Journal of Infrastructure, Policy and Development 2024, 8(12), 7153.
https://doi.org/10.24294/jipd.v8i12.7153

Article

Policy-driven innovations in fraud prevention: Developing an ARFLGB-XGBoost early warning model to mitigate online romance scams in telecommunication networks

Guancheng Chen[1], Wenzhuo Du[2], Yichen Shan[2], Anqi Wang[1], Liping Wang[1,*]

[1] School of Investigation, People's Public Security University of China, Beijing 100038, China
[2] School of Information and Network Security, People's Public Security University of China, Beijing 100038, China
*** Corresponding author:** Liping Wang, wangliping17@nudt.edu.cn

CITATION

Chen G, Du W, Shan Y, et al. (2024). Policy-driven innovations in fraud prevention: Developing an ARFLGB-XGBoost early warning model to mitigate online romance scams in telecommunication networks. Journal of Infrastructure, Policy and Development. 8(12): 7153. https://doi.org/10.24294/jipd.v8i12.7153

ARTICLE INFO

Received: 14 June 2024
Accepted: 25 July 2024
Available online: 30 October 2024

Abstract: To address the escalating online romance scams within telecom fraud, we developed an Adaptive Random Forest Light Gradient Boosting (ARFLGB)-XGBoost early warning system. Our method involves compiling detailed Online Romance Scams (ORS) incident data into a 24-variable dataset, categorized to analyze feature importance with Random Forest and LightGBM models. An innovative adaptive algorithm, the Adaptive Random Forest Light Gradient Boosting, optimizes these features for integration with XGBoost, enhancing early Online romance scams threat detection. Our model showed significant performance improvements over traditional models, with accuracy gains of 3.9%, a 12.5% increase in precision, recall improvement by 5%, an F1 score increase by 5.6%, and a 5.2% increase in Area Under the Curve (AUC). This research highlights the essential role of advanced fraud detection in preserving communication network integrity, contributing to a stable economy and public safety, with implications for policymakers and industry in advancing secure communication infrastructure.

Keywords: telecom network fraud; online romance scams; early warning model; XGBoost; innovative adaptive algorithm

1. Introduction

With the rapid advancement of global digital communication technology, various forms of online social networking have become an indispensable component of modern life, and it offers people a wider social realm and convenient channels of communication. However, this anonymous social environment also provides an avenue for criminal activities. Telecommunication fraudsters exploit the anonymity of online social networks, focusing their attention on these platforms and dating websites. They manipulate victims into engaging in fraudulent activities such as investments and money transfers by assuming false identities or posing as others. This type of scam, known as the Online Romance Scams (ORS), has caused substantial economic losses globally (Zhu et al., 2023). ORS scammers typically employ social media or dating applications to identify and gain the trust of their targets through the establishment of false intimate relationships. Once they obtain the victim's personal information, they can perpetrate more complex fraudulent activities. This form of fraud not only inflicts financial losses on the victims but also has a profound impact on their psychological well-being (Lazarus et al., 2022).

Incessant exploration is being conducted by researchers on leveraging machine learning techniques to enhance the accuracy of fraud pattern recognition. Lazarus et

al. (2023) comprehensively examined empirical literature and highlighted certain gaps and biases in preceding research arenas. Despite the preliminary empirical studies conducted by predecessors in this field, issues such as the absence of a scientific evaluation system and investigatory & early warning methods (Coluccia et al., 2020) persist. Particularly in the domain of ORS combat, the application of machine learning techniques has been overlooked, accompanied by a scarcity of case data support (Srivastava et al., 2019). Consequently, employing machine learning algorithms in fraud pattern recognition remains a viable research avenue.

XGBoost (Srivastava et al., 2019) is an enhanced machine learning algorithm derived from gradient boosting trees, and has garnered widespread application in the development of Telecom Fraud Crime and risk forecasting models. Numerous researchers have accomplished notable achievements in this domain. Tan et al. (2023) employed machine learning algorithms to financial fraud warnings by analyzing financial transaction data, enabling automatic identification and prediction of fraudulent behavior. Nanath and Olney (2023) leveraged machine learning algorithms for online recruitment fraud warnings by examining job postings and applicant feedback, automatically detecting and predicting fraudulent actions. Bahaghighat et al. (2023) utilized machine learning algorithms for high-precision phishing website detection by analyzing website content and structure, automatically identifying and anticipating phishing websites. Liu et al. (2022) employed machine learning algorithms for financial risk warnings by analyzing financial data and corporate governance structures, facilitating automatic identification and prediction of financial risks. Kamboj et al. (2023) employed machine learning techniques to identify and predict illegal bank accounts by analyzing account information and transaction patterns, respectively. Similarly, Liu et al. (2023) utilized machine learning algorithms to provide loan default warnings by examining lending data and borrower credit profiles. Zhao et al. (2023) employed similar methods to detect and forecast corporate crises by examining operational data and market conditions. Ashraf et al. (2023) developed high-risk road segment warnings by analyzing traffic data and road conditions, thereby identifying and predicting high-risk segments. Jiang et al. (2022) utilized blockchain 2.0 smart contract classification techniques to analyze smart contract content and operation status, generating automatic alerts and predictions of contract authenticity and security. Razavi et al. (2019) employed similar methods for power theft detection warnings by analyzing power data and user behavior, enabling automatic identification and prediction of power theft activities. Lastly, Murugan et al. (2023) utilized financial market data and company financial conditions to automatically identify and predict financial risks.

Previous research has offered valuable insights and practical guidance for employing the XGBoost algorithm in fraud and risk monitoring pattern recognition. Although preliminary empirical studies have been conducted by predecessors in this field, certain aspects such as the absence of an evaluation system from ORS Science, limited data volume, and a lack of detection and warning methods remain unaddressed. Furthermore, the robustness and generalization performance of XGBoost can be further enhanced in these studies. As such, future research should delve deeper into

the application of the XGBoost algorithm (Kolev, 2023) in the fraud pattern recognition field while simultaneously improving its algorithmic performance.

