Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Entropy-driven adaptive knowledge distillation with reliable candidate fusion for imbalanced binary classification

**Jihang Li**
Yulin University
**Jue Zhang**
130376359260163.com

Yulin University
**Hanye Liu**
Yulin University
**Ping Jiang**
Yulin University
**Lele Su**
Yulin University
**Hui Zhang**
Yulin University

**Article**

**Keywords:** Imbalanced binary classification, Knowledge distillation, Adaptive distillation, Reliable candidate fusion, Probability reliability

**Posted Date:** November 13th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11193122/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Entropy-driven adaptive knowledge distillation with reliable candidate fusion for imbalanced binary classification

*Jihang Li[1], Jue Zhang[1]*, Hanye Liu[1], Ping Jiang[1], Lele Su[1] & Hui Zhang[1]*

[1]School of Information Engineering, Yulin University, Yulin 719000, Shaanxi, China

*Email: 13037635926@163.com*

**Abstract**

Class imbalance can impair minority-class recognition and probability reliability, while conventional knowledge distillation often relies on fixed knowledge-transfer settings and a fixed prediction form. In this study, we propose entropy-driven adaptive knowledge distillation with reliable candidate fusion (EAKD-RCF) for imbalanced binary classification. EAKD-RCF consists of two main stages. First, adaptive distillation separates predictive uncertainty from teacher reliability: uncertainty determines the sample-specific distillation temperature, whereas training-fold-level reliability regulates the overall strength of knowledge transfer. Second, student, teacher, and fused predictions are treated as complementary candidates, and the final prediction form is selected using internal validation by jointly considering imbalanced classification performance and probability quality. The proposed method was evaluated on nine public datasets using repeated stratified five-fold cross-validation. EAKD-RCF achieved mean MCC, F1-score, and G-Mean values of 61.96%, 72.55%, and 79.07%, respectively, while maintaining competitive ECE and Brier scores of 7.97% and 10.30%. Stage-wise and ablation analyses further showed complementary contributions from adaptive distillation and candidate selection. These results demonstrate strong average classification performance across heterogeneous datasets while maintaining competitive probability reliability.

**Keywords:** Imbalanced binary classification, Knowledge distillation, Adaptive distillation, Reliable candidate fusion, Probability reliability

## Introduction

Class imbalance is a common challenge in binary classification tasks such as disease diagnosis, anomaly detection, credit fraud detection, and healthcare risk assessment. In these applications, minority-class samples are considerably less frequent than majority-class samples, although they often correspond to events of greater practical interest. Learning algorithms trained on such skewed data may therefore favor the majority class and fail to recognize rare but important cases[1]. Class imbalance can also affect the quality of predicted probabilities, leading to discrepancies between model confidence and observed outcomes[2]. This is particularly relevant in applications involving risk ranking, threshold-based decisions, or decision support, where both minority-class recognition and reliable probability estimates are important.

Existing approaches to class imbalance are commonly developed at the data, algorithm, or ensemble level. Data-level methods modify the class distribution through over-sampling or under-sampling, algorithm-level methods assign greater importance to minority-class samples during learning, and ensemble approaches combine multiple learners or balanced subsets[3-7]. Although these strategies can improve minority-class recognition, re-sampling may introduce artificial samples or discard potentially useful majority-class information, while fixed weighting strategies may not adequately account for differences in sample difficulty. These limitations motivate methods that improve minority-class learning without relying solely on modifications of the original data distribution.

Knowledge distillation (KD) provides an alternative perspective by transferring probabilistic information from one or more teacher models to a student model[8,9]. Compared with hard class labels, teacher probabilities provide additional information about prediction confidence and class relationships and can therefore offer useful supervision for difficult samples. Conventional KD, however, generally applies a fixed temperature and a fixed knowledge-transfer strength to all samples. Recent studies have consequently explored difficulty-aware distillation[10], adaptive-temperature strategies[11], class-conditional transfer[12], sample-wise reweighting[13], uncertainty-aware multi-expert distillation[14], and class-balanced knowledge transfer[15]. These studies demonstrate the potential of adapting the distillation process to different samples or classes.

A remaining issue is that predictive uncertainty and teacher reliability describe different properties of teacher knowledge. A high-entropy prediction may indicate a difficult boundary sample while still containing useful information, whereas a low-entropy prediction can be confidently incorrect. Treating uncertainty directly as a measure of reliability may therefore suppress informative hard samples or strengthen erroneous teacher signals. This problem can be more pronounced under class imbalance, where teacher models may themselves learn biased decision boundaries or provide poorly calibrated probabilities[16-19]. A more appropriate adaptive distillation strategy should therefore use prediction uncertainty to determine how teacher information is expressed, while assessing teacher reliability separately when determining how strongly that information should influence student learning.

A second issue arises after distillation. The teacher and distilled student may exhibit complementary prediction characteristics, and their relative advantages can vary across datasets. Consequently, always using the student output, the teacher output, or a predetermined combination of the two does not necessarily provide the most appropriate prediction under all data conditions. This problem becomes particularly relevant when both imbalanced classification performance and probability quality are considered, because improvements in discrimination do not always coincide with better probability estimates. Rather than imposing a single output rule, prediction forms derived from teacher and student outputs can therefore be treated as complementary candidates and evaluated using information that is independent of the final test data.

Based on these considerations, we propose entropy-driven adaptive knowledge distillation with reliable candidate fusion (EAKD-RCF) for imbalanced binary classification. The proposed method addresses two related stages of the prediction process. During knowledge transfer, predictive uncertainty is used to adapt the distillation temperature, whereas teacher reliability regulates the overall contribution of teacher knowledge, thereby avoiding the direct equivalence between uncertainty and reliability. After distillation, prediction candidates derived from teacher and student outputs are evaluated using an internal validation set, allowing the final prediction form to be determined by jointly considering imbalanced classification performance and probability quality. In this way, EAKD-RCF focuses on when and how teacher knowledge should be transferred and how complementary teacher–student predictions should be used after distillation, rather than relying on fixed knowledge-transfer and prediction settings.

Based on the above analysis, the contributions of this study are as follows:

- An entropy-driven adaptive knowledge distillation strategy is developed for imbalanced binary classification. The method separates predictive uncertainty from teacher reliability: uncertainty determines sample-specific probability smoothing, while reliability controls the overall contribution of teacher knowledge. This design allows difficult samples to retain useful teacher information without directly interpreting prediction confidence as knowledge reliability.

- A reliable candidate fusion strategy is introduced to avoid a fixed post-distillation prediction rule. Prediction candidates derived from teacher and student outputs are evaluated using only internal validation data, and the final prediction form is selected by considering both imbalanced classification performance and probability quality.
- EAKD-RCF is systematically evaluated on nine public imbalanced binary datasets under repeated stratified cross-validation. Comparisons with conventional classifiers, imbalance-learning approaches, and related knowledge-distillation methods are complemented by stage-wise analysis, component ablation, entropy–reliability analysis, architecture and parameter sensitivity analysis, and statistical tests to examine the effectiveness and robustness of the proposed approach.

