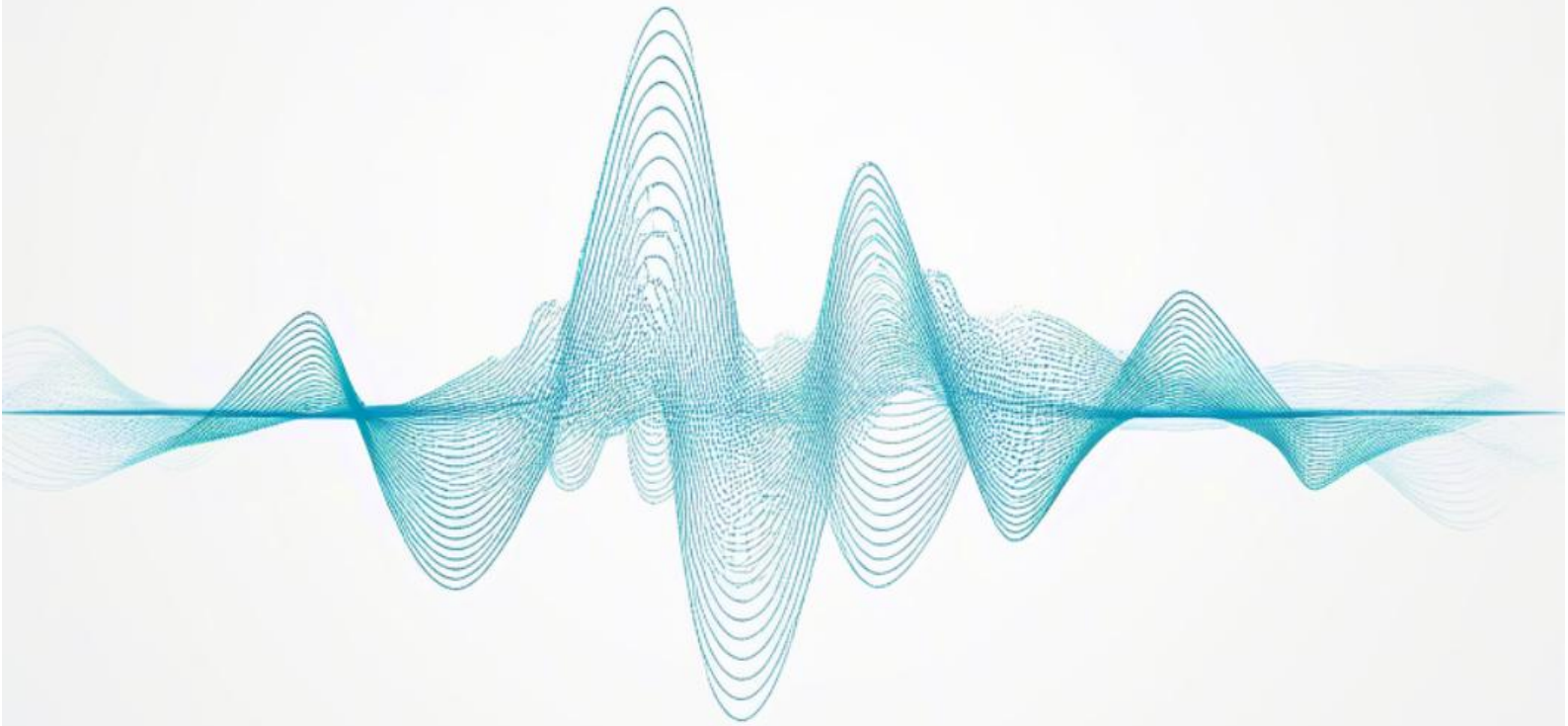
CSCHOLAR
Publishing INC

Volume 2 Number 1 2026

Journal of Computer Science and Digital Technology



https://jcsdt.cscholar.com ISSN: 3068-1448

Journal of Computer Science and Digital Technology

ISSN: 3068-1448

Volume 2, Issue 1, 2026

Twice a year

Cover Design: ConnectSix Scholar Publishing INC

Publishing Unit: ConnectSix Scholar Publishing INC

Publisher's website: http://www.cscholar.com/

Publisher's address:

6547 N Academy Blvd #2265

Colorado Springs CO 80918

US

Website of the journal *Journal of Computer Science and Digital Technology*:

https://jcsdt.cscholar.com/

Table of contents

Journal of Computer Science and Digital Technology, 2026, 2(1), 1-20
https://doi.org/10.71204/hefd5f67

Precise Identification of Multi-Regional Relative Poverty: A Two-Stage Knowledge-Distilled Adaptive Framework

Guanghuang Liu [1, *]

[1] School of Computer Science and Technology, Taiyuan Normal University, Shanxi 030619, China

*** Correspondence:**

Guanghuang Liu

lgh09263295@163.com

Received: 1 February 2026 / Accepted: 10 March 2026 / Published online: 11 March 2026

Abstract

As poverty reduction strategies shift comprehensively toward alleviating relative poverty, precisely identifying multidimensional relative poverty populations has become critical for social governance. However, existing data-driven models often face algorithmic bottlenecks—such as spatial heterogeneity, regional sample sparsity, and extreme category imbalance—when applied to complex scenarios characterized by vast territories and significant regional disparities. To address this, this study proposes a two-stage knowledge-distilled adaptive gradient boosting framework (TKDAF). First, in the prior knowledge extraction stage (Stage I), a base structure-regularized gradient boosting tree model is constructed. Combined with SHAP game-theoretic attribution, this stage quantifies and extracts the global objective weights of poverty-inducing features across regions. Second, in the spatial adaptive enhancement stage (Stage II), the S-DAGB (Spatial-Adaptive Distilled Gradient Boosting) core prediction model is introduced. It achieves deep integration of multiple regularization and feature enhancement mechanisms by incorporating: feature space nonlinear reconstruction based on prior knowledge, a dynamic category weighting mechanism based on effective sample size (ENS), and spatial adaptive optimization using the TPE Bayesian algorithm.Empirical results based on the 2020 China Family Panel Studies (CFPS) multidimensional dataset demonstrate that the S-DAGB model not only effectively overcomes the generalization bottleneck of deep tree models in the sample-constrained Northeast region (achieving 93.52% accuracy),but also significantly improves precision in regions with highly heterogeneous features and extreme class imbalance, such as central and western China. This effectively reduces wasteful allocation of poverty alleviation resources caused by false positive errors. This study provides an algorithmic solution that combines high accuracy with interpretability for precise identification of relative poverty in complex data distribution scenarios.

Keywords: Relative Poverty Prediction; Gradient-Boosted Trees; Knowledge Distillation; Spatial Heterogeneity; Extremely Imbalanced Classification; Bayesian Optimization

1. Introduction

As global poverty reduction efforts advance, China's poverty governance has historically shifted its focus from eliminating absolute poverty to alleviating relative poverty (Wang & Zeng, 2018). Unlike absolute poverty, which relies on a single income benchmark, relative poverty exhibits distinct multidimensionality, dynamism, and spatial heterogeneity. It encompasses not only economic deprivation but also the long-term accumulation of multidimensional disadvantages in education, health, living conditions, and assets (Wang et al., 2020; Sun et al., 2019). The Multidimensional Poverty Index (MPI), promoted by organizations such as the United Nations Development Programme (UNDP), provides a crucial framework for characterizing this form of poverty (Alkire & Foster, 2011). Against this backdrop, traditional identification methods relying on single thresholds or simple linear regression struggle to adequately meet the demand for precise policy interventions amid intersecting multidimensional characteristics.In recent years, machine learning algorithms such as Random Forest and Gradient Boosting Trees have demonstrated immense application potential in micro-level household feature extraction and poverty identification due to their exceptional ability to capture nonlinear feature interactions. They have gradually become a leading paradigm in computational social science (Shi et al., 2024; Jean et al., 2016).

However, existing data-driven relative poverty prediction models still face three severe algorithmic bottlenecks when applied to China's complex micro-survey data, characterized by vast geographical scope and significant regional disparities. The first challenge is regional data sparsity and overfitting risk. When predictions are narrowed to specific regions (such as central-western and northeastern China), the available sample size often shrinks dramatically. Complex deep ensemble tree models like XGBoost (Chen et al., 2016) and LightGBM (Ke et al., 2017) are highly susceptible to local noise interference in small-sample scenarios, severely limiting their generalization capabilities (Qian et al., 2022). Second is the extreme positive-negative class imbalance. After entering the phase of routine poverty governance, the poor population typically exhibits a long-tail distribution within the overall sample. Conventional loss functions, when optimizing such extremely imbalanced data, often sacrifice recall for the minority class (poverty samples) to achieve overall accuracy. Traditional resampling methods also tend to distort the original distribution (Chawla et al., 2002), making it difficult to align with the policy goal of "accurate identification."Finally, spatial heterogeneity in feature contributions and the absence of prior knowledge pose significant challenges. Authoritative development economics research indicates that core factors driving poverty exhibit pronounced structural and spatial imbalances across different regions (Zou & Fang, 2011; Ravallion & Chen, 2007). Existing end-to-end models predominantly explore raw feature spaces without prior constraints, not only reducing information extraction efficiency in small samples but also trapping models in a "black box" dilemma, lacking sociologically meaningful interpretability.

To address these challenges, this study proposes a Two-Stage Knowledge-Distilled Adaptive Gradient Boosting Framework (TKDAF).This framework innovatively decouples the prediction process into two stages: "prior knowledge extraction" and "spatial adaptive boosting." In Stage I, a base boosting tree model with strict structural regularization is constructed. The SHAP (Shapley Additive Explanations) game theory model is then employed to extract the attribution distribution of global impoverishment features as "prior knowledge" (Lundberg et al., 2017). In Stage II, the S-DAGB adaptive model is proposed, achieving high-precision, interpretable poverty identification through multi-mechanism fusion. The main contributions of this paper are as follows:

(1) We introduce a knowledge-driven nonlinear feature space reconstruction strategy. Using the SHAP prior weights extracted in Stage I as guiding factors, we reconstruct the input feature space through an adaptive scaling formula. This guides the underlying tree model to prioritize core impoverishment factors during tree splitting, significantly enhancing feature resolution in small-sample scenarios.

(2) We propose a dynamic weighting mechanism based on Effective Number of Samples (ENS). Addressing extreme data imbalance, it redefines the category penalty weight in the loss function by drawing from ENS theory (Cui et al., 2019), establishing a smoothing buffer layer within the sample space to effectively mitigate model bias toward majority-class dominant features.

(3) We designed a spatially adaptive regularization architecture based on TPE Bayesian optimization. By leveraging the Tree-like Parzen Estimator (TPE) (Bergstra et al., 2011) for joint optimization across multidimensional hyperparameter spaces, the model automatically adapts its underlying sampling engine and regularization terms according to regional data scale and heterogeneity, constructing a robust defense against overfitting.

2. Related Research and Model Framework

2.1. Relative Poverty Prediction and Multidimensional Feature Analysis

Since Alkire and Foster introduced the A-F dual-threshold measurement method, poverty metrics have expanded from a single economic dimension to encompass multidimensional relative poverty across health, education, and living standards (Alkire & Foster, 2011). Focusing on China's context, scholars have observed that regional disparities in resource endowments and economic development gradients result in strongly spatially unbalanced and multidimensionally dynamic trajectories of relative poverty (Zou & Fang, 2011; Ravallion & Chen, 2007). For instance, relative deprivation in the developed eastern regions manifests primarily as shortages of high-value assets and quality educational resources, whereas western and northeastern regions are more constrained by deficiencies in infrastructure and basic healthcare (Li et al., 2020). Traditional prediction methods based on a national-level perspective often obscure these structural regional disparities (Wang et al., 2022). Therefore, incorporating multidimensional feature analysis and designing algorithmic mechanisms at the model level that can adaptively

capture latent regional heterogeneity are crucial for enhancing the targeted identification of relative poverty.

2.2. Gradient Boosting Decision Trees (GBDT)

Ensemble learning possesses inherent advantages in handling complex tabular data (Breiman, 2001). Among these, Gradient Boosting Decision Trees (GBDT), a classic paradigm proposed by Friedman (Friedman, 2001), operates by iteratively accumulating multiple weak learners to progressively fit the negative gradient between the previous model's prediction and the true value.In recent years, with the widespread adoption of efficient frameworks like XGBoost (Chen et al., 2016) and LightGBM (Ke et al., 2017), gradient boosting algorithms have demonstrated significant engineering advantages in capturing nonlinear interactions among multidimensional features.However, when applied to micro-household survey data—particularly in sparsely sampled regions—conventional tree models tend to overfit local noise during node splitting due to insufficient information entropy constraints (Qian et al., 2022). Consequently, stringent structural regularization strategies (e.g.,L_2 leaf node penalties or dynamic feature sampling) are required to limit the complexity of individual trees, thereby ensuring generalization capability across regions.

2.3. Bayesian Hyperparameter Optimization Theory

Within complex machine learning ensemble frameworks, hyperparameter configuration directly determines a model's performance ceiling on specific data distributions. Traditional grid search not only incurs exponentially increasing computational costs but also struggles to locate global optima in continuous high-dimensional spaces. To address this challenge, Bayesian optimization theory based on probabilistic surrogate models emerged (Snoek et al., 2012). Among these, the Tree-like Parzen Estimator (TPE) algorithm proposed by Bergstra et al. constructs a probability distribution of the objective function across the hyperparameter space (Cui et al., 2019). By continuously computing the Expected Increment (EI) using historical evaluation results, it achieves an optimal balance between "Exploration" and "Exploitation."This mechanism enables deep tree models to break free from manual experience constraints, autonomously searching for optimal regularization configurations across varying data sparsity levels in different regions.

2.4. Overall Framework of the TKDAF Model

To effectively address challenges in multi-regional data—such as overfitting in small samples, extreme sample imbalance, and spatial heterogeneity—this study proposes a two-stage Knowledge Distillation Adaptive Gradient Boosting framework (TKDAF). This framework breaks the "black-box" paradigm of traditional end-to-end models by decoupling the prediction process into two stages: "knowledge extraction" and "adaptive enhancement"(overall architecture shown in Figure 1).

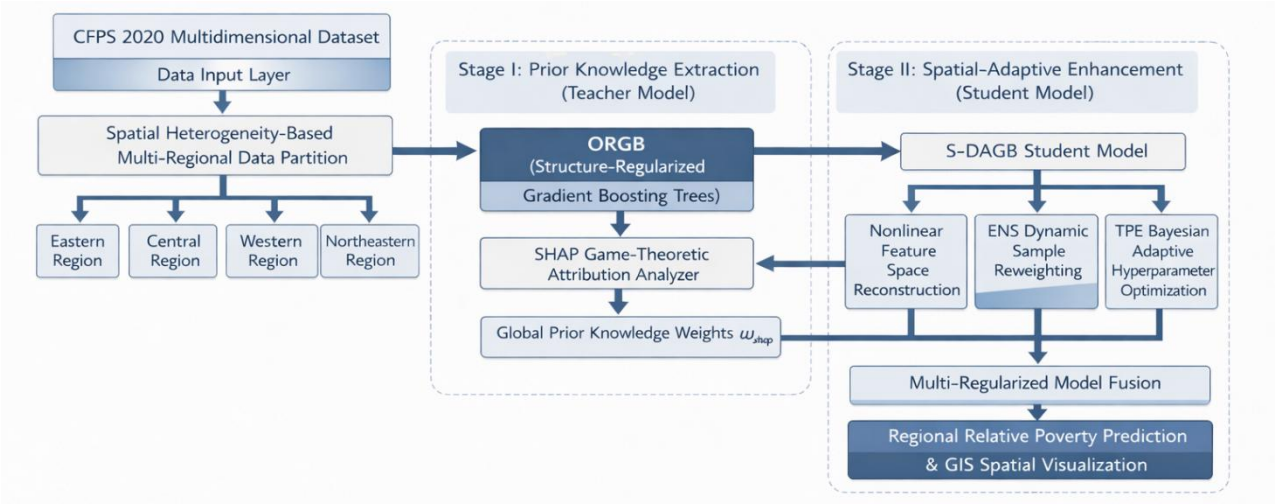


Figure 1. Overall framework architecture of TKDAF

(1) Stage I: Prior Knowledge Extraction Layer. The core objective of this stage is to extract objectively interpretable poverty patterns with sociological significance from geographically independent datasets. The framework first trains a strongly constrained basic objective regularized gradient boosting model (ORGB) on different geographic subsets.Upon model convergence, the SHAP game-theoretic attribution analyzer is applied to precisely quantify the marginal contribution of each multidimensional indicator to poverty status (Zou & Fang, 2011). These SHAP distributions, stripped of local noise, are converted into "global prior knowledge weights" and transferred losslessly to the next stage.

(2) Stage II: Spatially Adaptive Gradient Boosting (S-DAGB). This stage constitutes the core prediction engine—the S-DAGB model. Receiving raw regional data and Stage I prior knowledge, it achieves predictive performance leaps through three mechanisms: First, it performs nonlinear feature space reconstruction, adaptively scaling features using prior knowledge weights to amplify decision resolution for core impoverishment factors;Second, it introduces the ENS dynamic sample weighting mechanism (Ravallion & Chen, 2007) to reshape the category penalty weights in the loss function, countering gradient bias caused by long-tail imbalance (Lundberg & Lee, 2017). Finally, driven by TPE Bayesian adaptive optimization (Cui et al., 2019), the model automatically adapts to the data characteristics of the current region within a deep regularization space, outputting the final high-precision regional poverty identification results.

Unlike existing research paradigms in feature engineering, the knowledge distillation mechanism proposed in this study represents a significant methodological innovation. Current literature on interpretable machine learning (e.g., standard SHAP analysis) predominantly focuses on 'post-hoc explanation' —analyzing the black-box logic of trained models. The TKDAF framework innovatively shifts this approach by converting objective poverty attribution into 'pre-hoc prior knowledge.' Through establishing a nonlinear feature space reconstruction mechanism, the extracted global knowledge weights are seamlessly integrated into the underlying splitting logic of the student model (S-DAGB). This paradigm shift from 'passive interpretation' to 'active reconstruction' effectively guides tree models in small-sample regions to overcome noise traps, achieving truly knowledge-guided distillation.

Furthermore, conventional approaches to addressing extreme data imbalance and spatial heterogeneity predominantly rely on physical-level data resampling techniques (e.g., SMOTE algorithms), which may significantly distort the original multidimensional data's true spatial distribution. In contrast, this study introduces an ENS-based loss function reconstruction strategy within a spatially adaptive architecture, establishing a smoothing buffer layer through gradient optimization's mathematical foundation. The TPE engine-driven spatial adaptive framework dynamically activates underlying stochastic control and structural regularization constraints based on regional data sparsity (e.g., the sparse micro-samples in Northeast China versus the massive samples in the East). This non-invasive joint optimization solution fundamentally resolves the algorithmic bottleneck of traditional single-model approaches, which struggle to accommodate multi-regional heterogeneity while preserving the original spatial topology of regional data.

3. Research Methodology

3.1. Stage I: Extracting Prior Knowledge Based on Structured Regularized Gradient Boosting Trees

In the first stage, this study constructs an Objective Regularized Gradient Boosting (ORGB) model with structured regularization constraints to capture global objective patterns of poverty-inducing factors across different regions. Compared to directly feeding raw data into complex networks, using ORGB as a "teacher model" for preliminary data structure exploration effectively filters redundant noise.

To address the interpretability limitations of black-box models, this study incorporates the SHAP (Shapley Additive Explanations) game-theoretic attribution method to conduct in-depth analysis of ORGB predictions across regional datasets. For any sample's feature i , its SHAP attribution value ϕ_i is calculated as follows:

$$\phi_i(f,x)=\sum_{S\subseteq N\setminus\{i\}}\frac{|S|!(|N|-|S|-1)!}{|N|!}[f_x(S\cup\{i\})-f_x(S)]$$

Where N denotes the entire feature set, S represents the feature subset excluding i, and $f_x(S)$ is the model's predicted value on the subset S .

To obtain globally informative prior knowledge, we calculate the mean of the absolute SHAP values across all samples within a single region and normalize it to derive the global knowledge weight $w_{shap}^{(i)}$ for the i th feature in that region:

$$w_{shap}^{(i)}=\frac{\frac{1}{M}\sum_{j=1}^{M}|\phi_i^{(j)}|}{\sum_{k=1}^{K}\left(\frac{1}{M}\sum_{j=1}^{M}|\phi_k^{(j)}|\right)}$$

Where M denotes the total number of samples in the region, and K represents the total feature dimension. This weight precisely quantifies the actual contribution of each dimensional metric to poverty status in the physical world and is losslessly transferred as prior knowledge to Stage II.

3.2. Stage II: S-DAGB Spatially Adaptive Distilled Gradient Boosting Model

Guided by the prior knowledge extracted in Stage I, this stage introduces the full S-DAGB (Spatial-Adaptive Distilled Gradient Boosting) model. Through three core mechanisms, this model comprehensively addresses the bottlenecks of spatial heterogeneity and overfitting in small samples.

3.2.1. Unevenly Distributed Weight Reallocation Based on Effective Sample Size (ENS)

Relatively poverty data often exhibits extreme positive-negative class imbalance. Traditional resampling methods can distort the true data distribution, while simple inverse-frequency weighting may lead to model overfitting on outliers in the minority class. This study introduces the Effective Number of Samples (ENS) theory to redefine class weights. The ENS is defined as follows:

$$E_n=\frac{1-\beta^n}{1-\beta}$$

Where n represents the actual sample count for a specific class, and $\beta\in[0,1)$ denotes the decay coefficient (hyperparameter). As n increases, the new information provided by marginal samples diminishes progressively. Based on this, we make the weight α_c of each class in the loss function inversely proportional to its effective sample count and apply a normalization mapping:

$$W_c=\frac{1/E_{n_c}}{\sum(1/E_{n_i})}\times C$$

where C denotes the total number of classes. This mechanism penalizes majority class advantage while establishing a smoothing buffer layer, effectively suppressing gradient explosion in extremely imbalanced scenarios.

3.2.2. Knowledge-Guided Nonlinear Feature Space Reconstruction

To enhance the S-DAGB model's focus on key impoverishment factors during tree splitting, we propose a feature space reconstruction strategy based on Stage I prior knowledge. Through an adaptive scaling mechanism, the numerical distribution range of high-contribution features undergoes nonlinear stretching, as expressed by the following formula:

$$X_i'=X_i\times(1+w_{shap}^{(i)}\times\alpha)$$

where X_i represents the original input feature, $w_{shap}^{(i)}$ denotes the global prior weight passed from Stage I, and α is the feature scaling factor. This formula ensures stronger correlated features possess higher resolution during spatial partitioning, effectively preventing ineffective splits on noisy features in small-sample tree models.

3.2.3. TPE-Based Multi-Region Hyperparameter Adaptive Optimization and Regularization

Given the substantial differences in sample size and feature heterogeneity across Eastern, Central, Western, and Northeastern regions, a single parameter configuration cannot achieve global optimization. This study employs a Bayesian optimization algorithm based on the Tree-

structured Parzen Estimator (TPE) to jointly optimize the core regularization architecture and adaptive parameters of S-DAGB.

The optimization space encompasses not only the feature amplification factor (α) and ENS decay coefficient (β), but also deeply integrates structured regularization terms, including:

(1) Tree-structured regularization: Maximum tree depth (Depth) and minimum data in leaf nodes (Min Data In Leaf).

(2) Weight Penalty: Leaf node L_2 regularization coefficient (L_2Leaf Reg), strictly controlling model complexity.

(3) Dynamic Feature Sampling: Introduces hierarchical column sampling rate (Colsample By Level) to forcefully break excessive model dependence on any single dominant feature.

Simultaneously, the model incorporates a spatial scale adaptation mechanism at its core: - In regions with sufficient sample size (e.g., eastern areas), Bernoulli sampling is employed alongside subsampling regularization. - In sparsely sampled regions (e.g., northeast and central areas), the model automatically switches to Bayesian sampling mode and introduces the Bagging Temperature parameter to control randomness. This constructs multiple layers of defense against overfitting.

4. Experimental Results and Analysis

4.1. Dataset and Evaluation Metrics

This study conducts empirical analysis based on the 2020 China Family Panel Studies (CFPS) multidimensional dataset. To accurately capture the spatial heterogeneity of multidimensional relative poverty and validate the model's generalization capability under complex imbalanced distributions, the cleaned national sample is strictly divided into four independent subsamples according to geographical regions: Eastern Region (3,797 samples), Central Region (2,481 samples),Western Region (2,964 samples), and Northeastern Region (1,548 samples). The characteristics of each dataset are shown in Table 1 (using the Northeastern Region as an example).

It is particularly noteworthy that this dataset exhibits extreme spatial imbalance and scale heterogeneity. The available sample size for the Northeast region is exceptionally sparse, amounting to only 40.8% of the Eastern region's sample size.This severe "small-sample" distribution poses significant challenges to the overfitting resilience of complex machine learning models (such as ensemble tree algorithms) and provides an excellent experimental setting to validate the effectiveness of the S-DAGB adaptive regularization framework proposed in this paper. After data preprocessing and feature selection via analysis of variance (ANOVA), core dimensional features including household annual income, children's education, assets, and health status were retained for model input.