To effectively ORS cases using case data, this study develops a warning model based on the Adaptive Random Fusion Light Gradient Boost Machine (ARFLGB)-XGBoost model. The feature recognition analysis of ORS cases is conducted to identify appropriate characteristics for constructing the warning model. Secondly, drawing on the process of fraud detection and the XGBoost model-based warning model construction, we develop a hybrid machine learning framework for predicting ORS cases. Subsequently, we compare our framework with other fusion machine learning models to establish its superior predictive performance for ORS compared to other fraud detection fusion machine learning models. Lastly, based on the number of nodes in the average importance curve of feature importance, we determine important inflection points and changes in magnitude as primary and secondary features of ORS, providing target guidance and a research paradigm for future studies. Consequently, this research addresses the following gaps in the literature:

(1) Based on a comprehensive dataset of 1400 cases, compiled by the Investigation Academy of China People's Public Security University and the Shinan Branch of Qingdao Municipal Public Security Bureau, this study employs a combination of text analysis and feature extraction techniques to conduct empirical analysis utilizing the ORS (Pattern-Based Summarization) method. Each case consists of 24 independent variables and one dependent variable.

(2) The primary focus of this research is to optimize the feature importance of the XGBoost model by leveraging decision tree and gradient boosting machine learning algorithms. This is achieved through the integration of random forest and LightGBM algorithms, further enhancing feature importance. Additionally, an XGBoost fusion model based on an adaptive algorithm optimization approach is proposed.

(3) By deriving instructive conclusions from the feature importance output of the ARFLGB-XGBoost model, this study offers valuable insights for subsequent text extraction, sentiment analysis, and the integration of preventive measures with crackdown efforts in related fields. Furthermore, it provides guidance for fraud detection work in a professional and academic manner, adhering to the standards of the prestigious Nature Journal.

The novelty of this study lies in the proposal of an ARFLGB-XGBoost model for early warning of ORS, which has not been reported in the existing literature. The remainder of the paper is structured as follows: Section 2 provides an overview of the concept of regression-based ORS and its associated algorithms. Section 3 details the innovative framework and application method of the proposed model. Section 4 presents experimental settings based on case data and analyzes the performance of both the experiments and the model. Section 5 discusses the significance and importance of case data features. In Section 6, we summarize the research contributions and limitations presented in this paper.

2. Theoretical basis

This section provides a comprehensive literature review on the characteristics of ORS cases and explores the application of fusion models, namely XGBoost, Random

Forest, and LightGBM, as reported in relevant studies. Our research is primarily centered on employing fusion models for the early detection of ORS, identifying suitable algorithms for feature engineering, and optimizing algorithm structures based on the characteristics of the case data. Ultimately, we propose an adaptive fusion algorithm for feature importance.

2.1. Hazards of ORS

The pernicious impacts of ORS are exceedingly severe, primarily inflicting emotional distress, substantial financial losses, and a prevalence of criminal activities. Victims often exhibit a state of extreme passivity, accompanied by escalating levels of fraud (Whitty and Buchanan, 2016). Furthermore, some cases remain unreported or lack data, posing challenges in determining the precise extent of crimes involved. In recent years, scholars have employed diverse theoretical frameworks to examine ORS. Cross and Holt (2023) have proposed impression management theory as a cultural lens for analyzing public perception of ORS. Alternatively, Srivastava et al. (2019) have enhanced the precision and efficiency of sentiment classification by integrating naive Bayes and random forest machine learning algorithms with Twitter user data. This article presents a systematic review of literature spanning the past two decades to summarize the current state of research on ORS; however, apart from Lazarus et al. (2023) and Coluccia et al. (2020) studies on this topic have been identified to date. Consequently, it is evident that our understanding of ORS remains insufficient in terms of depth and key insights.

2.2. Improvement and application of XGBoost

The current research landscape is witnessing numerous scholarly endeavors aimed at optimizing XGBOOST, which have yielded a plethora of remarkable outcomes. Zhang et al. (2023) advanced the DS-XGBoost model, which is founded on the D-S evidence theory and XGBoost algorithm, for financial risk early warning. Yan et al. (2022) introduced OVR-XGBoost and OVO-XGBoost models, designed for multi-class prediction of theft crimes. Domashova et al. (2022) implemented machine learning models with fraud detection to identify abnormal bank transactions. Mohiuddin et al. (2023) proposed a weighted XGBoost model for network intrusion detection systems. Koc et al. (2021) employed the GA-XGBoost framework to predict disability status following construction worker accidents. Kolev (2023) introduced the XGB-COF model to address research challenges pertaining to enhancing material wear resistance and reducing friction.

2.3. Feature engineering for random forest and lightGBM identification

In the process of feature engineering, both Random Forest (RF) and LightGBM are proficient in identifying and exploiting patterns within the data. RF predicts by constructing and combining multiple decision trees, whereas LightGBM is founded on the gradient boosting decision tree algorithm. Both methods are capable of assessing the importance of features and determining which ones have the most significant impact on the dependent variable. Furthermore, they possess the ability to handle large-scale datasets and prevent overfitting. When implementing feature

engineering, it is crucial to fully exploit the merits of these two methods to enhance the predictive performance of models. Kamboj et al. (2023) have employed the XGBoost model combined with feature engineering processed by random forest to detect whether downloaded files contain malicious software. Wang and Thing (2023) concentrate on predicting default rates for P2P network loans and combine LightGBM with the XGBoost algorithm. Lao et al. (2023) have proposed an intelligent fault diagnosis solution for track switch machines in railway transportation, based on an improved version of LightGBM feature engineering (Lao et al., 2023).

2.4. Adaptive algorithm optimizing XGBoost model

Abbasimehr et al. (2023) proposed an improved XGBoost two-stage forecasting framework for energy demand forecasting. Cao et al. (2023) proposed an AM-Boost integrated learning model for fraud detection of financial transactions. Sha et al. (2022) proposed a new acoustic signal cavitation detection framework based on XGBoost and adaptive selection feature engineering to address valve cavitation. Afriyie et al. (2023) found that XGBoost model optimized based on random forest algorithm performed best in predicting and detecting credit card fraud. Mokbal et al. (2021) adopted XGBoost and advanced parameter optimization technology to propose a new cross-site scripting attack detection framework named XGBXSS.