## Methodology

### Research process

As shown in Fig. 1, this study follows a unified workflow for imbalanced binary classification, consisting of data preprocessing, model training and validation, and performance evaluation. First, all preprocessing operations are performed within the corresponding training partition, and the resulting parameters are subsequently applied to the associated validation or test data to avoid information leakage. Next, the proposed EAKD-RCF model is trained using the available training data, while internal validation data are used to determine the final prediction procedure without accessing the outer test fold. Finally, repeated stratified five-fold cross-validation is employed to evaluate the proposed method in terms of both imbalanced classification performance and probability quality. The model architecture and validation procedure are described in detail in the following

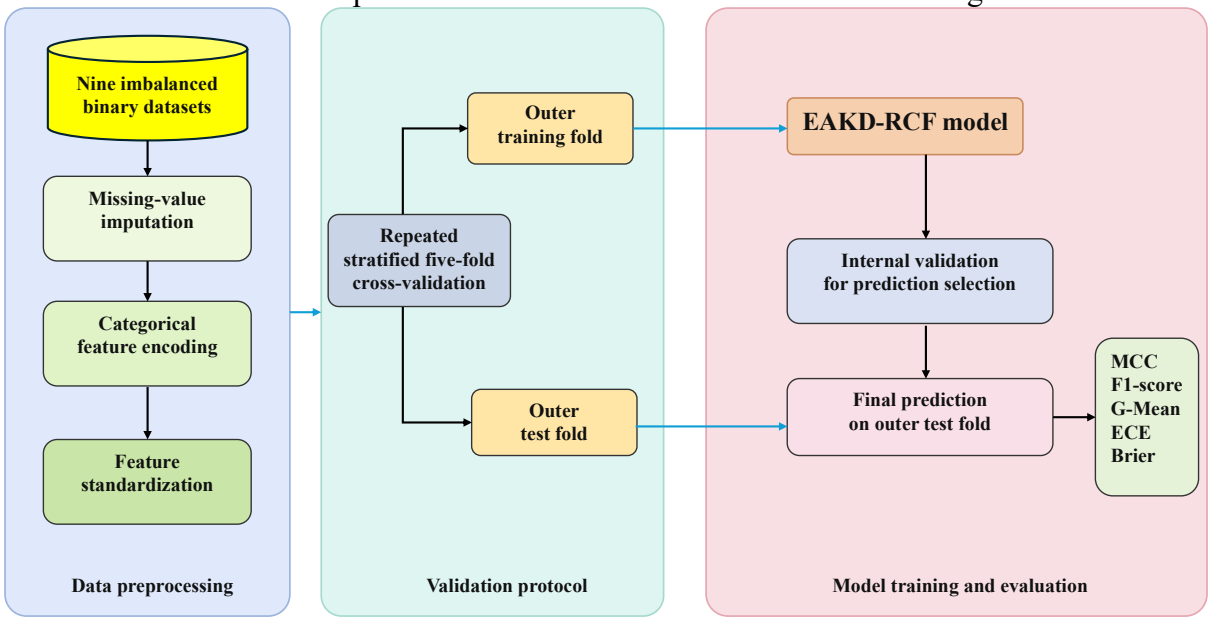



**Fig. 1.** EAKD-RCF prediction workflow.

## Data

This study uses nine publicly available imbalanced binary classification datasets: BCWO, Haberman, WDBC, WPBC, Pima, Sick, Mammographic, Heart, and Hepatitis. The datasets were obtained from current or archived distributions of the UCI Machine Learning Repository[20]. They differ substantially in sample size, feature dimensionality, and class distribution, providing heterogeneous conditions for evaluating the proposed method. As summarized in Table 1, the datasets contain between 155 and 2,800 samples, between 3 and 33 input features, and imbalance ratios ranging from 1.1596 to 15.3743. Sick represents the

most severely imbalanced case, whereas WPBC and Hepatitis provide relatively small-sample settings with limited minority-class observations.

| Dataset | Samples | Features | Majority class | Minority class | Imbalance ratio |
|---|---|---|---|---|---|
| BCWO | 698 | 10 | 457 | 241 | 1.8963 |
| Haberman | 306 | 3 | 225 | 81 | 2.7778 |
| WDBC | 569 | 30 | 357 | 212 | 1.6840 |
| WPBC | 198 | 33 | 151 | 47 | 3.2128 |
| Pima | 768 | 8 | 500 | 268 | 1.8657 |
| Sick | 2800 | 29 | 2629 | 171 | 15.3743 |
| Mammographic | 961 | 5 | 516 | 445 | 1.1596 |
| Heart | 303 | 13 | 164 | 139 | 1.1799 |
| Hepatitis | 155 | 19 | 123 | 32 | 3.8438 |

**Table 1.** Characteristics of the imbalanced binary classification datasets.

## Data preprocessing

*Missing value processing*

Missing values are handled separately according to feature type. For each training partition, missing numerical values are replaced by the median calculated from the available training samples, whereas missing categorical values are replaced by the corresponding training-set mode. The resulting imputation values are then applied unchanged to the associated validation or test data. Therefore, no information from the validation or test partition is used to determine the imputation parameters.

*Categorical feature encoding*

After missing-value imputation, categorical variables are transformed using one-hot encoding. The encoding categories are determined from the corresponding training data, and the same transformation is subsequently applied to the associated validation or test data.

*Feature standardization*

Numerical features may differ substantially in scale and are therefore standardized using the mean and standard deviation estimated from the corresponding training data. For a numerical feature $x$, the standardized value is calculated as

$$x' = \frac{x-\mu}{\sigma} \tag{1}$$

where $\mu$ and $\sigma$ denote the mean and standard deviation estimated from the training partition, respectively. The same parameters are subsequently used to transform the corresponding validation and test data.

## Model

*Teacher–student architecture*

In this study, a heterogeneous teacher–student architecture is constructed for imbalanced binary classification. As illustrated in Fig. 2, three heterogeneous classifiers, logistic regression (LR)[21], random forest (RF)[22], and XGBoost[23] , are used as teacher models. Because these models learn decision boundaries through different learning paradigms, their probability outputs provide complementary sources of supervision for subsequent knowledge distillation.

The student model is implemented as a multilayer perceptron (MLP) with two hidden layers containing 64 and 32 neurons, respectively. Each hidden layer uses ReLU activation followed by dropout with a rate of 0.1. A single sigmoid output unit produces the positive-class probability. During training, the student learns from the original class labels together with probabilistic supervision derived from the teacher models. The overall teacher–student architecture is shown in Fig. 2.

