negligible.

$$\begin{aligned} L_{\mathrm{KD}} =& \lambda \cdot L_{CE}(y, \hat{y}_s) + (1-\lambda) \cdot T^2 \\ & \cdot L_{KL}\left(\mathrm{Softmax}\left(\frac{\hat{y}_T}{T}\right), \mathrm{Softmax}\left(\frac{\hat{y}_s}{T}\right)\right) \end{aligned} \tag{8}$$

The final objective is a weighted combination of these two terms, where λ controls the trade-off between hard-label supervision and soft-label distillation. A larger temperature T produces smoother probability distributions, facilitating richer knowledge transfer from teacher to student. $L_{CE}(y, \hat{y}_s)$ denotes the Cross-Entropy loss between the student model's predictions $\hat{y}_s$ and the ground-truth labels y. $\hat{y}_T$ and $\hat{y}_s$ are the raw output logits from the teacher model and student model.

Figure 5 show the detailed process of rule distillation. For the Rule Distillation modeling procedure, it consists of the following steps and the corresponding pseudo code is provided in Algorithm 1.

Experiments

Datasets Description

We evaluated the proposed Rule Distillation model on 23 benchmark classification datasets from the UCI Machine Learning Repository, spanning domains such as medical diagnosis (e.g., Diabetes) and image recognition (e.g., Penbased). Model performance was assessed through three comparative experiments. Table 2 summarizes the datasets, where Samples denotes the number of instances, Features the number of attributes, and Categories the number of classes.

Training and Hyperparameters

We first analyzed convergence on the Iris dataset. Several optimizers and batch sizes were explored; Adam with batch size 64 yielded the best results. Convergence was sensitive to the initial learning rate: very large values impeded training of the BRB student, whereas very small values slowed convergence. We therefore set the learning rate to 0.002.

Dataset	Samples	Features	Categories
Tae	151	5	3
Iris	150	4	3
Bupa	345	6	2
Heart	270	13	2
Wdbc	569	30	2
Wine	178	13	3
Glass	214	9	6
Vowel	990	13	11
Magic	19020	10	2
Pima	768	8	2
Yeast	1484	8	10
Diabetes	393	8	2
Balance	625	4	3
Twonorm	7400	20	2
Phoneme	5404	5	2
Cleveland	297	13	5
Satimage	6435	36	6
Penbased	10992	16	10
Wisconsin	683	9	2
Transfusion	748	4	2
Page-Blocks	5472	10	5
Contraceptive	1473	9	3
Mammographic	830	5	2

Table 2: Basic information on the classification datasets.

The teacher (MLP) begins with a loss of about 1.1 and decreases to about 0.06 after 200 epochs. The student (BP-BRB) starts around 0.18 and rapidly converges toward zero, indicating substantially faster optimization under distillation. In test accuracy, the teacher remains below 60% for about 20 epochs and surpasses 80% around epoch 80, ultimately reaching about 96%. The student achieves about 78% early on and stabilizes above 96% after about 100 epochs, slightly outperforming the teacher. These results show that Rule Distillation accelerates convergence, improves generalization, and preserves interpretability through its rule-based structure. Details are shown in Figure 6.

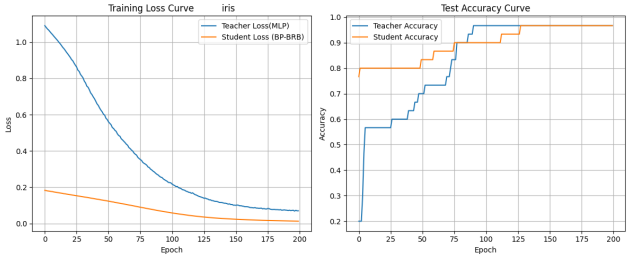

Figure 6: Training loss (left) and test accuracy (right) for Rule Distillation.

