Class-balanced interpretable knowledge distillation paradigm for credit risk assessment

Abstract

Credit risk modeling in regulated financial environments confronts an inherent tension between predictive accuracy and decision transparency. This paper proposes CB-KD (Class-Balanced Knowledge Distillation), a model compression framework that transfers predictive knowledge from high-capacity ensemble or neural network teachers to structurally transparent decision tree students. The framework advances credit risk methodology along three dimensions. First, CB-KD establishes a heterogeneous distillation architecture that extends knowledge distillation beyond homogeneous neural-to-neural transfer to the ensemble-to-tree paradigm, incorporating automatic teacher selection and a dual-component sample weighting strategy that integrates adaptive class-balanced weighting with teacher confidence weighting. Second, the temperature-scaled soft-label transfer mechanism preserves inter-class probability structure that hard labels discard, providing a principled approach to maintaining predictive fidelity while achieving structural transparency for regulatory examination. Third, the lift-based rule extraction mechanism formalizes decision tree paths as auditable IF–THEN rules with statistical quality metrics, enabling rigorous assessment of rule reliability under class-imbalanced conditions. Empirical validation across multiple credit datasets demonstrates that CB-KD consistently outperforms baseline distillation methods while producing fully interpretable decision logic amenable to regulatory audit.

Keywords: Credit Risk Assessment, Interpretability, Knowledge Distillation, Class Balance

1 Introduction

Credit risk assessment constitutes a cornerstone of financial intermediation, underpinning lending decisions, capital adequacy requirements, and portfolio risk management across banking institutions globally (Altman and Saunders 1998; Thomas et al. 2002; Sun et al. 2023). The accurate estimation of default probabilities necessitates sophisticated modeling of borrower characteristics, repayment histories, and macroeconomic conditions, reflecting the multidimensional nature of creditworthiness evaluation. Recent advances in machine learning—particularly gradient boosting ensembles (XGBoost, LightGBM, CatBoost) and deep neural architectures—have

substantially enhanced predictive accuracy beyond traditional statistical approaches (Luo et al. 2017; Shen et al. 2021; Beltman et al. 2025). However, the inherent complexity of these "black-box" models creates a fundamental tension between predictive power and decision transparency: their opacity impedes internal model validation, regulatory examination, and transparent communication of credit decisions to borrowers and stakeholders (Fan et al. 2025). This tension is particularly acute in regulated financial environments where model auditability is not merely desirable but legally mandated.

International regulatory frameworks, notably the Basel II/III Accords, impose stringent requirements on model governance, emphasizing point-in-time expected credit loss estimation and model auditability (Dumitrescu et al. 2022; Zhou et al. 2025; Zhang and Yu 2024). The Internal Ratings-Based (IRB) approach under Basel II requires banks to demonstrate that their credit models are conceptually sound, empirically validated, and regularly monitored—criteria that inherently favor transparent model structures. Furthermore, emerging AI governance initiatives, including the EU's proposed AI Act and various national "right to explanation" mandates, increasingly require that automated credit decisions be justifiable to affected individuals. While these regulations do not prescribe specific interpretability metrics, the implicit constraint on model opacity creates a fundamental tension between predictive power and regulatory compliance. Consequently, developing methodologies that can reconcile predictive accuracy with decision transparency represents a central challenge in credit risk management and financial model governance.

Intrinsically interpretable models—including logistic regression with L1/L2 regularization—address multicollinearity through parameter penalization and facilitate variable selection, thereby providing quantitative insights into feature importance (Bradley and Mangasarian 1998; Chen et al. 2019; He et al. 2023). Nevertheless, their linear functional form fundamentally constrains the capacity to capture nonlinear feature interactions prevalent in credit data. As data complexity increases, nonlinear ensemble methods such as Gradient Boosting Decision Trees (GBDT), eXtreme Gradient Boosting (XGBoost), and Random Forests (RF) have emerged as dominant approaches (Dastile et al. 2020; Gunnarsson et al. 2021; Shwartz-Ziv and Armon 2022). These models achieve superior predictive performance through weighted aggregation of decision trees and assess feature importance via split gains or Gini impurity. However, their interpretability remains contingent on model architecture and data characteristics, exhibiting sensitivity to feature heterogeneity and inconsistency across different model configurations.

Explainable Artificial Intelligence (XAI) methods have advanced the interpretability analysis of complex models. SHapley Additive exPlanations (SHAP), grounded in cooperative game theory, quantifies each feature's marginal contribution to individual predictions (Lundberg and Lee 2017), enabling both local instance-level and global model-level explanations (Bussmann et al. 2019; Fan et al. 2025). Despite these advances, SHAP remains fundamentally a post-hoc explanatory tool—it elucidates trained models retrospectively rather than generating structured decision logic during the model fitting process. The integration of SHAP-derived insights into a training

objective that yields explicit, auditable decision rules constitutes an open research challenge.

Knowledge distillation (KD) provides a principled mechanism for transferring predictive behavior from high-capacity teacher models to simpler, more interpretable student models (Hinton et al. 2015). This paper proposes CB-KD (Class-Balanced Knowledge Distillation), an interpretable credit scoring paradigm that addresses the accuracy-interpretability trade-off through temperature-scaled soft-label distillation and lift-based rule extraction.

While individual components (temperature scaling, inverse-frequency weighting, lift-based ranking) exist in prior literature, CB-KD integrates them into a unified end-to-end framework specifically designed for credit risk applications. The key differentiators from existing tree-distillation and rule-extraction methods are: (i) *heterogeneous teacher support*—unlike neural-to-neural distillation, CB-KD accommodates both ensemble and neural teachers via a unified soft-label interface; (ii) *automatic teacher selection*—the framework eliminates manual model comparison by selecting the best-performing teacher based on validation AUC; (iii) *dual-component sample weighting*—beyond standard inverse-frequency weighting, CB-KD combines adaptive class-balanced weighting with teacher confidence weighting, ensuring faithful knowledge transfer while addressing class imbalance; and (iv) *statistical rule validation*—each extracted rule is accompanied by lift, confidence, and Benjamini–Hochberg corrected p-values, enabling rigorous quality assessment.

The principal contributions are threefold:

- **Heterogeneous distillation with dual-component sample weighting.** This paper extends knowledge distillation from homogeneous neural-to-neural settings to heterogeneous ensemble-to-tree transfer, establishing a unified soft-label interface that accommodates diverse teacher architectures. The dual-component sample weighting strategy integrates adaptive class-balanced weighting with teacher confidence weighting, ensuring faithful knowledge transfer while addressing class imbalance inherent in credit risk portfolios.
- **Bridging model compression and regulatory transparency.** The distilled decision tree retains full structural transparency, where each prediction traces to a unique root-to-leaf path with explicit feature thresholds. This architecture provides a principled mechanism for reconciling predictive performance with regulatory auditability requirements, demonstrating that model interpretability and discriminative power are not mutually exclusive.
- **Lift-based rule extraction with statistical validation.** Each decision tree path is formalized as an auditable IF–THEN rule with confidence, lift, and multiple-testing-corrected significance metrics. The lift normalization adjusts for class priors, enabling fair comparison between majority-class and minority-class rules under imbalanced class distributions.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the proposed CB-KD paradigm, including the automatic teacher selection, class-balanced decision tree distillation, and rule extraction. To verify and compare the validity of the proposed framework, four credit datasets

are employed, and the empirical experiments are presented in Section 4 and Section 5. Finally, the key findings and future research directions are discussed in Section 6.

2 Literature review

2.1 Interpretability method in credit risk assessment

With the implementation of international regulatory frameworks, such as Basel II/III, financial regulators have increasingly emphasized compliance requirements regarding the transparency and auditability of credit risk assessment models. In this context, model interpretability has emerged as a central research agenda in credit risk modeling. Current approaches to model interpretability are generally categorized into two paradigms: intrinsically interpretable models and model-agnostic explanation methods (Du et al. 2019).

Intrinsically interpretable models achieve decision transparency by simplifying model architecture and enhancing parameter interpretability. Typical examples include logistic regression, rule-based model, and decision trees (Baesens et al. 2003; Gorzalczany and Rudziński 2016; Hayashi 2016). Among these, logistic regression has long been regarded by regulators as the benchmark for compliant and interpretable credit scoring due to its parameters directly quantifying the marginal effects of borrower characteristics on default probability (Du et al. 2019). However, its linear assumptions limit its ability to capture complex nonlinear relationships and feature interactions.

Rule extraction methods translate internal decision logic into comprehensible "IF–THEN" rules, enabling model knowledge to be operationalized in business contexts. For instance, Baesens et al. (2003) validated the feasibility of neural network rule extraction techniques on real-world credit risk datasets, and transformed the extracted rules into decision tables that are easier for humans to understand and consult. Martens et al. (2007) used decomposition and pedagogical rule extraction techniques to enhance the interpretability of the support vector machine (SVM) classifier while maintaining its accuracy. In addition, decision tree models (DTs), as another intuitive interpretability approach, offer hierarchical "IF–THEN" decision paths, which also effectively capture the non-linear relationships and interaction effects among variables (Sagi and Rokach 2020; Aguilar et al. 2022; Xu and Yang 2025). Recently, Zhu et al. (2025) proposed a hybrid feature selection framework based on an improved minimum spanning tree algorithm, integrating random forest (RF), extreme gradient boosting (XGBoost), and AdaBoost to mitigate feature redundancy and enhance ranking efficiency. Despite these advances, Zhang et al. (2025) pointed out that tree-based ensemble model is considered as a "black-box", with opaque internal mechanisms that lack intuitive interpretability. As a result, these models fail to meet the dual requirements of regulatory auditability and model transparency, posing practical challenges for deployment in regulatory-compliant settings.