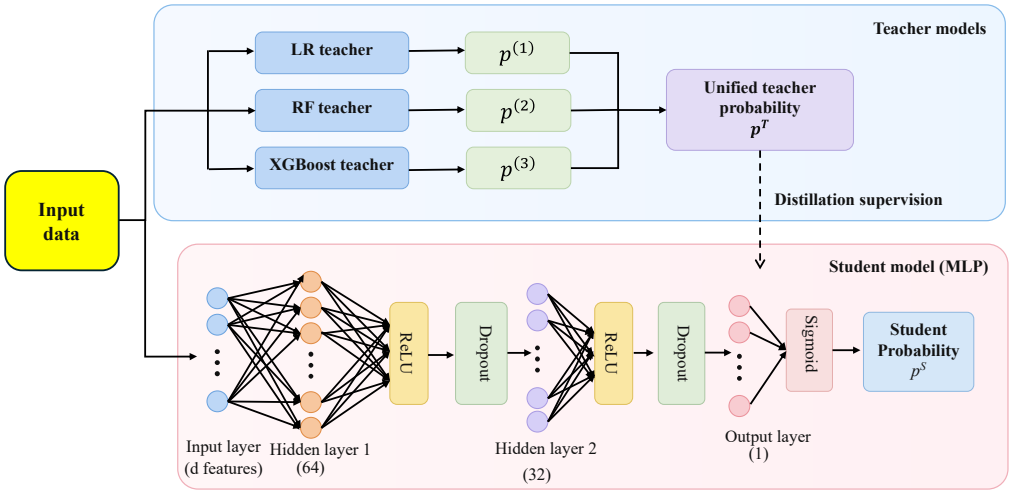



**Fig. 2.** Teacher–student architecture.

*Entropy-driven adaptive knowledge distillation with reliable candidate fusion (EAKD-RCF)*

EAKD-RCF is designed to address two related limitations of knowledge distillation under class imbalance: teacher information may not be equally informative for all samples, and a fixed prediction form may not fully exploit the complementary behavior of the teacher and distilled student. The proposed method therefore considers knowledge transfer and post-distillation prediction within a unified framework. During distillation, predictive uncertainty determines how teacher information is softened, whereas teacher reliability regulates its overall contribution to student learning. After distillation, prediction candidates derived from the teacher and student outputs are evaluated using internal validation data to determine the final prediction form. The overall architecture is illustrated in Fig. 3.

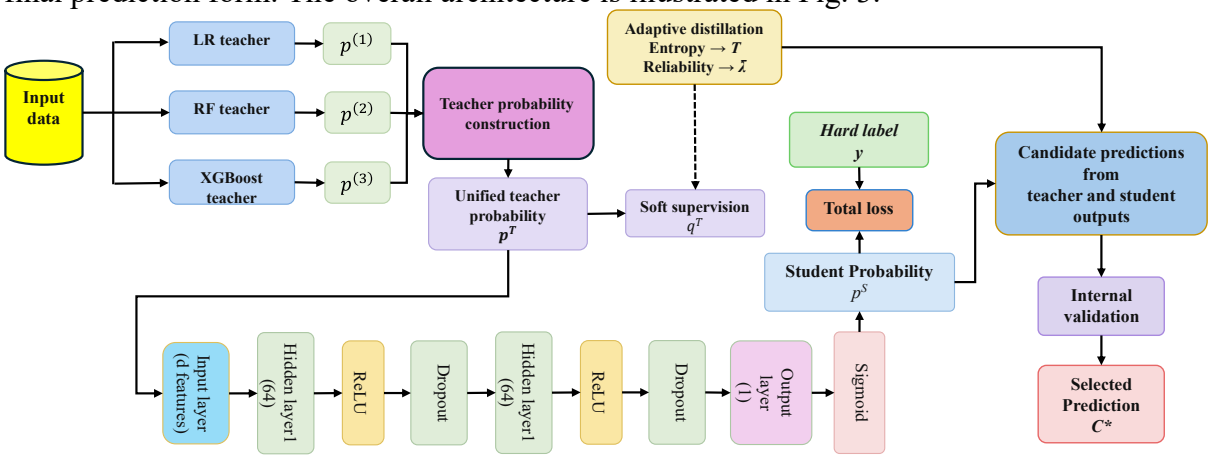



**Fig. 3.** Overall framework of EAKD-RCF.

To construct teacher supervision without introducing in-sample prediction bias, the three teacher models generate five-fold out-of-fold (OOF) probabilities on the training data. The resulting probabilities are subsequently calibrated using Platt scaling[24]. Because heterogeneous teachers may still differ in probability quality after calibration, their contributions are determined according to the corresponding Brier scores[25] , with smaller probability errors receiving larger weights. The calibrated teacher probabilities are then combined into a single teacher probability $p_i^T$ for sample $i$, as defined in Eq. (2). This procedure provides the student with a unified probabilistic target while reducing the influence of comparatively unreliable teacher estimates.

$$p_i^T = \sum_{k=1}^{K} \omega_k \, \tilde{p}_i^{(k)} \tag{2}$$

The teacher probability is then used as soft supervision during student training. EAKD-RCF distinguishes predictive uncertainty from teacher reliability rather than using a single quantity to control knowledge transfer. For each sample, uncertainty is quantified using normalized binary entropy, as given in Eq. (3), and mapped to a sample-specific distillation temperature through Eq. (4). Predictions with higher uncertainty are therefore softened more strongly, whereas more certain predictions retain sharper class information.

$$u_i = \frac{-p_i^T \ln p_i^T - (1-p_i^T)\ln(1-p_i^T)}{\ln 2} \tag{3}$$

and the corresponding sample-specific temperature is defined as

$$T_i = T_{\min} + (T_{\max} - T_{\min})u_i^{\gamma} \tag{4}$$

Predictive uncertainty, however, does not directly indicate whether the teacher prediction is correct. The contribution of teacher knowledge is therefore controlled separately using reliability estimated from the training data. Rather than assigning an additional reliability-dependent weight to every individual sample, the sample-level reliability estimates are summarized within the current training fold to obtain a common distillation coefficient. Consequently, uncertainty determines how teacher knowledge is expressed, whereas reliability determines how strongly that knowledge contributes to student learning. This separation avoids directly treating prediction confidence as knowledge reliability.

During student optimization, the original class labels and the adaptive teacher supervision are jointly used for training. Class-balanced weights[26] are applied to reduce the dominant influence of majority-class samples, and a lightweight Brier regularization term is included to constrain the student probability output. The resulting objective combines the supervised classification loss, the adaptive distillation loss, and the auxiliary probability regularization term, as defined in Eq. (5). The class-balanced weighting and Brier term are used as training regularization components rather than separate adaptation mechanisms.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left[w_{y_i}\left((1-\bar{\lambda})\mathcal{L}_{sup,i} + \bar{\lambda}T_i^2 D_{\mathrm{KL}}\left(q_i^T \parallel q_i^S\right)\right) + \mu_B\left(p_i^S - y_i\right)^2\right] \tag{5}$$

After distillation, the student and teacher ensemble provide complementary probability estimates. Because their relative advantages may vary across datasets and samples, EAKD-RCF does not impose a single teacher–student combination as the final prediction. Instead, prediction forms derived from the teacher and student outputs are organized as complementary candidates. A fixed teacher–student fusion is retained as the reference prediction, while an adaptive fusion and eligible teacher-based predictions are considered when sufficient validation information is available. This conditional construction allows the method to use additional prediction forms when they can be estimated reliably, while retaining the stable reference prediction for small or insufficient validation sets.

The final prediction form is determined exclusively from the internal validation data. Candidate evaluation jointly considers imbalanced classification performance, characterized by MCC, F1-score, and G-Mean, and probability quality, characterized by ECE and Brier score. The method does not simply select the candidate with the highest discrimination score: an alternative prediction is adopted only when its validation performance is sufficiently competitive without unacceptable deterioration in probability quality; otherwise, the fixed fusion is retained. The selected prediction form, probability calibration, and classification threshold are subsequently applied to the outer test data without using any information from that test fold during model or candidate selection.