Table 1. Descriptive statistics of relative poverty data by region

Region	Variable	Meaning	count	mean	std	min	0.25	0.50	0.75	max
Northeast	x1	Adult Education	1548	0.76	0.43	0.0	1.0	1.0	1.0	1.0
	x2	Children Education	1548	0.14	0.34	0.0	0.0	0.0	0.0	1.0
	x3	Adult Health	1548	2.61	1.35	1.0	1.0	3.0	3.0	5.0
	x4	Children Health	1548	3.86	0.53	1.0	4.0	4.0	4.0	4.0
	x5	Medical Insurance	1548	0.17	0.37	0.0	0.0	0.0	0.0	1.0
	x6	Cooking Fuel	1548	3.71	1.82	1.0	3.0	4.0	6.0	6.0
	x7	Drinking Water	1548	2.81	0.50	1.0	3.0	3.0	3.0	4.0
	x8	Housing	1548	2.62	0.77	1.0	3.0	3.0	3.0	3.0
	x9	Assets	1548	36803.64	80691.22	0.0	2.0×10^3	9.0×10^3	3.43×10^4	1.3×10^6
	x10	Annual household income	1548	62627.23	85312.97	0.0	2.6×10^4	5.0×10^4	8.0×10^4	2.6×10^6
	y0	Real number type ActualAnnual income	1548	26743.24	69852.63	0.0	1.0×10^4	2.0×10^4	3.33×10^4	2.6×10^6
	y1	Category-based actualAnnual income	1548	0.79	0.40	0.0	1.0	1.0	1.0	1.0

To comprehensively evaluate the recognition performance of the TKDAF framework and its comparison models, this study employs five core metrics: Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC). Their calculation formulas are as follows:

(1) Accuracy: The percentage of correctly classified samples relative to the total number of samples, measuring the overall predictive accuracy of the model.

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

(2) Precision: The ratio of samples predicted as positive by the model that are actually positive, focusing on false positives.

$$Precision=\frac{TP}{TP+FP}$$

(3) Recall: The percentage of true positives correctly predicted, primarily used to measure false negatives.

$$Recall=\frac{TP}{TP+FN}$$

(4) F1 score: The harmonic mean of precision and recall, used to evaluate the model's performance in positive class prediction.

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall}$$

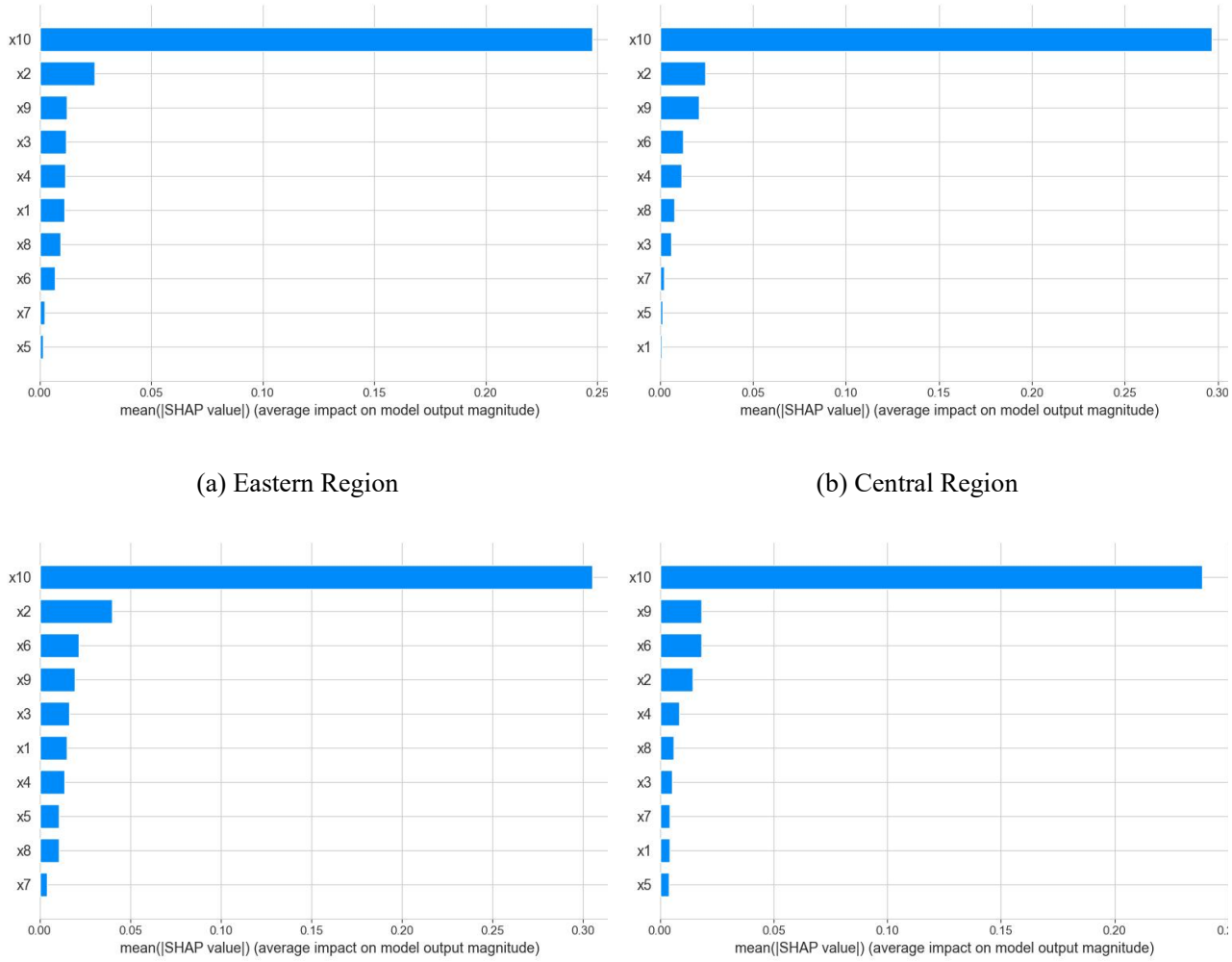
(5) AUC value: The area under the receiver operating characteristic (ROC) curve, measured as the area between the curve and the x-axis. It reflects the model's ability to distinguish between positive and negative classes. A value closer to 1 indicates better performance and greater model accuracy.

WhereTP represents correctly predicted poor samples,TN denotes correctly predicted non-poor samples,FP indicates false positive samples, andFN signifies false negative samples.

To ensure the objectivity of model evaluation and the reproducibility of results, this study adopted strict data partitioning and cross-validation strategies during the experimental design phase. For the independent datasets from four major regions, stratified sampling was used to divide the data into training and testing sets in a 80:20 ratio, ensuring that the initial distribution of positive and negative samples remained consistent before and after partitioning. Additionally, considering the extreme imbalance in data from the central and western regions, as well as the small sample size in the northeastern region, this study implemented stratified 10-fold cross-validation during both model training and TPE hyperparameter optimization. This strategy effectively mitigated the potential randomness in evaluations caused by single data splits, further enhancing the generalization and robustness of the S-DAGB model in complex multi-regional scenarios.

4.2. Stage I: SHAP-Based Extraction of Multidimensional Feature Prior Knowledge

In the first stage of the TKDAF framework, this study first independently trained a base gradient boosting model with structural regularization (Teacher Model) across four geographic subsets. The SHAP game theory model was then introduced to attribute and decompose the predictive contributions of each dimensional feature.The extracted Mean |SHAP value| intuitively reflects the objective weight of each feature in the global space. The SHAP contribution rankings across the four regions are shown in Figure 2.



(a) Eastern Region

(b) Central Region

(c) Western Region

(d) North-Eastern Region

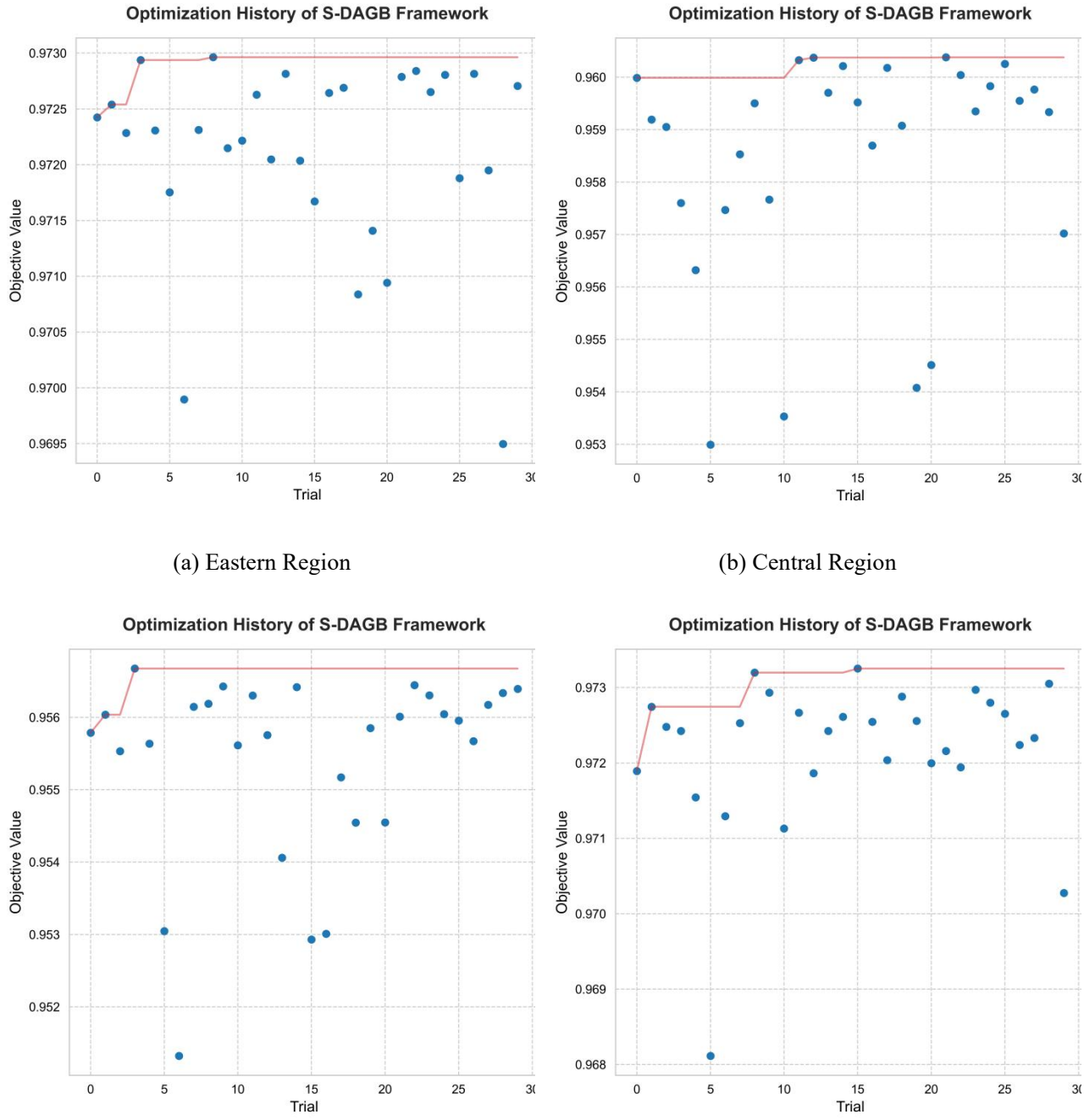
Figure 2. Comparison of mean absolute SHAP values across regions

The quantitative results in Figure 2 clearly reveal the spatial heterogeneity of the impoverishment logic:Although annual household income (x10) holds absolute dominance across all regions, developmental indicators like children's education (x2) and assets (x9) contribute significantly in eastern and central regions. Conversely, basic survival indicators such as cooking fuel (x6) and housing (x8) exhibit markedly elevated weights in western and northeastern regions.

Contrary to traditional sociological interpretations, the core purpose of extracting these distributions is to convert them into machine-readable "prior knowledge." By applying L_1 normalization to the absolute SHAP mean values corresponding to each bar chart, we generated specific prior knowledge weight vectors for the four regions: w_{shap} .After crossing the Stage I boundary, this vector is losslessly injected into the Stage II S-DAGB student model as a quantitative guidance factor for executing "feature space nonlinear reconstruction." This forces the underlying tree model to prioritize impoverishing features with high true physical significance during small-sample tree splitting, effectively shielding against local data noise.

4.3. Stage II: Bayesian Space Adaptive Optimization and Importance Analysis

To ensure the generalization capability of the S-DAGB framework across different regional datasets, this study introduces a Bayesian optimization algorithm based on the Tree-like Parzen Estimator (TPE) in Stage II. This algorithm performs joint optimization within a regularization space composed of feature reconstruction coefficients (Scale Multiplier), effective sample weighting decay rates (Beta), and multi-tree structure parameters.Figures 3 and 4 respectively illustrate the optimization history and hyperparameter importance distribution across four major geographic regions.



(a) Eastern Region

(b) Central Region

(c) Western Region

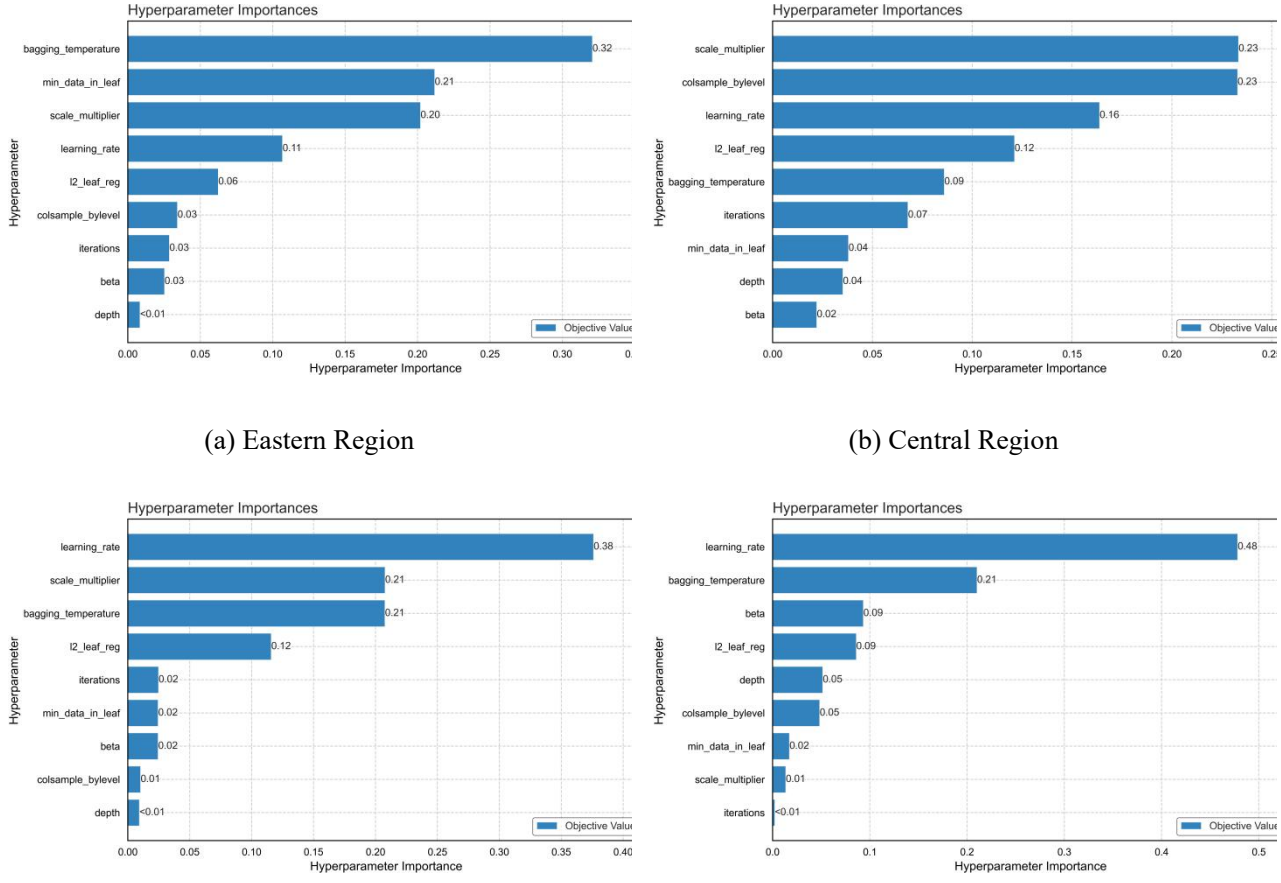
(d) North-Eastern Region

Figure 3. Comparison of optimization history across regions

(Note: Blue dots represent the objective value of a single trial, and the red line represents the best value envelope)

As shown in the optimization history trajectory of Figure 3, the TPE algorithm demonstrates exceptional search efficiency across all regions. The scatter points represent the objective function values from individual trials, while the red line denotes the envelope of the best values explored at each stage.It can be observed that after an initial exploration phase (approximately the first 5-10 trials), the red line exhibits a significant upward leap and converges rapidly to a suboptimal or optimal plateau region. This trajectory intuitively demonstrates that the TPE mechanism effectively reduces the computational overhead of traditional grid search while preventing the model from getting stuck in local optima.

Further examination of the hyperparameter importance bar chart in Figure 4 reveals that the S-DAGB model demonstrates a pronounced adaptive adjustment mechanism in response to data heterogeneity across different regions:



(a) Eastern Region

(b) Central Region

(c) Western Region

(d) North-Eastern Region

Figure 4. Comparison of hyperparameter importance across regions

(1) Validation of the Effectiveness of the A-priori Knowledge Reconstruction Mechanism: As shown by the bar charts for the eastern, central, and western regions in Figure 4, the core parameter scale_multiplier—which controls the strength of SHAP a-priori knowledge injection—exhibited exceptionally high decision weights, ranking third (0.20), first (0.23), and second (0.21)

in importance respectively.This intuitively indicates that in regions with relatively abundant sample information or significant feature distribution differences, the optimization engine adaptively tends to rely on the objective prior weights provided by Stage I. By amplifying the resolution of core impoverishment factors, it enhances overall prediction performance.

(2) Strong Regularization Constraints in Small-Sample Regions: Particularly noteworthy is the Northeast region, which has the most limited sample size (only 1,548 cases). As shown in Figure 4 (Northeast Region), the model does not rely solely on feature reconstruction. Instead, it assigns extremely high weights to the underlying randomness control parameters: bagging_temperature (0.21), dynamic sample weight decay rate beta (0.09), and leaf node penalty term l2_leaf_reg (0.09).This phenomenon clearly demonstrates that when confronted with sparse, small-sample data, the TPE optimization engine intelligently activates and enhances the underlying structural constraints of S-DAGB. This creates a smoothing buffer layer that effectively mitigates the overfitting risks commonly encountered in small-sample scenarios.

4.4. Multi-Region Performance Comparison and Analysis of S-DAGB

To comprehensively evaluate the predictive efficacy of the overall TKDAF framework, this study conducted comparative experiments between the full S-DAGB model and baseline models including Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), XGBoost, and LightGBM. Table 2 details the performance of each model across five evaluation metrics in four major regions, while Figure 5 plots the corresponding ROC curves.Experimental results demonstrate that S-DAGB exhibits superior overall performance in multi-region relative poverty prediction tasks:

(1) Model Generalization Analysis Under Small-Sample Constraints (Using Eastern and Northeast Regions as Examples). In the Eastern region with relatively rich sample features and the Northeast region with the most limited sample size, S-DAGB achieved optimal performance across multiple core metrics.Notably, in Northeast China, traditional tree models (e.g., CART decision trees) exhibited varying degrees of performance degradation on small datasets due to insufficient information constraints (CART accuracy reached only 89.47%).In contrast, S-DAGB leveraged prior knowledge injected in Stage I and combined it with regularization strategies selected via TPE. This approach not only maintained high stability but also elevated accuracy to 93.52% and achieved an F1-Score of 95.81%. This demonstrates the framework's effectiveness in mitigating the limitations of deep tree models' generalization capabilities under small-sample conditions.

(2) Precision Optimization Under Severely Imbalanced Data Distribution (Using Central and Western Regions as Examples). In regions like the central and western areas, characterized by high feature heterogeneity and extreme imbalance between positive and negative samples, the baseline teacher model ORGB demonstrates competitive performance in AUC. However, S-DAGB achieves more significant improvements in Precision and F1-Score. For instance, in the central region, S-DAGB achieves a precision of 93.91% (exceeding ORGB's 93.18%); in the western region, precision rises to 91.22%. In practical poverty alleviation applications, traditional prediction models often prioritize high recall, which can lead to elevated false positive rates and

result in inefficient allocation of poverty relief resources.By introducing an Effective Sample Weighting (ENS) mechanism, S-DAGB constructs a smoothing buffer layer at the loss function level, effectively suppressing model bias toward majority classes. This mechanism significantly reduces redundant false positives while maintaining high recall, better aligning with the policy objective of "accurate identification."

Table 2. Comparison Table of Regional Model Performance

Region	Model	Accuracy	Precision	Recall	F1	AUC
Eastern	LR	87.73%	88.64%	97.26%	93.02%	96.51%
	KNN	92.02%	94.10%	96.34%	95.18%	95.44%
	SVM	91.36%	94.83%	94.66%	94.71%	96.56%
	RF	92.02%	94.29%	96.11%	95.17%	96.39%
	CART	89.52%	93.98%	93.19%	93.56%	84.72%
	XGBoost	91.76%	94.33%	95.74%	95.01%	96.55%
	LightGBM	92.15%	94.59%	95.93%	95.25%	97.19%
	ORGB	93.39%	94.86%	97.31%	96.03%	96.59%
	S-DAGB	93.42%	95.17%	97.38%	96.30%	97.23%
Central	LR	88.02%	90.31%	94.97%	92.45%	94.88%
	KNN	89.92%	92.56%	94.39%	93.45%	92.65%
	SVM	89.68%	92.60%	93.96%	93.27%	95.22%
	RF	90.36%	92.77%	94.76%	93.73%	94.63%
	CART	86.13%	91.33%	90.42%	90.84%	81.61%
	XGBoost	89.96%	92.73%	94.25%	93.45%	94.94%
	LightGBM	89.96%	92.58%	94.40%	93.45%	95.49%
	ORGB	91.29%	93.18%	95.57%	94.35%	97.31%
	S-DAGB	91.45%	93.91%	94.90%	94.40%	97.01%
Western	LR	78.45%	80.04%	93.46%	86.03%	92.36%
	KNN	86.13%	88.97%	91.00%	89.96%	89.57%
	SVM	85.86%	87.65%	92.39%	89.93%	93.33%

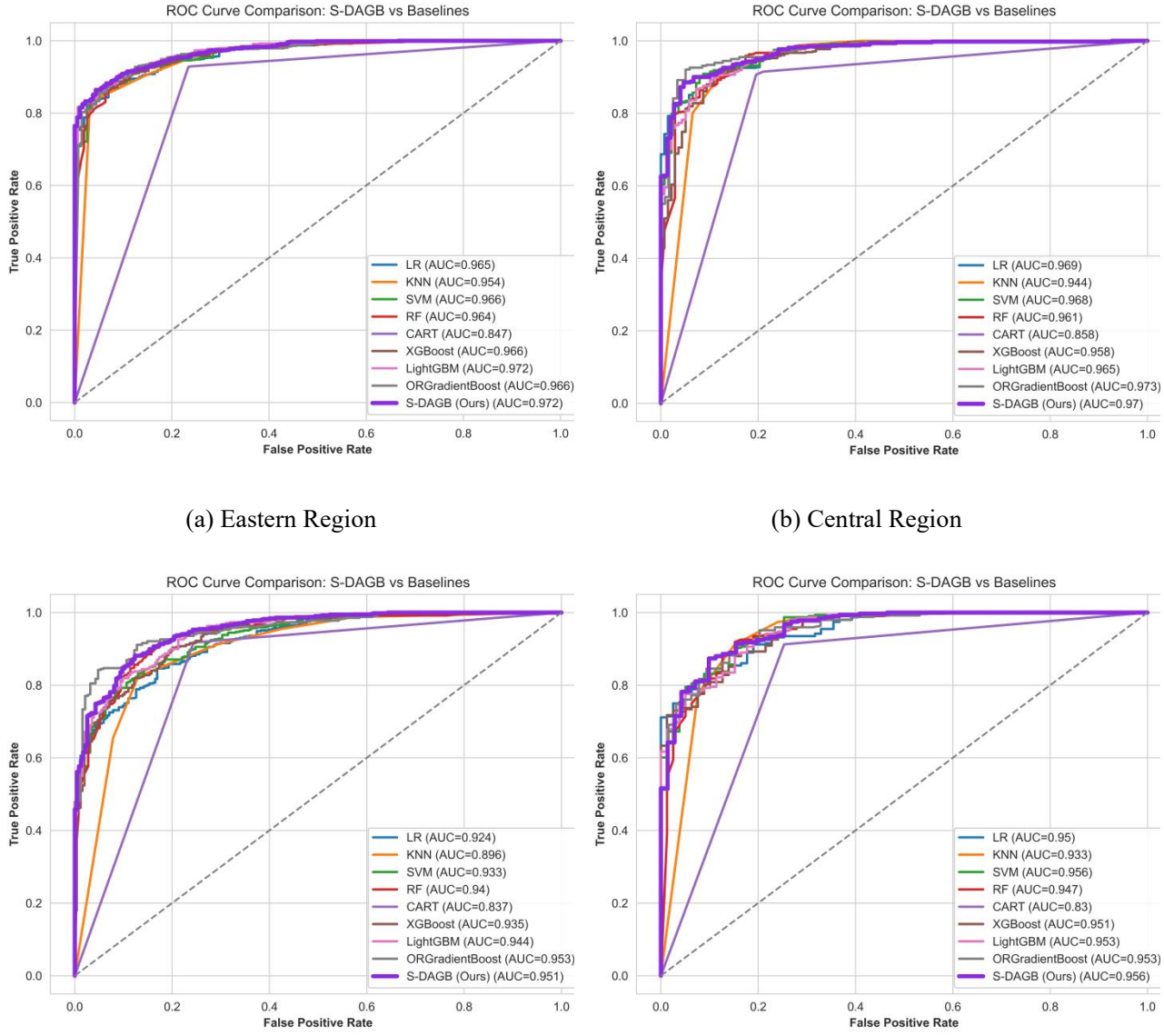
Region	Model	Accuracy	Precision	Recall	F1	AUC
	RF	88.12%	91.00%	91.69%	91.33%	94.02%
	CART	84.01%	89.12%	87.29%	88.16%	83.67%
	XGBoost	87.75%	90.48%	91.73%	91.07%	93.32%
	LightGBM	87.96%	90.82%	91.66%	91.20%	94.39%
	ORGB	89.07%	90.82%	93.49%	92.11%	95.26%
	S-DAGB	89.68%	91.22%	95.25%	92.24%	95.14%
Northeast	LR	87.08%	89.37%	95.94%	93.38%	95.01%
	KNN	92.31%	93.19%	97.48%	94.13%	93.31%
	SVM	92.50%	92.55%	98.54%	94.43%	95.59%
	RF	91.92%	93.51%	96.59%	92.31%	94.72%
	CART	89.47%	93.24%	93.58%	95.01%	83.03%
	XGBoost	90.63%	93.49%	94.80%	95.21%	95.13%
	LightGBM	91.09%	93.60%	95.30%	95.27%	95.26%
	ORGB	93.34%	93.83%	97.70%	95.44%	95.28%
	S-DAGB	93.52%	93.92%	97.78%	95.81%	95.64%

Comparing ROC curves across regions (Figure 5) further illustrates S-DAGB's predictive characteristics. While its curve does not absolutely envelop the baseline teacher model (ORGB) across all thresholds in extremely imbalanced areas like central and western China, it demonstrates two practical advantages:

(1) Remarkable curve smoothness and threshold robustness. Traditional ensemble tree models (e.g., CART and native XGBoost) often exhibit extreme probability distributions in ROC curves due to overfitting local noise when handling sparse or imbalanced data, resulting in "staircase-like" or "jagged" abrupt changes.S-DAGB effectively calibrates the continuity of output probabilities through the TPE depth regularization constraint introduced in Stage II and the ENS weight smoothing mechanism. This smooth ascent trajectory demonstrates the model's exceptional robustness to decision threshold fine-tuning, effectively mitigating drastic fluctuations in predictive performance.