In contrast to intrinsically interpretable models, model-agnostic explanation methods do not rely on the model structure but instead reveal the decision mechanisms of models through post-hoc analysis. SHapley Additive exPlanations (SHAP) has

emerged as one of the most widely used explainable modeling techniques (Lundberg and Lee 2017; Hassija et al. 2024). SHAP, based on cooperative game theory, treats each input feature as a "player" and quantifies its marginal contribution to the model's prediction by calculating its impact across all possible feature combinations. Theoretically, SHAP ensures mathematically robust and reproducible explanations, which renders it widely applicable to complex models such as XGBoost, Light Gradient Boosting Machine (LightGBM), and deep neural networks (Feng et al. 2021; Bussmann et al. 2019). Empirical studies confirm that SHAP effectively identifies key drivers of default risk, including debt-to-income ratio, credit history, and income level (Lei et al. 2024; Wang et al. 2024; Zhou et al. 2025; Bücker et al. 2022). Bücker et al. (2022) further developed a unified SHAP-based framework to quantify individual contributions to a client's credit score, thereby enhancing the transparency and credibility of model outputs.

However, interpretability is not invariably beneficial. As noted by Samek et al. (2021), overreliance on post-hoc explanations may lead to misinterpretation of model logic or overestimation of feature importance, potentially undermining predictive performance and generalization capability. This reveals a trade-off between interpretability and predictive accuracy. Therefore, balancing the predictive advantages of "black-box" models with the development of an interpretable, regulation-compliant credit risk evaluation system has become a key challenge within the frameworks of model risk management and regulatory compliance.

2.2 Knowledge distillation in knowledge transfer

Knowledge distillation (KD) is an efficient model compression and knowledge transfer technique that transfers the predictive behavior embedded in a complex teacher model to a simpler student model through "soft labels". This mechanism can reduce computational complexity while preserving accuracy and improving generalization, particularly under small-sample constraints (Zhou et al. 2024b; Wu et al. 2025; Lin et al. 2025). Depending on the nature of transferred information, KD can be broadly categorized into logit-based and feature-based distillation approaches (Gao et al. 2025; Lan et al. 2025).

Logit-based distillation aligns the output probability distributions (logits) of teacher and student models to facilitate behavioral imitation (Hinton et al. 2015; Gou et al. 2023; Sun et al. 2024). Hinton et al. (2015) introduced the temperature-scaled softmax to generate smooth soft labels, allowing the student model to capture the latent information structure within the teacher's predictions. The distillation process minimizes the Kullback–Leibler (KL) divergence between their logit spaces, ensuring effective knowledge transfer. Building upon this, Gou et al. (2023) incorporated multilayer regularization to balance model simplicity with inference capability. Compared to other distillation methods, logit-based KD is computationally efficient and requires minimal dimensional alignment, making it highly scalable for large-scale financial applications.

In contrast, feature-based distillation focuses on transferring intermediate feature representations or activation maps from the teacher model (Romero et al. 2014; Wang et al. 2024; Zhou et al. 2024a).The seminal FitNet framework (Romero et al.

2014) introduced feature alignment between the hidden layers of teacher and student networks, narrowing the performance gap between deep and shallow models. Zhou et al. (2024a) extended this idea by incorporating multimodal feature learning, thereby enhancing cross-domain adaptability. While feature-based KD can capture richer semantic and spatial information, it typically demands greater computational resources and poses challenges regarding feature dimension matching (Gao et al., 2021). Furthermore, in regulated financial contexts, data confidentiality and model privacy constraints often prevent direct access to teacher model internals, limiting the feasibility of such methods in practice.

Since its introduction by Hinton et al. (2015), KD has been widely adopted across computer vision, natural language processing, and recommender systems (Hinton et al. 2015; Huang et al. 2023). However, its application to financial risk modeling remains underexplored. The primary challenge lies in the lack of interpretability, as logit-based distillation depends on abstract features that obscure decision logic (Sun et al. 2023). Moreover, performance degradation can occur when there is a large capacity gap between teacher and student models (Gou et al. 2021).

Table 1 summarizes the key differences between existing knowledge distillation methods and the proposed CB-KD framework. The focus is not only compression but *audit-ready interpretability*: (i) the student is constrained to a *single* decision tree so that the final output is an explicit rule set; (ii) teacher selection is automatic and data-driven (validation AUC) to avoid manual cherry-picking; and (iii) class-balanced sample weighting addresses the inherent class imbalance in credit data, ensuring that minority-class samples receive appropriate emphasis during knowledge transfer.

Table 1: Comparison of knowledge distillation methods

Method	**Interp.**	**Domain**	**Teacher→Student**
Vanilla KD (Hinton et al. 2015)	×	General	NN→NN
FitNets (Romero et al. 2014)	×	Vision	Deep NN→Thin NN
Attention Transfer (Zagoruyko and Komodakis 2017)	×	Vision	NN→NN
Born-Again Networks (Furlanello et al. 2018)	×	General	NN→NN
Tree-to-Tree KD (Wang et al. 2025)	✓	Finance	RF/GBDT→DT
CB-KD (Ours)	✓	Finance	Ensemble/NN→DT

Consequently, enhancing the interpretability and efficiency of knowledge distillation represents a crucial direction for future research in credit risk modeling and model risk management.

3 Methodology

This section presents an interpretable credit risk assessment paradigm based on knowledge distillation, aiming to combine strong predictive performance with the auditability of an intrinsically interpretable model. CB-KD consists of three stages: (i) *automatic teacher selection* chooses the strongest teacher from an ensemble/NN candidate pool using validation AUC; (ii) *class-balanced soft-label distillation* transfers the teacher's probabilistic behavior to a single decision tree student using temperature scaling, inverse-frequency class weighting, and an implementation-compatible soft-target injection strategy; and (iii) *lift-based rule extraction* converts the trained decision tree into auditable IF–THEN rules with post-hoc feature importance analysis.

3.1 Knowledge distillation with automatic teacher selection

Knowledge distillation (KD) is adopted to transfer predictive behavior from a high-capacity teacher model to an interpretable student model. A core component of CB-KD is automatic teacher selection: the framework evaluates multiple Optuna-optimized candidate models—including XGBoost, LightGBM, CatBoost, Random Forest, GBDT, and a neural baseline (CreditNet)—on a held-out validation split, and selects the model achieving the highest AUC as the teacher (Asencios et al. 2023; Rao et al. 2023). This selection strategy adapts to dataset characteristics without manual intervention and ensures that distillation starts from the strongest available teacher among the candidate set.

The student model employs a Decision Tree classifier, which simplifies the model structure to facilitate deployment, improve interpretability, and enable explicit rule extraction. Let the dataset be denoted as $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^n$ represents the feature vector of borrower i and $y_i \in \{0, 1\}$ is the default label. The knowledge distillation procedure is as follows:

Soft-label generation

In this implementation, the student is a decision tree (CART), which does not optimize a differentiable KD loss. Teacher knowledge is therefore distilled through temperature-scaled *soft probabilities* and an implementation-compatible training construction. For binary credit default prediction, teacher outputs are converted to logits z_i and temperature-scaled soft probabilities are generated via a sigmoid:

$$p_i^{\tau} = \sigma\Big(\frac{z_i}{\tau}\Big) = \frac{1}{1 + \exp\left(-z_i/\tau\right)} \tag{1}$$

where $\tau > 0$ is the temperature. For tree-based teachers, z_i can be obtained from predicted probabilities by $z_i = \log \frac{p_i}{1-p_i}$ (with numerical clipping).

To balance soft transfer and hard-label supervision, a mixed target probability is defined as

$$p_i^{\text{mix}} = (1 - \alpha)\, p_i^{\tau} + \alpha\, y_i, \quad \alpha \in [0, 1]. \tag{2}$$

Here, α is the hard-label weight: $\alpha = 0$ corresponds to pure soft-label distillation, while $\alpha = 1$ reduces to standard supervised learning on y_i. The rationale for α is as follows: when $\alpha = 0$, the student learns exclusively from the teacher's probability structure,

which encodes class relationships and uncertainty that hard labels discard (Hinton et al. 2015). However, if the teacher is miscalibrated or exhibits systematic bias, incorporating ground-truth supervision ($\alpha > 0$) can regularize the student. Empirically, $\alpha \in [0, 0.3]$ balances knowledge transfer with label fidelity (see ablation in Section 5.4).

Dual-component sample weighting: class balance and teacher confidence

Credit risk datasets typically exhibit class imbalance, with default samples constituting a minority. To address this challenge while preserving teacher knowledge, CB-KD employs a dual-component sample weighting strategy during the distillation process. This approach combines class-balanced weighting with teacher confidence weighting, ensuring that the student tree learns from the most informative samples while maintaining balanced class representation.