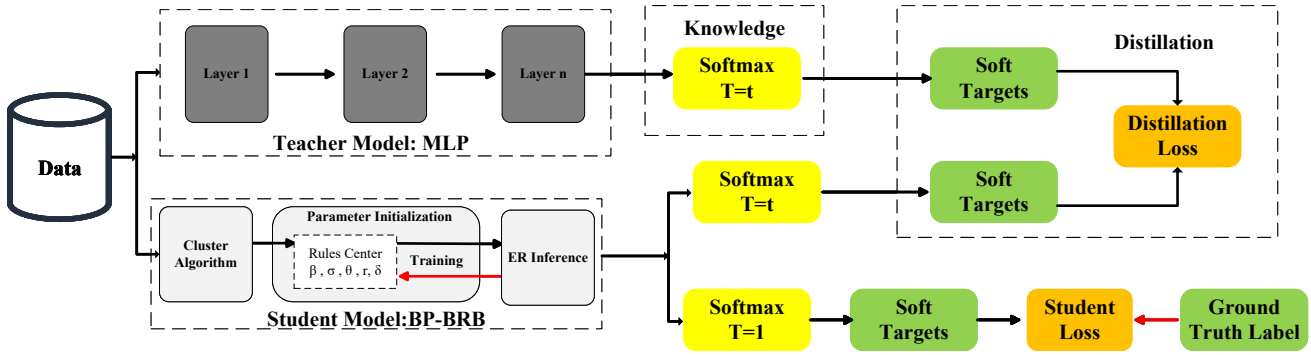

Figure 5: Rule Distillation detail framework.

Results and Analysis

For each dataset, we trained a Rule Distillation model and evaluated it with stratified 10-fold cross-validation (10-CV). Accuracy was used as the primary metric. Comparative baselines and reference results were taken from the literature (Yang et al. 2022; Yang et al. 2025).

Comparison of Rule Distillation with existing BRBSs: We compared the proposed model with nine improved BRB variants reported previously. Table 3 lists their basic characteristics and accuracies on representative datasets. As shown in Table 3, we conducted comparisons on 10 public datasets against eight representative BRB variants, including EBRB, DEA-EBRB, Micro-EBRB, DBSCAN-EBRB, CBRB, AFDRA-EBRB, RCPL-EBRB, and Optimzation-CBRB. These baselines encompass the major directions of recent BRB developments, spanning structural enhancements, inference mechanisms, and optimization strategies, thereby providing a comprehensive benchmark for evaluation.

Rule Distillation demonstrates a clear advantage across nearly all datasets. On ten datasets Tae, Wisconsin, Bupa, Wine, Page-Blocks, Balance, Contraceptive, Yeast, Mammographic, and Phoneme-it achieves the best performance. For instance, on the Bupa and Yeast datasets, the method reaches accuracies of 71.34% and 60.79%, respectively, representing improvements of approximately 11% and 18% over the second-best approaches. These results highlight the strong generalization ability of Rule Distillation in scenarios involving small-sample noisy data and high-dimensional complex data. Similarly, on the Page-Blocks and Balance datasets, Rule Distillation improves accuracy by more than 5%, underscoring its robustness on structured data.

It is worth noting that even in cases where traditional BRBS variants show comparable performance (e.g., Mammographic and Wisconsin datasets), Rule Distillation consistently achieves the highest accuracy. This suggests that the method effectively mitigates local optima during rule optimization and knowledge distillation. Overall, Rule Distillation ranks first on all ten datasets, yielding an average ranking score of 1.0-significantly outperforming existing methods and confirming its comprehensive advantages in both accuracy and stability.

Comparison of Rule Distillation with classical fuzzy rule-based system-related classifiers: The second set of experiments (Table 4) compares Rule Distillation with six classical fuzzy rule-based system (FRBS) classifiers on twelve widely used datasets. The six methods include: the structure learning algorithm on vague environment(SLAVE); fuzzy hybrid genetic-based machine learning (FH-GBML); steadystate genetic algorithm for extracting fuzzy classification rules from data(SGRED); fuzzy association rule-based classification method for high-dimensional problem (FARC-HD); classification with fuzzy association rules(CFAR); and the cumulative belief rule-based system (CBRBS). Notably, SGRED is an improved version of WM-FRBS designed for classification tasks. Table 4 reports the classification accuracies of these six FRBS-related classifiers across the twelve datasets. In each dataset, the number in parentheses indicates the ranking of the method, with smaller ranks denoting better performance.