(2) Robust performance in the low false positive rate (FPR) range. Along the left edge of the ROC curve (i.e., the locally high specificity range), S-DAGB's purple curve maintains a

consistently stable and steep upward trend.From the perspective of practical poverty governance, performance in this region is critical: it signifies that S-DAGB can robustly identify genuinely relatively impoverished individuals even under stringent conditions of strict false positive control (i.e., limiting non-impoverished households from occupying policy resources).This strategy of sacrificing a minimal portion of global AUC to achieve decision smoothness and local precision further validates the advanced capabilities of this two-stage knowledge distillation architecture in complex real-world scenarios.



(a) Eastern Region

(b) Central Region

(c) Western Region

(d) North-Eastern Region

Figure 5. ROC Curves of Different Machine Learning Models across the Four Regions

5. Conclusions and Outlook

Based on the 2020 China Family Panel Studies (CFPS) data, this study proposes the Two-stage Knowledge Distillation Adaptive Gradient Boosting Framework (TKDAF) and conducts

empirical research addressing spatial heterogeneity, small sample sizes, and extreme class imbalance in multi-regional relative poverty prediction. Key findings include:

(1) The effectiveness of extracting and reconstructing spatial heterogeneity prior knowledge was validated. By combining the base model with SHAP attribution, this study quantitatively extracted the marginal contributions of poverty-inducing features across regions. Introducing these prior weights into the second stage for feature space reconstruction effectively guided the model to focus on core poverty dimensions while reducing interference from redundant features. (2) Enhanced the generalization capability of deep tree models in small-sample scenarios. Addressing overfitting prone in sparsely sampled regions like Northeast China, the S-DAGB model leveraged TPE adaptive optimization and strong underlying regularization constraints (e.g., leaf node penalties) to effectively suppress overfitting on small datasets while maintaining high classification accuracy. (3) Achieved precision optimization under extremely imbalanced data distributions. Addressing the severe imbalance between positive and negative samples in central and western regions, the model reshapes the loss function by introducing an Effective Sample Number (ENS) dynamic weighting mechanism. Empirical results demonstrate that this mechanism significantly reduces false positive rates while maintaining high recall, better aligning with the practical demand for precise allocation of poverty alleviation resources.

Although this framework demonstrates strong predictive performance in empirical tasks, the following research directions remain open due to objective constraints: (1) Incorporating spatio-temporal dynamics: This study primarily relies on cross-sectional data, failing to fully capture the dynamic vulnerability of relative poverty. Future work could integrate multi-period panel data with spatio-temporal sequence prediction models to better track the temporal evolution of household poverty status. (2) Deep integration of multi-source heterogeneous data: Current input features are primarily confined to micro-level survey data. Subsequent research could explore integrating macro-level geographic information (e.g., nighttime light imagery, regional road networks) with multi-modal data to achieve more granular grid-based poverty identification. (3) Deepening the underlying knowledge distillation mechanism: The prior knowledge transfer in this paper remains at the shallow level of feature weight guidance. Future work could explore deeper knowledge distillation techniques like latent state alignment or logit matching to further expand performance boundaries while controlling model inference complexity.

Author Contributions:

Guanghuang Liu drafted the manuscript, implemented the coding, and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding:

This research received no external funding.

Institutional Review Board Statement:

Not applicable.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

The China Family Panel Studies (CFPS) data utilised in this research were collected and managed by the Institute of Social Science Surveys (ISSS), Peking University. This restricted-access dataset may be obtained by researchers submitting a formal application to ISSS. Details regarding data requests can be found on their official websites:http://www.isss.pku.edu.cn/cfps/.

Conflict of Interest:

The authors declare no conflict of interest.

References

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. Journal of Public Economics, 95(7-8), 476-487.

Bergstra, J., Bardenet, R., Bengio, Y., et al. (2011). Algorithms for hyper-parameter optimization. In Advances in Neural Information Processing Systems (pp. 2546-2554).

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., et al. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

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

Cui, Y., Jia, M., Lin, T. Y., et al. (2019). Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9268-9277).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

Jean, N., Burke, M., Xie, M., et al. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.

Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).

Li, X., Zhou, Y., & Chen, Y. (2020). Theory and methods for regional multidimensional poverty measurement. Acta Geographica Sinica, 75(4), 753-768.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Qian, Y., Wang, C., & Wang, J. (2022). Mutual information and decision tree algorithm for eliminating random consistency. Journal of Shanxi University (Natural Science Edition), 45(5), 1206-1215.

Ravallion, M., & Chen, S. (2007). China’s (uneven) progress against poverty. Journal of Development Economics, 82(1), 1-42.

Shi, Y., Ding, T., Qi, X., et al. (2024). An explainable model for relative poverty identification and early warning. Journal of Shanxi University (Natural Science Edition), 47(1), 155-165.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Advances in Neural Information Processing Systems (pp. 2951-2959).

Sun, J., & Xia, T. (2019). The evolution of China’s poverty alleviation strategy and post-2020 relative poverty governance. Chinese Rural Economy, (10), 98-111.

Wang, B., Luo, Q., Chen, G., et al. (2022). Differences and dynamics of multidimensional poverty in rural China from multiple perspectives analysis. Journal of Geographical Sciences, 32(8), 1383-1404.

Wang, S., & Zeng, X. (2018). Preliminary exploration of post-2020 poverty issues. Journal of Hohai University (Philosophy and Social Sciences Edition), 20(2), 7-13.

Wang, X., & Feng, H. (2020). China’s multidimensional relative poverty standards post-2020: International experience and policy orientations. Chinese Rural Economy, (3), 2-21.

Zou, W., & Fang, Y. (2011). A dynamic multidimensional study on poverty in China. Economic Research Journal, (12), 42-55.

Journal of Computer Science and Digital Technology, 2026, 2(1), 21-36
https://doi.org/10.71204/pyppdf95

RF-PBFT is an Improved PBFT Consensus Algorithm Based on the Random Forest Algorithm

Yingchen Xu [1,*], Ye Tian [1], Ying Jing [1], Fujiang Yuan [1]

[1] School of Computer Science and Technology, Taiyuan Normal University, Shanxi 030619, China

*** Correspondence:**

Yingchen Xu

274513634@qq.com

Received: 12 February 2026 / Accepted: 11 March 2026 / Published online: 12 March 2026

Abstract

The Practical Byzantine Fault-Tolerant (PBFT) consensus algorithm faces challenges such as random master node selection, high communication overhead, and a lack of incentive and penalty mechanisms, which undermine its efficiency and security in large-scale network environments. To address these issues, this paper proposes an improved PBFT consensus algorithm, RF-PBFT, which integrates a random forest-based node grouping mechanism with a dynamic reputation system. The algorithm leverages key behavioral data of nodes during the consensus process to train a random forest model, partitioning nodes into a consensus set and a candidate set. Nodes with superior predictive performance are selected to participate in the consensus process, thereby reducing redundant communication overhead. Furthermore, a differentiated reputation mechanism is established for the consensus group and the candidate group to encourage positive behavior and deter malicious actions. The master node is elected through a combination of prediction results and voting, enhancing system reliability and security. Simulation results demonstrate that, compared with traditional PBFT and other state-of-the-art variants, RF-PBFT significantly reduces communication overhead, decreases consensus latency, and improves throughput, validating its effectiveness in multi-node blockchain systems.

Keywords: Blockchain; PBFT; Random Forest Algorithm; Reputation Mechanism

1. Introduction

With the continuous iteration and innovation of digital technologies, cutting-edge technologies such as artificial intelligence, big data, and the Internet of Things are profoundly changing the industrial landscape and social structure. Against this backdrop, blockchain technology, due to its core characteristics such as decentralization, traceability, and tamper-proof nature, is gradually becoming an important pillar supporting the next generation of information infrastructure. Blockchain not only shows broad application prospects in fields such as finance, supply chain

management, healthcare, and intelligent manufacturing, but its underlying consensus algorithm is also considered crucial for ensuring system security and reliability (Veronese et al., 2009).

Consensus algorithms play a crucial role in coordinating distributed nodes to achieve a unified ledger state, directly impacting the performance, security, and scalability of a blockchain system. Currently, common consensus algorithms include Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), Raft, and Practical Byzantine Fault Tolerance (PBFT) (Vedant et al., 2022; Castro et al., 1999). Different algorithms present varying trade-offs in efficiency, security, and decentralization: PoW ensures security through hash calculations but suffers from high energy consumption and slow transaction confirmation speeds; PoS improves energy efficiency through stake allocation but carries the risk of stake concentration; DPoS increases transaction processing speed but is prone to centralization; Raft boasts low communication complexity and high efficiency but is highly dependent on network reliability and lacks fault tolerance (Li & Li, 2025; Madigan et al., 2012). In contrast, the PBFT algorithm achieves a relative balance between decentralization and performance. It can ensure consensus in a system with a certain number of Byzantine nodes without requiring massive computing resources, and is therefore widely used in consortium blockchains. PBFT's advantages lie primarily in its efficiency, low energy consumption, and strong Byzantine fault tolerance, making it particularly suitable for consortium and private blockchain environments with a limited number of nodes and relatively clear trust relationships. However, the PBFT algorithm faces several challenges, specifically in the following aspects:

(1) In large-scale network environments, the PBFT algorithm suffers a sharp decline in system throughput as the number of nodes increases dramatically, making it difficult to achieve efficient consensus processes in multi-node environments.

(2) The master node selection method is simplistic, typically determined by view number. This increases the probability of a malicious node being elected, severely impacting consensus efficiency.

(3) The lack of reward and punishment mechanisms reduces node participation in consensus, making it difficult to curb malicious behavior and compromising system stability and security.

To address the high complexity, random master node selection, and lack of reward and punishment mechanisms inherent in the PBFT consensus algorithm in large-scale blockchain networks, this paper proposes a PBFT consensus algorithm based on random forest grouping (FR-PBFT). First, to address the high communication overhead of PBFT in large-scale network scenarios, a random forest grouping algorithm is used to predict future node behavior. Based on the prediction results, nodes are divided into consensus and candidate node sets, selecting a subset of reliable nodes to participate in consensus and reducing unnecessary communication overhead. Secondly, to address the lack of a reward and punishment mechanism in PBFT, different reward and punishment mechanisms were set for the consensus node set and the candidate node set to ensure the enthusiasm of nodes to participate in consensus and voting. Finally, to address the issue of random selection of master nodes, a method of prediction results plus voting was adopted to elect reliable master nodes.

2. Related research

In recent years, the PBFT consensus algorithm has been extensively studied. Regarding scalability, many studies have proposed layered architectures and communication optimization schemes. Gueta et al.(2018)proposed SBFT (Scalable Byzantine Fault Tolerance) technology, which, with the help of collectors, threshold signatures and optimistic fast paths, achieves double the throughput and improves latency compared to traditional PBFT. In addition, Sukhwani et al. (2017)]analyzed the consensus process in permissioned blockchains through stochastic modeling, and confirmed the communication bottleneck that may occur when the node scale increases, emphasizing the necessity of PBFT optimization in large networks. HotStuff achieves linear communication complexity and responsiveness during network synchronization (Wu et al., 2025; Yin er al., 2019), making the consensus speed match the actual network latency; BFT-SMART improves system efficiency by improving communication strategies (Bessani et al., 2014). Li et al. (2020) designed a multi-layer PBFT model, which compresses the communication volume to a linear level through tree-like hierarchical grouping, but needs to balance cross-layer latency and fault tolerance performance. Na et al. (2022) proposed the DPNPBFT algorithm, which adopts a dual-master node power-sharing mechanism to improve the fault tolerance of the algorithm and is suitable for large-scale low-power scenarios such as the Internet of Things. Zhong et al. (2023) proposed an improved PBFT algorithm for secure knowledge transactions, ST-PBFT, which uses a grouping method based on consistent hashing to improve communication efficiency.

3. Overview of PBFT Consensus Algorithm and Random Forest Algorithm

3.1. PBFT Consensus Algorithm

The PBFT algorithm requires no staked equity or consumes competitive resources, resulting in low costs for malicious nodes. Its three-phase voting mechanism effectively eliminates interference from malicious nodes in the consensus process, possessing fault tolerance capabilities up to one-third of the total number of nodes. However, the current PBFT algorithm still has many areas requiring improvement, such as high communication overhead and strong randomness in master node selection, leading to low consensus efficiency. Therefore, most researchers have conducted in-depth research on master node selection methods and consensus consistency, striving to improve the consensus efficiency of the PBFT algorithm. In practical applications, such as transaction settlement in the financial sector and information sharing in supply chain management, there is an urgent need for efficient and reliable consensus algorithms. The PBFT algorithm has certain application potential in these scenarios due to its inherent characteristics, but existing problems severely restrict its effectiveness. Therefore, in-depth analysis of the PBFT algorithm and the proposal of practical and effective improvement strategies are of significant practical importance for further exploring the potential of blockchain technology and expanding its application boundaries. This is precisely the starting point and core direction of this paper. The consensus protocol execution process of the PBFT consensus algorithm is as follows:

The consensus protocol is the core of the PBFT (Practical Byzantine Fault Tolerance)

algorithm. In a traditional PBFT consensus network, nodes are mainly divided into three categories: client nodes, master nodes, and replica nodes (slave nodes). Client nodes are responsible for initiating consensus requests and receiving the final consensus results; master nodes receive client requests, number and sort them, and distribute the processed messages to replica nodes after signing; replica nodes are responsible for verifying the received consensus messages and determining the consistency of the consensus results. When a master node fails or malfunctions, the system initiates a view switching mechanism to elect a new master node within a specified time to continue the subsequent consensus process.

The execution process of the PBFT algorithm can be divided into five main stages: Request, Pre-prepare, Prepare, Commit, and Reply. The overall process is shown in Figure 1, and is detailed below:

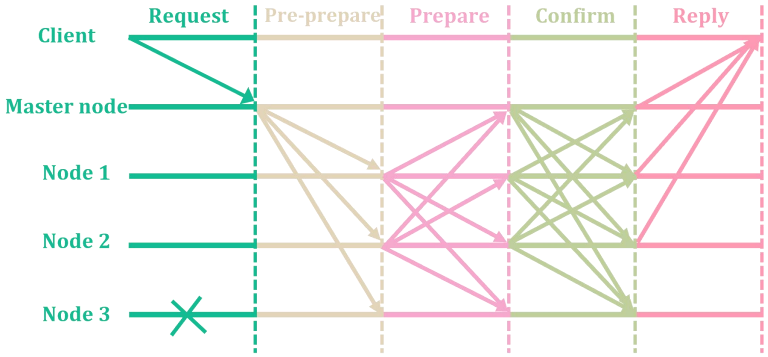


Figure 1. PBFT Algorithm Consensus Process

(1) Request Phase: The client sends a consensus request message to the master node, in the format <Request, m, t, c,σ_c>, where m represents the request content, t is the timestamp, c is the client identifier, and σ_c is the client signature.

(2) Pre-prepare Phase: After receiving the request, the master node numbers and sorts it, then generates a pre-prepare message <<Pre-prepare, v, n, D(m),σ_p>, m>, and sends it to all replica nodes. Here, v is the current view number, n is the message sequence number, D(m) is the message digest, and σ_p is the master node signature.

(3) Prepare Phase: After receiving the pre-prepare message, the replica nodes verify its validity and generate a prepare message <Prepare, v, n, D(m), b_i, σ_b> for broadcast, where b_i is the replica node identifier, and σ_b is the node's signature. In this phase, each replica node compares its generated preparation message with those sent by other replica nodes. When a node receives at least 2f+1 consistent preparation messages, it can proceed to the next phase. This phase helps screen for Byzantine malicious nodes in the network.

(4) Commit Phase: After the preparation conditions are met, each node generates a commit message <Commit, v, n, D(m), b_i, σ_b> and broadcasts it. When a node receives at least 2f+1 commit messages consistent with its own, it indicates that consensus has been largely reached, and it can proceed to the response phase.

(5) Reply (Response Phase): After confirmation in the commit phase, the node sends a response

message <Reply, v, n, σ_b, r> to the client, where r represents the result after executing the request. The client considers its request to have been successfully completed when it receives at least f+1 consistent response messages.

3.2. Random Forest Algorithm

Random Forest is a machine learning algorithm based on the Bagging ensemble strategy. Its core principle is to rely on multiple decorrelated decision trees for collaborative prediction. By leveraging a dual randomness mechanism, it overcomes the limitations of single decision trees and is now widely used in tasks such as classification and regression. The training process of Random Forest is shown in Figure 2.

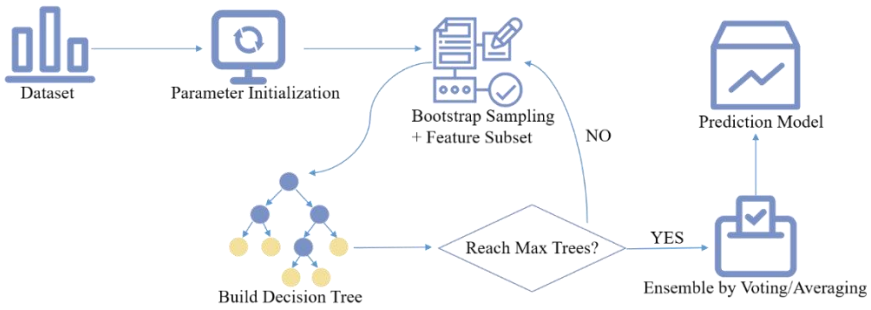


Figure 2. Random Forest Training Process

The training process of the Random Forest algorithm, and the specific implementation of each stage, are as follows:

(1) Input Dataset and Parameter Initialization

Random forest training first imports a dataset D containing sample feature vectors xi and labels yi. Then, configure the core parameters: number of decision trees T to balance complexity (few trees easily lead to underfitting, more trees increase cost); tree structure parameters (such as maximum depth d, minimum number of samples per leaf node) to control tree growth and avoid overfitting of a single tree; the number of features per tree m is set to $\lfloor \sqrt{d_{\text{total}}} \rfloor$ (d_{total} is the total number of features), and feature randomization enhances the independence between trees, allowing each tree to learn different feature subsets and adapt to multi-dimensional feature learning.

(2) Bootstrap Sample Sampling (Constructing Training Sets for Multiple Trees)

For the t-th tree (t = 1,...,T), n samples are drawn with replacement from D to generate Dt. Sampling characteristics: data diversity (different Dts have different characteristics, approximately 36.8% are out-of-bag samples used for OOB validation), and unbiasedness (sampling with replacement ensures that the training set distribution is consistent with the original data). This step provides differentiated data for each tree, constructs diverse base learners, supports subsequent ensemble learning (voting/averaging), and improves the model's generalization and fitting capabilities.

(3) Building a Single Decision Tree

In the building a single decision tree stage, a single decision tree is trained using a "sample subset Dt + feature subset Ft". The core logic revolves around greedy splitting and recursive termination. When performing greedy splitting, starting from the root node, the features in the feature subset Ft and the possible splitting thresholds are traversed, and the optimal splitting method is determined based on the task type. For classification tasks, the node purity is evaluated by calculating the Gini index, using the following formula:

$$\text{Gini}(S) = 1 - \sum_{c=1}^{C} \left(\frac{|S_c|}{|S|} \right)^2$$

(Where S is the current node sample set, C is the total number of classes, and S_c is the subset of samples belonging to class c in S) Select the split that minimizes the sum of the Gini coefficients of the child nodes to reduce sample uncertainty; for regression tasks, calculate the Mean Squared Error (MSE) to measure the dispersion of node samples, using the formula:

$$\text{MSE}(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y})^2$$

(where is the mean of the sample labels in S) Select the split that minimizes the MSE of the child nodes to improve the consistency of sample prediction. When the tree depth reaches the preset maximum depth max_depth, or the number of leaf node samples is less than min_samples_leaf, the recursion is terminated. In the classification task, the mode of the sample labels output by the leaf node (the category that appears most frequently) is used as the predicted value of the node.

(4) Determine if the maximum number of trees has been reached (Reach Max Trees).

Check if the number of decision trees built has reached the T set in the parameter initialization phase: if the number has not been reached, return to the "Bootstrap Sampling + Feature Subset Selection" step to continue building the next decision tree; if the number has been reached, enter the "Multi-Tree Integration" step to improve model performance by utilizing the diversity of multiple trees.

(5) Multi-tree ensemble (Ensemble by Voting/Averaging)

In the multi-tree ensemble stage, the prediction results of T decision trees need to be summarized and integrated according to the task type: For a regression task, let the predicted class of sample x by the t-th tree be: , The prediction results of all trees are summarized, and the class with the most votes is taken as the prediction result, which can be described by the formula:

$$H(x) = \arg\max_{c \in \{1,\ldots,C\}} \sum_{t=1}^{T} \text{ú}(h_t(x) = c)$$

(Where I is an indicator function, taking the value 1 when the condition is met and 0 otherwise), multi-tree results are fused through majority voting to reduce the impact of single-tree misjudgment; for regression tasks, let the predicted value of the t-th tree for sample x be

$h_t(x) \in R$,The arithmetic mean of all tree predictions is taken as the final result, and the consensus is:

$$H(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Mean aggregation is used to smooth the differences in multi-tree predictions and improve the stability of the results.

(6) Output Prediction Model

After completing the above steps, a random forest model is obtained, which is an ensemble of T decision trees through a "voting/averaging" process. When the model is applied, a new sample is input, each decision tree makes an independent prediction, and finally, the ensemble strategy outputs the final classification label or regression value, achieving effective prediction of unknown data.

4. RF-PBFT Algorithm Design

4.1. System Model

To address the shortcomings of the traditional PBFT consensus algorithm in node management and consensus efficiency, this system constructs a closed-loop architecture of "node grouping - master node election - consensus execution - reputation update". The core is to optimize the node grouping strategy through the random forest algorithm and combine it with the reputation mechanism to improve consensus security and efficiency. The specific system model structure is shown in figure 3.