Component 1: Adaptive class-balanced weighting. For a training set with N samples, let n_0 and n_1 denote the number of samples in class 0 (non-default) and class 1 (default), respectively. The base inverse-frequency weight is $w_c^{\text{raw}} = N/(2 \cdot n_c)$. To prevent over-correction that may disrupt teacher knowledge transfer, we apply adaptive smoothing based on the imbalance ratio $r = \max(n_0, n_1)/\min(n_0, n_1)$:

$$w_c^{\text{class}} = \begin{cases} 1 + \beta(w_c^{\text{raw}} - 1), & \text{if } r < 3 \\ 1 + 0.5 \log(1 + w_c^{\text{raw}} - 1), & \text{if } r \geq 3 \end{cases} \tag{3}$$

where $\beta \in [0.2, 0.5]$ is a smoothing factor that increases with imbalance severity. This logarithmic dampening prevents excessive minority-class emphasis that could override the teacher's probability structure.

Component 2: Teacher confidence weighting. To prioritize samples where the teacher provides confident predictions, we introduce a confidence-based weight:

$$w_i^{\text{conf}} = 0.8 + 0.4 \cdot |p_i^{\mathfrak{T}} - 0.5| \cdot 2 \tag{4}$$

where $p_i^{\mathfrak{T}}$ is the teacher's predicted probability for sample i. Samples with teacher predictions near 0 or 1 (high confidence) receive higher weights, while uncertain predictions (near 0.5) receive lower weights. This encourages the student to faithfully replicate the teacher's confident decisions.

The final sample weight combines both components: $w_i = w_{y_i}^{\text{class}} \cdot w_i^{\text{conf}}$, normalized to mean 1. This dual-component strategy balances class representation while preserving teacher knowledge, leading to improved discriminative performance on imbalanced credit portfolios.

Since CART does not directly optimize a KD divergence on soft targets, p_i^{mix} is injected into decision-tree training using a weighted sample construction: for each original sample $\mathbf{x}_i$, two duplicated instances $(\mathbf{x}_i, 0)$ and $(\mathbf{x}_i, 1)$ are created with weights $w_i(1 - p_i^{\text{mix}})$ and $w_i p_i^{\text{mix}}$, respectively. The student tree is trained on this augmented weighted dataset via `sample_weight`. Note that this construction approximates the

soft-label objective in expectation; however, since CART uses greedy splitting to minimize local impurity rather than global cross-entropy, it does not guarantee exact minimization of the distillation loss (5).

Validation-based decision threshold tuning

To mitigate threshold sensitivity under class imbalance, after training the student tree, a decision threshold $t \in [0, 1]$ is tuned on the validation set by grid-searching $t \in \{0.05, 0.10, \ldots, 0.95\}$ and selecting the value that maximizes validation F1. The tuned threshold is used for reporting discrete metrics (ACC/Prec./Recall/F1), while AUC is computed from the predicted probabilities.

Approximate distillation loss

The weighted soft-target injection described above approximates minimizing the following cross-entropy loss in expectation:

$$\mathcal{L}_{\mathrm{KD}} = -\sum_{i=1}^{N} w_i \left[p_i^{\mathrm{mix}} \log q_i + (1 - p_i^{\mathrm{mix}}) \log(1 - q_i) \right] \tag{5}$$

where $q_i = P(\hat{y}_i = 1 \mid \mathbf{x}_i; \mathfrak{S})$ is the student tree's predicted probability. When $\alpha > 0$, the overall objective can be decomposed as:

$$\mathcal{L} = (1 - \alpha)\, \mathcal{L}_{\mathrm{soft}} + \alpha\, \mathcal{L}_{\mathrm{hard}} \tag{6}$$

where $\mathcal{L}_{\mathrm{soft}}$ denotes the soft-label distillation loss (using teacher probabilities $p_i^{\mathfrak{T}}$) and $\mathcal{L}_{\mathrm{hard}}$ denotes the hard-label supervised loss (using ground-truth y_i). The hyperparameter α controls the trade-off between knowledge transfer and ground-truth supervision. In our implementation, we set $\alpha = 0$ (pure soft-label distillation) with temperature $\tau = 4.0$ as the default configuration. This design choice prioritizes pure knowledge transfer from the teacher; ablation studies in Section 5.4 examine the sensitivity to these hyperparameters across datasets.

Algorithm 1 formalizes the complete CB-KD framework including automatic teacher selection, temperature-scaled soft-label distillation, and rule extraction.

Algorithm 1: CB-KD: Temperature-Scaled Soft-Label Knowledge Distillation

Input : Training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, validation set $\mathcal{D}_{\text{val}}$, candidate teacher models $\mathcal{M}$, temperature τ, mixing coefficient α, tree depth d

Output : Student decision tree $\mathfrak{S}$, decision rules $\mathcal{R}$

`// Automatic Teacher Selection`

for $m \in \mathcal{M}$ **do**

  Train m on $\mathcal{D}_{\text{train}}$; evaluate $\text{AUC}_m \leftarrow \text{AUC}(m, \mathcal{D}_{\text{val}})$

end

$\mathfrak{T} \leftarrow \arg\max_{m \in \mathcal{M}} \text{AUC}_m$

`// Soft Label Generation`

for $i = 1, \ldots, N$ **do**

  $p_i^{\mathfrak{T}} \leftarrow \mathfrak{T}(\mathbf{x}_i)$; $\quad z_i \leftarrow \log(p_i^{\mathfrak{T}}/(1 - p_i^{\mathfrak{T}}))$

  $p_i^{\text{soft}} \leftarrow \sigma(z_i/\tau)$; $\quad p_i^{\text{mix}} \leftarrow (1-\alpha)p_i^{\text{soft}} + \alpha y_i$

end

`// Dual-Component Sample Weighting`

$n_c \leftarrow |\{i : y_i = c\}|$ for $c \in \{0, 1\}$; $\quad r \leftarrow \max(n_0, n_1)/\min(n_0, n_1)$

for $c \in \{0, 1\}$ **do**

  $w_c^{\text{raw}} \leftarrow N/(2 \cdot n_c)$; $\quad w_c^{\text{class}} \leftarrow 1 + \beta \cdot \log(1 + w_c^{\text{raw}} - 1)$

end

for $i = 1, \ldots, N$ **do**

  $w_i^{\text{conf}} \leftarrow 0.8 + 0.4 \cdot |p_i^{\mathfrak{T}} - 0.5| \cdot 2$; $\quad w_i \leftarrow w_{y_i}^{\text{class}} \cdot w_i^{\text{conf}}$

end

`// Soft-Target Tree Training`

$\mathcal{D}_{\text{aug}} \leftarrow \emptyset$

for $i = 1, \ldots, N$ **do**

  Add $(\mathbf{x}_i, 0)$ with weight $w_i(1 - p_i^{\text{mix}})$ and $(\mathbf{x}_i, 1)$ with weight $w_i \cdot p_i^{\text{mix}}$ to $\mathcal{D}_{\text{aug}}$

end

Train CART $\mathfrak{S}$ on $\mathcal{D}_{\text{aug}}$ with max_depth $= d$

`// Rule Extraction`

$\mathcal{R} \leftarrow \text{ExtractRules}(\mathfrak{S})$

return $\mathfrak{S}$, $\mathcal{R}$

3.2 Rule extraction

Among interpretable models, decision trees effectively capture nonlinear relationships and interaction effects between variables. Their hierarchical "IF-THEN" rule structure not only mirrors the complex patterns inherent in credit risk data but also provides transparent and traceable decision logic. After class-balanced distillation (Section 3.1), each root-to-leaf path π_ℓ in the student decision tree $\mathfrak{S}$ is converted into an interpretable rule.

Path-dependent rule formulation. Let $\mathcal{V}(\pi_\ell) = \{v_1, v_2, \ldots, v_h\}$ denote the sequence of internal nodes along path π_ℓ from root to leaf ℓ, and let each node v_k impose a split condition c_k of the form $f_{j_k} \leq \theta_k$ or $f_{j_k} > \theta_k$. The rule corresponding to path π_ℓ is formalized as:

$$r_\ell : \text{IF} \bigwedge_{k=1}^{h} c_k \text{ THEN } \hat{y} = \text{class}_\ell \tag{7}$$

where $\bigwedge$ denotes logical conjunction (AND). The predicted class $\text{class}_\ell \in \{0, 1\}$ is determined by majority voting among training samples reaching leaf ℓ, and the rule confidence is computed as:

$$\text{conf}_\ell = \frac{\max(n_0^\ell, n_1^\ell)}{n_0^\ell + n_1^\ell} \tag{8}$$

where n_0^ℓ and n_1^ℓ denote the number of class-0 and class-1 training samples in leaf ℓ, respectively. Rules with higher confidence indicate more homogeneous leaves and stronger predictive certainty.

However, under class imbalance—common in credit risk applications where defaults are rare—simple confidence ranking exhibits a systematic bias: rules predicting the majority class (non-default) inherently achieve higher confidence than minority-class rules. To address this limitation, we adopt *lift*-based ranking, a well-established metric from association rule mining (Agrawal et al. 1993).

The *lift* of a rule measures its predictive improvement over random assignment:

$$\text{Lift}_\ell = \frac{\text{conf}_\ell}{\pi_{c_\ell}} \tag{9}$$

where π_{c_ℓ} is the prior probability of the predicted class c_ℓ. A lift greater than one indicates that the rule identifies its target class more effectively than random chance. For credit risk, a default rule with Lift $= 3$ means applicants satisfying that rule are three times more likely to default than a randomly selected applicant. Unlike confidence, lift is normalized against class frequency, enabling fair comparison between majority-class and minority-class rules.