*Cross-validation*

K-fold cross-validation is widely used to evaluate model generalization by repeatedly training and testing a model on different subsets of the available data[27]. In this study, repeated stratified five-fold cross-validation is employed to obtain a robust evaluation while preserving the original class distribution as closely as possible.

For each repetition, the dataset is divided into five stratified folds. In each outer iteration, four folds are used as the outer training data and the remaining fold is retained exclusively for final testing. The procedure is repeated five times using different random seeds, resulting in 25 outer training–test evaluations for each dataset.

Within each outer iteration, 20% of the outer training data are further reserved as an internal validation set, while the remaining samples form the sub-training set. The sub-training data are used to construct the teacher models and train the distilled student, whereas the internal validation data are used to determine the prediction procedure, probability calibration, and classification threshold. After these settings have been determined, the teacher and student models are retrained using the complete outer training fold, and the selected prediction procedure is applied to the corresponding outer test fold.

The outer test fold is used only for final performance evaluation and does not participate in model training, teacher reliability estimation, prediction selection, probability calibration, or threshold determination. All compared methods use the same outer data partitions and evaluation protocol, and the final results are reported as the average performance over the 25 outer test evaluations.

*Evaluation indicators*

This study evaluates model performance from two complementary perspectives: imbalanced classification performance and probability quality. Matthews correlation coefficient (MCC)[28], F1-score, and geometric mean (G-Mean) are used to assess classification performance, whereas expected calibration error (ECE)[29] and Brier score are used to evaluate the reliability of predicted probabilities. MCC considers all four outcomes of the binary confusion matrix and is therefore suitable for evaluating classification performance under skewed class distributions. F1-score reflects the balance between precision and recall for the positive class, while G-Mean measures whether the classifier maintains balanced recognition of both classes. These metrics are defined as

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

$$\mathrm{F1} = \frac{2TP}{2TP+FP+FN} \tag{7}$$

$$\text{G-Mean} = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{TN+FP}} \tag{8}$$

where, $TP$, $TN$, $FP$, and $FN$ denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. In this study, the minority class is treated as the positive class. Higher MCC, F1-score, and G-Mean values indicate better imbalanced classification performance.

Probability quality is further evaluated using ECE and Brier score. ECE measures the discrepancy between predicted confidence and the observed positive-class frequency across probability bins and is defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \tag{9}$$

where $B_m$ denotes the $m$-th probability bin, $|B_m|$ is the number of samples in that bin, and $M$ is the total number of bins. The Brier score measures the mean squared difference between the predicted probability and the true binary outcome:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{10}$$

Lower ECE and Brier scores indicate better probability quality. Together, these five metrics allow the proposed method to be evaluated with respect to both minority-class discrimination and the reliability of its probability estimates.

*Comparison methods*

EAKD-RCF was compared with 14 representative methods covering conventional classifiers, imbalance-aware ensemble methods, and knowledge-distillation approaches. The conventional classifiers included logistic regression (LR)[21], random forest (RF)[22], and XGBoost[23]. Balanced Random Forest (BalancedRF)[30] and EasyEnsemble[31] were included as imbalance-aware ensemble baselines. The knowledge-distillation baselines included conventional KD[8], Yang-ATD[11], balanced knowledge distillation (BKD)[32], class-balanced distillation (CBD)[33], CalibKD[19], CCKD[12], SRW-KD[13], UMKD-UDD[14], and CB-DistillGrad[15].

All methods were evaluated using the same outer data partitions, preprocessing procedure, and evaluation protocol. For the knowledge-distillation methods, the same teacher set, student architecture, and training settings were used whenever applicable to ensure a consistent comparison. For Yang-ATD and UMKD-UDD, their core adaptive-temperature and uncertainty-aware knowledge-transfer principles were retained, while components specific to their original learning settings were adapted to the tabular binary-classification setting used in this study.

*Statistical analysis*

Statistical comparisons were conducted using the nine datasets as independent experimental blocks. For each evaluation metric, the Friedman test[34] was first applied to examine whether overall performance differences existed among the 15 evaluated methods. When pairwise comparisons were considered, EAKD-RCF was compared with each competing method using the Wilcoxon signed-rank test[35]. Holm correction[36] was applied separately within each metric to control for multiple comparisons. A significance level of 0.05 was used throughout the statistical analysis.

## Results

The experimental results are presented from five complementary perspectives. First, the overall classification performance and probability quality of EAKD-RCF are compared with those of the competing methods across the nine datasets. The contributions of different stages and components are then examined through stage-wise and ablation analyses. The relationship between predictive entropy and teacher reliability is subsequently analyzed, followed by experiments on student architecture and parameter sensitivity. Finally, statistical comparisons are conducted across the nine datasets.

## Performance evaluation

Table 2 summarizes the equal-weight mean performance of EAKD-RCF and the 14 comparison methods across the nine imbalanced binary classification datasets. MCC, F1-score, and G-Mean are used to characterize imbalanced classification performance, whereas ECE and Brier score reflect probability quality. The complete dataset-wise results are provided in Supplementary Table S1.

%

| Method | MCC | F1-score | G-Mean | ECE | Brier |
|---|---|---|---|---|---|
| LR | 59.39 | 69.84 | 77.48 | 8.26 | 10.74 |
| RF | 58.41 | 70.06 | 76.59 | 8.22 | 10.94 |
| XGBoost | 57.29 | 69.29 | 75.08 | 7.86 | 10.94 |
| BalancedRF | 58.78 | 69.87 | 76.24 | 7.93 | 10.93 |
| EasyEnsemble | 58.21 | 68.90 | 76.12 | 8.78 | 11.31 |
| KD | 61.01 | 71.36 | 78.29 | 7.97 | 10.23 |
| Yang-ATD | 60.40 | 70.97 | 77.53 | **7.71** | 10.38 |
| BKD | 61.71 | 71.97 | 78.70 | 8.11 | 10.25 |
| CBD | 61.89 | 72.23 | 78.31 | 7.86 | **10.19** |
| CalibKD | 60.22 | 70.58 | 77.23 | 7.96 | 10.40 |
| CCKD | 61.46 | 71.84 | 78.04 | 8.13 | 10.30 |
| SRW-KD | 61.70 | 71.90 | 78.51 | 8.13 | 10.21 |
| UMKD-UDD | 60.04 | 71.23 | 77.63 | 7.96 | 10.62 |
| CB-DistillGrad | 53.79 | 66.84 | 73.00 | 8.28 | 12.21 |
| EAKD-RCF | **61.96** | **72.55** | **79.07** | 7.97 | 10.30 |

**Table 2.** Mean performance of different methods across the nine imbalanced binary classification datasets.

As shown in Table 2, EAKD-RCF achieves mean MCC, F1-score, and G-Mean values of 61.96%, 72.55%, and 79.07%, respectively, representing the highest mean values among the evaluated methods for all three classification-oriented metrics. For probability quality, EAKD-RCF obtains an ECE of 7.97% and a Brier score of 10.30%. Although these values are not the lowest among all methods, they remain close to the lowest ECE obtained by Yang-ATD (7.71%) and the lowest Brier score obtained by CBD (10.19%). Thus, the improvement in imbalanced classification performance is not accompanied by a substantial deterioration in probability quality.