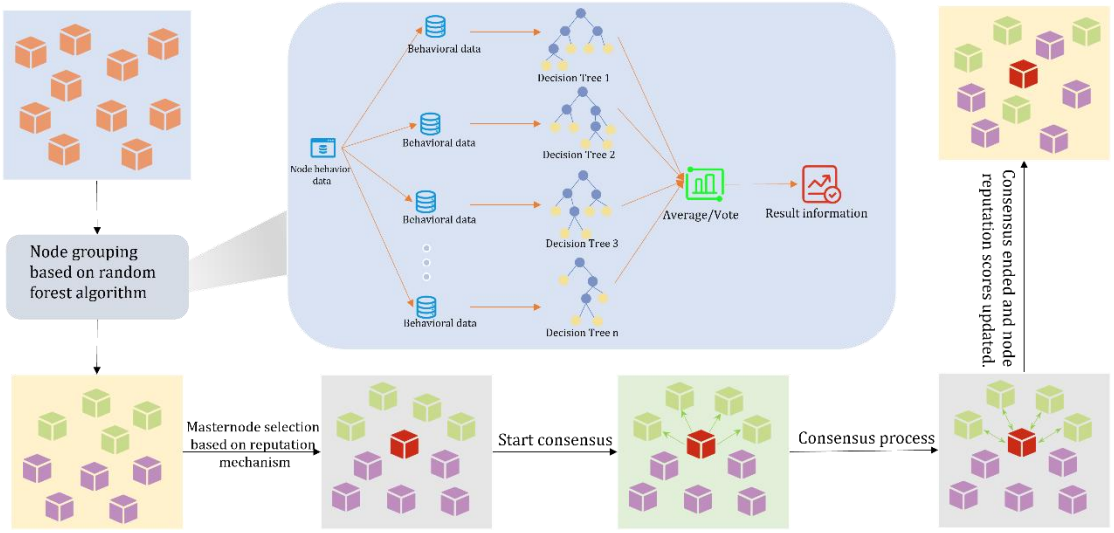


Figure 3. RF-PBFT system model diagram

In the RF-PBFT system model, the system takes historical node behavior data (including reputation records, interaction features, etc.) as input, deploys a random forest algorithm to construct an ensemble decision tree model, generates multiple training subsets through bootstrap sampling, trains decision trees independently in each subset, and integrates them by voting,

outputting a highly reliable consensus node group (participating in the core consensus) and a candidate node group for backup observation. On this basis, the master node election introduces a dynamic reputation mechanism, which integrates the grouping results (random forest predicted labels) with real-time behavior data (such as consensus response efficiency and consensus completion rate), selects nodes with high reputation values and stable behavior to be elected as the master node for this round of consensus. After consensus is completed, the system quantifies the node behavior performance (such as voting enthusiasm and consensus completion rate) and dynamically updates the reputation value (deduction for malicious behavior and reward for compliant behavior), and feeds the updated data back to the node grouping module as new training samples, forming a closed-loop iteration of "node grouping - master node election - consensus execution - reputation update", enabling the system to adapt to the dynamic changes in node behavior and continuously optimize the accuracy of the grouping strategy.

4.2. Reputation Mechanism

The reputation mechanism is a key indicator for measuring node behavior and a necessary condition for the grouping algorithm. This scheme distinguishes between the reputation mechanisms of consensus nodes and candidate nodes. Before the formal consensus process begins, all nodes have a default reputation score of 0. First, a three-round pre-consensus process is performed on the nodes. Based on the pre-consensus process, the node behavior data is obtained. After the RF random forest grouping model obtains the node behavior data, it predicts the future behavior of each node and assigns a corresponding group label. For example, the top 50% of nodes in terms of predicted behavior are labeled as "Consensus Group." The remaining nodes are in the candidate group. After grouping, the master node for this round is selected by voting, and the pre-consensus work is completed.

At the start of the consensus process, the performance of all nodes is recorded by the monitoring nodes. After each round of consensus, the node is assigned a reputation score based on its group's reputation mechanism. If a consensus node's reputation score is in the bottom half, it is designated as a candidate node and added to the candidate node set. At this point, candidate nodes are sorted according to their reputation scores, and the corresponding number of nodes are converted into consensus nodes. When a candidate node becomes a consensus node, its score is halved.

4.2.1. Consensus Node Reputation Mechanism

In this scheme, nodes are divided into consensus nodes and candidate nodes. These two types of nodes play different roles in the consensus process, hence their reputation score formulas also differ. The formula for measuring the reputation score of consensus node i is as follows:

$$\mathrm{Sc}_i = \alpha \cdot CE_i + \beta \cdot CR_i - \gamma \cdot E_{punish}^{-1} \cdot P_i(x)$$

Where Sc_i represents the overall evaluation score of consensus node i, with higher values indicating better performance. CE_i represents the node's consensus efficiency, reflecting its relative performance per unit time. CR_i represents the consensus completion rate, i.e., the proportion of nodes successfully participating in consensus. $\mathrm{E}_{\mathrm{punish}}$ is the anomaly penalty

coefficient, reflecting the negative impact of abnormal node behavior on consensus and imposing different levels of penalty based on the magnitude of the impact. The components of each formula will be introduced below.

(1) Consensus efficiency: Represents the relative rate at which a node reaches consensus per unit of time.

$$CE_i = \frac{C_i / T_i}{\frac{1}{N}\sum_{j=1}^{N}(C_j / T_j)}$$

Where C_i is the number of times node i successfully participates in consensus. T_i is the node's online time. N is the total number of consensus nodes in the system. The normalized ratio is expressed as: >1 for values above average, <1 for values below average.

(2) Consensus Completion Rate: Represents the success rate of a node in initiating or participating in consensus.

$$CR_{\mathrm{i}} = \frac{C_{\mathrm{i}}}{P_{\mathrm{i}}}$$

Where P_i is the total number of consensus rounds participated in by node i, and C_i is the number of consensus rounds successfully completed by node i.

(3) Abnormal penalty coefficient

$$E_{punish} = 1 - \frac{K_i}{P_i}$$

Here, E_{punish} is the anomaly penalty coefficient, K_i is the number of times node i triggers anomalies (such as duplicate proposals or data errors), and P_i represents the total number of consensus rounds participated in by node i. The logic of this coefficient is: the more anomalies, the heavier the penalty; if there are no anomalies, $E_{punish} = 1$, and the more anomalies, the closer it gets to 0.

(4) Consensus Penalty Scoring

$$P_i(x) = P_0 + \alpha \cdot \sqrt[3]{\frac{x}{\max_x}}$$

The core concept of this formula is that the penalty increases with the number of failures. P_0 is the base penalty, which is set to 0 in this scheme to avoid deducting points even when there are no failures. α is the penalty coefficient, which can be adjusted according to the scenario. X represents the cumulative number of times node i has failed to participate in consensus. Max_x represents the maximum cumulative number of times all nodes in the network have failed to successfully participate in consensus; the formula has undergone dynamic normalization to avoid extreme values.

4.2.2. Candidate Node Reputation Mechanism

(1) Response rate: The willingness and stability of a node to respond when requested for assistance by the system, representing the node's enthusiasm for participating in consensus.

$$R_{\mathrm{i}} = \frac{R_{\mathrm{i}}^{\mathrm{response}}}{R_{\mathrm{i}}^{\mathrm{request}}}$$

Where R_i is the node's response rate.

(2) Voting enthusiasm

When a candidate node casts its vote in support of the master node in each round of selection, it is considered to have successfully voted and will be awarded a certain reputation score. Otherwise, a certain reputation score will be deducted as a penalty.

$$CD_{i,k} = CD_{i,k-1} + UD; UD = \begin{cases} d, Vote\ successful \\ -2d, Vote\ failed \end{cases}$$

Where $CD_{i,k}$ and $CD_{i,k-1}$ are the reputation scores updated by node i after the consensus of the current round and the previous round, respectively. UD is the candidate node reputation score update factor. When a node successfully participates in voting in this round, it is given a certain reputation score d as a reward. When a node does not vote in this round, it is penalized by deducting 2d reputation scores. The default value of d in this scheme is 0.5.

4.3. Node Grouping and Master Node Selection

In the consensus process of the RF-PBFT algorithm, a detection node is first set up. This node does not participate in any consensus process; it only detects the behavior of other nodes. After the client sends a request to the master node, the behavior detection node in the system tracks the interaction behavior of nodes in the consensus process in real time (such as voting consistency, message response timeliness, etc.), and inputs the detected behavior data into the random forest prediction model after each round of consensus. The model has a high accuracy in predicting the future behavior of nodes after previous training. Based on the model prediction results, the top 50% of nodes can be included in the consensus node set. At this time, a node needs to be selected from theθconsensus nodes to serve as the master node for this round of consensus. However, relying solely on reputation score to select the master node is too simplistic and can easily lead to issues such as centralization and reduced system security. Therefore, this scheme improves the selection of the master node. After the node grouping is completed, all nodes vote for the consensus node, and the master node is elected according to the number of votes. The number of votes for node i is calculated as follows:

$$V_i = \sum_{y=1}^{N} S_y; S_y = \begin{cases} \theta, if\ node\ y\ supports\ node\ i \\ \beta, if\ node\ y\ does\ not\ support\ node\ i \end{cases}$$

In this paper, θ and β are set to 1 and -1, respectively.

4.4. Consistency Process

This paper's solution groups nodes based on behavioral data predictions, resulting in different permissions for each group. Nodes in the consensus node group participate in the consensus process, while nodes in the candidate node group passively receive the consensus message. This transforms the traditional PBFT's "all nodes participate in consensus" into "only reliable nodes participate, with the remaining nodes only responsible for message propagation," while simplifying the commit phase and avoiding unnecessary communication overhead. The improved consensus flowchart is shown in Figure 4.

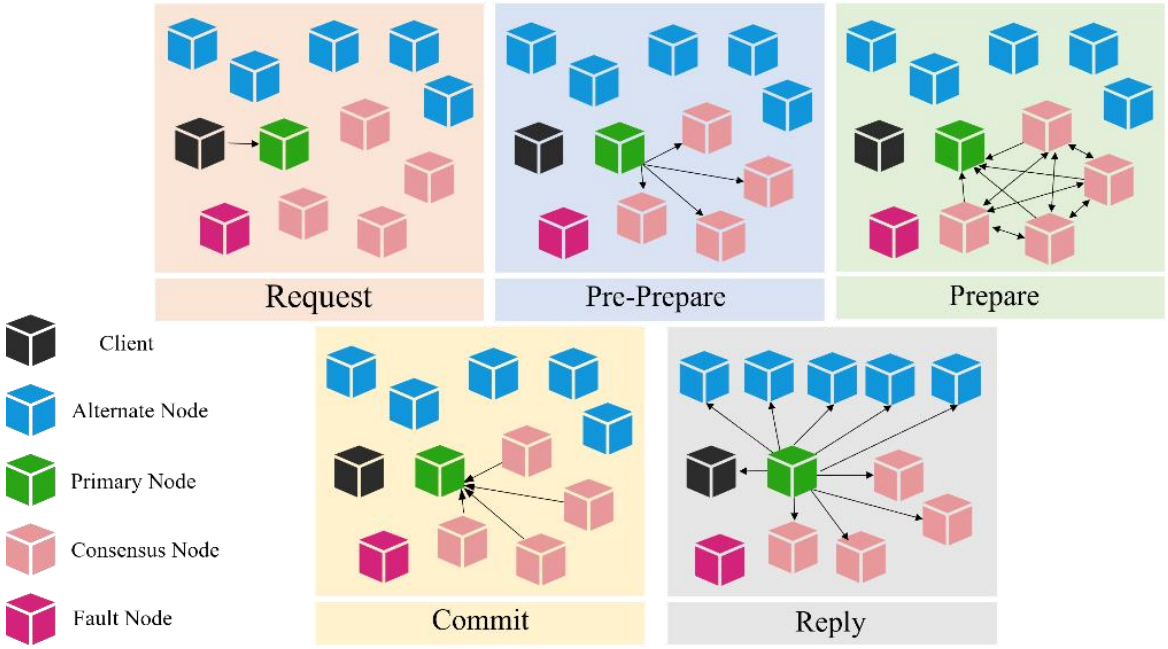


Figure 4. Consensus Process Improvement

5. Experimental Results and Analysis

The simulation experiment of the improved algorithm proposed in this paper used a 13th Gen Intel® Core™ i9-13900K 3.00GHz processor, Windows 11 operating system, 128GB of RAM, and Python 3.8 for code writing and implementation.

5.1. Classification Aaccuracy

This experiment compares the classification performance of three classification algorithms—Random Forest, Decision Tree, and Naive Bayes—based on a sample dataset of 1000 records. Accuracy and F1 score were used as evaluation metrics, and the results are shown in Table 1. The data shows that the Random Forest algorithm performs best in both metrics (Accuracy = 0.955, F1 Score = 0.957), with significantly higher classification accuracy than Decision Tree and Naive Bayes. The experiment demonstrates that the proposed grouping algorithm can better reduce the negative impact of malicious nodes on system consensus, thus improving system consensus efficiency and security.

Table 1. Performance Comparison Experiment of Classification Algorithms

Algorithm	Accuracy	F1 Score
Decision Tree	0.931	0.936
Naive Bayes	0.852	0.855
Random Forest Algorithm	0.955	0.957

5.2. Communication Overhead

Communication overhead refers to the total number of messages exchanged between nodes during the consensus process, and is an important indicator for judging the efficiency of a consensus algorithm.

Figure 5 compares the communication overhead of RF-PBFT, APBFT, and traditional PBFT. As shown in the figure, the amount of interactive messages increases with the number of nodes, but PBFT's increase is significantly greater than the other two, while RF-PBFT and APBFT's increases are relatively gradual. Furthermore, RF-PBFT's communication overhead is consistently significantly lower than that of similar algorithms. Therefore, RF-PBFT demonstrates a significant advantage in reducing network communication burden and is more suitable for blockchain networks composed of large numbers of nodes.

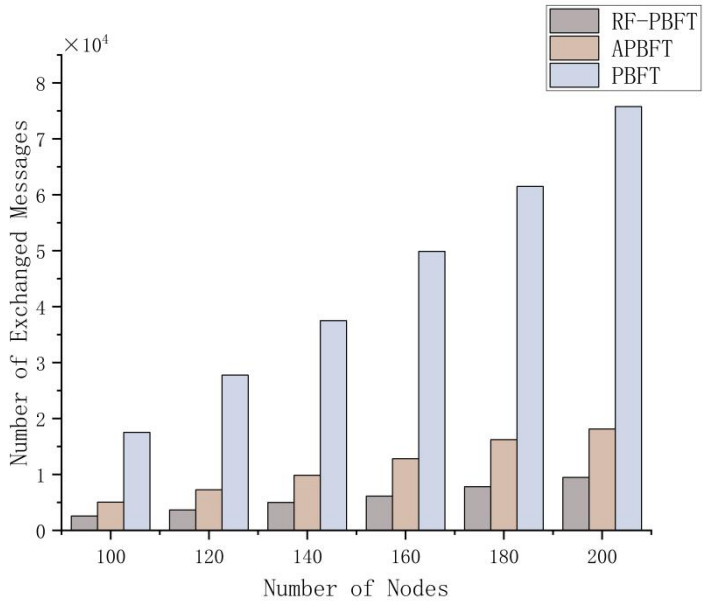


Figure 5. Comparison of communication overhead

5.3. Consensus Delay

Consensus latency is a crucial metric for measuring system performance, directly impacting system efficiency and user experience. Consensus latency refers to the time elapsed from when a client sends a request to when that request is fulfilled. Lower consensus latency among nodes in the system results in faster consensus being reached. The calculation of consensus latency is

shown in the following formula.

$$T = T_F - T_R$$

Where: T_F represents the time when the client receives f+1 identical reply messages, at which point the client's request has been successfully executed. T_R represents the time when the client sends the request to the master node.

With 100, 120, 140, 160, 180, and 200 nodes, 50 repeated experiments were conducted for each, and the average value was calculated. The final results are shown in Figure 6. When the number of nodes for the three algorithms is the same, the consensus latency of RF-PBFT and APBFT is significantly lower than that of traditional PBFT, and the advantage of this scheme is even more obvious compared to APBFT. It reduces some unnecessary communication overhead and demonstrates better consensus efficiency (figure 6).

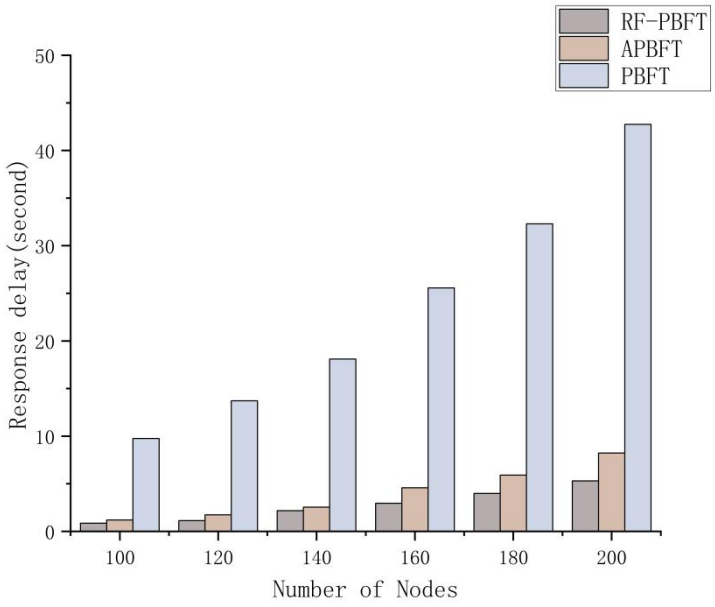


Figure 6. Consensus latency comparison

5.4. Throughput

Throughput (TPS), as a core indicator for measuring blockchain transaction processing capacity, reflects the number of transactions that complete consensus per unit of time and directly reflects the system's efficiency. The results of this latency test are shown in Figure 7: As the number of nodes increases, the TPS of RF-PBFT, APBFT, and traditional PBFT all show a downward trend. At the same node scale, RF-PBFT's transaction processing capacity advantage is particularly prominent: with 100 nodes, its TPS far exceeds that of PBFT; even when the number of nodes increases to 200, RF-PBFT still maintains a high throughput level.

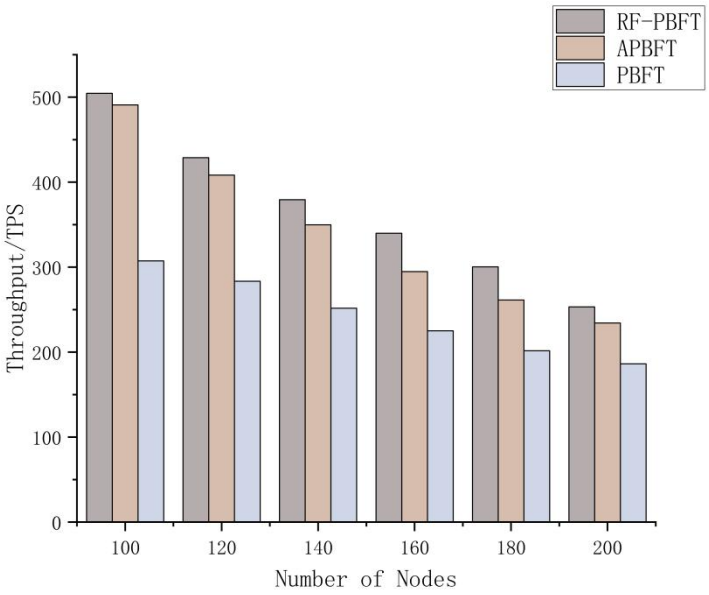


Figure 7. Throughput Comparison

5.5. Experiment Summary

This study focuses on the RF-PBFT consensus algorithm, conducting experiments on classification accuracy, communication overhead, consensus latency, and throughput. Comparison of grouping accuracy with random forest, decision tree, and Naive Bayes algorithms shows that the proposed RF-PBFT algorithm improves the reliability and security of system grouping and master node selection. In a multi-node environment of 100-200 nodes, this scheme groups nodes based on their behavioral data and introduces a reputation mechanism, ensuring reliable nodes participate in consensus, thus avoiding unnecessary communication overhead and improving system consensus efficiency. Simulation experiments demonstrate that the RF-PBFT scheme has significant advantages over PBFT and APBFT in multi-node scenarios in terms of communication overhead, consensus latency, and throughput.

6. Conclusion

This study proposes an improved PBFT consensus algorithm, FR-PBFT, which integrates random forest-based node grouping and a dual reputation mechanism to enhance consensus efficiency and security. The algorithm utilizes machine learning to predict node behavior and dynamically group nodes, effectively reducing communication complexity and mitigating the impact of malicious nodes. The introduction of the reputation system further incentivizes active node participation and effectively curbs malicious behavior. Experimental evaluations at different node scales demonstrate that FR-PBFT outperforms traditional PBFT and APBFT in terms of communication overhead, consensus latency, and throughput. This method provides a scalable and adaptive solution for consortium blockchains and private blockchains, making a theoretical and practical contribution to the field of distributed consensus. Future work will focus on further optimizing the real-time performance of this model and exploring its applicability in more dynamic and heterogeneous network environments.

Author Contributions:

All authors have read and agreed to the published version of the manuscript.

Funding:

This research received no external funding.

Institutional Review Board Statement:

Not applicable.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

Not applicable.

Conflict of Interest:

The authors declare no conflict of interest.

References

Bessani, A. N., Sousa, J., & Alchieri, E. A. P. (2014). State machine replication for the masses with BFT-SMART. In Proceedings of the 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN) (pp. 355 – 362). IEEE.

Castro, M., & Liskov, B. (1999). Practical Byzantine fault tolerance. In Proceedings of the 3rd USENIX Symposium on Operating Systems Design and Implementation (OSDI) (pp. 173 – 186).

Gueta, G. G., Abraham, I., Grossman, S., Malkhi, D., Pinkas, B., Reiter, M. K., Seredinschi, D.-A., Tamir, O., & Tomescu, A. (2018). SBFT: A scalable and decentralized trust infrastructure. arXiv preprint arXiv:1804.01626.

Li, W., Feng, C., Zhang, L., Xu, H., Cao, B., & Imran, M. (2020). A scalable multi-layer PBFT consensus for blockchain. IEEE Transactions on Parallel and Distributed Systems, 32(5), 1146 – 1160. https://doi.org/10.1109/TPDS.2020.3038142

Li, Z., Wang, J., & Li, Y. (2025). An improved PBFT consensus algorithm based on reputation and gaming. The Journal of Supercomputing, 81(1), 323.

Madigan, D. J., Litvin, S. Y., Popp, B. N., Carlisle, A. B., Farwell, C. J., & Block, B. A. (2012). Tissue turnover rates and isotopic trophic discrimination factors in the endothermic teleost, Pacific bluefin tuna (Thunnus orientalis). PLOS ONE, 7(11), e49220. https://doi.org/10.1371/journal.pone.0049220

Na, Y., Wen, Z., Fang, J., Zhang, X., & Wang, J. (2022). A derivative PBFT blockchain consensus algorithm with dual primary nodes based on separation of powers (DPNPBFT). IEEE Access, 10, 76114 – 76124.

Sukhwani, H., Martínez, J. M., Chang, X., Trivedi, K. S., & Rindos, A. (2017). Performance modeling of PBFT consensus process for permissioned blockchain network (Hyperledger

Fabric). In Proceedings of the 36th IEEE Symposium on Reliable Distributed Systems (SRDS) (pp. 253–255). IEEE.

Vedant, A., Yadav, A., Sharma, S., Thite, O., & Sheikh, A. (2022). A practical Byzantine fault tolerance blockchain for securing vehicle-to-grid energy trading. In Proceedings of the Global Energy Conference (GEC) (pp. 1–6). IEEE.

Veronese, G. S., Correia, M., Bessani, A. N., & Lung, L. C. (2009). Spin one's wheels? Byzantine fault tolerance with a spinning primary. In Proceedings of the 28th IEEE International Symposium on Reliable Distributed Systems (SRDS) (pp. 135–144). IEEE.

Wu, X., Wang, Z., Li, X., Ding, J., Tian, J., & Liu, Y. (2025). DBPBFT: A hierarchical PBFT consensus algorithm with dual blockchain for IoT. Future Generation Computer Systems, 162, 107429.

Yin, M., Malkhi, D., Reiter, M. K., Gueta, G. G., & Abraham, I. (2019). HotStuff: BFT consensus in the lens of blockchain. In Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC) (pp. 347–356).

Zhong, W., Feng, W., Huang, M., Du, D., & Li, Z. (2023). ST-PBFT: An optimized PBFT consensus algorithm for intellectual property transaction scenarios. Electronics, 12(2), 325.

Journal of Computer Science and Digital Technology, 2026, 2(1), 37-50
https://doi.org/10.71204/ew979106

Night UAV Vehicle Detection Based on Infrared-Visible Fusion and Improved YOLO11

Hongxia Su [1, *]

[1] School of Computer Science and Technology, Taiyuan Normal University, Shanxi 030619, China

*** Correspondence:**

Hongxia Su

2862201740@qq.com

Received: 25 February 2026 / Accepted: 15 March 2026 / Published online: 16 March 2026

Abstract

To address the challenges of vehicle detection accuracy in nighttime UAV aerial photography scenarios caused by low light conditions, strong noise, and dense small object distribution, this paper proposes an improved YOLO11 vehicle detection method based on multi-modal fusion. First, a collaborative enhancement strategy combining CLAHE and Gamma correction is applied to preprocessing nighttime visible light images, effectively restoring vehicle texture details in dark areas. Second, the enhanced visible light images are fused with infrared images through fixed-weight weighted fusion (infrared weight α=0.7), fully leveraging the complementary advantages of both modalities. Finally, the MANet module is introduced on top of YOLO11n to enhance backbone multi-scale feature extraction capabilities, while the ADFPN-DASI module is designed to collaboratively optimize neck multi-scale feature representation. Experiments conducted on the DroneVehicle dataset demonstrate that the proposed method achieves an mAP50 of 80.53%, representing an improvement of 11.45 percentage points over the YOLO11n baseline. The method maintains lightweight advantages in parameter and computational resources, outperforming mainstream comparison methods such as YOLOv5n.