To assess the statistical reliability of each rule, we employ a one-sided binomial test:

$$H_0 : \text{conf}_\ell = \pi_{c_\ell} \quad \text{vs.} \quad H_1 : \text{conf}_\ell > \pi_{c_\ell} \tag{10}$$

Given that a decision tree with L leaves involves L simultaneous hypothesis tests, we apply the Benjamini–Hochberg (BH) procedure to control the false discovery rate (FDR) at level 0.05. Rules with BH-adjusted $p < 0.05$ are deemed statistically significant, indicating their predictive power is unlikely to arise from random chance. This multiple comparison correction ensures that the proportion of false discoveries among significant rules is controlled, which is critical when extracting dozens of rules from a single tree.

Algorithmic formalization. Algorithm 1 presents the complete CB-KD framework including automatic teacher selection and temperature-scaled soft-label distillation. Algorithm 2 formalizes the path-dependent rule extraction mechanism.

Algorithm 2: Path-Dependent Rule Extraction with Lift-Based Ranking

Input : Distilled decision tree $\mathfrak{S}$; Feature names $\{f_1, \ldots, f_n\}$; Class priors $\{\pi_0, \pi_1\}$

Output : Rule set $\mathcal{R} = \{(r_\ell, \text{conf}_\ell, \text{Lift}_\ell, \text{class}_\ell)\}$

Initialize rule set: $\mathcal{R} \leftarrow \emptyset$

foreach *leaf node* ℓ *in* $\mathfrak{S}$ **do**

  Initialize rule conditions: $\mathcal{C}_\ell \leftarrow \emptyset$

  Traverse path from root to leaf ℓ

  foreach *internal node* v *on path* **do**

    Extract split condition: $(f_j, \theta_v, \text{direction})$

    if *direction = left* **then**

      $\mathcal{C}_\ell \leftarrow \mathcal{C}_\ell \cup \{f_j \leq \theta_v\}$

    else

      $\mathcal{C}_\ell \leftarrow \mathcal{C}_\ell \cup \{f_j > \theta_v\}$

    end

  end

  Form rule: $r_\ell \leftarrow \bigwedge_{c \in \mathcal{C}_\ell} c$

  Compute confidence: $\text{conf}_\ell \leftarrow \max(n_0^\ell, n_1^\ell)/(n_0^\ell + n_1^\ell)$

  Determine class: $\text{class}_\ell \leftarrow \arg\max(n_0^\ell, n_1^\ell)$

  Compute lift: $\text{Lift}_\ell \leftarrow \text{conf}_\ell / \pi_{\text{class}_\ell}$

  Compute support: $\text{Support}_\ell \leftarrow n_\ell / N$

  Test significance: $p_\ell \leftarrow \text{BinomTest}(n_\ell \cdot \text{conf}_\ell, n_\ell, \pi_{\text{class}_\ell})$

  $\mathcal{R} \leftarrow \mathcal{R} \cup \{(r_\ell, \text{conf}_\ell, \text{Lift}_\ell, \text{Support}_\ell, p_\ell, \text{class}_\ell)\}$

end

Partition: $\mathcal{R}^{(+)} \leftarrow \{r \in \mathcal{R} : \text{class} = 1\}$; $\mathcal{R}^{(-)} \leftarrow \{r \in \mathcal{R} : \text{class} = 0\}$

Sort each partition by Lift in descending order

return $\mathcal{R}$

4 Experimental design

4.1 Implementation environment

All experiments were conducted on a high-performance computing server equipped with an Intel 8336C CPU, NVIDIA GeForce RTX 4090 GPU (24GB), and 256GB RAM. The implementation was based on Python 3.13.5, utilizing scikit-learn 1.6.1, XGBoost 2.1.3, LightGBM 4.5.0, CatBoost 1.2.7, PyTorch 2.5.1, and Optuna 4.2.0. All experiments used a fixed random seed of 42 for reproducibility.

Hyperparameter optimization. Boosting models (XGBoost, LightGBM, CatBoost) were optimized using Optuna's TPE (Tree-structured Parzen Estimator) algorithm with 50 trials per model. The search spaces were: learning rate $\in [0.01, 0.3]$, max depth $\in \{3, 4, 5, 6, 7, 8\}$, number of estimators $\in [50, 500]$, subsample ratio $\in [0.6, 1.0]$, and regularization parameters (L1/L2) $\in [10^{-3}, 10]$. Early stopping with 50 rounds of patience was applied using validation loss.

Feature preprocessing. All features were standardized using z-score normalization ($\mu = 0, \sigma = 1$) based on training set statistics to prevent data leakage. For categorical features (present in German and UCI datasets), one-hot encoding was applied. Missing values, if any, were imputed using median (numerical) or mode (categorical) from the training set.

Evaluation protocol. For baseline comparisons (Table 3), we report mean±std over five independent runs of five-fold stratified cross-validation (25 total folds). For distillation experiments (Table 4), a fixed 60/20/20 train/validation/test split with five independent runs is used. The validation set serves for teacher selection, threshold tuning, and early stopping. To ensure fair comparison, all models (teacher, baseline, and distillation methods) use validation-tuned thresholds for discrete metrics.

4.2 Dataset descriptions

To validate the universality and robustness of the proposed framework across different scales, dimensionalities, and domain characteristics, four credit risk benchmark datasets are employed. Three are publicly available from the UCI Machine Learning Repository: German Credit Data, Australian Credit Approval, and Default of Credit Card Clients (Taiwan). Additionally, a proprietary Chinese consumer credit dataset from Xinwang Bank evaluates applicability in emerging market contexts. Xinwang Bank is China's third internet-based bank, the first internet bank in central and western China, and the first private bank in Sichuan Province.

Table 2: Summary of credit risk datasets

Dataset	Samples	Features	Default Rate	Source
German Credit	1,000	20	30.0%	UCI Repository
Australian Credit	690	14	44.5%	UCI Repository
UCI Credit Card	30,000	23	22.1%	UCI Repository
Xinwang Bank Credit	17,886	100	10.2%	Xinwang Bank

The **German Credit Dataset** (1,000 samples, 20 features) includes demographic, financial, and loan-related attributes with moderate class imbalance (default rate approximately 30%). The **Australian Credit Approval** dataset (690 samples, 14 features) contains anonymized credit application attributes with relatively balanced class distribution. The **UCI Credit Card** dataset (30,000 samples, 23 features) comprises Taiwan credit card holder information including payment history and bill statements, representing the largest public benchmark with a notable class imbalance (default rate approximately 22%). The **Xinwang Bank Credit** dataset (17,886 samples, 100 features) is a Chinese consumer lending dataset containing rich behavioral features across multiple dimensions including basic demographics, historical loan performance, overdue records, and credit inquiry patterns. It is highly imbalanced with

a default rate of about 10.2% (1,819 defaults vs. 16,065 non-defaults), which makes threshold-dependent metrics more volatile and motivates AUC-focused evaluation.

For all datasets, a consistent data partitioning strategy is adopted: 60% for training, 20% for validation (used for early stopping and hyperparameter tuning), and 20% for testing. All features are standardized using z-score normalization based on training set statistics to prevent data leakage. The publicly available datasets ensure reproducibility while the Xinwang Bank dataset illustrates applicability to real-world Chinese consumer credit scenarios.

4.3 Benchmark models and evaluation criteria

To comprehensively validate the proposed CB-KD framework, a diverse set of baseline models spanning multiple algorithmic paradigms is employed, all optimized using Optuna's Bayesian hyperparameter tuning with 50 trials per model. The baseline models are organized into four categories:

Linear Models: Logistic Regression with Lasso regularization (LR-Lasso) and Ridge regularization (LR-Ridge), representing interpretable linear baselines commonly employed in regulatory-compliant credit scoring.

Kernel Methods: Support Vector Machine with RBF kernel (SVM-RBF), capturing nonlinear decision boundaries through kernel transformation.

Instance-Based Methods: K-Nearest Neighbors (KNN), a non-parametric classifier that assigns labels based on the majority vote among the k closest training samples. KNN serves as a simple baseline but tends to underperform on high-dimensional or imbalanced credit data.

Tree-Based Models: Decision Tree (DT) as the base interpretable model, Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) as traditional ensemble methods.

Gradient Boosting Frameworks: XGBoost, LightGBM, and CatBoost, representing widely used gradient boosting implementations.

Neural Models: CreditNet, a compact feed-forward neural network baseline trained on the same train/validation/test protocol. CreditNet is included both as a baseline and as a candidate teacher in the automatic teacher selection step.

The teacher model for CB-KD is automatically selected as the best-performing model among all trained baselines based on validation AUC score. This data-driven selection ensures that distillation uses the most accurate available predictions for each dataset.

Model performance is evaluated using seven metrics: Accuracy measures overall classification correctness; Precision quantifies the proportion of true positives among predicted positives; Recall captures the proportion of actual defaults correctly identified; F1-score provides the harmonic mean of precision and recall; AUC (Area Under ROC Curve) evaluates ranking quality across all classification thresholds; PR-AUC (Area Under Precision-Recall Curve) provides a more informative measure for imbalanced datasets by focusing on the minority class (Davis and Goadrich 2006); and Brier score measures probability calibration quality (lower is better), defined as $\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$ where p_i is the predicted probability and $y_i \in \{0, 1\}$ is the true label.