As shown in Table 4, Rule Distillation achieves either the best or near-best performance on the vast majority of datasets. On several benchmarks-including Heart, Wdbc, Pima, Phoneme, Twonorm, Magic, Wine, Cleveland, and Satimge-it consistently obtains the highest accuracy. In particular, on the Magic and Twonorm datasets, Rule Distillation improves upon the second-best methods by approximately 3% and 0.5%, respectively, demonstrating clear advantages in large-scale and medium-difficulty classification tasks.

On high-dimensional datasets such as Wdbc and Wine, Rule Distillation achieves accuracies of 96.84% and 98.33%, markedly surpassing existing methods and validating its effectiveness in high-dimensional feature spaces. On medical datasets such as Pima and Cleveland, the method again delivers the best results, indicating robust reasoning ability even in the presence of noisy data and class imbalance.

Notably, on the Iris and Penbased datasets, Rule Distillation matches the best-performing approaches, achieving accuracies of 96.00% and 99.14%, respectively. This highlights the method's stability on standard small-scale classification tasks. Overall, Rule Distillation ranks first on 11 out

Indicator/ Dataset	Improved BRBS								Rule Distillation
	EBRB	DEA-EBRB	Micro-EBRB	DBSCAN	CBRB	AFDRA	RCPL	Opt-CBRB	
Tae	49.01(6)	46.36(8)	44.37(9)	46.39(7)	50.99(2)	49.67(5)	50.97(3)	50.33(4)	**54.25(1)**
Wisconsin	95.9(5)	95.46(7)	95.46(7)	55.26(8)	95.61(6)	96.19(3)	95.91(4)	96.34(2)	**96.78(1)**
Bupa	57.97(4)	57.97(4)	57.97(4)	57.97(4)	57.97(4)	57.97(4)	58.26(3)	60.29(2)	**71.34(1)**
Wine	96.32(4)	80.9(7)	95.84(5)	97.87(2)	96.63(3)	96.63(3)	74.73(8)	93.82(6)	**98.33(1)**
Page-Blocks	89.88(5)	89.84(6)	89.84(6)	89.88(5)	89.89(4)	89.91(3)	80.6(7)	90.61(2)	**96.86(1)**
Balance	89.44(3)	84.16(7)	80.32(8)	89.44(3)	84.8(5)	90.56(2)	84.64(6)	85.6(4)	**91.69(1)**
Contraceptive	47.73(3)	46.23(6)	45.49(7)	43.86(8)	46.37(4)	47.73(3)	46.3(5)	48.34(2)	**55.94(1)**
Yeast	33.89(6)	38.81(3)	31.33(9)	33.02(7)	32.35(8)	34.43(5)	38.41(4)	42.05(2)	**60.79(1)**
Mammographic	77.64(9)	80.84(3)	79.88(6)	79.57(8)	79.76(7)	80.48(4)	80.34(5)	81.08(2)	**82.77(1)**
Phoneme	70.65(4)	70.65(4)	70.65(4)	70.65(4)	82.42(2)	70.65(4)	70.65(4)	75.19(3)	**84.73(1)**
Average rank	4.9	5.5	6.5	5.6	4.5	3.6	4.9	2.9	**1**