Keywords: Nighttime Vehicle Detection; Drone Aerial Photography; Multimodal Fusion; YOLO11; Feature Pyramid Network

1. Introduction

In recent years, the widespread adoption of drones (UAVs) and high-resolution aerial imaging technology has demonstrated significant value in target detection based on aerial imagery for urban traffic monitoring, intelligent traffic management, and disaster response (Mittal et al., 2020). However, nighttime drone aerial imaging faces severe challenges such as extreme low-light conditions, strong noise, and loss of target details, making it a major research challenge. Li et al. (2024) also highlighted the prominent issue of accuracy degradation in nighttime scenarios. Sun

et al. (2022) constructed the DroneVehicle dataset, which includes 28,439 pairs of RGB-infrared images covering various urban scenes during both day and night, providing a crucial multimodal benchmark for related research.

In image enhancement, CLAHE effectively improves the visibility of details in visible light images under low-light conditions through local adaptive contrast equalization (Persiya et al., 2025). When used in conjunction with Gamma correction, it further restores the recognizability of textures in dark areas (Ma et al., 2023). Yuan et al. (2023) also validated the practical value of CLAHE in preprocessing nighttime low-light scenarios for drones. To compensate for the limitations of visible light modality at night, infrared-visible fusion technology has become a key breakthrough (Ma et al., 2019). ESM-YOLO (Zhang et al., 2024) achieves high accuracy in small target detection for aerial remote sensing through its bilateral excitation fusion module, while IV-YOLO (Tian et al., 2024) reaches a 74.6% mean absolute precision (mAP) on UAV remote sensing datasets.

In the field of detection networks, first-stage algorithms like the YOLO series (Redmon and Farhadi, 2018)are better suited for nighttime drone scenarios due to their real-time performance advantages. YOLO11 (Khanam and Hussain, 2024)introduced the C3k2 module and C2PSA attention mechanism, with its nano version containing only about 2.6 million parameters, making it suitable for edge deployment. YOLOv12 (Tian et al., 2025)further incorporated area attention mechanism and R-ELAN, surpassing previous CNN-dominated frameworks in balancing speed and accuracy. Wang et al.achieved significant improvements in small object detection accuracy on the VisDrone dataset through enhancements to YOLOv8 (Wang et al., 2023). Luo et al. (2025) proposed NOC-YOLO, which reached a mAP50 of 79.5% on the DroneVehicle dataset. Feng et al. (2024) introduced Hyper-YOLO, integrating supergraph computation and MANet backbone into the detection framework, achieving a 12% mAP improvement over YOLOv8-N on the COCO dataset. However, existing methods still exhibit notable limitations in dense small object detection at night and cross-modal fusion.

To address this, this paper proposes an improved YOLO11 vehicle detection method for nighttime scenes in the DroneVehicle dataset, with the following key contributions:

(1) The CLAHE+Gamma collaborative enhancement strategy is applied to pre-process nighttime visible light images, effectively restoring vehicle texture details in dark areas.

(2) The enhanced visible light image and infrared image are fused by fixed weight weighted fusion (infrared weight 0.7) to give full play to the complementary advantages of the two modes;

(3) The self-developed ADFPN (Aggregated Diffusion Feature Pyramid Network) integrates the HCFNet-based DASI module and MANet backbone, achieving significant improvements in nighttime dense small object detection accuracy while maintaining its lightweight advantages.

2. Methodology Design

The proposed method is structured in three phases as shown in Figure 1: visible light image preprocessing, infrared-visible fusion, and an enhanced YOLO11 detection network.

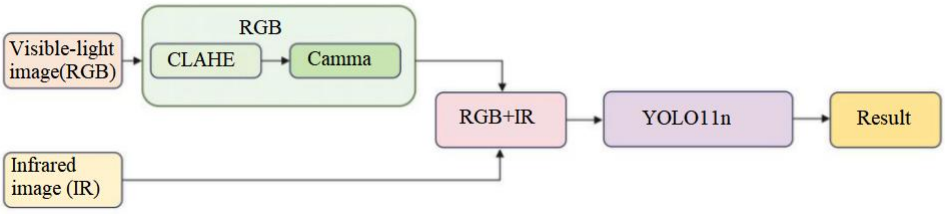


Figure 1. Overall framework of this paper

2.1. CLAHE+Gamma Visible Light Image Enhancement

To address the low contrast and loss of details in dark areas of nighttime visible light images, this paper adopts a collaborative enhancement strategy combining CLAHE and Gamma correction.

The strategy comprises two key processing steps: First, the CLAHE module converts the image to the LAB color space and applies adaptive histogram equalization (LUT) solely to the L channel (8×8 sub-block size, clipLimit=2.0) to suppress noise while enhancing local contrast and restoring dark area texture details. Second, the Gamma correction module (γ=0.7) performs efficient nonlinear brightness mapping through a lookup table (LUT) to specifically brighten dark areas, further improving vehicle contour recognition. Through this two-step collaborative processing, the contours and texture details of vehicles in nighttime dark areas are effectively restored, providing high-quality input for subsequent infrared-visible fusion.

2.2. Infrared-Visible Weighted Image Fusion

To address the complementary limitations of visible light modality in nighttime imaging and infrared modality's insufficient texture representation, this study proposes a fixed-weight fusion strategy. The approach leverages the inherent complementarity between the two modalities: infrared images remain stable in low-light conditions while reliably capturing vehicle thermal radiation characteristics, whereas CLAHE+Gamma-enhanced visible light images preserve rich texture and structural details. The fusion formula is as follows:

$$I_{fused}=\alpha\cdot\ I_{IR}+(1-\alpha)\cdot I_{RGB\_enhanced}$$

Here, IIR denotes infrared images, IRGB_enhanced represents enhanced visible light images, and α is the infrared weight. Ablation experiments on the DroneVehicle validation set demonstrated that the model achieves optimal performance (79.14% mAP50) with α=0.7, indicating that infrared modality should be given higher fusion weight in nighttime scenes to fully leverage its light-insensitivity advantage while retaining 30% of visible light texture information for further precision improvement. The final fused images are input into the subsequent detection network as three-channel inputs.

2.3. The Improved YOLO11 Algorithm

To address challenges in nighttime UAV aerial photography—such as target detail loss due to low light, severe degradation of visible light single-modality performance, and high false-negative rates caused by dense small object distribution—this study proposes an improved algorithm for

nighttime dense small object detection by enhancing YOLO11n through multi-modal fusion. The enhanced network architecture is illustrated in Figure 2.

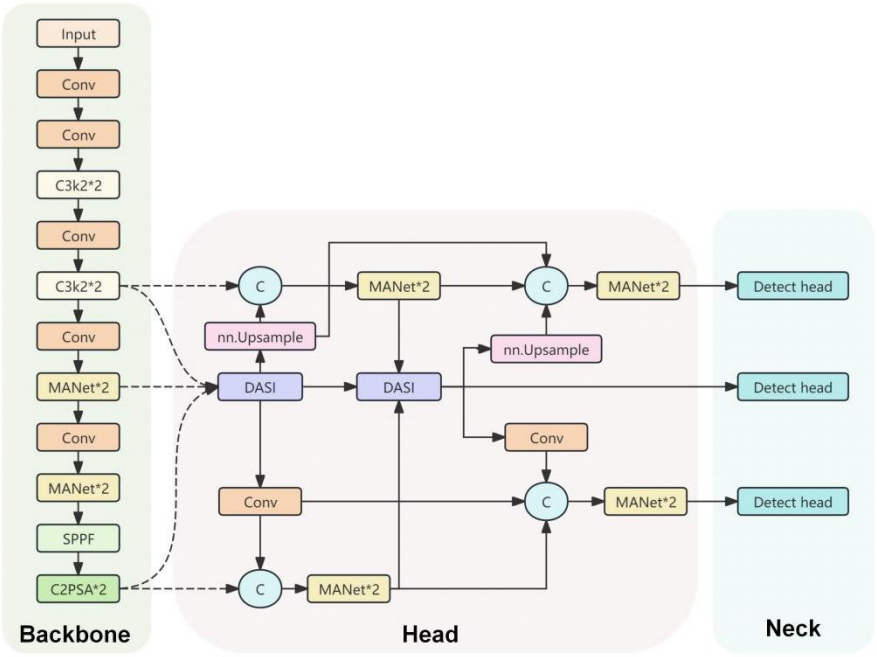


Figure 2. Network Architecture Diagram

The algorithm primarily consists of two enhancement modules:

(1) MANet Module: To address the backbone network's limited capability in extracting multi-scale features of small targets at night, this paper introduces the MANet (Mixed Aggregation Network) module, replacing some C3k2 feature extraction units in the YOLO11n backbone. MANet simultaneously receives multi-scale feature maps as inputs and employs a hybrid aggregation mechanism to achieve cross-scale adaptive fusion of multi-sensory field features. Features at each scale are first segmented through channel splitting and then fed into the C2f Block for local feature extraction. Subsequently, the HyperC2 Block utilizes the Hyperedge mechanism to enable cross-scale semantic transfer, allowing each detection scale to perceive contextual information from other scales. The final output combines features from all scales with the original input through channel dimension concatenation. This cross-scale hybrid aggregation mechanism significantly enhances the network's multi-scale feature representation capability for densely distributed small targets in complex nighttime backgrounds, effectively mitigating target detail loss caused by low illumination.

(2) ADFPN-DASI Module: To address the issues of insufficient cross-scale information transfer and strong background noise interference in traditional feature pyramid networks, this paper designs the ADFPN-DASI module. Based on a bidirectional feature pyramid architecture, this module replaces the fixed fusion structure in traditional FPN with an improved DASI (Dimension-aware Selective Integration) node derived from HCFNet. In the top-down path, the DASI node simultaneously receives three-scale inputs (P3, P4, P5), performing adaptive weighting of multi-scale features in both channel and spatial dimensions to selectively enhance

effective feature responses while suppressing background noise. In the bottom-up path, the second DASI node performs cross-scale aggregation on enhanced features from each scale, completing secondary feature diffusion through symmetric sampling and splicing operations to ensure semantic-rich features are fully distributed across all detection scales. The coordinated operation of the two DASI nodes enables each detection scale to obtain context-rich feature representations through dimension-aware adaptive fusion, maintaining lightweight advantages while collaboratively improving multi-scale detection capabilities for dense small targets in nighttime scenarios.

2.3.1. MANet

To address the limitations of backbone networks in nighttime aerial photography for multi-scale feature extraction of small targets, this study introduces the MANet (Mixed Aggregation Network) module to replace certain C3k2 feature extraction units in the YOLO11n backbone. Unlike the original C3k2 module that extracts features locally at a single scale, MANet employs a cross-scale hybrid aggregation mechanism to simultaneously model semantic relationships across multiple detection scales, making it more suitable for feature enhancement of densely distributed small targets under nighttime low signal-to-noise ratio conditions.

As shown in Figure 3, the MANet's overall processing workflow consists of three distinct phases.

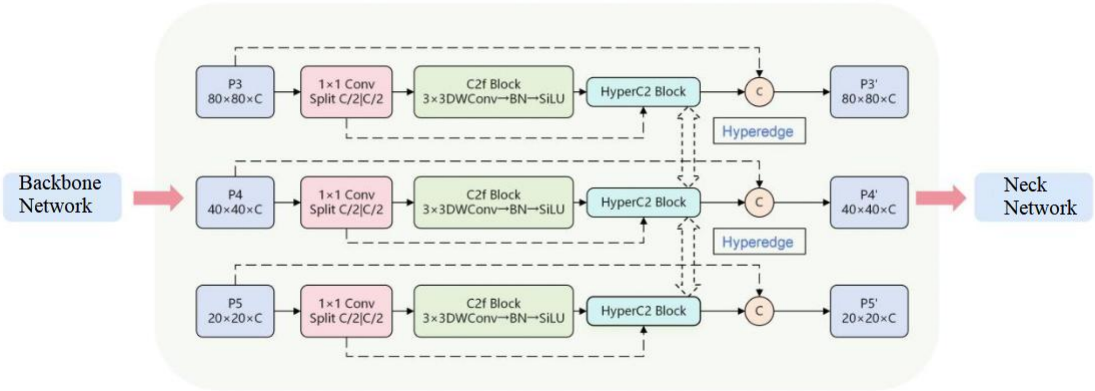


Figure 3. MANet structure diagram

(1) Multi-scale Input and Channel Splitting: MANet processes three-scale feature maps (P3: 80×80×C, P4: 40×40×C, P5: 20×20×C) simultaneously. Each scale's features undergo 1×1 convolution for channel splitting (Split C/2 | C/2), dividing the feature channels into two streams—one for local feature extraction and the other for cross-scale semantic transfer. This approach maintains complete feature information across scales while controlling computational load.

(2) Local Feature Extraction and Cross-Scale Semantic Transfer: The segmented features are fed into the C2f Block for local feature extraction, which employs a 3×3 depthwise separable convolution (3×3DWConv→BN→SiLU) to efficiently capture local details across scales while reducing parameter size and maintaining receptive field coverage. The extracted features are then

processed by the HyperC2 Block, which establishes explicit cross-scale semantic connections between P3, P4, and P5 scales through the Hyperedge mechanism. This enables each scale to perceive and integrate contextual semantic information from other scales. This mechanism is particularly critical in nighttime scenes—when target responses become weak due to low illumination, semantic supplements from other scales effectively compensate for feature incompleteness.

(3) Feature Fusion and Output: After cross-scale aggregation, the enhanced features from each scale are concatenated (concat) with the original input features along the channel dimension, forming an enhanced feature representation that integrates local details and global semantics. The final output consists of three multi-scale feature maps (P3 ', P4', P5') for subsequent use by the neck network and detection head.

Through the cross-scale hybrid aggregation mechanism, MANet enables the model to simultaneously integrate local detail information and global semantic context across three detection scales. This significantly enhances the feature representation capability for densely distributed small targets in complex nighttime backgrounds, thereby reducing false negatives and improving overall detection accuracy.

2.3.2. ADFPN-DASI

The ADFPN-DASI framework employs a bidirectional feature fusion pathway, replacing the fixed fusion structure of traditional FPNs with a DASI (Dimension-Aware Selective Integration) module enhanced by HCFNet (Zhu et al., 2021) as the core aggregation node. Unlike conventional FPNs that rely solely on simple upscaling and concatenation for cross-scale feature fusion, the DASI node performs adaptive weighting of multi-scale features simultaneously across channel and spatial dimensions. It selectively amplifies target-relevant feature responses while actively suppressing noise interference from complex nighttime backgrounds, making it particularly suitable for multi-scale detection of dense small targets in low-light conditions.

The ADFPN-DASI workflow comprises two distinct approaches: a top-down and a bottom-up methodology, as illustrated in Figure 4.

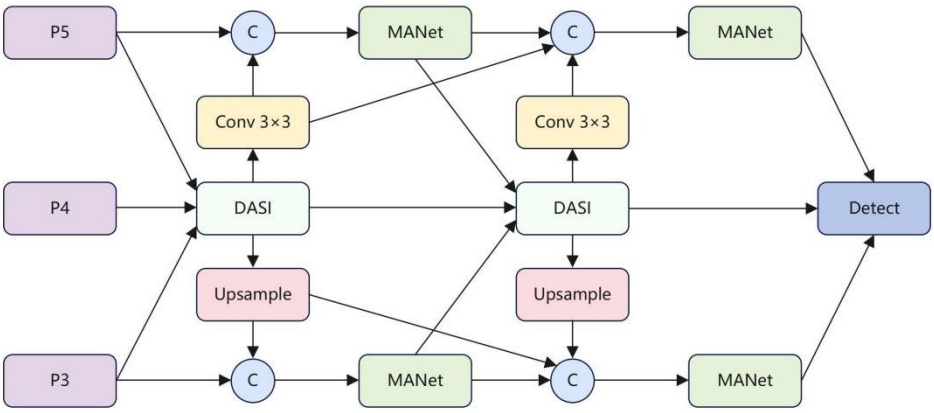


Figure 4. ADFPN-DASI structural diagram

(1) Top-down pathway: The P5 feature map undergoes down-sampling via Convolutional Downsampling to align spatial resolution, while the P3 feature map is upsampled to a uniform scale. The three feature streams (P3, P4, P5) then converge into the first DASI node. The DASI node first evaluates the importance of features across channels and applies adaptive weighting to filter redundant channel responses. It then selectively activates each feature map position in the spatial dimension, highlighting target region features to achieve active background noise suppression during fusion. The aggregated features are fed into the MANet module for further cross-scale local feature extraction, yielding preliminary enhanced feature representations at each scale.

(2) Bottom-up Path: The feature enhancement outputs from the top-down path are reintegrated into the second DASI node for secondary dimension-aware adaptive fusion. Symmetrically aligned with the first DASI node, this path undergoes corresponding sampling and stitching operations to fully integrate low-level detail information with high-level semantic information, completing secondary feature diffusion. The final output consists of three bidirectional path-enhanced feature maps (P3, P4, P5), which are respectively fed into corresponding scale detection heads for target localization and classification.

Two DASI nodes collaboratively handle cross-scale aggregation in bidirectional paths, forming a complete bidirectional enhancement loop that combines "top-down semantic transfer with bottom-up detail supplementation." This ensures each detection scale receives context-rich feature representations through dimension-aware adaptive fusion. The design proves particularly critical in nighttime dense small-object scenarios: the top-down path supplements global semantic information for small-scale targets, while the bottom-up path preserves fine spatial localization details for large-scale features. Their synergy significantly enhances multi-scale detection capabilities for nighttime dense small objects.

3. Experiments and Result Analysis

3.1. Introduction to the Dataset

The experiment was conducted using the DroneVehicle dataset, which contains 28,439 registered RGB-infrared image pairs. These images cover diverse urban scenarios including roads, parking lots, and residential areas, with multiple lighting conditions (daytime and nighttime). The dataset is labeled with five categories: car, truck, bus, van, and freight-car. In this study, we follow the official classification to separate the training and validation sets.

3.2. Experimental Environment and Evaluation Indicators

The hardware and software configurations used in the experiment are shown in Table 1. To comprehensively evaluate model performance, this study employs five evaluation metrics: Precision, Recall, mAP, Parameters (M), and GFLOPs. Precision measures the proportion of predicted positive samples that are actually positive, while Recall indicates the proportion of true positive samples correctly detected. mAP represents the average AP value across all categories. Specifically, we use mAP@0.5 and mAP@0.5:0.95 as metrics, corresponding to average

precision at IoU thresholds of 0.5 and 0.5–0.95 (with a step size of 0.05). Parameters (in megabytes) and GFLOPs (in gigaflops) reflect the model's storage overhead and computational inference load, respectively, serving as key indicators for measuring model lightweighting.

The model was trained for 250 epochs with a batch size of 64 and an input resolution of 640×640 pixels. The SGD optimizer was adopted with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005. Early stopping with a patience of 100 epochs was applied to prevent overfitting.

Table 1. Experimental Environment Configuration

environment	Parameter configuration
operating system	Windows11
CPU	Intel® Xeon® Gold 5418Y processor, 10 cores
GPU	NVIDIA GeForce RTX4090
exploitation environment	PyCharm
compiling environment	Python 3.10
Deep learning framework	Pytorch 2.2.1

3.3. Ablation Experiment

Table 2 presents the detection performance of YOLO11n under various input modalities and fusion weight configurations, demonstrating the effectiveness of preprocessing enhancement and fusion strategies.

Table 2. Experimental results of different modalities and fusion weight ablation

Model configuration	P	R	mAP50	mAP50-95	GFLOPs	Parameters
YOLO11n+RGB	71.51	65.12	69.08	43.70	6.3	2.58 million
YOLO11n+IR	77.56	74.56	78.79	59.05	6.3	2.58 million
YOLO11n+ fusion (α=0.5)	76.73	75.17	78.49	58.25	6.3	2.58 million
YOLO11n+ fusion (α=0.6)	76.96	74.83	78.83	58.73	6.3	2.58 million
YOLO11n+ fusion (α=0.7)	77.41	75.34	79.14	59.15	6.3	2.58 million
YOLO11n+ fusion (α=0.8)	76.99	75.55	78.78	59.17	6.3	2.58 million

Table 2 reveals the following conclusions: The single infrared modality (78.79%) significantly outperforms the single visible RGB modality (69.08%), with a 9.71 percentage point improvement, demonstrating the core advantage of infrared modality in nighttime scenes. Weighted fusion outperforms single RGB modality across all weight configurations, achieving optimal mAP50 (79.14%) when the infrared weight α=0.7. This indicates that infrared information should dominate in nighttime scenes while retaining some visible texture information to further enhance accuracy. The fusion scheme (79.14%) surpasses the single infrared modality (78.79%), proving that the visible images enhanced by CLAHE+Gamma indeed provide incremental information for fusion.

Table 3 presents the ablation results of three enhancements—MANet, FDPN, and DASI—introduced progressively on the optimal fusion input (α=0.7).

Table 3. Results of Ablation Experiment for Network Modules

Model configuration	P	R	mAP50	mAP50-95	GFLOPs	Parameters
YOLO11n+ fusion (α=0.7)	77.41	75.34	79.14	59.15	6.3	2.58 million
+MANet	77.61	77.19	80.33	60.41	8.4	3.78 million
+ADFPN-DASI	78.46	75.39	79.51	59.37	7.4	2.68 million
+MANet+ADFPN-DASI	79.12	76.01	80.53	60.28	8.6	3.28 million

As shown in Table 3, the introduction of MANet increased mAP50 by 1.19 percentage points (79.14%→80.33%) and recall rate by 1.85 percentage points, demonstrating that the hybrid aggregation mechanism effectively enhances the backbone's feature extraction capability for small nighttime targets. The addition of ADFPN-DASI improved mAP50 by 0.37 percentage points (79.14%→79.51%), achieving effective multi-scale feature optimization with only an increase of approximately 100,000 parameters, highlighting its lightweight characteristics. When both modules are used simultaneously, the optimal mAP50 (80.53%) is achieved, representing a 1.39 percentage point improvement over the fusion baseline and an 11.45 percentage point enhancement compared to the YOLO11n+RGB baseline, validating the synergistic effects of each module.

3.4. Comparison of Mainstream Algorithms

Table 4 compares the final method proposed in this paper with mainstream detection models of comparable scale on the DroneVehicle dataset.

Table 4. Comparison of Experimental Results with Mainstream Methods

model	P	R	mAP50	mAP50-95	GFLOPs	Parameters
YOLOv5n	76.71	75.13	78.49	58.39	5.8	2.18 million
YOLOv8n	77.03	74.55	78.21	58.38	6.8	2.69 million
YOLOv12n	76.63	76.25	79.32	59.28	5.8	2.51 million
YOLOv13n	76.79	74.32	78.09	58.72	6.1	2.45 million
YOLO11n (baseline)	71.51	65.12	69.08	43.70	6.3	2.58 million
Methodology of this paper	79.12	76.01	80.53	60.28	8.6	3.28 million

As shown in Table 4, our proposed method outperforms mainstream baseline models (e.g., YOLOv5n, YOLOv8n, YOLOv12n, and YOLOv13n) in both mAP50 and mAP50-95 metrics, achieving a 1.21 percentage point improvement over the closest competitor YOLOv12n. While the proposed method requires more parameters and GFLOPs than MANet before its integration, these metrics remain within a lightweight range, ensuring deployability on edge devices in drones. These results conclusively validate the effectiveness of our multimodal fusion and network optimization strategy.

3.5. Visualization Results

To further validate the effectiveness of our proposed method, Figures 5-8 systematically compare YOLO11n, YOLOv12n, and our method through visual detection in multiple typical application scenarios, clearly demonstrating their performance differences under various challenging conditions.

Figure 5 demonstrates a relatively uniform distribution of scene targets under moderate lighting conditions, which facilitates focused evaluation of the modeling capabilities of various methods in target confidence assessment. Comparative results reveal that our proposed method achieves significantly higher confidence scores across all detection targets compared to YOLO11n and YOLOv12n. This phenomenon indicates that the CLAHE+Gamma collaborative enhancement preprocessing and the integration of the MANet attention mechanism effectively strengthen the model's feature representation capability in target regions, enabling the network to complete target localization and classification with higher confidence.

Figure 6 depicts a scene with numerous targets exhibiting varying degrees of occlusion and scale variations. Experimental results demonstrate that YOLOv12n exhibits significant missed detection in this scenario, failing to effectively retrieve some small and edge targets. While YOLO11n shows some improvement, its detection completeness remains inadequate. The proposed method achieves target detection quantities comparable to or even surpassing YOLO11n, with a markedly reduced missed detection rate. These results conclusively validate the positive

impact of the ADFPN-DASI multi-scale feature fusion module in enhancing small target perception capabilities and suppressing missed detection.

Figure 7 depicts a complex background with coexisting multi-category objects, demanding high classification discrimination capability from the model. Comparative analysis reveals that both YOLO11n and YOLOv12n exhibit certain category misclassifications, misclassifying background interference or adjacent objects. In contrast, our method demonstrates the lowest false detection rate and superior category prediction accuracy. This indicates that through multimodal feature fusion and attention enhancement, the model achieves enhanced inter-category discrimination, enabling more precise differentiation of objects across diverse categories in complex scenarios.

Figure 8 presents a typical nighttime drone aerial photography scenario with high object density and significant overlap, representing one of the most challenging conditions. Experimental results demonstrate that both YOLO11n and YOLOv12n exhibit varying degrees of confidence degradation and missed detection, indicating overall unstable performance. In contrast, our proposed method achieves more accurate object localization and consistently reliable detection results. This demonstrates that the infrared-visible weighted fusion strategy effectively compensates for information loss in low-light nighttime visibility. By leveraging the infrared modality's high sensitivity to thermal radiation targets, the model significantly improves detection performance in dense and complex nighttime environments.