The need for further exploration in the academic community regarding ORS research is evident, particularly in text analysis and feature extraction where case data is scarce. Machine learning technology and the construction of warning models, often overlooked when applying machine learning warning model construction techniques to other types of fraud detection, should be given due consideration. In optimizing the XGBoost model algorithm, scholars predominantly focus on data cleaning, while the optimization of output feature importance is overlooked. Among existing optimization algorithms, most scholars utilize only a single algorithm for parameter tuning, neglecting the application of multi-model techniques to optimize the XGBoost model. Moreover, the absence of more universally adaptive technologies in research on multi-model fusion is notable.

3. Algorithm design

The ARFLGB-XGBoost framework, as illustrated in **Figure 1**, is introduced in this section. The case data for ORS comprises 24 independent variables and 1 dependent variable. These 24 independent variables are categorized into four groups and fed into the Random Forest and LightGBM models. Feature engineering is employed to obtain feature importance and classification metrics individually for each model. Subsequently, an adaptive algorithm consolidates the classification metrics from Random Forest and LightGBM to generate weighted feature importance, which is then integrated into the XGBoost model. Finally, the performance of the ARFLGB-XGBoost is assessed using a validation set.

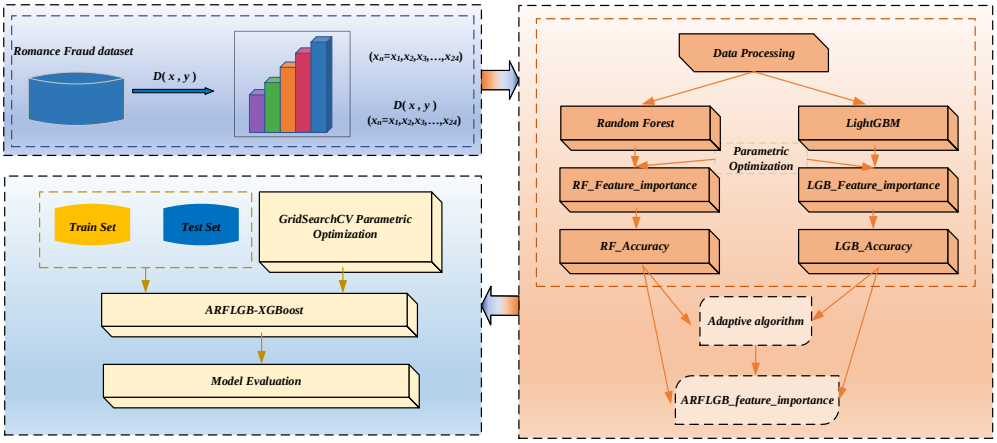

Figure 1. ARFLGB-XGBoost design framework.

3.1. Random forest method

The Random Forest (RF) algorithm was proposed in 2001 (Nti and Somanathan, 2024), building upon the random decision forest method developed at Bell Labs. By constructing multiple weak learners known as Classification and Regression Trees (CART), RF forms robust learners to tackle classification or regression prediction tasks effectively (Musbah et al., 2022). This algorithm demonstrates exceptional accuracy in handling binary classification problems, making it well-suited for big data and high-dimensional feature datasets while providing valuable insights into the significance of each evaluation factor. Moreover, RF offers the advantage of assessing feature relevance during category determination and preventing overfitting during prediction. The RF diagram is depicted in **Figure 2**.

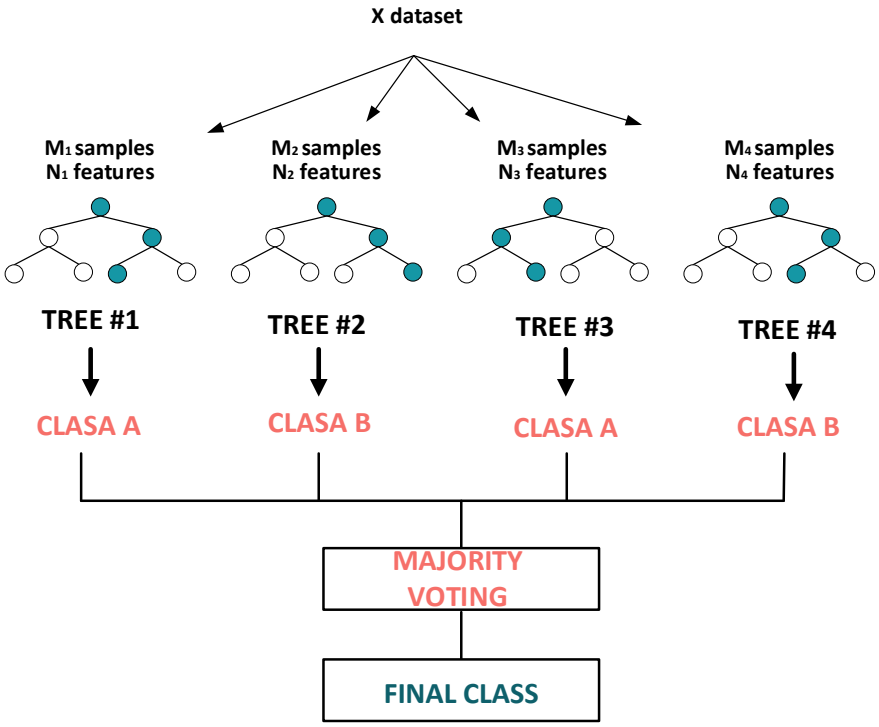

Figure 2. Random Forest framework.