Table 3: Comparison of Rule Distillation with existing BRBSs

Dataset	SLAVE	FH-GBML	SGRED	FARC-HD	CFAR	CBRBS	Rule Distillation
Heart	71.36(6)	75.93(4)	73.21(5)	84.44(1)	82.22(2)	80.37(3)	**84.44(1)**
Wdbc	92.33(4)	92.26(5)	90.68(6)	95.25(3)	–	96.49(2)	**96.84(1)**
Pima	73.71(5)	75.26(3)	73.37(6)	75.66(2)	65.11(7)	74.09(4)	**76.55(1)**
Phoneme	76.41(5)	79.66(4)	75.55(6)	82.14(3)	70.65(7)	82.42(2)	**84.73(1)**
Twonorm	86.99(5)	85.97(6)	73.98(7)	95.28(3)	91.66(4)	96.77(2)	**97.32(1)**
Magic	73.96(5)	81.30(3)	72.06(6)	84.51(2)	64.84(7)	80.61(4)	**87.22(1)**
Iris	94.44(3)	94.00(4)	94.89(2)	96.00(1)	90.67(5)	96.00(1)	**96.00(1)**
Wine	89.47(7)	92.61(5)	91.88(6)	94.35(3)	93.24(4)	96.63(2)	**98.33(1)**
Cleveland	48.82(7)	53.51(5)	51.59(6)	55.24(3)	53.88(4)	55.89(2)	**57.62(1)**
Satimage	81.69(4)	74.72(6)	77.10(5)	87.32(3)	–	90.46(2)	**90.89(1)**
Penbased	81.16(3)	50.45(5)	67.93(4)	96.04(2)	36.43(6)	99.14(1)	**99.14(1)**
Vowel	71.11(4)	67.07(5)	65.83(6)	71.82(3)	–	98.38(1)	**93.54(2)**
Num No.1	0	0	0	2	0	3	**11**
Average rank	4.8	4.6	5.1	2.4	5.1	2.2	**1.1**

Table 4: Comparison of Rule Distillation with classical FRBS-related classifiers

of 12 datasets, with an average ranking score of 1.1, providing strong evidence of its universality and robustness across datasets of varying scales and levels of difficulty.

Comparison of Rule Distillation with classical ML-related classifiers: To further validate the effectiveness of the proposed model, a third experiment compares Rule Distillation with several classical machine learning classifiers: Artificial Neural Networks (ANN) and k-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Trees (DT), and Support Vector Machines (SVM).

Table 5 summarizes the classification accuracies and rankings across 23 public datasets. The results show that Rule Distillation consistently outperforms traditional machine learning methods on most datasets. Among the 23 benchmarks, it achieves the best performance on 17 datasets, with an average rank of 1.3-substantially better than all competing approaches.

On several challenging datasets such as Glass, Balance, Contraceptive, and Yeast, Rule Distillation demonstrates particularly strong performance. For example, on the Glass dataset, it reaches an accuracy of 88.79%, exceeding Decision Trees and KNN by more than 20 percentage points. On Balance and Contraceptive, accuracies increase to 91.69% and 55.94%, respectively, underscoring the method's adaptability to complex decision boundaries and imbalanced data.