Compared with conventional KD, EAKD-RCF improves mean MCC, F1-score, and G-Mean by 0.95, 1.19, and 0.78 percentage points, respectively. Similar differences are observed relative to adaptive and uncertainty-aware distillation approaches. Compared with Yang-ATD, the corresponding improvements are 1.56, 1.58, and 1.54 percentage points, whereas improvements of approximately 1.92, 1.32, and 1.44 percentage points are observed relative to UMKD-UDD. In contrast, the probability-quality metrics do not follow the same ordering across methods, indicating that discrimination performance and probability quality are not necessarily optimized simultaneously.

The detailed results in Supplementary Table S1 further show that the relative performance of the methods varies across datasets, and EAKD-RCF does not dominate every individual dataset or metric. Therefore, the higher mean MCC, F1-score, and G-Mean should be interpreted as evidence of strong overall classification performance across heterogeneous datasets rather than uniform superiority on every individual task.

## Stage-wise and ablation analysis

As shown in Table 3, the student model alone achieves mean MCC, F1-score, and G-Mean values of 55.73%, 68.29%, and 75.02%, respectively. After entropy-driven adaptive knowledge distillation, the three metrics increase to 57.61%, 68.80%, and 75.70%. The ECE decreases from 8.52% to 7.80%, while the Brier score decreases from 11.44% to 10.70%. These results show that introducing teacher probability information improves the student model compared with hard-label training alone.

When fixed teacher–student fusion is applied, the mean MCC, F1-score, and G-Mean further increase to 59.99%, 70.66%, and 77.33%, respectively. The complete EAKD-RCF

obtains 61.96%, 72.55%, and 79.07%, representing improvements of 1.98, 1.89, and 1.74 percentage points over fixed fusion. The corresponding ECE and Brier score are 7.97% and 10.30%, compared with 8.05% and 10.45% for fixed fusion. Overall, the stage-wise results show a progressive improvement from the original student model to adaptive distillation and then to the final prediction stage.

The ablation results in Table 3 provide a further comparison of the main design choices. Removing the adaptive temperature reduces MCC, F1-score, and G-Mean from 61.96%, 72.55%, and 79.07% to 61.57%, 72.11%, and 78.61%, respectively. When the reliability-controlled distillation strength is removed, the corresponding values decrease to 61.57%, 71.95%, and 78.33%. Removing the candidate pool produces a larger reduction, with MCC, F1-score, and G-Mean decreasing to 60.63%, 71.25%, and 77.73%. By contrast, removing the Brier regularization leads to relatively small changes in the three classification metrics, while the Brier score changes from 10.30% to 10.34%. These results indicate that the candidate-selection stage has a comparatively larger effect on the final classification performance, whereas the Brier term mainly provides an auxiliary constraint on probability quality.

%

| Method | MCC | F1-score | G-Mean | ECE | Brier |
|---|---|---|---|---|---|
| **Stage-wise variants** | | | | | |
| Student only | 55.73 | 68.29 | 75.02 | 8.52 | 11.44 |
| EAKD student | 57.61 | 68.80 | 75.70 | 7.80 | 10.70 |
| EAKD-RCF (fixed fusion) | 59.99 | 70.66 | 77.33 | 8.05 | 10.45 |
| Full EAKD-RCF | 61.96 | 72.55 | 79.07 | 7.97 | 10.30 |
| **Component ablations** | | | | | |
| w/o adaptive temperature | 61.57 | 72.11 | 78.61 | 7.93 | 10.32 |
| w/o global reliability KD strength | 61.57 | 71.95 | 78.33 | 8.06 | 10.33 |
| w/o Brier regularization | 61.50 | 72.37 | 79.05 | 7.96 | 10.34 |
| w/o candidate pool | 60.63 | 71.25 | 77.73 | 7.95 | 10.33 |

**Table 3.** Stage-wise and ablation results of EAKD-RCF.

Figure 4 further examines the effect of the final prediction stage at the dataset level. As shown in Fig. 4a, the MCC differences between the complete EAKD-RCF and the simplified prediction variants vary across the nine datasets. The complete method provides clear improvements on several datasets, whereas the differences are small on some datasets and a simplified prediction form can occasionally perform similarly or better. This variation indicates that the benefit of the final prediction form is dataset dependent rather than uniform.

Figure 4b shows the prediction forms selected during the 225 outer-fold evaluations. Fixed fusion, adaptive fusion, and teacher-related candidates account for 35.6%, 12.4%, and 52.0% of the selected outputs, respectively. WPBC and Hepatitis more frequently retain fixed fusion, whereas teacher-related candidates are selected more often on Pima, Sick, and Heart. WDBC and Mammographic show more mixed selection patterns. Taken together, the dataset-level results show that no single prediction form is consistently preferred across all datasets, which supports selecting the final prediction from complementary candidates using internal validation.