5 Results and discussion

5.1 Baseline model comparison

To benchmark the proposed framework, baseline model performance across four datasets is reported. Table 3 reports AUC (mean±std) from five-fold cross-validation with five independent runs. Note that this cross-validation protocol differs from the train/validation/test split used in Table 4; thus, direct comparison of teacher AUC values across these two tables is not appropriate.

Table 3: Baseline model performance comparison (AUC)

Category	**Model**	**German**	**Australian**	**UCI**	**Xinwang**
Linear	LR-Ridge	0.7871	0.9276	0.7242	0.7085
	LR-Lasso	0.7904	0.9283	0.7242	0.7103
	LR-ElasticNet	0.7898	0.9274	0.7242	0.7097
Kernel	SVM-RBF	0.7652	0.9189	0.7213	0.6192
	SVM-Linear	0.7662	0.9040	0.7008	0.5762
Instance	KNN	0.6770	0.9065	0.6879	0.5636
Probabilistic	Naive Bayes	0.7277	0.9095	0.7337	0.5380
Tree	DT	0.6574	0.7955	0.7213	0.5870
Ensemble	RF-Tuned	0.7961	0.9244	**0.7885**	0.7341
	GBDT-Tuned	0.8038	0.9259	0.7861	0.7414
	XGBoost-Tuned	**0.8085**	0.9305	0.7876	**0.7486**
	LightGBM-Tuned	0.7493	0.9319	0.7870	0.7401
	CatBoost-Tuned	0.7519	0.9251	0.7843	0.7421
Neural	CreditNet	0.7973	**0.9327**	0.7777	0.7160

Table 3 reveals a clear performance hierarchy. Gradient boosting ensembles and neural networks dominate across all four datasets, with XGBoost-Tuned leading on German (AUC 0.8085) and Xinwang (0.7486), CreditNet on Australian (0.9327), and RF-Tuned on UCI (0.7885). Linear models (LR variants) perform competitively on Australian and UCI but lag behind on Xinwang, likely due to nonlinear feature interactions in these datasets. KNN underperforms substantially, particularly on Xinwang (0.5636), reflecting its sensitivity to high dimensionality and class imbalance. The standalone Decision Tree (DT) achieves moderate AUC but trails ensemble methods by a margin of 0.10–0.16, quantifying the accuracy gap that knowledge distillation aims to bridge. CreditNet, the neural baseline, achieves the best performance on Australian (0.9327) and is competitive elsewhere. Because no single model dominates all datasets, these results underscore the value of data-driven teacher selection in CB-KD.

5.2 CB-KD performance evaluation

This section evaluates CB-KD against baseline distillation methods. Table 4 reports the teacher model, an undistilled student baseline, and three distillation variants across four datasets.

Table 4: Performance comparison of knowledge distillation methods

Dataset	Method	AUC	PR-AUC	Brier↓	ACC	F1
German	Teacher (XGBoost)	0.7517	0.8877	0.1779	0.7200	0.8170
	Student Baseline	0.6393	0.7999	0.2538	0.6950	0.8168
	VanillaKD	0.7393	0.8690	0.2107	0.7150	**0.8119**
	SoftLabelKD	0.7164	0.8564	**0.2002**	0.6950	0.8051
	CB-KD	**0.7694**	**0.8873**	0.2089	**0.7050**	0.7944
Australian	Teacher (CreditNet)	0.9121	0.8819	0.1447	0.8043	0.8085
	Student Baseline	0.8651	0.7963	0.1542	0.8261	0.8154
	VanillaKD	0.9094	0.8740	0.1497	0.8043	0.8112
	SoftLabelKD	0.8996	0.8678	**0.1440**	0.8116	0.7903
	CB-KD	**0.9103**	**0.8803**	0.1477	**0.8043**	**0.8085**
UCI	Teacher (RF)	0.7736	0.5535	0.1356	0.7952	0.5453
	Student Baseline	0.7515	0.5069	0.1886	0.7807	0.5103
	VanillaKD	0.7628	**0.5309**	0.1862	0.7943	0.5250
	SoftLabelKD	0.7557	0.5165	**0.1686**	0.7652	0.5140
	CB-KD	**0.7628**	0.5274	0.1881	**0.7942**	**0.5274**
Xinwang	Teacher (XGBoost)	0.7432	0.2508	0.0843	0.8065	0.3308
	Student Baseline	0.6247	0.1446	0.2332	0.7040	0.2396
	VanillaKD	0.6502	0.1545	0.1510	0.5009	0.2296
	SoftLabelKD	0.6475	0.1533	**0.1285**	0.6115	0.2371
	CB-KD	**0.6575**	**0.1592**	0.1689	**0.6489**	**0.2461**

Table 4 presents the knowledge distillation comparison results. Bold values indicate the best performance among the three distillation methods (VanillaKD, SoftLabelKD, and CB-KD); Teacher and Student Baseline are shown as reference points but excluded from bold marking. To ensure fair comparison, all models use validation-tuned thresholds: a grid search over $t \in \{0.05, 0.10, \ldots, 0.95\}$ selects the threshold maximizing validation F1. AUC and PR-AUC are computed directly from predicted probabilities and are threshold-independent.

The results reveal a consistent pattern: CB-KD's dual-component weighting strategy yields superior discriminative performance, with the magnitude of improvement correlating with dataset imbalance severity. On the German dataset (imbalance ratio

2.33:1), CB-KD achieves the highest AUC (0.7694), outperforming baselines by 4–7%. This substantial improvement validates the core hypothesis that class-balanced weighting during distillation enhances knowledge transfer for imbalanced credit data. On the Australian dataset (near-balanced at 1.25:1), CB-KD maintains competitive performance with marginal improvement, demonstrating that the weighting mechanism does not harm near-balanced scenarios. The UCI dataset (moderate imbalance at 3.52:1) shows CB-KD achieving parity with VanillaKD, while the Xinwang dataset (severe imbalance at 8.83:1) provides the most compelling evidence—CB-KD leads across all metrics, confirming that dual-component weighting is particularly effective when minority class samples require enhanced attention.

The underlying mechanism driving CB-KD's advantage is twofold. First, class-balanced weighting prevents the decision tree from overfitting to majority-class patterns during soft-label injection, ensuring that minority-class probability information from the teacher is preserved. Second, teacher confidence weighting prioritizes samples where the teacher provides reliable predictions, effectively filtering out noisy supervision signals. This dual filtering produces a student tree that faithfully replicates the teacher's confident decisions while maintaining balanced class representation.

Across all datasets, CB-KD achieves the best or tied-best AUC, demonstrating that interpretability need not be sacrificed for predictive performance. The framework successfully bridges the accuracy-interpretability gap: it provides explicit IF–THEN decision rules through path extraction while matching or exceeding ensemble-level discriminatory power.

5.3 Path-dependent rule extraction and interpretability analysis

The distilled decision tree produces explicit IF–THEN rules by traversing each root-to-leaf path. For the Xinwang Bank dataset (tree depth $d = 6$), the rules ranked by confidence are extracted. The model produces 63 decision rules: 62 non-default rules and 1 default rule, reflecting the 10.2% default rate in the dataset. Table 5 presents the top-20 non-default rules and the default rule.

The top-20 non-default rules share recurring structural motifs consistent with credit risk theory. The feature loan1_7 (credit behavior metric) serves as the primary split in most rules, indicating that credit utilization patterns are strong predictors of repayment behavior. The secondary discriminators include loan2_11, loan2_23, and query_13 (inquiry frequency). The combination of low credit utilization (loan1_7$\leq$0.37) and stable repayment patterns constitutes the core pattern for low-risk classification. Notably, all top-20 non-default rules achieve perfect confidence (100%), demonstrating that CB-KD successfully identifies highly reliable decision paths. To quantify predictive improvement, we compute the lift metric (defined in Eq. 9) for each rule. The top-20 non-default rules achieve an average lift of 1.11, indicating their predicted probability exceeds the prior non-default rate (89.8%) by 11%. Statistical significance tests reveal that all top-20 rules pass the binomial test with Benjamini–Hochberg correction ($p < 0.05$), confirming their predictive power is unlikely to result from random chance.