In large-scale multi-class tasks-including Satimage, Pen-

Dataset	KNN	NB	DT	SVM	ANN	Rule Distillation
Heart	74.81(5)	83.70(2)	77.41(4)	55.56(6)	82.22(3)	**84.44(1)**
Wdbc	95.96(3)	92.97(5)	93.32(4)	62.74(6)	96.31(2)	**96.84(1)**
Pima	68.83(6)	77.27(1)	71.82(5)	75.45(3)	73.77(4)	**76.55(2)**
Phoneme	79.44(5)	76.05(6)	86.42(1)	84.49(3)	80.98(4)	**84.73(2)**
Twonorm	94.86(5)	97.83(1)	85.12(6)	97.69(2)	96.96(4)	**97.32(3)**
Magic	74.79(4)	72.69(5)	81.17(3)	65.88(6)	83.73(2)	**87.22(1)**
Iris	94.67(3)	94.00(4)	94.67(3)	97.33(1)	97.33(1)	**96.00(2)**
Wine	97.19(2)	96.63(4)	92.13(5)	44.38(6)	97.17(3)	**98.33(1)**
Cleveland	55.56(3)	54.88(4)	56.57(2)	53.87(5)	52.53(6)	**57.62(1)**
Satimage	86.67(2)	75.39(6)	82.34(5)	82.71(4)	85.61(3)	**90.89(1)**
Penbased	55.90(5)	85.68(3)	74.35(4)	13.71(6)	94.39(2)	**99.14(1)**
Vowel	15.25(6)	67.07(4)	46.97(5)	88.48(2)	84.14(3)	**93.54(1)**
Diabetes	74.09(4)	76.30(1)	73.82(5)	65.10(6)	75.39(3)	**76.03(2)**
Transfusion	76.20(4)	75.40(5)	78.34(2)	75.27(6)	76.34(3)	**79.95(1)**
Glass	66.36(5)	48.60(6)	66.82(4)	68.69(2)	67.76(3)	**88.79(1)**
Mammographic	79.04(6)	82.41(3)	83.98(1)	79.88(5)	80.60(4)	**83.25(2)**
Tae	–	34.44(3)	49.01(2)	34.37(4)	–	**54.25(1)**
Wisconsin	–	65.01(3)	91.51(2)	65.01(3)	–	**96.78(1)**
Bupa	–	57.97(2)	57.39(3)	57.97(2)	–	**71.34(1)**
Page-Blocks	–	89.78(3)	89.44(4)	89.82(2)	–	**96.86(1)**
Balance	–	45.76(4)	59.52(2)	53.60(3)	–	**91.69(1)**
Contraceptive	–	42.70(3)	43.65(2)	42.70(3)	–	**55.94(1)**
Yeast	–	31.22(3)	56.81(2)	31.20(4)	–	**60.79(1)**
Num No.1	0	3	2	1	1	**17**
Average rank	4.25	3.5	3.3	3.9	3.1	**1.3**

Table 5: Comparison of Rule Distillation with classical ML-related classifiers.

based, and Page-Blocks-Rule Distillation again achieves the highest accuracies (90.89%, 99.14%, and 96.86%, respectively), highlighting its strong generalization capability in high-dimensional and multi-class settings. Notably, on certain medical datasets (e.g., Pima and Mammographic), its performance remains close to the best methods, consistently ranking within the top two. This indicates its reliability in medical diagnostic tasks.

Overall, Rule Distillation exhibits superior classification accuracy and stability across diverse scenarios, confirming its potential value for interpretable learning and real-world applications.

Conclusion and Future Work

This paper proposes a Rule Distillation framework that integrates knowledge distillation with Belief Rule Bases (BRB), thereby achieving a fusion of neural networks and symbolic reasoning. By employing an MLP as the teacher model and leveraging its soft predictive outputs to guide the parameter learning of the BRB student model, the framework enables the student to achieve performance comparable to, or even surpassing, the teacher model, while preserving an interpretable rule structure.

The proposed method not only enhances the adaptability and trainability of BRBs but also offers a novel approach for developing interpretable reasoning systems with human-AI collaborative capabilities. Future work will explore extending this framework to more complex tasks and incorporating transfer learning to improve performance in data-scarce scenarios. Moreover, the model can easily accommodate domain-specific constraints or controllable rules-or example, by fixing certain rule weights or setting upper and lower bounds-thus providing a practical pathway toward controllable and trustworthy AI.

Acknowledgments

The authors would like to appreciate all participants of peer review. This work was supported in part by National Natural Science Foundation of China under Grant 62176142 and 62406051 and in part by Taishan Scholar under Grant tsqn202507224.

References

Badhon, B.; Chakrabortty, R. K.; Anavatti, S. G.; and Vanhoucke, M. 2025. IRAF-BRB: An explainable AI framework for enhanced interpretability in project risk assessment. *Expert Systems with Applications*, 127979.

Bi, W.; Gao, F.; Zhang, A.; and Bao, S. 2022. A framework for extended belief rule base reduction and training with the greedy strategy and parameter learning. *Multimedia Tools and Applications*, 81(8): 11127–11143.

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 535–541.

Calzada, A.; Liu, J.; Wang, H.; and Kashyap, A. 2014. A new dynamic rule activation method for extended belief rule-based systems. *IEEE Transactions on knowledge and data engineering*, 27(4): 880–894.