3.2. LightGbm method

LightGBM is a novel algorithm developed by Microsoft Research Asia that enhances the gradient boosting framework and GBDT model. It combines multiple weak regression trees into a single powerful regression tree in a linear manner (Zheng et al., 2023), while significantly reducing time complexity through the Histogram decision tree optimization algorithm. Additionally, LightGBM adopts the Leaf-Wise leaf growth strategy with depth limitation, enabling efficient parallel training, feature parallelism, and fast processing of large-scale data. By addressing scalability and running speed limitations of traditional boosting algorithms, LightGBM supports parallel learning to greatly reduce training time and computational costs. The LightGBM model employs histogram optimization for feature representation and employs the Leaf-wise strategy to enhance model accuracy, as illustrated in **Figures 3** and **4**.

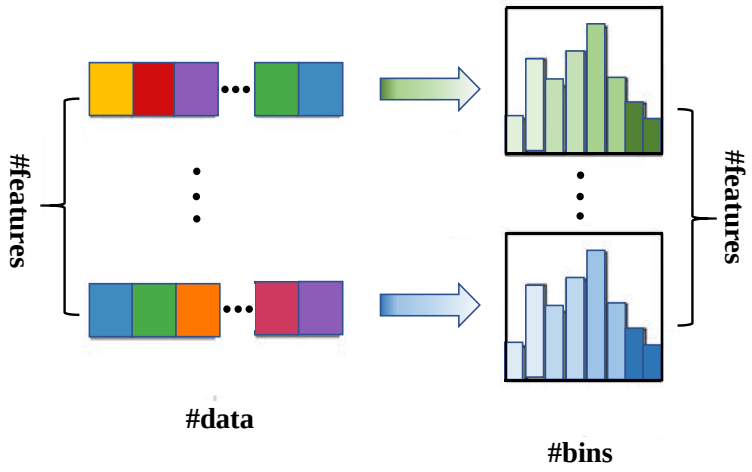

Figure 3. Histogram optimization.

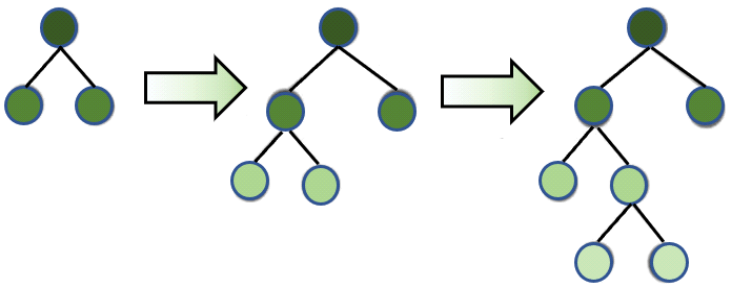

Figure 4. Leaf-wise growth strategy.

3.3. XGBoost method

XGBoost is an optimization algorithm that combines base functions and weights to enhance data fitting. Due to its outstanding generalization capability, scalability, and computational efficiency (Gogineni et al., 2023), XGBoost has garnered attention in the fields of statistics, data mining, and machine learning. For a dataset with n instances and m dimensions, the XGBoost model can be expressed as Equation (1) (Zhou et al., 2023). When building an XGBoost model, it is crucial to identify the optimal parameters by minimizing the objective function to achieve the best model.

This objective function consists of error terms and model complexity terms, as illustrated in Equations (2)–(4).

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F(i = 1,2,\cdots,n) \tag{1}$$

$$Obj = L + \Omega \tag{2}$$

$$L = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{3}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

Among them, $F = \{f(x) = \omega_{q(x)}\}(q: R^m \to \{1,2,\cdots,T\}, \omega \in R^T)$ denotes the set of CART structures. q represents the tree structure where samples are mapped to leaf nodes, Tstands for the number of leaf nodes, ω denotes the real-valued scores of leaf nodes, γT refers to the L_1 regularization term, $\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ corresponds to the L_2 regularization term, $\widehat{y_i}^{(t)}$ signifies the prediction result of the model in the *t*-th iteration, and $f_t(x_i)$ indicates the new function added in the *t*-th iteration.

In the course of refining the training process through the utilization of training data, the original model remains unaltered while incorporating novel base learners to incrementally decrease the discrepancy between predicted and actual values (Qian et al., 2020), thereby mitigating model bias. The training procedure is depicted in Equations (5) and (6).

$$\widehat{y_i}^{(t)} = \widehat{y_i}^{(t-1)} + f_t(x_i) \tag{5}$$

$$Obj^{(t)} \approx \sum_{i=1}^{n} [y_i - (\widehat{y_i}^{(t-1)} + f(x_i))]^2 + \Omega \tag{6}$$

To facilitate a rapid search for parameters minimizing the objective function, we conducted a second-order Taylor expansion on the objective function $Obj^{(t)}$, resulting in an approximate objective function, as illustrated in Equation (1). It can be discerned that this objective function is solely reliant on the first and second derivatives of the error function, thereby deriving the objective function portrayed in Equation (2).

$$Obj^{(t)} \approx \sum_{i=1}^{n} [(y_i - \hat{y}^{(t-1)})^2 + 2(y_i - \widehat{y_i}^{(t-1)}) f_t(x_i) - h_i f_t^2(x_i) + \Omega \tag{7}$$

$$Obj^{(t)} \approx \sum_{i=1}^{n}[g_i w_{q(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2] + \lambda T + \frac{1}{2}\sum_{j=1}^{T} w_j^2$$
$$= \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)\omega_j^2] + \gamma T \quad (8)$$

Assuming the structure of the decision tree component q is known, the optimal ω_j can be determined by employing the objective function, yielding the optimal value of the objective function. This problem can be generalized as finding the minimum value of a quadratic function, as illustrated in Equations (9) and (10). $Obj^{(t)}$ can serve as a scoring function for evaluating model performance, where lower $Obj^{(t)}$ values indicate better model effectiveness. By recursively employing the aforementioned tree-building method, numerous regression tree structures can be generated, enabling the search for the optimal tree structure using $Obj^{(t)}$ scores. Integration of these optimized trees into existing models allows for the construction of highly optimized XGBoost models. The algorithmic workflow diagram of XGBoost models is visualized in **Figure 5**.

$$w_j^* = \frac{-\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \quad (9)$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \quad (10)$$