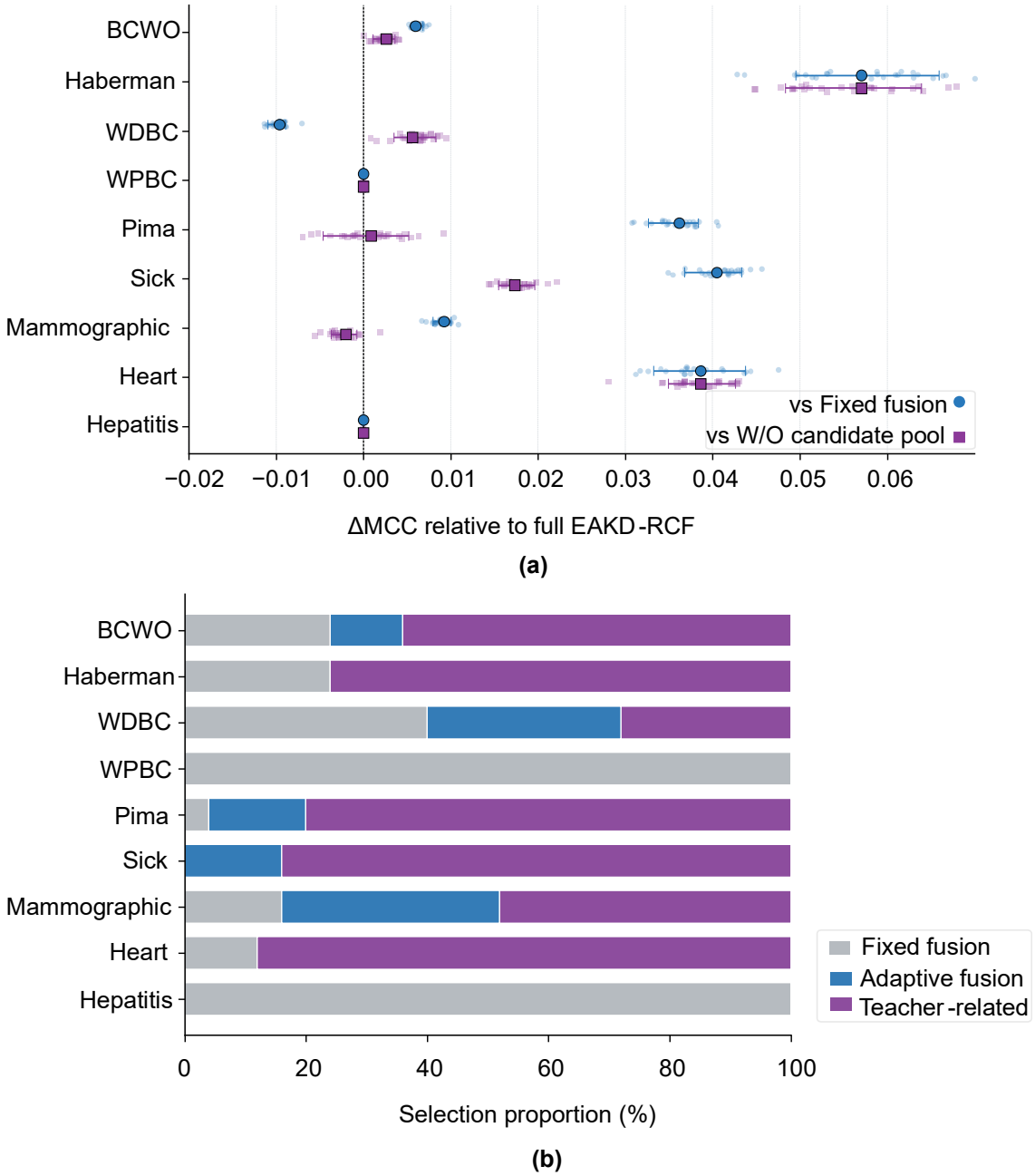



**Fig. 4.** Dataset-level performance differences and candidate-selection behavior of EAKD-RCF. (a) MCC differences between the full method and simplified prediction variants across the nine datasets. Positive values indicate higher MCC for the full EAKD-RCF. (b) Selection proportions of fixed fusion, adaptive fusion, and teacher-related candidates across outer-fold evaluations.

## Entropy–reliability analysis

Figure 5 examines the relationship between teacher predictive entropy, prediction correctness, and true-class reliability. As shown in Fig. 5a, among the 67,732 teacher predictions included in the analysis, 33,022 (48.75%) are low-entropy and correctly classified, whereas 843 (1.24%) are low-entropy but incorrectly classified. In the high-entropy group, 26,434 predictions (39.03%) remain correct, while 7,433 (10.97%) are incorrect. Although correct predictions account for a larger proportion of the low-entropy group, the coexistence of low-entropy errors and a large number of high-entropy correct predictions indicates that predictive entropy alone cannot uniquely determine prediction correctness.

Figure 5b further illustrates this relationship at the sample level. Correct predictions are mainly distributed in regions with higher true-class reliability, whereas incorrect predictions are more concentrated at lower reliability values. However, the two groups overlap substantially along the predictive-entropy axis. In particular, some low-entropy predictions still exhibit low true-class reliability, while some high-entropy predictions retain relatively high reliability. These results indicate that predictions with similar uncertainty levels can correspond to different levels of teacher reliability, and therefore predictive entropy should not be used directly as a substitute for teacher reliability.

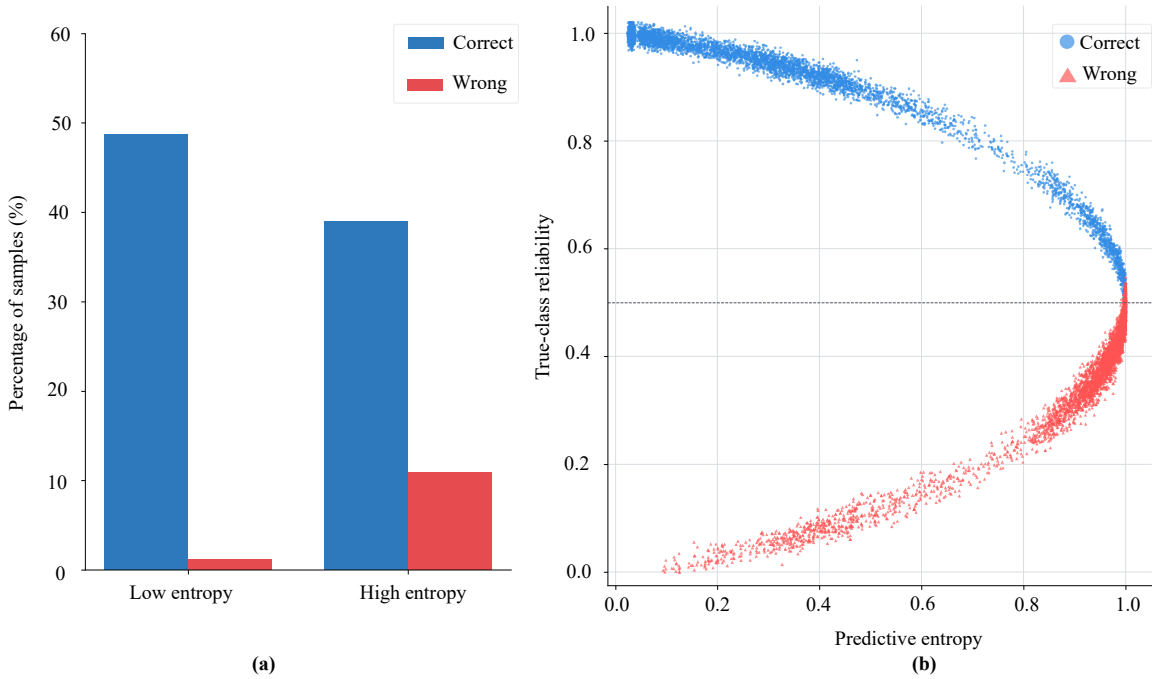



**Fig. 5.** Relationship between teacher predictive entropy, prediction correctness, and reliability. (a) Distribution of correct and incorrect predictions in the low- and high-entropy groups. (b) Sample-level relationship between predictive entropy and true-class reliability for correct and incorrect teacher predictions.

## Robustness and sensitivity analysis

Figure 6a compares the performance of EAKD-RCF using four different student architectures. The linear student obtains mean MCC, F1-score, and G-Mean values of 61.44%, 72.25%, and 78.76%, respectively. The corresponding values for MLP-32 are 62.09%, 72.79%, and 79.23%, while the predefined MLP-64-32 architecture achieves 61.96%, 72.55%, and 79.07%. Increasing the network size to MLP-128-64 results in 61.32%, 72.01%, and 78.65%. The maximum differences among the four architectures are only 0.77, 0.78, and 0.58 percentage points for MCC, F1-score, and G-Mean, respectively. Although MLP-32 produces slightly higher mean values, the differences are small, indicating that the reported performance is not restricted to a specific student architecture. The MLP-64-32 architecture is retained as the predefined default configuration rather than being selected according to the test results.

Figure 6b further shows the sensitivity of EAKD-RCF to the lower and upper RCF mixing bounds. Across the nine tested combinations, the mean MCC ranges from approximately 61.85% to 61.98%, corresponding to a maximum variation of only 0.13 percentage points. The predefined setting achieves an MCC of 61.96% and lies close to the highest value observed within the tested range. Thus, moderate changes in the RCF mixing

bounds have only a limited effect on the final classification performance. Additional sensitivity analyses for the entropy exponent $\gamma$, maximum knowledge-distillation strength $\lambda_{\max}$, Brier regularization weight $\mu_B$, reliability threshold $\tau$, and the joint $\gamma - \tau$ setting are provided in Supplementary Fig. S1. These additional results show larger performance changes when $\gamma$, $\lambda_{\max}$, or $\tau$ move farther from their predefined values, whereas changes in the Brier regularization weight have a comparatively smaller effect.