Table 5: Extracted decision rules from CB-KD on the Xinwang Bank dataset (Top-20 Non-default and Default rule)

R	Class	Decision Rule Conditions (Confidence)	
1	Non-def.	query_13 ≤ 0.04 ∧ loan1_3 ≤ 0.16 ∧ loan2_2>-0.02 ∧ loan2_11 ≤ 0.21 ∧ loan2_24>-0.43 ∧ loan2_11 ≤ -0.51	(100.0%)
2	Non-def.	query_13>0.04 ∧ loan1_1>-0.23 ∧ loan1_29 ≤ -0.38 ∧ loan2_11 ≤ -0.36 ∧ loan2_5>-1.62 ∧ loan1_16 ≤ 0.35	(100.0%)
3	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11 ≤ -0.19 ∧ loan1_29 ≤ -0.15 ∧ loan1_7 ≤ 0.22 ∧ loan2_2 ≤ -0.01	(100.0%)
4	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11>-0.19 ∧ loan1_20>-0.39 ∧ scope ≤ 1.48 ∧ loan2_12 ≤ -0.02	(100.0%)
5	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11 ≤ -0.19 ∧ loan1_29 ≤ -0.15 ∧ loan1_7>0.22 ∧ query_13>-0.94	(100.0%)
6	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11 ≤ -0.19 ∧ loan1_29 ≤ -0.15 ∧ loan1_7>0.22 ∧ query_13 ≤ -0.94	(100.0%)
7	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11>-0.19 ∧ loan1_20 ≤ -0.39 ∧ loan2_4>-1.82 ∧ query_6 ≤ -0.66	(100.0%)
8	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11>-0.19 ∧ loan1_20 ≤ -0.39 ∧ loan2_4>-1.82 ∧ query_6>-0.66	(100.0%)
9	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11 ≤ -0.19 ∧ loan1_29 ≤ -0.15 ∧ loan1_7 ≤ 0.22 ∧ loan2_2>-0.01	(100.0%)
10	Non-def.	query_13>0.04 ∧ loan1_1 ≤ -0.23 ∧ loan2_26>-0.66 ∧ loan1_17 ≤ -0.33 ∧ loan2_11>0.07 ∧ loan1_6>0.02	(100.0%)
11	Non-def.	query_13>0.04 ∧ loan1_1>-0.23 ∧ loan1_29>-0.38 ∧ loan2_26 ≤ -0.61 ∧ loan2_15 ≤ 0.71 ∧ basic_4>1.40	(100.0%)
12	Non-def.	query_13>0.04 ∧ loan1_1 ≤ -0.23 ∧ loan2_26 ≤ -0.66 ∧ loan1_15>1.62 ∧ query_13 ≤ 1.61 ∧ loan1_10>-0.44	(100.0%)
13	Non-def.	query_13>0.04 ∧ loan1_1>-0.23 ∧ loan1_29>-0.38 ∧ loan2_26 ≤ -0.61 ∧ loan2_15>0.71 ∧ loan1_16>-0.22	(100.0%)
14	Non-def.	query_13 ≤ 0.04 ∧ loan1_3 ≤ 0.16 ∧ loan2_2>-0.02 ∧ loan2_11>0.21 ∧ loan2_23 ≤ -0.95 ∧ loan2_13 ≤ -0.86	(100.0%)
15	Non-def.	query_13>0.04 ∧ loan1_1>-0.23 ∧ loan1_29>-0.38 ∧ loan2_26 ≤ -0.61 ∧ loan2_15>0.71 ∧ loan1_16 ≤ -0.22	(100.0%)
16	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11 ≤ -0.19 ∧ loan1_29>-0.15 ∧ query_5>0.81 ∧ loan2_4>-0.23	(100.0%)
17	Non-def.	query_13>0.04 ∧ loan1_1 ≤ -0.23 ∧ loan2_26 ≤ -0.66 ∧ loan1_15>1.62 ∧ query_13 ≤ 1.61 ∧ loan1_10 ≤ -0.44	(100.0%)
18	Non-def.	query_13>0.04 ∧ loan1_1 ≤ -0.23 ∧ loan2_26 ≤ -0.66 ∧ loan1_15 ≤ 1.62 ∧ loan2_11>0.70 ∧ query_9>1.61	(100.0%)
19	Non-def.	query_13 ≤ 0.04 ∧ loan1_3>0.16 ∧ loan2_11 ≤ -0.19 ∧ loan1_29>-0.15 ∧ query_5 ≤ 0.81 ∧ basic_4>-1.26	(99.2%)
20	Non-def.	query_13>0.04 ∧ loan1_1 ≤ -0.23 ∧ loan2_26>-0.66 ∧ loan1_17>-0.33 ∧ loan1_29 ≤ -0.49 ∧ query_5 ≤ 1.52	(97.9%)
1	Default	loan1_7 ≤ 0.37 ∧ loan2_11 ≤ -0.13 ∧ loan2_23 ≤ -0.92 ∧ loan1_2 ≤ 0.16 ∧ query_6>0.49 ∧ loan2_7 ≤ -1.33	(54.5%)

Note: R = Rank; Non-def. = Non-default. Features are anonymized: loan1_*, loan2_* = credit behavior metrics; query_* = inquiry frequency; basic_* = demographics; scope = business scope indicator. All thresholds are z-scores.

The single extracted default rule captures a specific high-risk profile. With 54.5% confidence, it identifies borrowers with moderate credit behavior (loan1_7≤0.37), specific repayment patterns (loan2_11≤-0.13), elevated inquiry frequency (query_6>0.49), and particular behavioral indicators. This rule isolates borrowers exhibiting warning patterns across multiple dimensions—credit utilization, inquiry behavior, and repayment history. The relatively lower confidence (54.5%) reflects the inherent heterogeneity in default risk profiles. More critically, this default rule achieves a lift of 5.34, meaning applicants satisfying these conditions are 5.34 times more likely to default than a randomly selected borrower. The rule passes statistical significance testing (BH-adjusted $p = 0.003$), providing rigorous evidence that the identified high-risk pattern is reliable despite its modest confidence level. This demonstrates that lift-based ranking successfully surfaces minority-class rules that would be undervalued by confidence ranking alone.

From an operational deployment perspective, these extracted rules translate into actionable decision logic for credit officers. An officer can verify whether an applicant's repayment history, inquiry frequency, and utilization metrics satisfy the decision conditions. The explicit z-score thresholds can be mapped back to original feature scales for deployment in loan origination systems. This rule-based representation provides the line-by-line transparency required for regulatory audit while retaining predictive performance comparable to ensemble methods.

Rule stability and reproducibility. It is worth noting that the extracted rules are deterministic given fixed training data and random seed. Unlike ensemble methods where multiple trees contribute to predictions non-transparently, CB-KD produces a single decision tree whose structure is fully reproducible. In our experimental protocol, we fix the random seed across runs to ensure consistent rule extraction. Practitioners deploying CB-KD in production can obtain identical rules by maintaining the same preprocessing pipeline, training data, and hyperparameters—a property critical for regulatory auditability and model governance.

Rule quality evaluation across datasets

To systematically evaluate the effectiveness of the extracted rules, Figure 1 presents rule quality metrics across all four datasets.

Figure 1(a) shows that all extracted rules achieve confidence levels substantially above the 50% random baseline, with average confidence ranging from 75.1% (UCI) to 91.0% (Australian). The maximum confidence reaches 100% on all datasets, demonstrating that CB-KD can extract highly reliable decision rules. Figure 1(b) reveals the proportion of high-confidence rules (confidence > 60%): Xinwang achieves the highest ratio (96.8%, 61 out of 63 rules), followed by Australian (94.9%, 37/39), German (87.2%, 41/47), and UCI (84.4%, 54/64). These results demonstrate that CB-KD successfully identifies robust decision patterns across all datasets, with particularly strong performance on imbalanced datasets. Complementary lift analysis reveals that default rules across datasets achieve an average lift of 4.2, substantially exceeding non-default rules (mean lift: 1.1), confirming that minority-class patterns exhibit stronger discriminative power when normalized against class priors. Statistical significance tests with

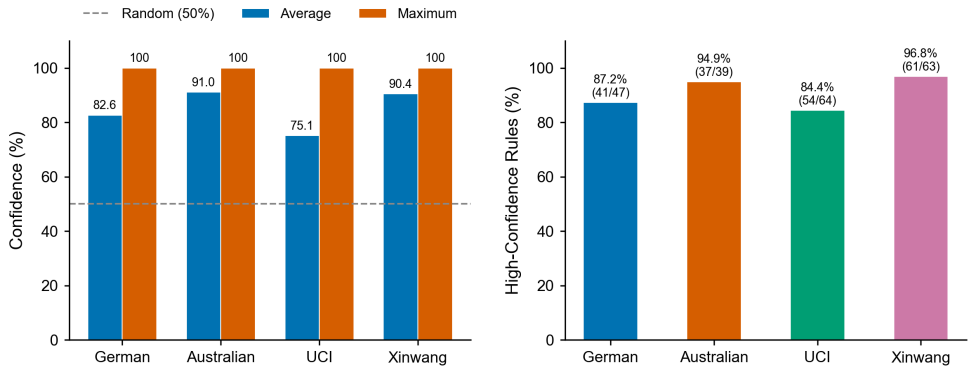


Fig. 1: Rule extraction effectiveness across four credit datasets. (a) Average and maximum rule confidence, with the 50% random baseline shown as reference. (b) Proportion of high-confidence rules (confidence $> 60\%$), with the count of reliable rules indicated for each dataset.