The comprehensive visual comparison results demonstrate that the proposed method outperforms the baseline in multiple metrics including confidence level, recall rate, classification accuracy, and nighttime robustness, further validating the effectiveness of the synergistic effects among the improved modules.

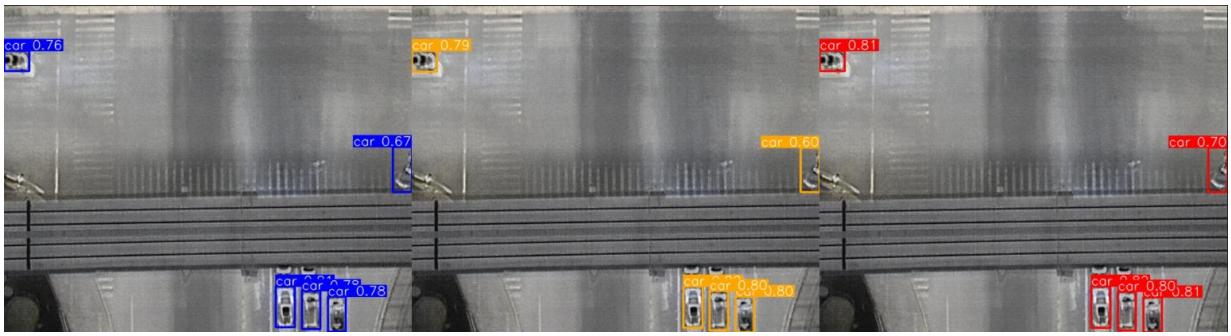


Figure 5. Visualizes Figure

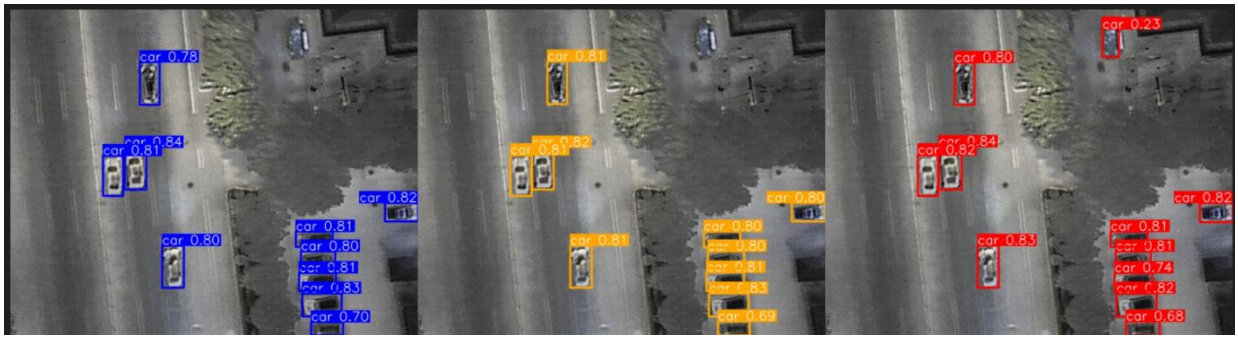


Figure 6. Visualizes Figure

Figure 7. Visualizes Figure

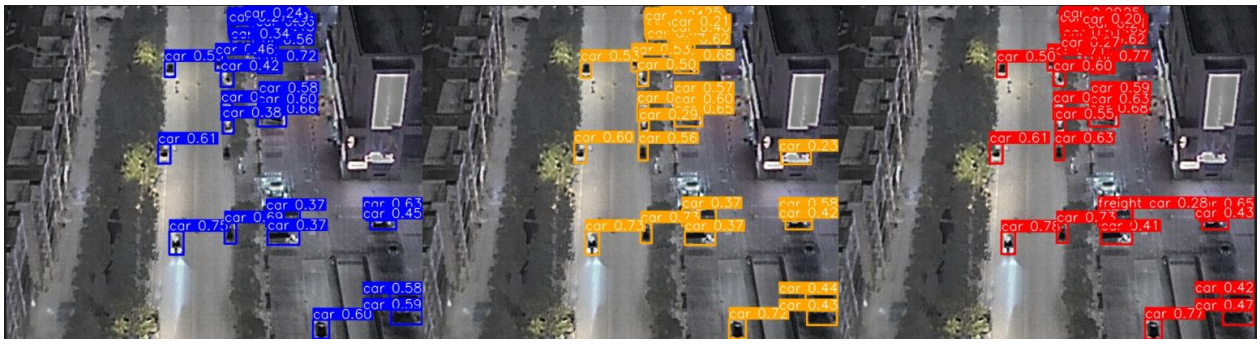


Figure 8. Visualizes Figure

4. Conclusion

This study addresses key challenges in nighttime UAV vehicle detection, including low illumination, severe noise, and multi-modal data fusion difficulties. We propose a multi-modal fusion-enhanced detection method: image preprocessing adopts a CLAHE-Gamma collaborative enhancement strategy to improve low-light contrast and detail visibility; a weighted fusion mechanism fuses infrared (thermal radiation sensitivity) and visible (rich texture-semantic information) images for modal complementarity; the network integrates the MANet attention mechanism and ADFPN-DASI feature pyramid module to enhance multi-scale feature aggregation and cross-layer interaction, boosting small/dense target detection.

Experiments on the DroneVehicle dataset show the method achieves an mAP50 of 80.53%, 11.45 percentage points higher than the baseline, validating its effectiveness. Ablation experiments confirm the independent and synergistic contributions of the enhancement module, fusion strategy, and network improvements.

Limitations include: fixed-weight fusion fails to adapt to dynamic lighting in real aerial scenarios, leading to performance bottlenecks; unoptimized computational complexity restricts deployment on resource-constrained edge platforms. Future work will focus on lightweight deployment and diverse dataset construction to enhance robustness and practicality.

Author Contributions:

All authors have read and agreed to the published version of the manuscript.

Funding:

This research received no external funding.

Institutional Review Board Statement:

Not applicable.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

Not applicable.

Conflict of Interest:

The authors declare no conflict of interest.

References

Feng, Y., Huang, J., Du, S., et al. (2024). Hyper-YOLO: When visual object detection meets hypergraph computation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(4), 2388–2401.

Khanam, R., & Hussain, M. (2024). YOLO11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.

Li, X., Li, X., et al. (2024). A survey of object detection for UAVs based on deep learning. Remote Sensing, 16(1), 149.

Ma, J., Ma, Y., & Li, C. (2019). Infrared and visible image fusion methods and applications: A survey. Information Fusion, 45, 153–178.

Ma, M., Wang, H., & Wang, J. (2023). An underwater image enhancement algorithm based on improved MSRCR–CLAHE fusion. Infrared Technology, 45(1), 23–32.

Mittal, P., Singh, R., & Sharma, A. (2020). Deep learning-based object detection in low-altitude UAV datasets: A survey. Image and Vision Computing, 104, 104046.

Persiya, J., & Sasithradevi, A. (2025). Synergistic fusion: An integrated pipeline of CLAHE, YOLO models, and advanced super-resolution for enhanced thermal eye detection. PLOS ONE, 20(7), e0328227.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Sun, Y., Cao, B., Zhu, P., et al. (2022). Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(10), 6700–6713.

Tian, D., Yan, X., Zhou, D., et al. (2024). IV-YOLO: A lightweight dual-branch object detection network. Sensors, 24(19), 6181.

Tian, Y., Ye, Q., & Doermann, D. (2025). YOLOv12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524.

Wang, G., Wang, G., et al. (2023). UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. Sensors, 23(16), 7190.

Yuan, Z., Zeng, J., Wei, Z., et al. (2023). CLAHE-based low-light image enhancement for robust object detection in overhead power transmission system. IEEE Transactions on Power Delivery, 38(3), 2240 – 2243.

Zhang, Q., Qiu, L., Zhou, L., et al. (2024). ESM-YOLO: Enhanced small target detection based on visible and infrared multi-modal fusion. In Proceedings of the Asian Conference on Computer Vision (pp. 1454 – 1469).

Zhang, Y., Dai, Z., Pan, C., et al. (2025). NOC-YOLO: An exploration to enhance small-target vehicle detection accuracy in aerial infrared images. Infrared Physics & Technology, 149, 105905.

Zhu, P., Wen, L., Du, D., et al. (2021). Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11), 7380 – 7399.

Journal of Computer Science and Digital Technology, 2026, 2(1), 51-70
https://doi.org/10.71204/vwpczb65

A Discussion on Ethics, Trust and Security Governance in the Evolution of Artificial Intelligence Technology

Shufeng Wang [1, *]

[1] School of Computer Science and Technology, Taiyuan Normal University, Shanxi 030619, China

*** Correspondence:**

Shufeng Wang

18235230439@163.com

Received: 10 March 2026 / Accepted: 27 March 2026 / Published online: 30 March 2026

Abstract

With the rapid advancement and widespread deployment of artificial intelligence , issues related to ethics, trust, and security have become increasingly prominent, posing significant challenges to its sustainable development. This paper aims to systematically investigate these intertwined challenges by constructing a structured analytical framework encompassing three core dimensions: ethics, trust, and security.Methodologically, the study analyzes the evolutionary trajectory and application contexts of AI, and synthesizes key risk patterns and governance concerns. It examines major ethical issues, including algorithmic bias, data privacy, responsibility attribution, and value alignment, and further explores the underlying tension between technological rationality and social values.In addition, the paper proposes a multi-level trust formation mechanism based on technological reliability, institutional assurance, and public cognition, emphasizing that trust emerges from the joint interaction of technical systems and governance structures. At the security level, it highlights the complexity and cross-domain nature of AI risks, and advocates for a full life-cycle governance framework integrating risk assessment, technical auditing, and continuous supervision.The study's key contribution lies in integrating ethical, trust, and security perspectives into a unified analytical framework, and in proposing a collaborative governance approach involving governments, enterprises, and international stakeholders.

Keywords: Security Risk Management; AI Ethics; Algorithmic Fairness; Explainable AI

1. Introduction

Over the past few decades, artificial intelligence technology has undergone leapfrog development and, through technological breakthroughs such as deep learning, big data and computing power improvement, has gradually moved from laboratory research to large-scale application, and has penetrated into many key fields such as medical care, finance, transportation,

education and entertainment. The spread of AI technology has not only significantly improved the efficiency of social operation and resource allocation capabilities, but also profoundly reshaped the human-computer interaction mode and decision-making mechanism. However, while the widespread application of artificial intelligence has brought significant economic and social benefits, it has also triggered a series of complex and deep-seated social problems (Russell et al., 2021). Ethical issues have become a core topic of widespread discussion globally. From the perspective of value norms and social governance, the decision-making process of AI systems often lacks transparency, and structural biases or discrimination mechanisms may be embedded in the data collection, processing and algorithm modeling process. Such technical biases may not only damage individual privacy and personal dignity, but may also exacerbate social inequality and resource allocation imbalance, thus posing a serious challenge to the principle of social fairness. In this context, building a systematic, operable and universally binding artificial intelligence ethical framework has become an important issue in the global governance system. At the same time, the rapid iteration and widespread application of artificial intelligence technology are constantly impacting the existing social norm system and ethical boundary structure. As the participation of intelligent systems in the decision-making process continues to deepen, the traditional responsibility allocation mechanism and rights and obligations framework built with human subjects as the core are facing structural challenges (Batool et al., 2025). These deep-seated theoretical propositions urgently need to be systematically responded to and institutionalized within the framework of ethical philosophy, jurisprudence and public governance.

From the perspective of social trust mechanism, trust constitutes the basic condition for the widespread application of artificial intelligence in social embedding. Its formation and maintenance directly affect the social acceptance and sustainable development capability of AI system. Although artificial intelligence has shown significant advantages in efficiency optimization and decision support, due to the complexity of algorithm models and the opacity of decision paths, the public often lacks a full understanding of its operating logic and risk control mechanism, which in turn affects the degree of trust in the system. Therefore, how to build a trustworthy artificial intelligence system with verifiability and accountability by enhancing the interpretability of algorithms, improving the transparency of operation and clarifying the responsibility allocation mechanism has become an important direction for current technology governance and institutional innovation. At the level of security governance, the risks faced by artificial intelligence systems also show a trend of increasing complexity and diversification. With the continuous expansion of algorithm capabilities and application scenarios, problems such as malicious attacks, adversarial sample interference, data tampering and model loss of control are constantly emerging, and their potential impact may affect public safety, economic stability and even national security (McCormack et al., 2024). Especially under the dual effect of "black box" characteristics and adaptive learning mechanism, the unpredictability of some system behaviors has been significantly enhanced, increasing the difficulty of risk assessment and responsibility definition. Therefore, while continuously promoting technological innovation and industrial upgrading, constructing a systematic safety assessment system and a multi-level risk prevention and control mechanism to ensure the controllability and reliability of artificial intelligence systems has become a critical issue that society must address. This study aims to systematically

explore the ethical, trust, and security issues in artificial intelligence, analyze how these issues intertwine and influence each other, and provide useful insights and references for academic research and practical applications in related fields.

2. Basic Concepts and Theoretical Foundations of Artificial Intelligence

2.1. Definition of Artificial Intelligence

Artificial Intelligence, as an important branch of computer science, has been evolving and expanding with technological development since its proposal in the mid-20th century. Generally speaking, artificial intelligence refers to a series of theories, methods and technologies that simulate, extend or expand human intelligent activities through computer systems. Its core goal is to enable machines to have the ability to perceive the environment, understand information, learn from experience, reason and make decisions, and act autonomously (Li et al., 2024). The formal proposal of the concept of artificial intelligence is usually traced back to the Dartmouth Conference in 1956. At this conference, John McCarthy first proposed the term "Artificial Intelligence" and defined it as "the science and engineering of how to make machines exhibit human-like intelligent behavior", emphasizing that through algorithm design and program implementation, machines can exhibit human-like cognitive abilities in specific tasks. At the same time, Alan Turing, in the earlier Turing Test, explored from a behaviorist perspective whether machines can exhibit intelligent reactions that are indistinguishable from humans, providing an important testable theoretical basis for artificial intelligence research.

From the perspective of the degree of intelligence realization, artificial intelligence is usually divided into weak artificial intelligence (Weak AI) and strong artificial intelligence (Strong AI). Weak artificial intelligence, also known as narrow-domain artificial intelligence, mainly focuses on achieving high efficiency in specific tasks or application scenarios, such as image recognition, speech recognition and natural language processing. Most practical application systems currently belong to this category. In contrast, strong artificial intelligence refers to intelligent systems that possess general cognitive abilities comparable to or even surpassing those of humans, capable of transfer learning, autonomous understanding and creative thinking between different tasks. Although strong artificial intelligence is still in the theoretical exploration stage, its concept is of great significance in the research on artificial intelligence ethics, philosophy and future technological development. Overall, artificial intelligence is not a single technology or method, but a typical interdisciplinary research field that integrates the theoretical foundations of multiple disciplines such as computer science, mathematics, cognitive science, neuroscience and philosophy (Kattnig et al., 2024). Therefore, in academic research, the definition of artificial intelligence should not only focus on the algorithm and system implementation at the technical level, but also systematically understand and analyze it from multiple dimensions such as the essence of intelligence, behavioral performance and decision-making mechanism, so as to more comprehensively reveal the development connotation and research value of artificial intelligence.

2.2. Main research content of artificial intelligence

As a highly comprehensive interdisciplinary field, artificial intelligence covers multiple levels

from basic theory to applied technology. Around the core goal of building an intelligent system with perception, learning, reasoning and decision-making capabilities, artificial intelligence has gradually formed several interrelated research directions, mainly including machine learning, deep learning, natural language processing, computer vision, knowledge representation and reasoning, and reinforcement learning. These research directions are intertwined in theory and method, and together constitute the main framework of the artificial intelligence technology system (Sistla, 2024). Among them, machine learning, as an important theoretical foundation of artificial intelligence, takes data-driven as its core paradigm. Its basic idea is to enable computer systems to automatically identify potential patterns and complete prediction or decision-making tasks through training and modeling of sample data. Compared with the traditional method of relying on explicit rule programming, machine learning emphasizes the optimization of parameters and the abstraction of patterns in the process of data input and model training, thereby significantly improving the system's adaptive ability and generalization performance. Based on the differences in learning mechanisms and data structures, machine learning can usually be divided into supervised learning, unsupervised learning and reinforcement learning. Through its high flexibility and adaptability in data modeling, pattern recognition, and decision optimization, machine learning not only provides systematic algorithmic tools and modeling methods for artificial intelligence research, but also gradually forms a sustainable and expandable technical framework and methodological foundation, making it an indispensable core technical support in the design and implementation of current intelligent systems.

Based on the continuous development of the theoretical system of machine learning, deep learning, as an important branch, has significantly improved the ability to represent and process complex data by constructing multi-layer neural network structures. The core advantage of deep learning is that it can automatically learn hierarchical feature representations from large-scale datasets, thereby reducing the dependence on manual feature engineering and enabling it to achieve breakthrough results in high-dimensional and unstructured data processing tasks. Typical model structures include convolutional neural networks, recurrent neural networks, and Transformer, which has been widely used in recent years. The development of deep learning has not only promoted the rapid progress of image recognition, speech recognition and natural language processing, but also marked the important transformation of the artificial intelligence research paradigm from "feature engineering driven" to "representation learning driven" (Vercelli, 2024). At the same time, knowledge representation and reasoning, as an important research direction in the early development of artificial intelligence, focuses on how to represent knowledge in a formal way and make inferences through logical rules, providing an important theoretical foundation for applications such as expert systems and semantic networks. Although data-driven methods have gradually become dominant in recent years, knowledge representation and symbolic reasoning still have irreplaceable value in interpretable artificial intelligence and complex decision-making systems. Overall, the core research content of artificial intelligence includes not only data-driven learning algorithm systems, but also cognitive mechanisms centered on knowledge expression and logical reasoning, and extends to multiple intelligence levels such as perception, cognition and decision-making. The various research directions form a relationship of mutual support and synergistic evolution in theory and technology, jointly promoting the

transformation of artificial intelligence from dedicated systems for single tasks to comprehensive intelligent systems with multiple functions (Lainjo, 2024). With the continuous deepening of interdisciplinary integration, the research boundaries of artificial intelligence are continuously expanding to more forward-looking fields such as multimodal learning, human-machine collaboration and general intelligence, providing a more solid theoretical foundation and methodological support for the continuous innovation and system upgrading of intelligent technologies.

2.3. The theoretical foundation of artificial intelligence

The development of artificial intelligence not only relies on continuous breakthroughs in engineering technology, but is also deeply rooted in the theoretical system of multidisciplinary integration. As a comprehensive discipline, the theoretical foundation of artificial intelligence involves a wide range of research fields such as mathematical theory, statistical learning theory, information theory, control theory and cognitive science. These theories together constitute an important supporting framework for the design of artificial intelligence algorithms and the construction of intelligent systems. In this system, mathematical foundation is regarded as the most core theoretical pillar. Among them, linear algebra provides key formal tools for matrix operations and feature representation in neural networks. Vector space theory, feature decomposition methods and matrix transformation mechanisms constitute important mathematical foundations for the computational structure of deep models. At the same time, probability theory and mathematical statistics provide theoretical basis for uncertainty modeling and statistical inference, and play an important role in the method system of Bayesian inference, hidden Markov models and probabilistic graphical models (Sharma, 2023). In addition, optimization theory lays the algorithmic foundation for model parameter learning. Among them, gradient descent and convex optimization methods have become the core technical paths in deep learning model training. From a theoretical perspective, the learning process in artificial intelligence can essentially be regarded as an optimization problem in a high-dimensional parameter space. Its convergence and stability largely depend on the support of rigorous and systematic mathematical theory.

The development of artificial intelligence not only relies on continuous breakthroughs in engineering technology, but is also deeply rooted in the theoretical system of multidisciplinary integration. As a comprehensive discipline, the theoretical foundation of artificial intelligence involves a wide range of research fields such as mathematical theory, statistical learning theory, information theory, control theory and cognitive science. These theories together constitute an important supporting framework for the design of artificial intelligence algorithms and the construction of intelligent systems. In this system, mathematical foundation is regarded as the most core theoretical pillar. Among them, linear algebra provides key formal tools for matrix operations and feature representation in neural networks. Vector space theory, feature decomposition methods and matrix transformation mechanisms constitute important mathematical foundations for the computational structure of deep models. At the same time, probability theory and mathematical statistics provide theoretical basis for uncertainty modeling and statistical inference, and play an important role in the method system of Bayesian inference, hidden Markov

models and probabilistic graphical models (Taeihagh, 2025). In addition, optimization theory lays the algorithmic foundation for model parameter learning. Among them, gradient descent and convex optimization methods have become the core technical paths in deep learning model training. From a theoretical perspective, the learning process in artificial intelligence can essentially be regarded as an optimization problem in a high-dimensional parameter space. Its convergence and stability largely depend on the support of rigorous and systematic mathematical theory.

3. Key Technological Advances in Artificial Intelligence

3.1. Machine Learning

Machine learning is one of the most core branches of artificial intelligence. Its basic idea is to enable computer systems to automatically learn the potential patterns in data without explicit rule programming through data-driven methods, and to complete prediction, classification or decision-making tasks accordingly. Unlike traditional algorithms that rely on manual rule design, machine learning emphasizes that the model continuously improves its performance through parameter optimization and pattern abstraction during sample training. Its essence can be regarded as the process of approximating an unknown function under given data distribution conditions. From a methodological perspective, the construction of machine learning models usually includes three key links: model selection, parameter estimation and performance evaluation. Model selection mainly involves the determination of hypothesis space and control of model complexity. Parameter estimation relies on optimization algorithms to solve the objective function. Performance evaluation is carried out by quantitative analysis of the model's predictive ability through methods such as cross-validation, error index or confusion matrix (Corrêa et al., 2023). In high-dimensional data environments, models are prone to overfitting. Therefore, regularization methods, model pruning and early stopping strategies are widely used to improve the generalization ability and stability of the model, thereby ensuring the reliable performance of the machine learning system on unknown data.

From the perspective of learning paradigms, machine learning can generally be divided into three basic types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning trains a model by minimizing the error between the predicted value and the true value under the condition of given input-output sample pairs. Its typical applications include classification and regression problems. Representative algorithms include linear regression, logistic regression, support vector machine, and neural network. Its theoretical basis mainly comes from the principle of risk minimization and statistical learning theory. Unsupervised learning achieves pattern discovery and feature representation by mining the internal structure of data under the condition of lack of label information. Typical methods include cluster analysis, principal component analysis, and autoencoders. These methods can reveal the potential distribution structure or low-dimensional representation of data and play an important role in tasks such as data preprocessing, anomaly detection, and feature extraction. Reinforcement learning is a learning paradigm based on interaction mechanism. Its core idea is to continuously

optimize the behavior strategy through continuous interaction between the agent and the environment during the trial and error process (Xu et al., 2024). Reinforcement learning is typically built on the Markov decision process framework, aiming to maximize long-term cumulative rewards through iterative updates of the value function or policy function. Representative methods include Q-learning, policy gradient methods, and deep reinforcement learning, and have achieved significant results in fields such as game theory decision-making, robot control, and autonomous driving.

With the exponential growth of data scale and the continuous improvement of computing power, the machine learning method system is gradually evolving from traditional models with shallow structures to multi-layered, highly complex deep architectures. This evolution not only significantly enhances the expressive power of the model, but also enables the algorithm to more effectively characterize the complex nonlinear feature relationships in high-dimensional data. In this process, ensemble learning methods effectively reduce the variance and bias of a single model by constructing multiple base learners and strategically combining them, achieving synergistic optimization of prediction performance and model stability, and showing strong generalization ability and robustness in structured data analysis tasks (Ricciardi et al., 2025). At the same time, the continuous improvement of large-scale distributed computing frameworks and parallel training mechanisms provides important technical support for the efficient training and large-scale deployment of machine learning algorithms in massive data environments. Relying on distributed storage systems, parameter server architectures and high-performance computing resources, the computational bottleneck in the model training process is significantly alleviated, thereby promoting the widespread application of machine learning in industrial scenarios. Overall, machine learning has gradually built a key technical link between data resources and intelligent applications, playing a fundamental and pivotal role in data value mining and intelligent decision support systems.

3.2. Deep Learning

Deep learning, as an important branch of machine learning (its structural relationship is shown in Figure 1), has become one of the core driving forces for breakthroughs in artificial intelligence technology in recent years. Its basic idea lies in constructing a multi-layered nonlinear neural network structure to achieve hierarchical representation and automatic feature extraction of complex data. Unlike traditional machine learning methods that rely on manually designed features, deep learning emphasizes an end-to-end training mechanism, enabling the model to automatically learn high-level abstract features driven by large-scale data, thereby significantly improving the performance of complex pattern recognition tasks. Structurally, deep learning models typically consist of an input layer, multiple hidden layers, and an output layer. The hidden layers use nonlinear activation functions to map features layer by layer, allowing the model to gradually learn more abstract and higher-level representations. According to the general approximation theorem, multi-layered neural networks can theoretically approximate any continuous function, providing an important mathematical basis for the powerful expressive capabilities of deep learning models. During model training, parameters are typically calculated using the backpropagation algorithm, and weights are continuously updated using stochastic

gradient descent and its improved optimization algorithms to minimize the loss function and improve the model's predictive performance.