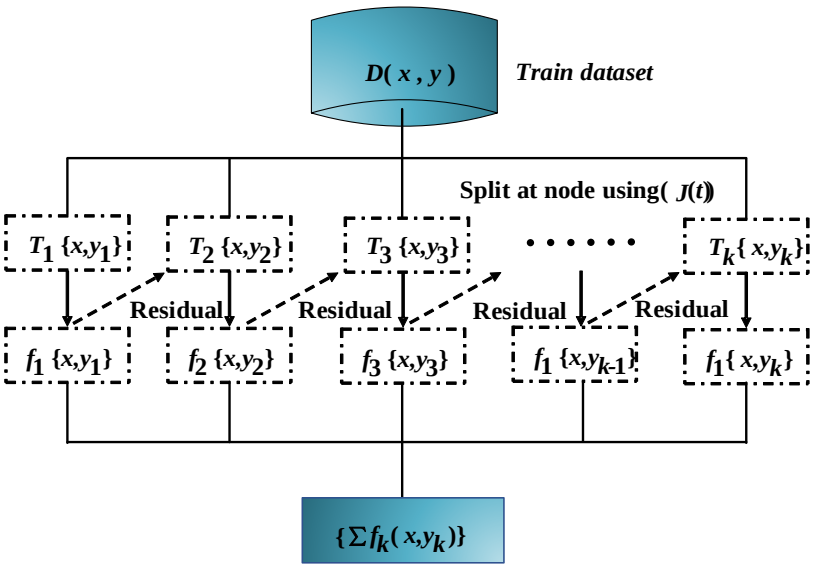

Figure 5. XGBoost algorithm framework.

3.4. Adaptive fusion construction

STEP1: The process of calculating Feature Importance (FI) utilizing the random forest algorithm begins with Mutual Information (MI) (Wang et al., 2022), a metric that quantifies the reduction in impurity when a given feature is employed as the

splitting criterion in each decision tree (Breiman, 2001). Subsequently, this value is averaged across all decision trees, resulting in the calculation of RF-FI, which is represented by Equation (11).

$$\text{RF\_FI} = \frac{1}{N_{trees}} \sum_{t=1}^{N_{trees}} \sum_{s \in S_{tj}} \triangle I(s,t) \tag{11}$$

Among them, N_{trees} denotes the number of trees, S_{tj} refers to all nodes s that employ the j-th feature for splitting in the t-th tree, and $\triangle I(s,t)$ signifies the impurity reduction induced by this node's split. The specific calculation process for FI is presented in **Table 1**.

Table 1. RF-FI computing framework.

STEP1: Random Forest feature importance calculation process	
	Input: The number of trees N_{trees}, the reduction of the impurity caused by node segmentation of each tree $\triangle I(s,t)$
	Output: RF_Feature_importance;
1	**For** n: $1 \rightarrow N_{\text{trees}}$ **do**
2	**for** All nodes of the tree **do**
3	RF_Feature_importance = RF_Feature_importance + $\triangle I(s,t)$
4	**end**
5	**end**
6	RF_Feature_importance = RF_Feature_importance /N_{trees}

STEP2: Calculation of LightGBM Feature Importance (Tianyu et al., 2019). The feature importance outputted by LightGBM can be computed using information gain. The total information gain obtained by splitting based on this feature is represented as Equation (12).

$$LG\_FI = \sum_{s \in Sj} \triangle I(s) \tag{12}$$

The set S_j represents all the nodes that are partitioned using feature X_j, while $\triangle I(s)$ denotes the information gain resulting from the split at node s. The specific calculation process of LG-FI is illustrated in **Table 2**.

Table 2. LightGBM-FI computing framework.

STEP2: Process of calculating feature importance in LightGBM	
	Input: Segmentation of nodes, information gain or reduction of impurity $\triangle I(s)$; Output: LGB_Feature_importance;
1	**for** All nodes of the tree do
2	LGB_Feature_importance = LGB_Feature_importance + $\triangle I(s)$
3	**end**

The feature importance evaluated by Accuracy as the criterion was obtained through the optimization of adaptive algorithms, which is specifically demonstrated in Equations (13)–(17).

$$\text{RF\_AC\_weight} = \frac{\text{RF\_Accuracy}}{\text{RF\_Accuracy} + \text{LGB\_Accuracy}} \tag{13}$$

$$\text{LGB\_AC\_weight} = \frac{\text{LGB\_Accuracy}}{\text{RF\_Accuracy} + \text{LGB\_Accuracy}} \tag{14}$$

$$\text{ARF\_AC\_FI} = \text{RF\_FI} \times \frac{\text{RF\_Accuracy}}{\text{RF\_Accuracy} + \text{LGB\_Accuracy}} \tag{15}$$

$$\text{ALGB\_AC\_FI} = \text{LGB\_FI} \times \frac{\text{RF\_Accuracy}}{\text{RF\_Accuracy} + \text{LGB\_Accuracy}} \tag{16}$$

$$\text{ARFLGB\_AC\_FI} = \text{ARF\_AC\_FI} + \text{ALGB\_AC\_FI} \tag{17}$$

Among them, RF_AC_weight denotes the Random Forest feature importance weight computed based on the Accuracy metric, whereas LGB_AC_weight represents the LightGBM feature importance weight derived from Accuracy. The calculation methods for Accuracy, Precision-Recall, Recall, F1 score, and AUC will be expounded in detail in Section 4.

Step 3: Determine the optimal performance metric and apply the adaptive algorithm to feature importance based on Precision, Recall, F1, and AUC. This will result in ARFLGB_PR_FI, ARFLGB_RE_FI, ARFLGB_F1_FI, and ARFLGB_AUC_FI. Additionally, for testing purposes, apply ARFLGB_AC_FI to the Input XGBoost. The most effective optimization metric will be determined through this process as outlined in **Table 3**. Detailed explanations of the calculation methods for Accuracy, Precision, Recall, F1 and AUC will be provided in Section 4 (Tang et al., 2022).