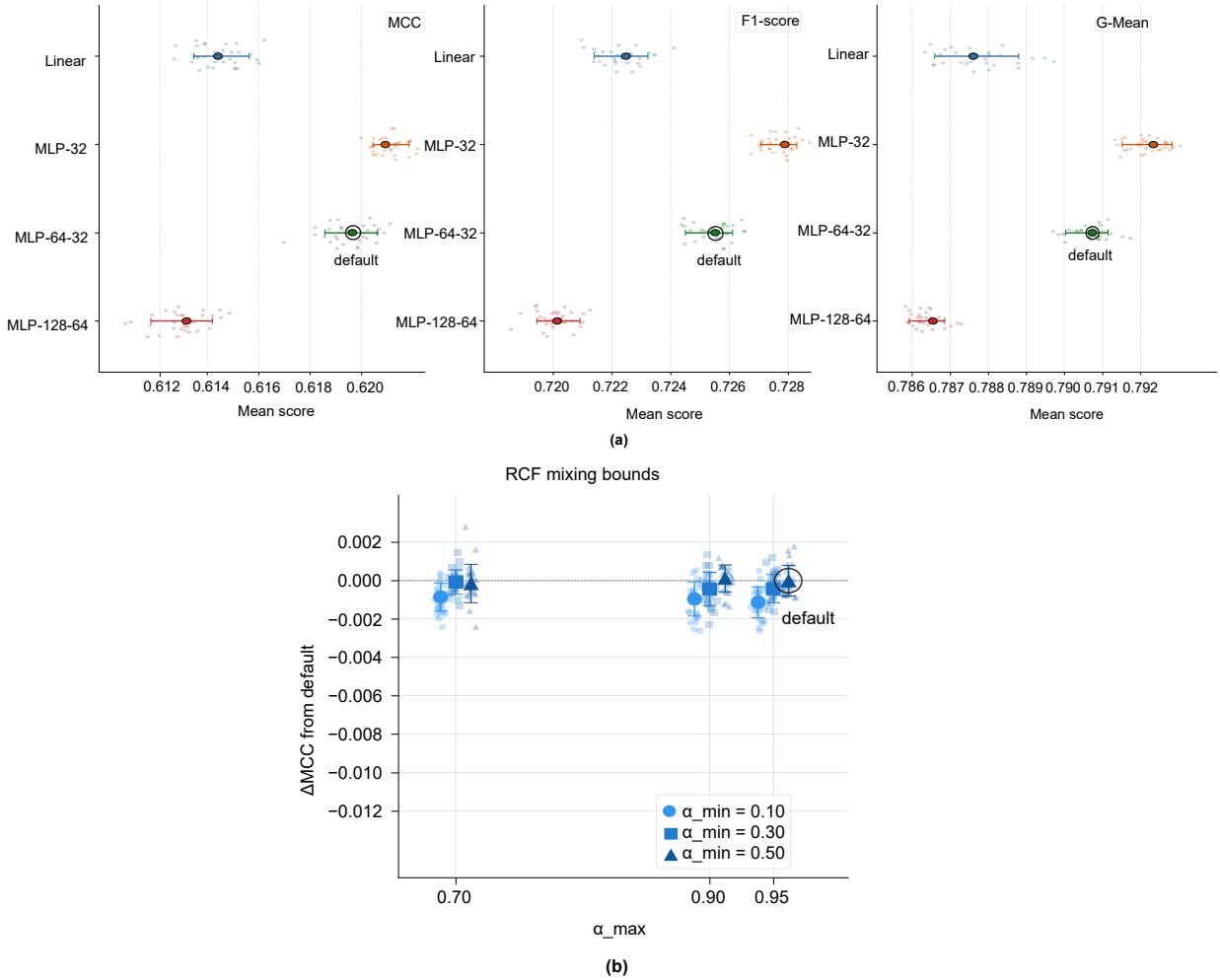



**Fig. 6.** Robustness and sensitivity of EAKD-RCF. (a) Mean MCC, F1-score, and G-Mean obtained using four student architectures. (b) Changes in MCC under different RCF mixing-bound settings relative to the predefined default configuration.

## Statistical comparison

The Friedman test was further used to compare the 15 methods across the nine datasets. Significant overall differences were observed for MCC ($\chi^2 = 37.30, p = 0.000664$), F1-score ($\chi^2$=35.03, $p = 0.001453$), G-Mean ($\chi^2 = 30.36, p = 0.006816$), and Brier score ($\chi^2 = 53.87, p = 1.35 \times 10^{-6}$), whereas no significant overall difference was observed for ECE ($\chi^2 = 10.46, p = 0.728177$).

Pairwise Wilcoxon signed-rank tests with Holm correction were subsequently performed between EAKD-RCF and the comparison methods. None of the corrected pairwise comparisons reached the 0.05 significance level, with the smallest adjusted $p$-value being 0.0547. These results indicate that, although overall differences exist among the evaluated methods for the main classification metrics and Brier score, the present nine-dataset evaluation does not provide sufficient evidence to claim that EAKD-RCF significantly outperforms every individual baseline. The statistical findings therefore support the more

conservative interpretation that EAKD-RCF achieves strong average classification performance and broad cross-dataset competitiveness rather than uniform statistical superiority.

## Discussion

Imbalanced binary classification remains challenging because minority-class recognition and probability reliability can be affected simultaneously by skewed class distributions. In this study, we proposed EAKD-RCF to improve classification performance while maintaining reliable probability estimates. The method uses predictive entropy to adapt knowledge transfer during distillation and employs validation-driven candidate selection to avoid relying on a fixed teacher–student prediction form.

Experimental results on nine datasets show that EAKD-RCF achieves mean MCC, F1-score, and G-Mean values of 61.96%, 72.55%, and 79.07%, respectively, which are the highest average classification results among the evaluated methods. Its ECE and Brier scores are 7.97% and 10.30%, indicating competitive probability quality. Although EAKD-RCF does not obtain the best result on every dataset or metric, the overall results demonstrate strong average classification performance across different data distributions.

Further analysis provides an explanation for this behavior. The entropy–reliability results show that low-entropy predictions can still be incorrect, whereas high-entropy predictions can remain correct, indicating that predictive uncertainty should not be directly treated as teacher reliability. This distinction is also relevant to previous findings showing that teacher calibration and probability quality can influence the effectiveness of knowledge transfer during distillation[16][18,19]. In addition, different datasets prefer different final prediction forms, suggesting that a fixed fusion strategy is not always suitable. These observations support separating uncertainty from reliability during distillation and selecting among complementary prediction candidates using internal validation.

In future studies, EAKD-RCF can be evaluated on more diverse datasets and extended to multi-class imbalance and long-tailed learning. More efficient candidate-selection and probability-calibration strategies can also be explored to further improve its applicability and computational efficiency.