Chang, L.; Zhou, Y.; Jiang, J.; Li, M.; and Zhang, X. 2013. Structure learning for belief rule base expert system: A comparative study. *Knowledge-Based Systems*, 39: 159–172.

Cong-Lem, N.; Soyoof, A.; and Tsering, D. 2025. A systematic review of the limitations and associated opportunities of ChatGPT. *International Journal of Human–Computer Interaction*, 41(7): 3851–3866.

Fu, Y.-G.; Lin, X.-Y.; Fang, G.-C.; Li, J.; Cai, H.-Y.; Gong, X.-T.; and Wang, Y.-M. 2024. A novel extended rule-based system based on K-Nearest Neighbor graph. *Information Sciences*, 662: 120158.

Fu, Y.-G.; Ye, J.-F.; Yin, Z.-F.; Chen, L.-J.; Wang, Y.-M.; and Liu, G.-G. 2021. Construction of EBRB classifier for imbalanced data based on Fuzzy C-Means clustering. *Knowledge-based systems*, 234: 107590.

Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born again neural networks. In *International conference on machine learning*, 1607–1616. PMLR.

Gao, F.; Zhang, A.; Bi, W.; and Ma, J. 2021. A greedy belief rule base generation and learning method for classification problem. *Applied Soft Computing*, 98: 106856.

Gutiérrez-Urzúa, F.; Freddi, F.; and Tubaldi, E. 2025. Seismic risk and failure modes assessment of steel BRB frames under earthquake sequences. *Structural Safety*, 115: 102598.

Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; and Hussain, A. 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1): 45–74.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, G.; He, W.; Sun, C.; Zhu, H.; Li, K.; and Jiang, L. 2023. Hierarchical belief rule-based model for imbalanced multi-classification. *Expert Systems with Applications*, 216: 119451.

Kabir, S.; Islam, R. U.; Hossain, M. S.; and Andersson, K. 2020. An integrated approach of belief rule base and deep learning to predict air pollution. *Sensors*, 20(7): 1956.

Leyva, H.; Bojórquez, J.; Bojórquez, E.; Reyes-Salazar, A.; Carrillo, J.; and López-Almansa, F. 2021. Multi-objective seismic design of BRBs-reinforced concrete buildings using genetic algorithms. *Structural and Multidisciplinary Optimization*, 64(4): 2097–2112.

Li, J.; Shi, B.; Cui, E.; Wei, H.; and Zheng, Q. 2024. Teaching MLP more graph information: A three-stage multitask knowledge distillation framework. *arXiv preprint arXiv:2403.01079*.

Liu, J.; Martinez, L.; Calzada, A.; and Wang, H. 2013. A novel belief rule base representation, generation and its inference methodology. *Knowledge-based systems*, 53: 129–141.

Lu, X.; and Lee, J. J. 2025. Knowledge Distillation Decision Tree for Unravelling Black-Box Machine Learning Models. *The New England Journal of Statistics in Data Science*, 1–24.

Qian, B.; Wang, Q.-Q.; Hu, R.; Zhou, Z.-J.; Yu, C.-Q.; and Zhou, Z.-G. 2019. An effective soft computing technology based on belief-rule-base and particle swarm optimization for tipping paper permeability measurement. *Journal of Ambient Intelligence and Humanized Computing*, 10(3): 841–850.

Ren, T.-Y.; Ye, F.-F.; Yang, L.-H.; Liu, J.; and Wang, Y. 2021. Dynamic Rule Activation Method Based on Activation Factor for Extended Belief Rule-based Systems. In *2021 16th international conference on intelligent systems and knowledge engineering (ISKE)*, 82–86. IEEE.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; and Gatta, C. 2014. Yoshua 426 Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 473.

Sun, R. 1995. Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, 75(2): 241–295.

Werbos. 1988. Backpropagation: Past and future. In *IEEE 1988 International Conference on Neural Networks*, 343–353. IEEE.