Table 3. Adaptive algorithm construction.

STEP3: An adaptive algorithm based on Accuracy index as an example	
1	Input: RF_Accuracy, LGB_Accuracy, RF_FI, LG_FI; Output: ARFLGB_AC_FI;
2	RF_AC_weight = RF_Accuracy/ (RF_Accuracy + LGB_Accuracy)
3	LGB_AC_weight = LGB_Accuracy/ (RF_Accuracy + LGB_Accuracy)
4	ARF_AC_FI = RF_AC_weight × RF_FI
5	ALGB_AC_FIt = LGB_AC_weight × LGB_FI
6	ARFLGB_AC_FI = ARF_AC_FI + ALGB_AC_FI

STEP4: To compute the adaptive algorithm based on Accuracy as a metric, we establish the feature importance weights of both the Random Forest and LightGBM as the ratio of their Accuracy to the aggregate Accuracy. By means of this optimization process, the feature importance of both Random Forest and LightGBM are amplified by their respective adaptive weights (Ma et al., 2022). The fused model's output

feature importance, which is derived from the optimization of the adaptive algorithm, is determined by the aggregate of the optimized output feature importance from both RF and LightGBM. The adaptive optimal index search process is shown in **Table 4**.

Table 4. Adaptive optimal index search.

STEP4: Search for the best ORS adaptive metrics		
	Input: Data set D consisting of ARFLGB_AC_FI, ARFLGB_PR_FI, ARFLGB_RE_FI, ARFLGB_F1_FI, ARFLGB_AUC_FI; Output: The best adaptive index;	
1	MAX = D[0]	
2	for	n:1 → N−1 **do**
3		**if** D[n] > MAX
4		MAX = D[n]
5	end	
6	Output MAX against the deserved index	
7	end	

The Accuracy metric in Step 3 is calculated using the adaptive algorithm idea. We will calculate adaptive algorithms separately for Precision, Recall, F1, and AUC metrics to obtain dataset D. The feature importance output from the fusion model of these five adaptive algorithm outputs will serve as input for the XGBoost. We will select the optimal model that yields the best classification results. The evaluation metric for this ARFLGB-XGBoost fusion model will be based on the chosen adaptive algorithm metric.

4. Case study

4.1. Data compilation

Table 5. 24 ORS independent variable classifications.

Indicators	**NO.**	**Meaning of the indicator**
Crime trail characteristics (CTC)	F_1	Replacement of communication equipment more than 5 times
	F_2	Failed to pay more than 3 times in the “Transaction” stage.
	F_3	Whether the number of times of changing IP is more than 3 times
	F_4	Whether the transaction amount exceeds 1000 RMB
Crime Evidence Characteristics (CEC)	F_5	Communication Equipment Features
	F_6	Characteristics of transaction dissuasion
	F_7	Changing IP Characteristics
	F_8	Characteristics of the current transaction
	F_9	Photographic evidence
	F_{10}	False identification
	F_{11}	Email, chat records
	F_{12}	Transfer records, transaction records
	F_{13}	Video evidence

Table 5. (*Continued*).

Indicators	NO.	Meaning of the indicator
Geographic characteristics of offences (GCO)	F_{14}	Level of economic education
	F_{15}	Level of updating of technical means of crime
	F_{16}	Changes in the social and economic environment
	F_{17}	Popularity of publicity and education by relevant departments
	F_{18}	Enforcement efforts of public security organs
	F_{19}	Level of people's awareness of network security
	F_{20}	Number of Internet dating ORS cases in the same month
Factors affecting the number of offences (FANO)	F_{21}	IP Replacement Features
	F_{22}	Language Features
	F_{23}	Personalised labels
	F_{24}	Common IP Locations

The data employed in this study was compiled from a collective effort between the Investigation Academy of China People's Public Security University and the Nanshan Branch of Qingdao City, Shandong Province, China, from 2016 to 2023. The agency's collection contains 1400 ORS cases, which were then subjected to textual analysis and feature extraction techniques. The cleaning process led to the identification of 24 ORS independent variables, which were subsequently categorized into four dimensions, as illustrated in **Table 5**.

The cases were meticulously modeled and purged of any inconsistencies, ensuring the accuracy and reliability of the data. In accordance with the distinct features of ORS, we delineated four dimensions of indicators: crime clue features, crime evidence features, regional characteristics of criminals, and factors influencing the frequency of crimes. Section 4.5 will divulge further details regarding the specific features incorporated within these dimensions. To assess the efficacy of the optimized XGBoost model (G.S. et al., 2021), we employed five metrics: AUC (Area Under Curve) (Belly et al., 2023), Accuracy, Precision, Recall, and F1 score. The formulas for calculating these five metrics are enumerated in Equations (18)–(25).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{18}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Recall} \tag{22}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{23}$$

$$\mathrm{I}(P_{\text{Positive sample}}, P_{\text{Negative sample}}) = \begin{cases} 1, P_{\text{Positive sample}} > P_{\text{Negative sample}} \\ 0.5, P_{\text{Positive sample}} = P_{\text{Negative sample}} \\ 0, P_{\text{Positive sample}} < P_{\text{Negative sample}} \end{cases} \tag{24}$$

$$\begin{aligned} \mathrm{AUC} &= P(P_{\text{Positive sample}} > P_{\text{Negative sample}}) \\ &= \frac{\sum \mathrm{I}(P_{\text{Positive sample}}, P_{\text{Negative sample}})}{M * N} \end{aligned} \tag{25}$$

4.2. Parametric search

Following an extensive grid search optimization of the XGBoost algorithm, we adopted a combined approach of cross-validation and grid search (Nti and Somanathan, 2024). Eventually, through a 5-fold cross-validation process, we determined the optimal parameters for each machine learning model subsequent to the grid search optimization (Wang et al., 2022). The process and iterative outcomes of parameter search are presented in **Tables 6** and **7**, respectively.

Table 6. Index parameter search.