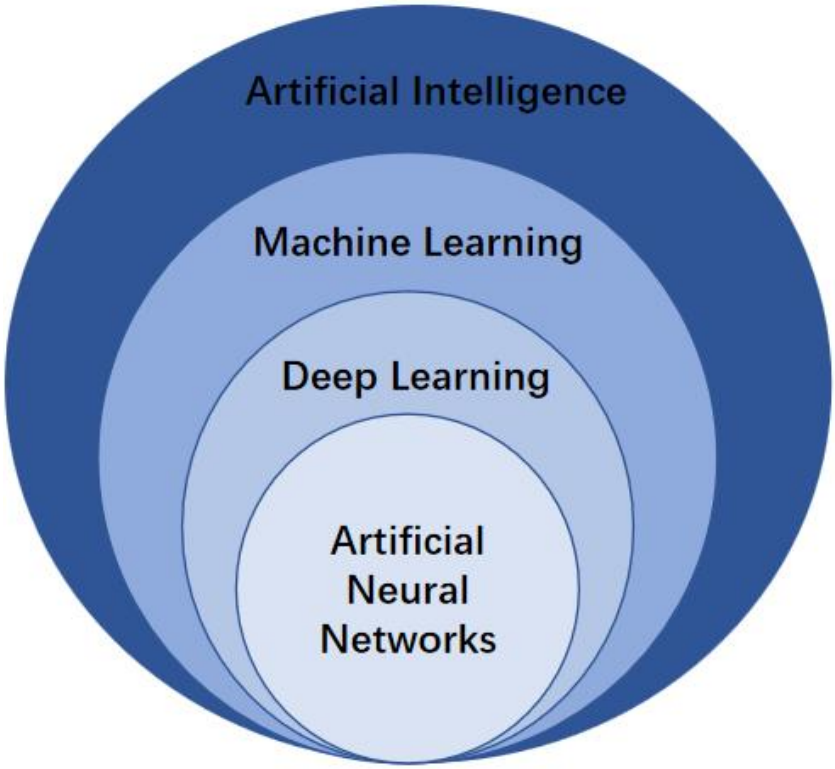


Figure 1. Hierarchical framework diagram of artificial intelligence technology

In terms of specific model structure, Convolutional Neural Network (CNN) is an important representative of the breakthroughs achieved by deep learning in the field of computer vision. CNN can effectively extract local spatial features through convolution operation and weight sharing mechanism, while significantly reducing the number of model parameters, thereby improving training efficiency and enhancing the generalization ability of the model. Therefore, this structure has shown excellent performance in tasks such as image classification, object detection and medical image analysis. On the other hand, Recurrent Neural Network (RNN) is mainly used to process sequential data with time dependence. It realizes the modeling of context information through recurrent connection structure and plays an important role in tasks such as natural language processing and speech recognition (Whig, 2025). However, traditional RNN is prone to gradient vanishing and gradient explosion problems during long sequence training. To solve this limitation, researchers have proposed improved structures such as Long Short-Term Memory Network (LSTM) and Gated Recurrent Unit (GRU), which enhance the model's ability to express long-term dependencies by introducing gating mechanism, thereby significantly improving the sequence modeling effect.

In recent years, the Transformer architecture based on attention mechanisms has become a significant milestone in the development of deep learning. Unlike RNNs, which rely on sequential computation, Transformers achieve global dependency modeling through self-attention, significantly improving parallel computing efficiency and model training speed. This architecture has spurred the development of large-scale pre-trained models in natural language processing and

propelled artificial intelligence research into a new phase centered on large-scale models. The success of the Transformer architecture has not only changed the traditional sequence modeling paradigm but also laid an important foundation for multimodal learning and cross-domain transfer learning. Overall, deep learning, relying on multi-layered neural network structures and automated feature representation mechanisms, has significantly enhanced the task processing capabilities of artificial intelligence systems in complex scenarios and high-dimensional data environments. With continuous innovation in model architecture, improved computing resources, and the development of cross-modal fusion technologies, deep learning is continuously driving the evolution of artificial intelligence from dedicated systems for single tasks to multifunctional integrated intelligent systems, and will play a sustained and far-reaching role in the future development towards higher levels of autonomy and general intelligence.

3.3. Large Model and Data-Driven Approach

With the continuous evolution of deep learning technology, artificial intelligence has gradually entered a development stage with "large-scale models" as its core feature. The so-called large model usually refers to a deep neural network model with a massive number of parameters, trained on large-scale data and capable of cross-task generalization. Its basic process is shown in Figure 2. Such models usually rely on the training paradigm of "pre-training and fine-tuning", are pre-trained in multi-domain corpora or multi-modal data environments, and are adaptively optimized on specific tasks to achieve wide application transfer. From the perspective of technical path, the development of large models is largely based on self-supervised learning mechanisms. By designing pre-training tasks that do not require manual annotation, the model can learn general representations in massive unlabeled data. For example, in the field of natural language processing, the model can perform language modeling through context prediction or mask word prediction, thereby mastering language structure and semantic rules; in the field of computer vision, feature representation learning can be achieved through image reconstruction or contrastive learning (Kashefi et al., 2024). Meanwhile, the development of large-scale models also exhibits a significant scale effect; that is, as the size of model parameters, the size of training data, and computing resources expand simultaneously, model performance shows a relatively stable upward trend. This pattern is often referred to as the law of scale, which reveals the functional relationship between model performance and parameter size, data size, and computing resources. Scaling not only enhances the model's ability to express complex patterns but also promotes cross-task transfer capabilities, enabling the model to demonstrate a certain degree of generalization ability even on tasks without explicit training. This ability is considered an important indicator of artificial intelligence moving towards a higher level of intelligence.

At the application level, the rise of large models has promoted the development of multi-domain technology integration and cross-modal learning. Large models based on a unified architecture can process multiple types of data such as text, images, speech and even video at the same time, realize the collaborative modeling and unified representation of multimodal information, and thus significantly improve the comprehensive perception and understanding capabilities of intelligent systems in complex scenarios. For example, through joint training of text and images, the model can complete tasks such as visual question answering, image and text

generation and scene understanding, demonstrating highly integrated information processing capabilities. However, at the theoretical and practical level, the rapid expansion of large models has also triggered widespread attention to the interpretability, stability and security of models. Due to the huge scale of model parameters, its internal representation structure and decision-making mechanism are often difficult to explain intuitively, which may bring potential challenges in high-risk application scenarios. At the same time, the bias that may exist in the training data may also be amplified by the model, thus causing fairness and ethical risks.

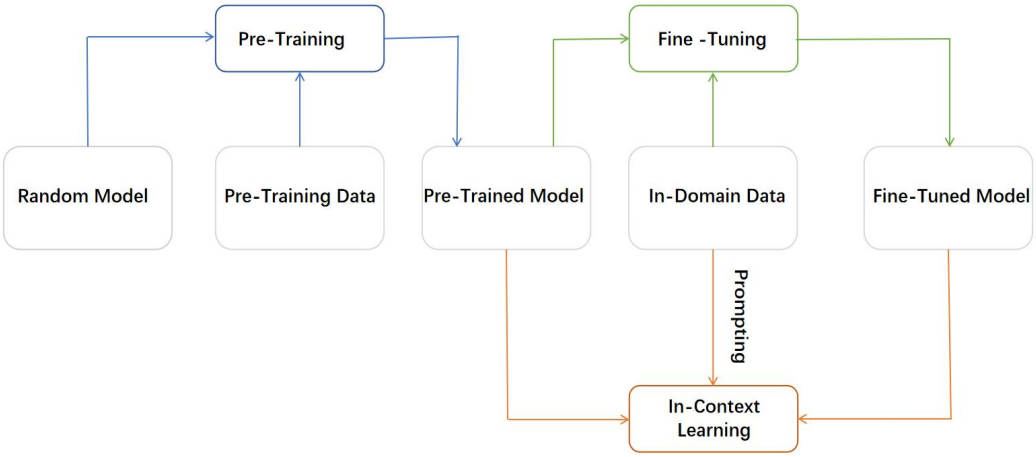


Figure 2. Data training flowchart

Therefore, while continuously promoting the expansion of model scale, how to achieve efficient training, model compression and knowledge transfer (such as knowledge distillation) and secure alignment has become an important direction of current research. Overall, the establishment of large-scale models and data-driven paradigms marks the entry of artificial intelligence into a new stage of development (Lahusen et al., 2024). With the continuous expansion of parameter scale and the widespread application of self-supervised learning mechanisms, intelligent systems are gradually demonstrating the ability to integrate and transfer knowledge across domains, laying an important foundation for exploring higher levels of general intelligence. However, this development path is also accompanied by significant resource consumption and governance challenges, including rapidly increasing computing power requirements, rising pressure on data compliance and security management, and the accumulation of potential ethical risks. Therefore, optimizing resource utilization efficiency and controlling risks while improving model performance will become a key issue that urgently needs attention in the future evolution of large-scale model systems.

4. Application areas of artificial intelligence

4.1. Medical Field

The application of artificial intelligence technology in the medical field is considered to be one of the most important directions with social value and development potential. With the continuous growth of medical data scale and the significant improvement of computing power, artificial intelligence has shown broad development prospects in disease screening, auxiliary diagnosis,

treatment decision support and health management. Its core goal is to improve the efficiency of medical services and the accuracy of diagnosis and treatment through data-driven intelligent analysis, thereby optimizing the allocation of medical resources and promoting the development of precision medicine. In the field of medical image analysis, artificial intelligence technology has made significant progress. Medical image data usually has the characteristics of large scale, high dimension and complex structure. The traditional manual interpretation method that relies on physician experience is not only inefficient, but also inevitably has certain subjective differences. Deep learning models based on convolutional neural networks can automatically complete image feature extraction and hierarchical representation learning, realize accurate identification and segmentation of lesion areas, and show high accuracy and stability in multiple disease detection tasks (Li et al., 2024). By constructing a large-scale labeled medical image dataset and carrying out multi-center clinical verification, artificial intelligence systems have, to a certain extent, the ability to assist in early disease screening and improve diagnostic consistency, thereby significantly improving the objectivity and standardization of clinical image analysis.

Artificial intelligence also plays a crucial role in clinical decision support. Intelligent systems based on Clinical Decision Support Systems (CDSS) can integrate multi-source information, including electronic medical records, laboratory test indicators, and historical case data, to comprehensively assess patient health and predict risks. Leveraging the pattern recognition and statistical inference capabilities of machine learning models, these systems can model and analyze disease progression trends, readmission risks, and individualized treatment strategies, thereby improving the accuracy and foresight of clinical decisions. Simultaneously, AI demonstrates significant value in drug development and precision medicine. Traditional drug development is often characterized by long cycles and high costs, while AI-based molecular structure prediction and virtual screening technologies can significantly shorten candidate drug screening time. For example, using deep learning models to predict protein structures and molecular interactions can improve drug design efficiency. Furthermore, by systematically analyzing genomic data and constructing individualized risk prediction models, AI can extract key features from massive amounts of genomic data, providing crucial support for individualized treatment plans in precision medicine.

Despite the broad application prospects of artificial intelligence in the medical field, its practical promotion and in-depth application still face multiple challenges. From a data governance perspective, medical data often contains highly sensitive personal privacy information. Therefore, ensuring patient privacy and data security while achieving data sharing and value mining is a crucial issue that must be balanced. The tension between data openness and privacy protection makes the design of systems and the construction of technical protection mechanisms particularly critical. Furthermore, regarding model performance and algorithmic fairness, the representativeness and quality of training data directly affect the system's generalization ability and stability. If the data sources have issues such as regional concentration, unbalanced sample structure, or potential biases, the model's performance may differ across different populations or clinical scenarios, leading to uneven treatment outcomes and the risk of algorithmic bias. Therefore, building a multi-center, diverse, and high-quality medical data system is a fundamental

condition for improving the reliability of AI-powered medical systems. Overall, against the backdrop of the deepening application of AI in medicine, achieving synergistic advancement of technological innovation and institutional governance has become a key prerequisite for its sustainable development. By establishing robust data security mechanisms, ethical standards, and regulatory frameworks, and promoting the deep integration of artificial intelligence technology with medical knowledge and multi-source data, AI is evolving from an early single-task auxiliary tool into a comprehensive intelligent support system integrating multiple modules. This is driving the continuous transformation of healthcare services towards data-driven, refined, and personalized approaches. With the optimization of algorithm performance, the improvement of data governance systems, and the continuous strengthening of interdisciplinary collaboration mechanisms, AI will play a more fundamental and strategic supporting role in the modern healthcare system, providing continuous impetus for improving healthcare quality and innovating healthcare models.

4.2. Education Sector

The application of artificial intelligence technology in the field of education is profoundly changing the traditional teaching model and learning methods. With the continuous enrichment of digital educational resources and the widespread popularity of online learning platforms, educational data is gradually showing characteristics of scale and diversification. Against this background, artificial intelligence provides important technical support for personalized teaching, intelligent assessment and educational management optimization by systematically analyzing and modeling learning behavior data. Its core goal is to improve teaching efficiency and learning effect, thereby promoting the development of a learner-centered precision education model. In terms of specific applications, personalized learning systems have become one of the important directions of artificial intelligence education applications[18]. Traditional classroom teaching usually adopts a uniform pace and standardized teaching content, which is difficult to fully respond to the differentiated needs of students in terms of knowledge base, learning pace and cognitive style. Relying on machine learning methods and knowledge tracking models, intelligent teaching systems can analyze students' learning behavior data and answer performance in real time, assess their knowledge mastery, and dynamically adjust the learning content structure and task difficulty accordingly. This adaptive learning mechanism based on data modeling helps to improve the matching degree between teaching resources and learning needs, thereby enhancing learning motivation and improving overall learning efficiency.

Building upon this foundation, the application value of Intelligent Tutoring Systems (ITS) in educational practice is becoming increasingly prominent. These systems achieve continuous monitoring and dynamic feedback of the learning process by constructing student models, domain knowledge models, and teaching strategy models. By leveraging artificial intelligence algorithms to analyze student error types, learning paths, and knowledge gaps, the system can generate targeted explanations and prompts, thus simulating personalized tutoring to some extent. Related research indicates that intelligent tutoring systems demonstrate good teaching effects in mathematics, programming, and language learning, and have potential advantages in alleviating the problem of uneven distribution of educational resources. Furthermore, artificial intelligence

also shows significant application prospects in teaching evaluation and educational management. For example, automatic scoring technology, combined with natural language processing and deep learning models, performs semantic understanding and score prediction on subjective question answers, effectively improving assessment efficiency and scoring consistency. Simultaneously, by clustering and trend prediction of learning behavior data, educational management departments can identify potential learning risk groups and implement targeted interventions. This data-driven educational decision-making model helps optimize resource allocation and improve the scientific level of educational governance.

Despite the vast potential of artificial intelligence in education, its practical application still faces numerous challenges. From a data governance perspective, learning behavior data often involves issues of personal privacy and the protection of minors' information; therefore, strict adherence to relevant laws, regulations, and ethical norms is essential during data collection, storage, and use. Regarding algorithmic fairness, model bias can negatively impact learning evaluation results, especially in high-risk exams or critical assessment scenarios, making it crucial to enhance model transparency and interpretability. Furthermore, the deep integration of AI technology may affect teachers' professional roles and teaching autonomy; therefore, constructing a teaching model centered on human-machine collaboration has become a key research direction. From an overall development perspective, AI is driving the transformation of education models from traditional "uniform teaching" to a learner-centered "personalized learning" system. Leveraging data analysis and intelligent modeling technologies, education systems can more accurately identify learning needs, assess teaching effectiveness, and optimize resource allocation. In the future, with the further integration of algorithmic innovation and educational theory, and the continued deepening of interdisciplinary collaboration mechanisms, AI is expected to play a more profound and systematic role in promoting educational equity, improving teaching quality, and building a lifelong learning system.

4.3. Financial Sector

Artificial intelligence technology is increasingly widely used in the financial field and has gradually become an important technological foundation for promoting the digital transformation of the financial industry. The financial industry has the characteristics of large data scale, high transaction frequency and strong risk sensitivity, which provides rich application scenarios for the deployment and optimization of intelligent algorithms. By systematically modeling and analyzing structured and unstructured financial data, artificial intelligence has shown significant advantages in risk management, credit assessment, financial market prediction and intelligent services, thereby improving the efficiency of financial services and the scientific nature of decision-making. In the field of risk control and credit assessment, machine learning models have become an important tool for financial institutions to improve risk identification capabilities and pricing accuracy. By comprehensively analyzing historical transaction data, user behavior characteristics and external credit information, relevant algorithms can construct a credit scoring system and predict the probability of default. Compared with traditional rule-based or statistical regression methods, models such as gradient boosting trees, random forests and deep neural networks have stronger expressive power and can characterize complex nonlinear relationships between

variables, thereby significantly improving risk prediction performance[19]. In personal credit and micro and small enterprise financing scenarios, this type of technology not only expands the coverage of financial services, but also promotes the refinement and differentiation of risk pricing. However, during model training, issues such as data structure imbalance, sample bias, or inappropriate feature selection may lead to systematic shifts in the scoring results. Therefore, it is necessary to strengthen the model validation mechanism and the algorithm fairness evaluation system to improve the reliability of risk assessment results.

In the fields of financial market forecasting and quantitative investment, artificial intelligence methods also demonstrate broad application prospects. Relying on technologies such as time series analysis, deep learning, and reinforcement learning, models can extract potential structural features from historical price fluctuations and macroeconomic indicators, providing data support for asset allocation optimization and trading strategy construction. For example, models based on recurrent neural networks or Transformer structures have strong pattern-capturing capabilities in financial time series modeling, helping to improve the accuracy of short-term price fluctuation predictions; while reinforcement learning methods achieve a dynamic balance between return objectives and risk constraints by simulating market environments and strategy iteration processes. Meanwhile, in the field of robo-advisory, robo-advisory systems provide investors with personalized asset allocation solutions by building customer profiles and risk preference models. The system can automatically generate investment portfolio recommendations and make dynamic adjustments based on the user's financial situation, investment objectives, and risk tolerance, thereby reducing service costs while improving investment decision-making efficiency. Furthermore, artificial intelligence also plays a crucial role in fraud prevention and anomaly detection. By monitoring and recognizing transaction behavior in real time, machine learning models can identify potential fraudulent activities and abnormal transaction behaviors. Among them, graph neural networks and other methods have shown significant advantages in the analysis of complex transaction network structures, which helps to reveal hidden relationships and risk propagation paths, and significantly enhances the security and stability of financial system operations in electronic payment and cross-border transaction scenarios.

Despite significant progress in the financial sector, artificial intelligence still faces numerous challenges in its application and promotion. From a data governance perspective, financial data often contains highly sensitive information, thus data security and privacy protection must be a primary focus during data collection, storage, and sharing. In terms of regulatory compliance, the interpretability of model decision-making results is crucial for key business operations such as credit approval and risk pricing; therefore, relevant models need high transparency and auditability to meet regulatory compliance requirements. Furthermore, the stability and robustness of algorithmic models need further improvement under extreme market conditions or systemic shocks to prevent the accumulation of potential systemic risks. From an overall development trend perspective, AI is driving the financial industry to gradually shift from a traditional experience-driven decision-making model to a decision-making system centered on data analysis and intelligent algorithms. Through efficient processing and in-depth mining of large-scale financial data, AI technology has significantly enhanced risk management capabilities and service

efficiency, and provided a key technological foundation for the formation of new financial service models. In the future, with the continuous improvement of the regulatory framework and the sustained enhancement of algorithmic capabilities, AI will play a more fundamental and strategic role in promoting the sound operation of the financial system, optimizing resource allocation, and promoting inclusive finance.

5. Ethical and Social Impact

With the widespread application of artificial intelligence technology in multiple fields such as medicine, finance, education, industry and public governance, its social impact is becoming increasingly prominent. While significantly improving production efficiency and optimizing social governance structure, the development of artificial intelligence has also triggered a series of complex and profound ethical and social issues. These issues not only involve the security, reliability and controllability of the technology system itself, but also further relate to social fairness, privacy protection, human subject status and the stability of social value system. Figure 3 systematically shows the core ethical issues involved in the social application of artificial intelligence and their interrelationships. By constructing an analytical framework covering dimensions such as data privacy protection, algorithm fairness, responsibility attribution, social structural changes and value alignment, we can more systematically understand the inherent logical connection between the artificial intelligence risk generation mechanism and governance path. Among them, data governance issues are particularly critical[20]. Artificial intelligence systems usually rely on massive data resources during model training and performance optimization. These data often contain sensitive content such as personal identity information, behavioral trajectories and health records. In the process of data collection, storage, sharing and cross-domain circulation, if there is a lack of sound security protection and compliance mechanisms, it is easy to trigger risks such as privacy leakage, data abuse and network attacks. While technologies such as differential privacy, federated learning, and homomorphic encryption offer solutions for privacy protection to some extent, the core challenge in the governance of artificial intelligence remains how to ensure the reasonable use of data value while protecting individual privacy rights.

Meanwhile, the fairness of algorithmic decision-making has gradually become an important topic in AI ethics research. The prediction results of AI models largely depend on the distribution structure and sample characteristics of the training data. When the data itself has structural biases or insufficient sample representativeness, the model may solidify or even amplify existing inequalities during the learning process, leading to potential discriminatory consequences in scenarios such as credit approval, recruitment screening, or judicial assistance decision-making. To mitigate this problem, researchers have proposed various technical approaches, including data resampling, constraint optimization, and the construction of fairness indicators, to reduce the negative impact of algorithmic bias. However, defining fairness standards in diverse social contexts and translating them into actionable model optimization objectives still faces significant theoretical and practical challenges. Furthermore, the rapid development of AI technology is profoundly impacting labor structures and the social division of labor. The widespread application

of automated and intelligent systems in manufacturing, service, and management sectors risks the replacement of some repetitive and rule-based jobs, leading to a restructuring of employment and changes in skill requirements. While technological progress creates new job types and industrial opportunities, structural unemployment and skills mismatch may exacerbate social inequality in the short term. Therefore, guiding the workforce to complete skills transformation and career upgrading through education system reform, vocational skills training, and public policy adjustments has become one of the important issues in social governance in the era of artificial intelligence.

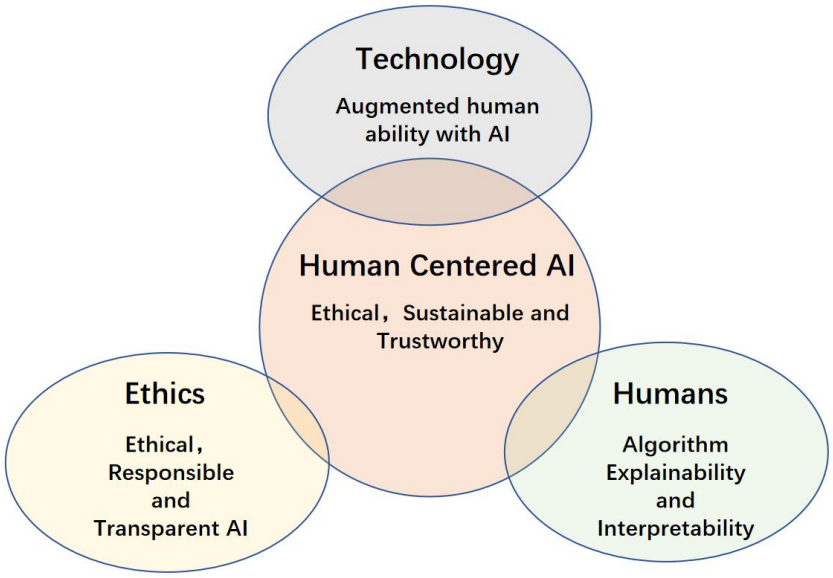


Figure 3. The core ethical issues of artificial intelligence and their interrelationships

At a deeper institutional and value level, the widespread application of artificial intelligence systems also brings challenges in terms of responsibility definition and technology governance. In particular, large-scale models based on deep learning often exhibit significant "black box" characteristics, with their internal decision-making logic difficult to explain intuitively. When system judgments erroneous or cause real-world harm, the division of responsibility among algorithm developers, system deployment organizations, and end-users still lacks unified standards, which is particularly prominent in high-risk application scenarios such as autonomous driving and medical assisted diagnosis. Therefore, strengthening research on model interpretability , improving decision-making transparency, and establishing a clear legal responsibility framework have become crucial links in promoting the standardized development of AI. Simultaneously, AI technology itself has a distinct dual nature: on the one hand, it can enhance social governance capabilities and information production efficiency; on the other hand, it can also be used for activities such as generating false information, spreading deepfakes,and automated cyberattacks, impacting the public opinion environment and social trust systems. With the development of generative models, the threshold for producing false content continues to decrease, increasing the difficulty of governance. At a more macro level, as AI systems gradually acquire autonomous decision-making and content generation capabilities, the boundaries of human-machine relationships are constantly being reshaped. How to ensure that the behavior of AI systems aligns with human social values and norms has become an important direction for

value alignment research. Current research attempts to improve the ethical consistency of model outputs by introducing human feedback mechanisms and reinforcement learning alignment strategies. However, due to the cultural differences and context-dependent nature of values, constructing a value alignment mechanism that is both universal and dynamically adaptable remains an important topic that needs to be explored in depth in the study of artificial intelligence ethics.