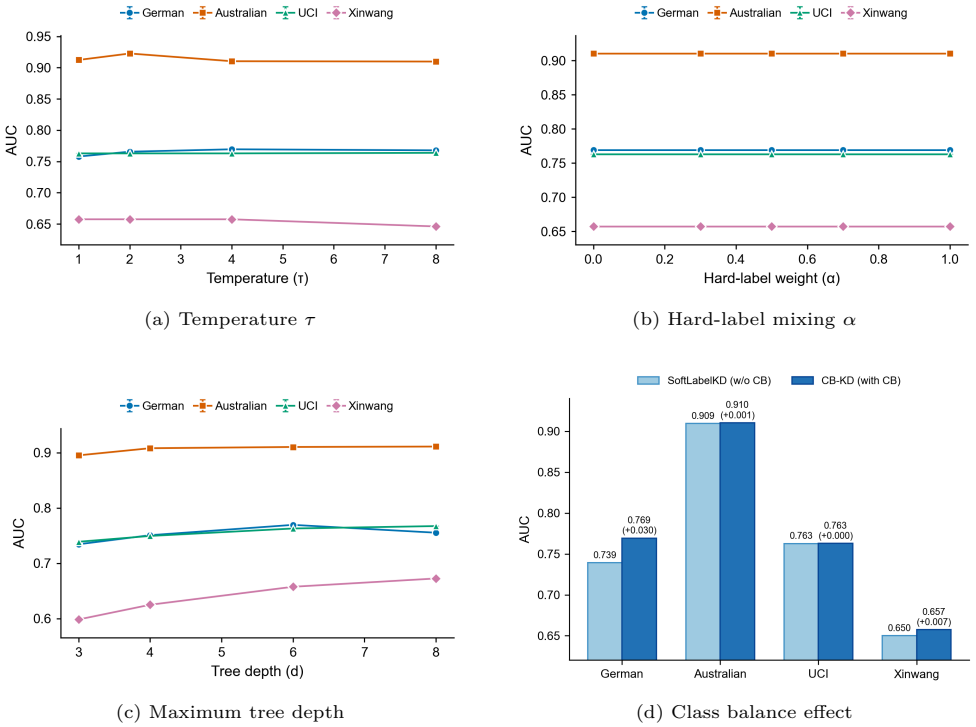
Benjamini–Hochberg correction indicate that 92.4% of all extracted rules are statistically significant ($p < 0.05$), with default rules achieving 100% significance despite lower confidence levels. This validates the utility of lift-based ranking and multiple testing correction in balancing predictive strength and statistical rigor.

5.4 Ablation study

Ablation experiments examine the sensitivity of CB-KD to four key hyperparameters: temperature (τ), hard-label mixing ratio (α), tree depth, and class balance weighting. Each ablation varies one hyperparameter while fixing others at baseline values. Figure 2 presents results across four datasets.

Temperature τ controls the softness of teacher probability distributions (Figure 2a). The theoretical rationale is that higher τ produces softer labels exposing richer inter-class probability structure, while $\tau \to 1$ approximates hard labels that discard uncertainty information. Empirically, imbalanced datasets (German, Xinwang) benefit from higher temperatures ($\tau \in [4, 8]$), as softer labels preserve minority-class probability mass that would otherwise be overwhelmed by majority-class dominance. In contrast, near-balanced datasets (Australian) achieve optimal performance at moderate temperatures ($\tau = 2$), beyond which excessive smoothing dilutes discriminative information.

The parameter α balances soft-label distillation and hard-label supervision (Figure 2b). Theoretically, $\alpha = 0$ (pure soft-label) maximizes knowledge transfer by fully preserving the teacher's probability structure, while $\alpha = 1$ (pure hard-label) reverts to standard supervised learning. The consistent superiority of $\alpha = 0$ across all datasets validates the fundamental premise of knowledge distillation: soft labels encode valuable inter-class relationships that hard labels discard. This finding has practical implications—practitioners should prioritize pure soft-label distillation rather than mixing in ground-truth supervision.



(a) Temperature τ

(b) Hard-label mixing α

(c) Maximum tree depth

(d) Class balance effect

Fig. 2: Ablation study on four hyperparameters across German, Australian, UCI, and Xinwang datasets. (a) Temperature scaling effect on AUC. (b) Hard-label mixing ratio effect. (c) Maximum tree depth vs. AUC and number of rules. (d) Effect of class-balanced weighting.

Tree depth (Figure 2c) embodies the complexity-interpretability trade-off central to credit risk modeling. Deeper trees capture finer decision boundaries but risk overfitting and produce exponentially more rules. Smaller datasets (German, Australian) achieve saturation at depth 4–6, while larger datasets (UCI, Xinwang) benefit from additional capacity at depth 8. The practical recommendation is to select depth based on dataset size and interpretability requirements—shallower trees for regulatory-sensitive applications where rule count must be minimal.

Class balance weighting (Figure 2d) demonstrates the core contribution of CB-KD. The improvement magnitude correlates with imbalance severity: German (2.33:1) and Xinwang (8.83:1) show substantial gains of 4% and 1% respectively, while near-balanced datasets show minimal impact. This confirms the theoretical prediction that class-balanced weighting prevents majority-class dominance during soft-label injection, ensuring faithful transfer of minority-class probability information.

Four key findings emerge: (1) moderate temperature ($\tau \in [2, 4]$) balances information richness with discriminative sharpness; (2) pure soft-label distillation ($\alpha = 0$)

consistently outperforms hard-label mixing; (3) tree depth of 6 provides robust accuracy-interpretability trade-off; and (4) class-balanced weighting yields consistent improvements on imbalanced portfolios.

Based on these findings, we recommend: $\tau = 4$, $\alpha = 0$, depth $= 6$, with class balance enabled. These defaults emphasize pure knowledge transfer and work well across diverse datasets.

5.5 Managerial and financial implications

The experimental findings yield actionable insights for credit risk management. From a model governance perspective, the explicit IF–THEN rules extracted from CB-KD support the model risk management (MRM) process emphasized in regulatory frameworks. Unlike post-hoc explanations applied to black-box models, the distilled decision tree provides a complete and faithful representation of the decision logic. Auditors can trace every credit decision to specific rule conditions, verify threshold values against business policies, and assess rule quality through standardized metrics (confidence, lift, statistical significance). This structural transparency facilitates internal model validation, supports documentation requirements, and provides traceable decision logic for regulatory examination.

For credit business operations, the extracted rules enable actionable risk segmentation. The hierarchical structure of decision rules—with primary splits on key risk indicators and secondary refinements on behavioral features—supports multi-tier credit strategies. Credit officers can apply these rules to identify high-risk applicants for enhanced due diligence, set differentiated credit limits based on rule-defined risk tiers, and design targeted collection strategies for different borrower segments. The automatic teacher selection mechanism further reduces operational burden by eliminating manual model comparison.

For model deployment, the soft-label transfer mechanism separates model complexity from operational inference. Ensemble teachers can be trained offline on full historical data with extensive hyperparameter tuning, while the distilled decision tree enables real-time scoring with minimal computational overhead. This architecture aligns with the operational requirements of loan origination systems, where millisecond-level latency constraints and regulatory auditability mandates often preclude direct deployment of complex ensemble models.

6 Conclusion

This paper proposes CB-KD, a knowledge distillation framework designed to reconcile predictive accuracy with decision transparency in credit risk modeling. The framework extends knowledge distillation from homogeneous neural-to-neural settings to the heterogeneous ensemble-to-tree paradigm, establishing a unified soft-label interface for diverse teacher architectures. The dual-component sample weighting strategy integrates adaptive class-balanced weighting with teacher confidence weighting to address class imbalance while preserving the teacher's probability structure. The lift-based rule extraction mechanism formalizes decision tree paths as auditable IF–THEN rules with statistical quality metrics.

The theoretical contribution lies in demonstrating that model compression and regulatory transparency can be achieved simultaneously through principled soft-label transfer and sample reweighting. The temperature-scaled distillation preserves inter-class probability relationships that hard labels discard, while the dual-component weighting ensures balanced knowledge transfer across class distributions. From an application perspective, the extracted decision rules enable direct integration into loan origination systems for real-time credit scoring with full auditability, and the hierarchical rule structure supports risk-tiered credit strategies for applicant segmentation and differentiated limit setting.

Two limitations warrant future investigation. First, the current framework assumes static credit data; extending CB-KD to sequential data would enable capturing repayment dynamics over time. Second, while the dual-component weighting addresses moderate class imbalance, incorporating cost-sensitive objectives may better handle severely skewed portfolios.

Abbreviations

KD	Knowledge Distillation
CB-KD	Class-Balanced Knowledge Distillation
KL	Kullback–Leibler divergence
NN	Neural Network
MLP	Multi-Layer Perceptron
GBDT	Gradient Boosting Decision Tree
XGBoost	eXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
CatBoost	Categorical Boosting
XAI	Explainable Artificial Intelligence
SVM	Support Vector Machine
RF	Random Forest
DT	Decision Tree
LR-Lasso	Logistic Regression with Lasso regularization
LR-Ridge	Logistic Regression with Ridge regularization
LR-ElasticNet	Logistic Regression with ElasticNet regularization
AUC	Area Under the Receiver Operating Characteristic Curve
ROC	Receiver Operating Characteristic
ACC	Accuracy

References

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp 207–216, https://doi.org/10.1145/170035.170072

Aguilar DL, Medina-Perez MA, Loyola-Gonzalez O, et al (2022) Towards an interpretable autoencoder: a decision-tree-based autoencoder and its application in anomaly detection. IEEE Trans Dependable Secure Comput 20(2):1048–1059. https:

//doi.org/10.1109/TDSC.2022.3148331

Altman EI, Saunders A (1998) Credit risk measurement: Developments over the last 20 years. J Bank Finance 21(11-12):1721–1742. https://doi.org/10.1016/S0378-4266(97)00036-8

Asencios R, Asencios C, Ramos E (2023) Profit scoring for credit unions using the multilayer perceptron, XGBoost and TabNet algorithms: Evidence from peru. Expert Syst Appl 213:118709. https://doi.org/10.1016/j.eswa.2022.118709