## Conclusion

This study proposed EAKD-RCF for imbalanced binary classification. The method distinguishes predictive uncertainty from teacher reliability during adaptive knowledge distillation and uses internal validation to select among complementary teacher–student prediction forms. Experimental results on nine public datasets show that EAKD-RCF achieves mean MCC, F1-score, and G-Mean values of 61.96%, 72.55%, and 79.07%, respectively, while maintaining competitive ECE and Brier scores.

The results show that EAKD-RCF provides a competitive knowledge-distillation approach for imbalanced binary classification and can improve overall classification performance without substantial deterioration in probability quality. Future research can further evaluate the method on more diverse datasets and extend it to multi-class and long-tailed learning scenarios.

## Data availability

The nine benchmark datasets analysed in this study are publicly available. Eight datasets are available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/), and the Pima Indians Diabetes dataset was obtained from a historical public distribution of the UCI Machine Learning Repository. Dataset-specific source links and data-preparation instructions

are provided in the project repository at https://github.com/lijihang666/EAKD-RCF under DATA_SOURCES.md. No new datasets were generated during this study.

## References


[1] Cemernek, D., Siddiqi, S. & Kern, R. Effects of class imbalance countermeasures on interpretability. *IEEE Access* **12**, 45342–45358 (2024).

[2] Lin, F., Dai, Y. & Liu, Y. On probability estimation for unbalanced classification. *J. Stat. Theory Pract.* **20**, 57 (2026).

[3] Han, M., Li, C. P., Li, A. et al. Imbalanced stream classification algorithm based on weighted and dynamic selection. *Comput. Eng. Appl.* [in Chinese] **61**, 79–95 (2025).

[4] Meng, D. X., Yao, Y. F. & Yang, J. Ensemble classification algorithm for imbalanced data with dynamic undersampling based on micro-clusters. *Comput. Eng. Appl.* [in Chinese] **62**, 110–121 (2026).

[5] Liu, J. X., Li, D. W., Ren, L. J. et al. Multi-stage hybrid feature selection algorithm for imbalanced medical data. *Comput. Eng. Appl.* [in Chinese] **61**, 158–169 (2025).

[6] Li, Y. & Liu, S. Optimizing CatBoost with improved black-winged kite algorithm for fraud detection of imbalanced health insurance data. *Comput. Eng. Appl.* [in Chinese] **62**, 382–393 (2026).

[7] Nikpour, B., Rahmati, F., Mirzaei, B. et al. A comprehensive review on data-level methods for imbalanced data classification. *Expert Syst. Appl.* **295**, 128920 (2026).

[8] Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at https://arxiv.org/abs/1503.02531 (2015).

[9] Gou, J., Yu, B., Maybank, S. J. et al. Knowledge distillation: a survey. *Int. J. Comput. Vis.* **129**, 1789–1819 (2021).

[10] Ham, G., Cho, Y., Lee, J. H. et al. Difficulty level-based knowledge distillation. *Neurocomputing* **606**, 128375 (2024).

[11] Yang, S., Yang, X., Ren, J. et al. Adaptive temperature distillation method for mining hard samples' knowledge. *Neurocomputing* **636**, 129745 (2025).

[12] Li, L., Li, X. C., Ye, H. J. et al. Enhancing class-imbalanced learning with pre-trained guidance through class-conditional knowledge distillation. In *Proceedings of the 41st International Conference on Machine Learning* **235**, 28204–28221 (PMLR, 2024).

[13] Yu, T., Zhao, X., An, Y. et al. Knowledge distillation dealing with sample-wise long-tail problem. In *Computer Vision—ACCV 2024*, 2354–2370 (Springer, 2025).

[14] Tong, S., Gao, S., Liu, K. et al. Uncertainty-aware multi-expert knowledge distillation for imbalanced disease grading. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025*, 624–634 (Springer, 2025).

[15] Vemuri, S. M., Gundrapally, A., Xia, Z. et al. CB-DistillGrad: class-balanced distillation and gradient conflict resolution for low power edge AI applications. *IEEE Access* **14**, 3040–3061 (2026).

[16] Fan, W. S., Lu, S., Li, X. C. et al. Revisit the essence of distilling knowledge through calibration. In *Proceedings of the 41st International Conference on Machine Learning* **235**, 12882–12894 (PMLR, 2024).

[17] Hebbalaguppe, R., Baranwal, M., Anand, K. et al. Calibration transfer via knowledge distillation. In *Computer Vision—ACCV 2024*, 513–530 (Springer, 2025).

[18] Kim, S., Park, S., Lee, J. et al. The role of teacher calibration in knowledge distillation. *IEEE Access* **13**, 115548–115557 (2025).

[19] Mishra, I., Sethu, V. K. & Mishra, D. Beyond accurate distillation: calibrated knowledge distillation for reliable predictions. *IEEE Trans. Artif. Intell.* **7**, 1702–1714 (2026).

[20] Kelly, M., Longjohn, R. & Nottingham, K. The UCI Machine Learning Repository. *UCI Machine Learning Repository* https://archive.ics.uci.edu (2023).

[21] Cox, D. R. The regression analysis of binary sequences. *J. R. Stat. Soc. Series B Stat. Methodol.* **20**, 215–232 (1958).

[22] Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

[23] Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).

[24] Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers* (eds. Smola, A. J., Bartlett, P. L., Schölkopf, B. & Schuurmans, D.) 61–74 (MIT Press, 1999).

[25] Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).

[26] Cui, Y., Jia, M., Lin, T.-Y., Song, Y. & Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9268–9277 (2019).

[27] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1137–1145 (1995).
[28] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
[29] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* **70**, 1321–1330 (PMLR, 2017).
[30] Chen, C., Liaw, A. & Breiman, L. Using random forest to learn imbalanced data. Technical Report 666, Department of Statistics, University of California, Berkeley (2004).
[31] Liu, X.-Y., Wu, J. & Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. B Cybern.* **39**, 539–550 (2009).
[32] Zhang, S. Y., Chen, C., Hu, X. Y. et al. Balanced knowledge distillation for long-tailed learning. *Neurocomputing* **527**, 36–46 (2023).
[33] Iscen, A., Araujo, A., Gong, B. & Schmid, C. Class-balanced distillation for long-tailed visual recognition. In *Proceedings of the British Machine Vision Conference* 165 (2021).
[34] Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32**, 675–701 (1937).
[35] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80–83 (1945).
[36] Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

## Code availability

The source code and scripts used to reproduce the experiments, statistical analyses, tables, and figures in this study are publicly available at https://github.com/lijihang666/EAKD-RCF.

## Funding

This work was supported by the National Natural Science Foundation of China (Nos. 6226070934 and 62167009), the Key Research and Development Program of Shaanxi Province (No. S2023-YF-YBSF-0889), the Yulin Industry–University–Research Project (No. 2023-CXY128), the Yulin "Light of Science and Technology" Talent Program (No. 2023KJZG06), and the High-Level Talent Program of Yulin University (Nos. 22GK02 and 22GK04).

## Author contributions

J.L. conceived the study and developed the methodology and software. J.Z. supervised the study and contributed to the experimental design. J.L. performed the experiments and analysed the results. H.L., P.J., L.S. and H.Z. contributed to data analysis and manuscript revision. J.L. drafted the manuscript. All authors reviewed and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationEAKDRCF.pdf