6. Future Development Trends

With the continuous optimization of algorithm model architecture, the exponential improvement of computing power and the continuous expansion of data resources, artificial intelligence is gradually entering a new development stage with large-scale pre-trained models, cross-modal fusion and intelligent agent systems as the core features. Unlike the early development path that mainly relied on the performance improvement of a single algorithm, the current evolution of artificial intelligence reflects a comprehensive trend of systematic integration, cross-domain collaboration and parallel development of technology and system. From the perspective of overall technology, the rise of large-scale pre-trained models marks the transformation of artificial intelligence from a task-specific paradigm to a general framework. By implementing self-supervised learning on massive data, the model can obtain cross-task transfer and multi-scenario adaptation capabilities. The "pre-training-fine-tuning" paradigm significantly reduces the dependence on manually labeled data and improves the generalization performance of the model to a certain extent. The future development of models will pay more attention to structural optimization and parameter efficiency, and improve resource utilization through model compression, knowledge distillation and efficient computing power scheduling mechanisms. At the same time, the integration of reinforcement learning methods and symbolic reasoning mechanisms is also regarded as an important direction for improving the logical reasoning and complex planning capabilities of the system. Although true general artificial intelligence is still in the exploratory stage, the development of cross-task transfer capabilities and multimodal representation learning has provided a key technological foundation for it.

At the data processing and cognitive level, multimodal fusion is gradually becoming an important path for artificial intelligence to improve its environmental understanding capabilities. Real-world information often exists in the form of interwoven text, images, voice, video, and data from multiple sensors, making it difficult for a single-modal model to achieve a comprehensive and accurate environmental representation. By constructing a unified semantic representation space and achieving cross-modal alignment, AI systems can perform comprehensive perception, information integration, and reasoning decision-making in complex situations, thereby improving the overall robustness and adaptability of the system. In highly complex application scenarios such as medical diagnosis, autonomous driving, and intelligent manufacturing, multimodal collaboration mechanisms not only significantly improve decision-making accuracy but also promote interdisciplinary integration and innovation. Meanwhile, with the widespread deployment of IoT devices and smart terminals, traditional cloud-centric centralized computing models are gradually facing latency and bandwidth bottlenecks. Against this backdrop, edge

computing and distributed intelligent systems have become important development directions. By deploying some inference tasks to network edge nodes, data transmission pressure can be effectively reduced and the system's real-time response capability improved. Furthermore, with the help of distributed training mechanisms such as federated learning, multiple devices can achieve collaborative model updates while ensuring data privacy, thereby gradually building an intelligent computing infrastructure that operates collaboratively across the cloud, edge, and terminal.

In terms of human-machine relationships and social governance, the development trend of artificial intelligence is gradually shifting from simple automation replacement to a collaborative intelligence model centered on enhancing human capabilities. This paradigm emphasizes that AI, as a cognitive support and decision-making aid, forms a complementary collaborative relationship with humans. For example, in highly specialized fields such as medicine, law, and scientific research, AI systems can undertake data analysis, pattern recognition, and knowledge discovery tasks, while value judgments and final decisions are still made by human agents. To achieve efficient human-machine collaboration, system design needs to further enhance interpretability, interactivity, and context awareness, improve user trust through transparency and feedback mechanisms, and enhance the system's understanding of human intentions and social context. Furthermore, from a sustainable development perspective, as model scale continues to expand, the energy consumption resulting from training and deployment is increasingly attracting attention. Through technical approaches such as model pruning, parameter sharing, knowledge distillation, and low-power hardware architecture design, computational resource consumption can be reduced while maintaining performance stability, promoting the development of green AI. At the institutional and governance level, future AI systems will place greater emphasis on standardization and ethically embedded design. By introducing risk assessment mechanisms and value alignment strategies during model development, ethical principles will be reflected upfront in the technical process. Simultaneously, international cooperation and the development of technical standards systems will become important trends. By establishing unified evaluation frameworks and compliance mechanisms, the safety, transparency, and social credibility of AI systems will be improved, thereby achieving a long-term and stable dynamic balance between technological innovation and social responsibility.

Author Contributions:

All authors have read and agreed to the published version of the manuscript.

Funding:

This research received no external funding.

Institutional Review Board Statement:

Not applicable.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

Not applicable.

Conflict of Interest:

The authors declare no conflict of interest.

References

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. AI and Ethics, 5(3), 3265 – 3279.

Corrêa, N. K., Galvão, C., Santos, J. W., et al. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns, 4(10), 100857.

Kashefi, P., Kashefi, Y., & Ghafouri Mirsaraei, A. H. (2024). Shaping the future of AI: Balancing innovation and ethics in global regulation. Uniform Law Review, 29(3), 524 – 548.

Kattnig, M., Angerschmid, A., Reichel, T., et al. (2024). Assessing trustworthy AI: Technical and legal perspectives of fairness in AI. Computer Law & Security Review, 55, 106053.

Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. Scientific Reports, 14(1), 20752.

Lainjo, B. (2024). The role of artificial intelligence in achieving the United Nations sustainable development goals. Journal of Sustainable Development, 17(5), 30.

Li, Y., Wu, B., Huang, Y., et al. (2024). Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust. Frontiers in Psychology, 15, 1382693.

McCormack, L., & Bendechache, M. (2024). Ethical AI governance: Methods for evaluating trustworthy AI. arXiv preprint arXiv:2409.07473.

Ricciardi Celsi, L., & Zomaya, A. Y. (2025). Perspectives on managing AI ethics in the digital age. Information, 16(4), 318.

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th US ed.). Pearson.

Sharma, S. (2023). Trustworthy artificial intelligence: Design of AI governance framework. Strategic Analysis, 47(5), 443 – 464.

Sistla, S. (2024). AI with integrity: The necessity of responsible AI governance. Journal of Artificial Intelligence & Cloud Computing, 3(5), 2 – 3.

Taeihagh, A. (2025). Governance of generative AI. Policy and Society, 44(1), 1 – 22.

Vercelli, A. (2024). United Nations, artificial intelligences and regulations: Analysis of the General Assembly AI resolutions and the final report of the advisory body on AI (short paper). In Proceedings of BEWARE@AI (pp. 99 – 106).

Whig, D. P. (2025). Ethical AI governance: A framework for ensuring transparency, fairness, and accountability. Journal of Healthcare AI and ML, 12, 12.

Xu, J., Lee, T., & Goggin, G. (2024). AI governance in Asia: Policies, praxis and approaches. Communication Research and Practice, 10(3), 275–287.

Journal of Computer Science and Digital Technology, 2026, 2(1), 71-79
https://doi.org/10.71204/fef1s997

Design and Implementation of a Non-Contact Magnetic Sensing and Signal Processing System for High-Precision Robot Joint Encoders

Chaohui Zhang [1, *], Lujie Ren [1], Jianming Mao [1], Haijun Lei [1], Dengcheng Lu [1], Shishui Zhou [1]

[1] Hopo Technology (Ningbo) Co., Ltd., Ningbo 315000, China

*** Correspondence:**

Chaohui Zhang

zhangchaohui7828@163.com

Received: 21 October 2025 / Accepted: 9 May 2026 / Published online: 12 May 2026

Abstract

Robot joint encoders are critical sensing components in industrial robots, where positioning accuracy, dynamic response, and environmental adaptability directly affect motion-control performance and operational reliability. However, conventional encoder technologies still face challenges related to signal instability under harsh industrial conditions, insufficient high-speed response capability, and limited long-term reliability. To address these issues, this paper presents the design and implementation of a non-contact magnetic sensing and signal-processing system for high-precision robot joint encoders. The proposed system integrates incremental magnetic grid design, tunnel magnetoresistance (TMR) sensing technology, adaptive signal conditioning, multi-dimensional error compensation, and high-speed signal-processing algorithms into a unified encoder framework. A dual-reading-head magnetic grid structure and dynamic gain compensation mechanism are introduced to improve signal stability under vibration, temperature variation, oil contamination, and electromagnetic interference conditions. In addition, adaptive filtering and real-time interpolation algorithms are employed to enhance dynamic response capability and suppress nonlinear measurement errors during high-speed robot motion. The developed encoder system supports multiple industrial communication interfaces, including RS422, SSI, and CANopen, enabling flexible integration with industrial robot control systems. Experimental results demonstrate that the proposed encoder achieves repeatability accuracy within ±1 μm and a frequency response of 8000 kHz while maintaining stable operation under harsh industrial environments. Engineering deployment in industrial robots and precision manufacturing equipment further verifies the robustness, reliability, and practical applicability of the proposed system. The research provides an effective technical solution for high-precision robotic motion sensing and intelligent industrial measurement systems.

Keywords: Non-Contact Magnetic Sensing; Robot Joint Encoder; Signal Processing; Adaptive Error Compensation; TMR Sensor; Robotic Motion Control

1. Introduction

Against the background of the upgrading of the intelligent manufacturing industry, industrial robots have been widely used in automotive manufacturing, precision machining, high-end equipment and other fields, and the requirements for joint positioning accuracy, dynamic response speed and adaptability to harsh working conditions are constantly rising (Wang et al., 2024). As the "nerve ending" of robot motion control, the joint encoder is responsible for converting the angular displacement and angular velocity signals of the joint into electrical signals to provide accurate position feedback for the control system. Its performance directly affects the motion accuracy and operational reliability of robots (Elecfans, 2025).

At present, high-end robot joint encoders in the market mostly rely on imports, facing problems such as technological monopoly, high cost and unstable supply cycles, which seriously restrict the independent development of China's industrial robot industry (Intelligent Manufacturing, 2025). Although traditional optical encoders can achieve high accuracy, they have defects such as weak resistance to pollution, vibration and temperature change, making it difficult to adapt to the harsh working conditions on industrial sites; contact encoders suffer from serious wear, short service life and high maintenance costs, failing to meet the demand for long-term stable operation of robots (Elecfans, 2025).

With the advantages of strong anti-interference ability, resistance to harsh environments, simple structure and long service life, non-contact magnetic measurement technology has become the core technical direction for the research and development of robot joint encoders (Sohu, 2026). With more than 10 years of R&D experience in the field of industrial precision control, the author focuses on the development of incremental magnetic grids and robot joint encoders, and carries out technical research centering on magnetic grid array design, sensor signal processing and engineering adaptation of joint encoders. A series of technological breakthroughs and engineering achievements have been made targeting the industry's existing technical difficulties such as signal stability under harsh working conditions and dynamic response in high-speed motion. Combined with R&D practices, this paper elaborates on the application research of non-contact magnetic measurement technology in robot joint encoders, providing references for the R&D and engineering application of similar products and boosting the domestic substitution process of core components of industrial robots in China.

2. Core Technical Requirements and Existing Difficulties of Robot Joint Encoders

2.1. Core Technical Requirements

The core function of robot joint encoders is to achieve accurate positioning and high-speed dynamic response of joint motion. Combined with the application scenarios of industrial robots, their core technical requirements mainly include three aspects: first, accuracy requirements, the joint positioning accuracy of industrial robots needs to reach the micron level, and the repeatability accuracy should be controlled within ±1 μm to meet the operation requirements of precision machining, assembly and other scenarios; second, dynamic response requirements, it needs to adapt to high-speed motion scenarios of robot joints, with a frequency response of more

than 8000 kHz to ensure no delay and no distortion in signal transmission; third, environmental adaptability requirements, it needs to withstand harsh working conditions such as oil contamination, dust, vibration and extreme temperature changes on industrial sites to ensure the stability of signal transmission and product service life (Taobao Digital Network, 2026).

In addition, from the perspective of industrial development, encoder products also need to have the characteristics of strong engineering adaptability, controllable cost and mass production to achieve large-scale application, promote domestic substitution, and help customers reduce costs, improve efficiency and achieve independent and controllable technology (Intelligent Manufacturing, 2025).

2.2. Existing Technical Difficulties

Combined with industry practices and R&D experience, the core technical difficulties in the current R&D of robot joint encoders mainly focus on three aspects:

First, poor signal stability under harsh working conditions. Oil contamination and dust on industrial sites affect signal transmission; vibration causes relative position offset between the magnetic grid and the sensor; extreme temperature changes lead to changes in the magnetic field strength of the magnetic grid and increased sensor drift, resulting in signal distortion, increased measurement errors, and even signal loss (Sohu Mobile, 2026).

Second, insufficient dynamic response under high-speed motion. When robot joints move at high speed, the relative motion speed between the magnetic grid and the sensor is fast, leading to drastic changes in magnetic field signals. Traditional signal processing algorithms are difficult to capture and analyze signals quickly, resulting in signal delay and distortion, affecting the dynamic response performance of encoders (Wang et al., 2024).

Third, insufficient engineering adaptability. The structures of robot joints of different models and scenarios vary greatly. The installation size, interface type and signal output mode of encoders need to be accurately adapted to the joint structure, while taking into account the consistency of mass production and cost control. The traditional R&D model is difficult to achieve accurate adaptation and large-scale production (Original Force Document, 2025).

In addition, the dependence on imports of core devices such as high-end TMR sensing chips and signal processing ASICs also restricts the performance improvement and domestic substitution process of encoder products (Sohu Mobile, 2026).

3. Encoder R&D Scheme Based on Non-Contact Magnetic Measurement Technology

Taking non-contact magnetic measurement technology as the core, this paper proposes a targeted R&D scheme around three key links: magnetic grid array design, sensor signal processing and engineering adaptation of joint encoders, breaking through the above technical difficulties and achieving dual improvement in encoder performance and engineering application.

3.1. Magnetic Grid Array Design

The magnetic grid array is the basis for encoders to achieve accurate measurement, and its design quality directly affects measurement accuracy and signal stability. Combined with the accuracy requirements and harsh working condition adaptation requirements of robot joint encoders, this paper adopts nano-scale magnetic material etching technology to design high-precision incremental magnetic grid arrays. The specific scheme is as follows:

First, selection and preparation of magnetic grid materials. Amorphous alloy strips with high magnetic permeability and high stability are selected as the magnetic grid substrate, which have good resistance to temperature change and vibration, and can adapt to the extreme temperature change environment of -40 ℃ to 120 ℃ (Elecfans, 2025). Laser etching technology is used to form an equally spaced N/S pole alternating magnetic field array on the substrate, with the grid pitch controlled at 100 μm. Precise calibration ensures that the periodic consistency error of the grid pitch is ≤ ±2 μm, laying a foundation for accurate measurement (Taobao Digital Network, 2026).

Second, optimization of magnetic grid array structure. Aiming at the relative position offset between the magnetic grid and the sensor caused by vibration, a magnetic grid array structure with symmetrical layout of double reading heads is designed. The distance between the two reading heads is 1/4 of the nominal grid pitch, forming a four-phase orthogonal sampling structure, which can effectively offset the measurement error caused by relative position offset and avoid signal loss (Original Force Document, 2025). Meanwhile, an integrated injection molding packaging process is adopted to package the magnetic grid array in a 6061-T6 aluminum alloy shell, with a silicone shock-absorbing layer filled inside, improving the vibration and impact resistance of the magnetic grid, with an impact resistance strength of up to 1000 g (Taobao Digital Network, 2026).

Third, magnetic field strength calibration. High-precision magnetic field calibration equipment is used to accurately calibrate the magnetic field strength of the magnetic grid array, ensuring that the uniformity error of magnetic field strength is ≤ ±3%, avoiding signal distortion caused by uneven magnetic field strength and improving measurement accuracy (Sohu Mobile, 2026).

3.2. Sensor Signal Processing Technology

Sensor signal processing is the core to improve encoder accuracy and dynamic response performance. Aiming at problems such as signal distortion under harsh working conditions and insufficient dynamic response under high-speed motion, this paper designs a high-precision and high-response signal processing system, including three modules: signal conditioning, error compensation and high-speed analysis.

First, signal conditioning module. A differential magnetoresistive sensor (TMR) is used as the signal acquisition element, which has the characteristics of high sensitivity, strong anti-interference ability and low power consumption, with an operating current of only 95 mA, and can effectively capture weak magnetic field changes of the magnetic grid array (Taobao Digital Network, 2026). A dynamic gain compensation circuit is designed at the front end of the sensor, with an operating frequency covering 10 Hz to 8000 kHz, which can automatically suppress low-

frequency drift and high-frequency common-mode noise, ensuring that the signal amplitude fluctuation is controlled within ±3% and improving signal stability under harsh working conditions (Taobao Digital Network, 2026). Meanwhile, a four-layer PCB stack design is adopted, the distance between the signal layer and the ground layer is controlled at 80 μm, and key analog traces are fully wrapped with a shielded ground network, with a common-mode rejection ratio (CMRR) of up to 95 dB, effectively resisting strong electromagnetic interference on industrial sites (Taobao Digital Network, 2026).

Second, error compensation module. A multi-dimensional error compensation model is established for measurement errors caused by temperature drift, magnetic grid periodic error and installation error. The ambient temperature is collected in real time through a temperature sensor, and the temperature drift error is dynamically compensated by a linear interpolation algorithm; combined with the non-linear compensation table pre-stored at the factory (16 correction coefficients stored per millimeter), the magnetic grid periodic error and installation error are statically compensated, suppressing the overall non-linear error of the system to within ±0.5 μm (Taobao Digital Network, 2026). Meanwhile, an adaptive SG (ASG) filtering algorithm is introduced to dynamically adjust the filtering length according to the working conditions of variable speed and positive/negative rotation of robot joints, effectively suppressing instantaneous angular velocity signal errors and improving the adaptability of error compensation (Wang et al., 2024).

Third, high-speed signal analysis module. A high-performance signal processing chip is used to optimize the signal analysis algorithm, controlling the signal processing delay within 50 μs, achieving a high frequency response of 8000 kHz, which can accurately capture the magnetic field signal changes during high-speed motion of robot joints and ensure no delay and no distortion in signal analysis (Sohu Mobile, 2026). Meanwhile, RS422 differential signal output is adopted to realize 20-meter long-distance signal transmission, adapting to the complex installation scenarios of industrial robots (Elecfans, 2025).

3.3. Engineering Adaptation Design of Joint Encoders

To achieve large-scale application and multi-scenario adaptation of encoders, engineering adaptation design is carried out from three aspects: installation structure, interface type and mass production.

First, installation structure adaptation. Aiming at the structural differences of robot joints of different models, a modular installation structure is designed, adopting a hollow shaft magnetic ring design with coaxiality error < 0.1 mm. The installation size can be flexibly adjusted according to the joint size without major modification of the joint structure, reducing adaptation costs (Sohu Mobile, 2026). Meanwhile, the installation and positioning structure is optimized, adopting a combined positioning method of positioning pins and locking bolts to ensure installation accuracy and reduce the impact of installation errors on measurement accuracy (Original Force Document, 2025).

Second, interface and signal output adaptation. Multiple types of interfaces (such as RS422, SSI, CANopen) are designed, and the signal output mode can be flexibly switched according to

customer needs to adapt to different robot control systems (Elecfans, 2025). Meanwhile, an expansion interface is reserved to facilitate subsequent function upgrade and debugging, improving product compatibility and scalability.

Third, mass production adaptation. The production process is optimized, adopting automated assembly and calibration equipment to realize automated production of magnetic grid array preparation, sensor packaging, signal calibration and other links, improving the consistency of mass production and reducing production costs (Taobao Digital Network, 2026). A complete quality inspection system is established to comprehensively test the accuracy, frequency response and environmental adaptability of each encoder, ensuring that the product pass rate reaches more than 99.5% to meet the demand for large-scale application.

4. Experimental Verification and Engineering Application

4.1. Performance Experimental Verification

To verify the performance of the developed encoder, a special experimental platform is built, and experimental verification is carried out from three aspects: accuracy, frequency response and environmental adaptability.

First, accuracy experiment. A high-precision laser interferometer is used as the standard measuring equipment to test the repeatability accuracy of the encoder. The experimental results show that the repeatability accuracy of the encoder is stable within ±1 μm, and the non-linear error is < ±0.5 μm, which is better than the ±2~3 μm level of similar products in the industry, meeting the precision positioning requirements of robot joints (Taobao Digital Network, 2026).

Second, frequency response experiment. A high-speed motion simulation platform is used to simulate the high-speed motion scenario of robot joints and test the dynamic response performance of the encoder. The experimental results show that the frequency response of the encoder can reach 8000 kHz, and the signal-to-noise ratio remains > 45 dB at a linear speed of 20 m/s, with no delay and no distortion in signal transmission, which can adapt to the high-speed motion requirements of robot joints (Taobao Digital Network, 2026).

Third, environmental adaptability experiment. The encoder is placed in simulated harsh working conditions such as oil contamination, dust, vibration and extreme temperature changes for 2000 hours of continuous operation to test signal stability. The experimental results show that in an environment with oil mist concentration of 50 mg/m³, the signal amplitude fluctuation of the encoder is ≤ ±3%; in the environment of 1000 g impact and -40 ℃~120 ℃ temperature change, it can still output signals stably without signal loss or distortion, and its environmental adaptability is more than 3 times that of traditional optical encoders (Elecfans, 2025).

4.2. Engineering Application Cases

Based on the above R&D achievements, the developed incremental magnetic grid robot joint encoder has been implemented in multi-scenario engineering applications, applied in automotive

manufacturing, precision machining, high-end equipment and other fields. Some typical application cases are as follows:

Case 1: Adaptation of automotive parts manufacturing robots. The welding robot joints of an automotive parts factory originally used imported optical encoders, which had problems of weak oil resistance and high maintenance costs. After replacing with the encoder developed in this paper, the equipment downtime was reduced by 65%, the annual maintenance cost was reduced by 280,000 yuan, and the positioning accuracy was improved to ±1 μm, meeting the requirements of precision welding (Elecfans, 2025).

Case 2: Adaptation of precision machining robots. The robot joints of a precision machine tool need to achieve high-speed and high-precision positioning. With 8000 kHz frequency response and ±1 μm repeatability accuracy, the encoder developed in this paper effectively improves the motion control accuracy of the robot, reducing part machining errors by 30% and improving production efficiency by 25% (Taobao Digital Network, 2026).

Case 3: Adaptation of robots operating in harsh environments. The joints of a mining machinery robot work in an environment with much dust, large vibration and severe temperature changes, and traditional encoders cannot work stably. Through the optimized magnetic grid structure and signal processing technology, the encoder developed in this paper achieves long-term stable operation with a service life of more than 100,000 hours, helping customers achieve stable equipment operation, cost reduction and efficiency improvement (Elecfans, 2025).

Up to now, the developed encoder products have been applied in more than 10 scenarios, replacing more than 300 sets of imported products, helping customers reduce procurement costs by more than 40%, effectively promoting the domestic substitution of robot joint encoders and boosting the independent and controllable development of technology in the industrial robot field (Intelligent Manufacturing, 2025).

5. Conclusion and Prospect

5.1. Conclusion

Combined with the author's more than 10 years of R&D experience in the field of industrial precision control, taking non-contact magnetic measurement technology as the core and focusing on the development of incremental magnetic grids and robot joint encoders, this paper proposes a complete R&D scheme aiming at technical difficulties such as signal stability under harsh working conditions, insufficient dynamic response in high-speed motion and poor engineering adaptability. Through the optimized design of magnetic grid array, high-precision signal processing technology and modular engineering adaptation design, the encoder performance has been significantly improved. The main conclusions are as follows:

The incremental magnetic grid array designed with nano-scale laser etching technology and symmetrical layout of double reading heads effectively improves the accuracy, vibration resistance and temperature change resistance of the magnetic grid, laying a foundation for accurate measurement of encoders;

The signal processing system designed based on differential magnetoresistive sensors and multi-dimensional error compensation algorithms achieves performance breakthroughs of ±1 μm repeatability accuracy and 8000 kHz frequency response, solving the problems of signal distortion under harsh working conditions and insufficient dynamic response in high-speed motion;

The modular engineering adaptation design realizes the accurate adaptation of encoders to different types of robot joints. Through automated production processes, production costs are reduced, the consistency of mass production is improved, and the multi-scenario landing application of products is promoted;

Experimental verification and engineering application show that the developed encoder has excellent performance, strong environmental adaptability and controllable cost, which can effectively support the high-reliability positioning of robot joints, promote the domestic substitution of core components, and help customers reduce costs, improve efficiency and achieve independent and controllable technology.

5.2. Prospect

With the development of industrial robots toward lightweight, high-precision and intelligentization, the performance requirements for joint encoders will be further improved. In the future, further research will be carried out around the following directions: first, continuously optimize the magnetic grid array design and signal processing algorithms to further improve the accuracy and dynamic response performance of encoders, striving to achieve ±0.5 μm repeatability accuracy and 10000 kHz frequency response; second, increase the R&D efforts of domestic core devices, break the dependence on imports of core devices such as TMR sensing chips and signal processing ASICs, further reduce production costs and enhance the core competitiveness of products (Sohu Mobile, 2026); third, expand the application scenarios of encoders, develop special encoder products for special scenarios such as collaborative robots and medical robots, and promote the application of non-contact magnetic measurement technology in more precision control fields; fourth, combine industrial internet technology to realize condition monitoring and remote debugging of encoders, improve the intelligent level of products, and support the intelligent upgrading of industrial robots.

Author Contributions:

All authors have read and agreed to the published version of the manuscript.

Funding:

This research received no external funding.

Institutional Review Board Statement:

Not applicable.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

Not applicable.

Conflict of Interest:

The authors declare no conflict of interest.

References

Elecfans. (2025, September 3). Magnetic incremental encoder: The “invisible champion” in the field of industrial precision control.

Intelligent Manufacturing. (2025). Research and application of domestic industrial internet equipment control systems. Intelligent Manufacturing, (8), 45–50.

Original Force Document. (2025, December 3). Rotary encoder based on double reading head magnetic grid ruler.

Sohu Mobile. (2026, February 27). Working mechanism of magnetic encoder: Magnetic steel, sensing unit, and signal-processing link.

Taobao Digital Network. (2026, January 9). ROQ 42 encoder technical disassembly: Why high-resolution magnetic grid feedback is worth 7999 yuan.

Wang, H., Guo, Y., & Yin, X. (2024). Error compensation method for industrial robot incremental grating encoder. Journal of Vibration and Shock, 43(22), 225–231.