STEP5: The highest accuracy rate is the highest accuracy rate of the model;	
	Input: dataset D, hyperparameters and corresponding value ranges; Output: the best combination of parameters, the highest accuracy of the model.
1	Construct a parameter grid consisting of the corresponding values of each hyperparameter.
2	Load the data into XGBoost;
3	**for** for all parameter grids **do**
4	Calculate the accuracy rate corresponding to the parameter grid;
5	Save the accuracy and the corresponding parameter combination;
6	**end**
7	**for** for all accuracies and parameter combinations **do**
8	Get the highest accuracy and its corresponding parameter combination:
9	**end**
10	Output The highest accuracy rate is the highest accuracy rate of the model;
11	Output The parameter combination corresponding to the highest accuracy is the best parameter combination;
12	**end**

Table 7. Grid search parameters.

	Parameter	Value range	Value result
XGBoost	n_estimators	[80, 100, 120, 160, 200]	120
	max_depth	[2, 4, 6, 8, 10, 12]	8
	learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3]	0.01
	min_child_weight	[1, 3, 5, 7]	1
	gamma	[0, 0.1, 0.2, 0.3, 0.4]	0
	colsample_bytree	[0.75, 0.8, 0.85]	0.85
	subsample	[0.75, 0.8, 0.85]	0.8
	reg_alpha	[1×10^{-5} , 0.01, 0.1, 1, 100]	1×10^{-5}
LightGBM	n_estimators	[80, 100, 120, 160, 200]	200
	max_depth	[2, 4, 6, 8, 10, 12, 20]	20
	learning_rate	[0.01, 0.04, 0.1, 0.2, 0.3]	0.04
	colsample_bytree	[0.01,0.1,0.5,1,2]	1
	subsample	[0.8,0.9,1.1.1 1.2]	1
	reg_alpha	[1×10^{-5} , 0.01, 0.1, 1, 100]	1×10^{-5}
	min_split_gain	[0, 0.01, 0.1, 1, 2]	0
	reg_lambda	[0, 0.01, 0.1, 1, 2]	0
Random Forest	n_estimators	[80, 100, 120, 160, 200]	200
	max_depth	[2, 4, 6, 8, 10, 12, 20]	10
	max_leaf_nodes	[0.01, 0.1, 0.5, 1, 5, 10, 20, 30, 50]	50
	min_samples_leaf	[None, 1, 5, 10, 15, 30, 50]	1
	min_samples_split	[0.01, 0.1, 0.5, 1, 5, 10, 20, 30, 50]	2
DecisionTree	max_depth	[2, 4, 6, 8, 10, 12, 20]	10
	min_samples_split	[0.01, 0.1, 0.5, 1, 5, 10, 20, 30, 50]	50
	min_samples_leaf	[None, 1, 5, 10, 15, 30, 50]	5
	min_impurity_decrease	[1×10^{-5} , 0.01, 0.1, 1, 100]	1×10^{-5}
	max_leaf_nodes	[25.50, 75, 100]	75
SVM	C	[1×10^{-5} , 0.01, 0.1, 1, 100]	1
	cache_size	[50, 100, 150, 200]	200
	coef0	[0, RBF, Poly]	0
	decision_function_shape	[ovo, ovr, None]	ovr
	degree	[None, 0.01, 0.1, 1, 2, 3]	3
	gamma	[auto, RBF, Poly, Sigmoid]	auto
	kernel	[Linear, RBF, Poly, Sigmoid]	RBF
	max_iter	[1, −1]	−1
	probability	[True, False]	FALSE
	shrinking	[True, False]	TRUE
	tol	[1×10^{-3}]	0.001
	verbose	[True, False]	FALSE

4.3. Improved evaluation of ARFLGB-XGBoost

To evaluate the memory improvement of ARFLGB-XGBoost (Kumar, 2023), its accuracy was simulated and analyzed, and the confusion matrix of the ARFLGB-XGBoost model was obtained, as shown in **Figure 6**.

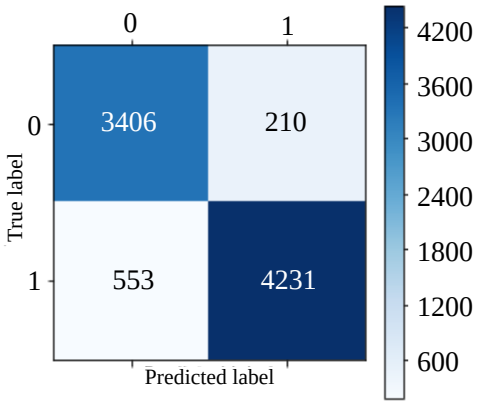

Figure 6. ARFLGB-XGBoost confusion matrix.

The optimised ARFLGB-XGBoost model was employed for 8400 classification predictions via 5-fold cross-validation. As illustrated in **Figure 6**, the numbers denote the number of true responses for each class target value. Based on this confusion matrix, it can be discerned that the ARFLGB-XGBoost model exhibits a high classification accuracy between correctly predicted positives and correctly predicted negatives.

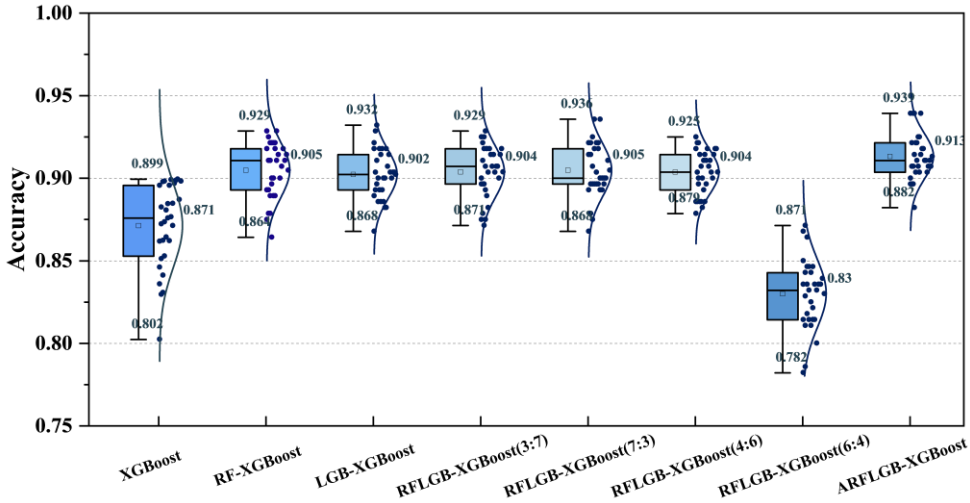

Figure 7. Comparison of ARFLGB-XGBoost model accuracy.