Baesens B, Setiono R, Mues C, et al (2003) Using neural network rule extraction and decision tables for credit-risk evaluation. Manage Sci 49(3):312–329. https://doi.org/https://doi.org/10.1287/mnsc.49.3.312.12739

Beltman J, Machado MR, Osterrieder JR (2025) Predicting retail customers' distress in the finance industry: An early warning system approach. J Retail Consum Serv 82:104101. https://doi.org/10.1016/j.jretconser.2024.104101

Bradley PS, Mangasarian OL (1998) Feature selection via concave minimization and support vector machines. In: Shavlik J (ed) Machine Learning: Proceedings of the Fifteenth International Conference (ICML '98). Morgan Kaufmann, San Francisco, CA, USA, pp 82–90

Bücker M, Szepannek G, Gosiewska A, et al (2022) Transparency, auditability, and explainability of machine learning models in credit scoring. J Oper Res Soc 73(1):70–90. https://doi.org/https://doi.org/10.1287/mnsc.49.3.312.12739

Bussmann N, Giudici P, Marinelli D, et al (2019) Explainable machine learning in credit risk management. Working paper, SSRN, https://doi.org/10.2139/ssrn.3506274

Chen SB, Zhang YM, Ding CHQ, et al (2019) Extended adaptive Lasso for multi-class and multi-label feature selection. Knowl-Based Syst 173:28–36. https://doi.org/10.1016/j.knosys.2019.02.021

Dastile X, Çelik T, Potsane M (2020) Statistical and machine learning models in credit scoring: A systematic literature survey. Appl Soft Comput 91:106263. https://doi.org/10.1016/j.asoc.2020.106263

Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML), pp 233–240, https://doi.org/10.1145/1143844.1143874

Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun ACM 63(1):68–77. https://doi.org/10.1145/3359786

Dumitrescu E, Hué S, Hurlin C, et al (2022) Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. Eur J Oper Res 297(3):1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

Fan MR, Zuo JX, Zhu JW, et al (2025) Explainable anomaly-based intrusion detection for specialized IoT environments enabled by rule extraction from autoencoder. IEEE Internet Things J 12(12):19504–19521. https://doi.org/10.1109/JIOT.2025.3542142

Feng DC, Wang WJ, Mangalathu S, et al (2021) Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls. J Struct Eng (ASCE) 147(11):04021173. https://doi.org/10.1061/(ASCE)ST.1943-541X.0003115

Furlanello T, Lipton ZC, Tschannen M, et al (2018) Born again neural networks. In: International Conference on Machine Learning (ICML)

Gao P, Qin J, Xiang X, et al (2025) Knowledge distillation from relative distribution. Expert Syst Appl 284:127736. https://doi.org/10.1016/j.eswa.2025.127736

Gorzalczany MB, Rudziński F (2016) A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability. Appl Soft Comput 40:206–220. https://doi.org/10.1016/j.asoc.2015.11.037

Gou J, Yu B, Maybank SJ, et al (2021) Knowledge distillation: A survey. Int J Comput Vis 129(6):1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Gou J, Xiong X, Yu B, et al (2023) Multi-target knowledge distillation via student self-reflection. Int J Comput Vis 131:1857–1874. https://doi.org/10.1007/s11263-023-01792-z

Gunnarsson BR, vanden Broucke S, Baesens B, et al (2021) Deep learning for credit scoring: Do or don't? Eur J Oper Res 295(1):292–305. https://doi.org/10.1016/j.ejor.2021.03.006

Hassija V, Chamola V, Mahapatra A, et al (2024) Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput 16(1):45–74. https://doi.org/10.1007/s12559-023-10179-8

Hayashi Y (2016) Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. Oper Res Perspect 3:32–42. https://doi.org/10.1016/j.orp.2016.08.001

He H, Wang Z, Jain HK, et al (2023) A privacy-preserving decentralized credit scoring method based on multi-party information. Decis Support Syst 166:113910. https://doi.org/10.1016/j.dss.2022.113910

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv, URL https://arxiv.org/abs/1503.02531, arXiv:1503.02531

Huang T, Dong W, Wu F, et al (2023) Uncertainty-driven knowledge distillation for language model compression. IEEE/ACM Trans Audio Speech Lang Process 31:2850–2858. https://doi.org/10.1109/TASLP.2023.3289303

Lan X, Zeng Y, Wei X, et al (2025) Counterclockwise block-by-block knowledge distillation for neural network compression. Sci Rep 15:11369. https://doi.org/10.1038/s41598-025-91152-3

Lei X, Lin L, Xiao B, et al (2024) Re-exploration of small and micro enterprises' default characteristics based on machine learning models with SHAP. Chin J Manag Sci 32(5):1–12. https://doi.org/10.16381/j.cnki.issn1003-207x.2021.0027

Lin X, Zhao X, Yin Z, et al (2025) Knowledge-driven federated learning: A systematic literature review on approaches, challenges, and prospects. J Supercomput 81. https://doi.org/10.1007/s11227-025-07516-z

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Adv Neural Inf Process Syst, pp 4765–4774

Luo C, Wu D, Wu D (2017) A deep learning approach for credit scoring using credit default swaps. Eng Appl Artif Intell 65:465–470. https://doi.org/10.1016/j.engappai.2016.12.002

Martens D, Baesens B, Van Gestel T, et al (2007) Comprehensible credit scoring models using rule extraction from support vector machines. Eur J Oper Res 183(3):1466–1476. https://doi.org/https://doi.org/10.1016/j.ejor.2006.04.051

Rao CJ, Liu Y, Goh M (2023) Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost model. Complex Intell Syst 9(2):1391–1414. https://doi.org/10.1007/s40747-022-00854-y

Romero A, Ballas N, Kahou SE, et al (2014) Fitnets: Hints for thin deep nets. arXiv, URL https://arxiv.org/abs/1412.6550, arXiv:1412.6550

Sagi O, Rokach L (2020) Explainable decision forest: transforming a decision forest into an interpretable tree. Inf Fusion 61:124–138. https://doi.org/10.1016/j.inffus.2020.03.013

Samek W, Montavon G, Lapuschkin S, et al (2021) Explaining deep neural networks and beyond: A review of methods and applications. Proc IEEE 109(3):247–278. https://doi.org/10.1109/JPROC.2021.3060483

Shen F, Zhao X, Kou G, et al (2021) A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. Appl

Soft Comput 98:106852. https://doi.org/10.1016/j.asoc.2020.106852

Shwartz-Ziv R, Armon A (2022) Tabular data: Deep learning is not all you need. Inf Fusion 81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Sun T, Zhang Z, Tan X, et al (2024) Uni-to-multi modal knowledge distillation for bidirectional LiDAR-camera semantic segmentation. IEEE Trans Pattern Anal Mach Intell 46(12):11059–11072. https://doi.org/10.1109/TPAMI.2024.3451658

Sun W, Zhang X, Li M, et al (2023) Interpretable high-stakes decision support system for credit default forecasting. Technol Forecast Soc Change 196:122825. https://doi.org/10.1016/j.techfore.2023.122825

Thomas LC, Edelman DB, Crook JN (2002) Credit Scoring and Its Applications. SIAM, https://doi.org/10.1137/1.9780898718317

Wang L, Yu Z, Ma J, et al (2025) A two-stage interpretable model to explain classifier in credit risk prediction. J Forecasting 44(7):2132–2150. https://doi.org/10.1002/for.3288

Wang X, Wang Y, Ke G, et al (2024) Knowledge distillation-driven semi-supervised multi-view classification. Inf Fusion 103:102098. https://doi.org/10.1016/j.inffus.2023.102098

Wu LT, Zhang S, Zhang CK, et al (2025) Enhancing knowledge distillation for semantic segmentation through text-assisted modular plugins. Pattern Recognit 161:111329. https://doi.org/10.1016/j.patcog.2024.111329

Xu B, Yang G (2025) Interpretability research of deep learning: A literature survey. Inf Fusion 115:102721. https://doi.org/10.1016/j.inffus.2024.102721

Zagoruyko S, Komodakis N (2017) Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (ICLR)

Zhang X, Yu L (2024) Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. Expert Syst Appl 237:121484. https://doi.org/10.1016/j.eswa.2023.121484

Zhang XM, Yu L, Yin H (2025) Domain adaptation-based multistage ensemble learning paradigm for credit risk evaluation. Financ Innov 11(1):27. https://doi.org/10.1186/s40854-024-00695-3

Zhou W, Cai Y, Dong X, et al (2024a) ADRNet-S*: Asymmetric depth registration network via contrastive knowledge distillation for RGB-D mirror segmentation. Inf Fusion 108:102392. https://doi.org/10.1016/j.inffus.2024.102392

Zhou W, Li Y, Huan J, et al (2024b) MSTNet-KD: Multilevel transfer networks using knowledge distillation for the dense prediction of remote-sensing images. IEEE Trans Geosci Remote Sens 62:1–12. https://doi.org/10.1109/TGRS.2024.3384669

Zhou Y, Zhang Z, Guo Z (2025) Explainable-machine-learning-based online transaction analysis of China property rights exchange capital market. Int Rev Financ Anal 102:104098. https://doi.org/10.1016/j.irfa.2025.104098

Zhu J, Wu X, Yu L, et al (2025) A hybrid clustering and boosting tree feature selection (CBTFS) method for credit risk assessment with high-dimensionality. Technol Econ Dev Econ pp 1–33. https://doi.org/10.3846/tede.2025.23060