Yang, J.-B.; Liu, J.; Wang, J.; Sii, H.-S.; and Wang, H.-W. 2006. Belief rule-base inference methodology using the evidential reasoning approach-RIMER. *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans*, 36(2): 266–285.

Yang, L.-H.; Liu, J.; Ye, F.-F.; Wang, Y.-M.; Nugent, C.; Wang, H.; and Martínez, L. 2022. Highly explainable cumulative belief rule-based system with effective rule-base modeling and inference scheme. *Knowledge-Based Systems*, 240: 107805.

Yang, L.-H.; Wang, Y.-M.; Lan, Y.-X.; Chen, L.; and Fu, Y.-G. 2017. A data envelopment analysis (DEA)-based method for rule reduction in extended belief-rule-based systems. *Knowledge-Based Systems*, 123: 174–187.

Yang, L.-H.; Ye, F.-F.; Liu, J.; and Wang, Y.-M. 2023. Belief rule-base expert system with multilayer tree structure for complex problems modeling. *Expert Systems with Applications*, 217: 119567.

Yang, L.-H.; Yu, D.-N.; Ye, F.-F.; Hu, H.; and Ye, Q. 2025. A Novel Modeling Approach for Cumulative Belief Rule-Base With Joint Optimization and Rule Synthesis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

You, Y.; Sun, J.; Chen, Y.-w.; Niu, C.; and Jiang, J. 2021. Ensemble belief rule-based model for complex system classification and prediction. *Expert Systems with Applications*, 164: 113952.

You, Y.; Sun, J.; Guo, Y.; Tan, Y.; and Jiang, J. 2022. Interpretability and accuracy trade-off in the modeling of belief rule-based systems. *Knowledge-Based Systems*, 236: 107491.

Zhang, A.; Gao, F.; Yang, M.; and Bi, W. 2020. A new rule reduction and training method for extended belief rule base based on DBSCAN algorithm. *International Journal of Approximate Reasoning*, 119: 20–39.

Zhang, J.; Liu, W.; Zhang, M.; and Long, Y. 2025. Effects of BRB on cyclic behavior of resilient RAC composite frames with ultra-high strength steel bars. *Engineering Structures*, 332: 120059.

Zhenjie, Z.; Xiaobin, X.; Peng, C.; Xudong, W.; Xiaojian, X.; and Guodong, W. 2021. A novel nonlinear causal inference approach using vector-based belief rule base. *International Journal of Intelligent Systems*, 36(9): 5005–5027.

Zhou, Z.; Ming, Z.; Wang, J.; Tang, S.; Cao, Y.; Han, X.; and Xiang, G. 2023. A Novel Belief Rule-Based Fault Diagnosis Method with Interpretability. *Computer Modeling in Engineering & Sciences (CMES)*, 136(2).

Zisad, S. N.; Chowdhury, E.; Hossain, M. S.; Islam, R. U.; and Andersson, K. 2021. An integrated deep learning and belief rule-based expert system for visual sentiment analysis under uncertainty. *Algorithms*, 14(7): 213.

Algorithm 1: Pseudocode of Rule Distillation modeling procedure

1. Teacher Model (MLP)

```
teacher := MLP(input_dim, hidden_dim, num_classes)
train(teacher, train_loader, loss=CrossEntropy)
soft_labels := softmax(teacher(train_X))
```

2. Build Rule Center (FCM)

```
rule_centers := []
while class c do
    samples := x — y=c
    centers := FCM(samples, K=clusters_per_class)
    rule_centers.append(centers)
end while
```

3. Student Model(BRB)

```
student := BRB(rule_centers, num_classes)
```

4. Training

```
while (xb, soft_yb) in (train_loader_with_softlabels) do
    pred := student(xb)
    loss := distill_loss(pred, soft_yb)
    update(student, loss)
end while
```

5. evaluate

```
acc_teacher := evaluate(teacher, test_loader)
acc_student := evaluate(student, test_loader)
return acc_teacher, acc_student
```