After conducting 30 repeated experiments, we performed classification on the original data and obtained the accuracy results for both the basic XGBoost model and the improved XGBoost model. As depicted in **Figure 7**, RF-XGBoost and LGB-XGBoost exhibited significant enhancements in accuracy, indicating that opting for RF and LightGBM models for feature engineering was a correct decision. However, the classification performance of RFLGB-XGBoost based on a certain fusion proportion did not surpass or even fell below that of RF-XGBoost and LGB-XGBoost

models, suggesting that adopting RFLGB-XGBoost model with fixed fusion proportion is not advisable.

Conversely, ARFLGB-XGBoost model fused by an adaptive algorithm outperformed the basic Xgboost model in terms of accuracy with its maximum value reaching 0.939 and upper quartile at 0.921; both values are higher than those achieved by RFLGB-XGBoost mode based on a certain fusion proportion. The median value is identical to RF-XGBoost mode but higher than other fusion modes; however, ARFLGB-XGBoost mode's maximum value, upper quartile, lower quartile, and minimum are all superior to those of RF-XGBoost; where minimum exceeds 0.88, lower quartile is approximately equal to 0.9, upper quartile is approximately equal to 0.92; indicating that this mode exhibits excellent overall classification accuracy.

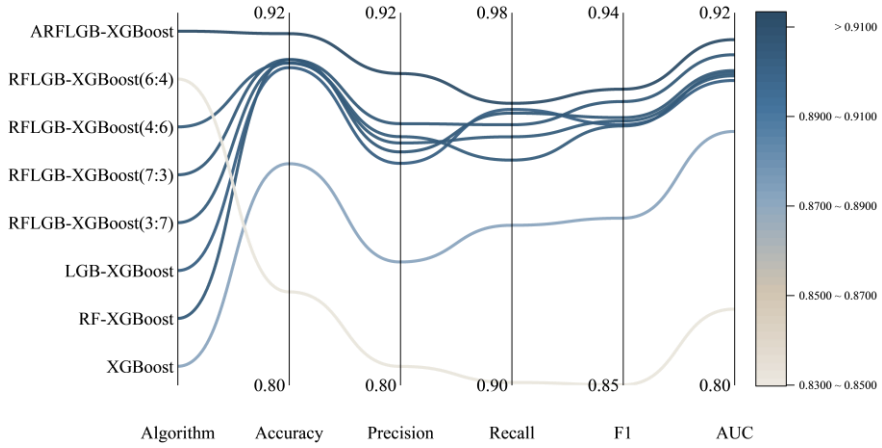

Figure 8. Comparison between ARFLGB-XGBoost model and optimized model.

According to **Figure 8**, the average values of AUC, F1, Recall, Precision, and Accuracy for the basic XGBoost model and the improved XGBoost model are as follows. From the graph, the ARFLGB-XGBoost model shows significant improvements in all five metrics. However, the RFLGB-XGBoost model based on a certain proportion of fusion does not surpass RF-XGBoost and LGB-XGBoost models completely; in fact, it may even perform worse than the basic XGBoost model. Therefore, we cannot adopt the RFLGB-XGBoost model based on a certain proportion of fusion. On the other hand, the ARFLGB-XGBoost model obtained through adaptive algorithm fusion performs significantly better than the RFLGB-XGBoost model based on a certain proportion of fusion in all five metrics mentioned above and far exceeds the performance of XGBoost basic model. The average value of Accuracy in ARFLGB-XGBoost classification results is 0.913; F1 score is 0.922; AUC score is 0.911; Recall score is 0.962. All these metrics are above 0.9 which indicates that this model greatly improves upon XGBoost's basic classification performance and possesses excellent overall classification capability.

5. Discussions

5.1. Significance analysis of ORS features

In conducting predictive analysis, we examined the output characteristics of the RF and LightGBM (Yang et al., 2021), which had been optimized by adaptive

algorithms. To ensure the accuracy of our analysis, we conducted thirty replicates of the experiments and meticulously documented the outcomes of each replicate as depicted in **Figures 9** and **10**. Ultimately, we leveraged the Output feature importance of the ARFLGB-XGBoost model to elucidate this process. These analytical findings not only contribute to a deeper understanding of the model's working mechanism but also facilitate its optimization for enhanced prediction accuracy (Ribeiro et al., 2016).

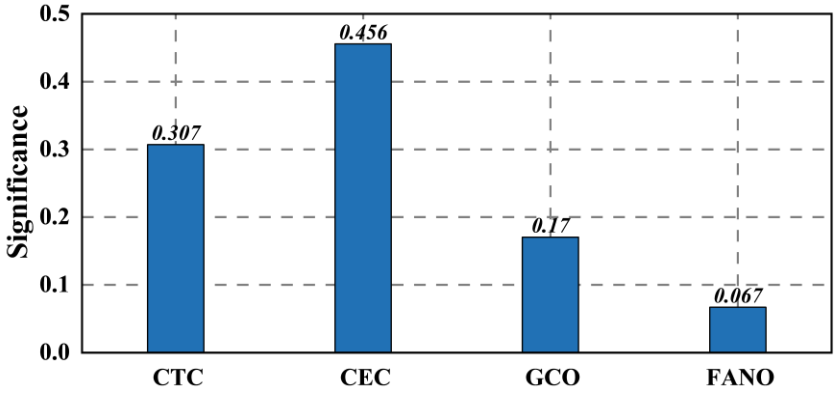

Figure 9. Primary characteristics feature significance.

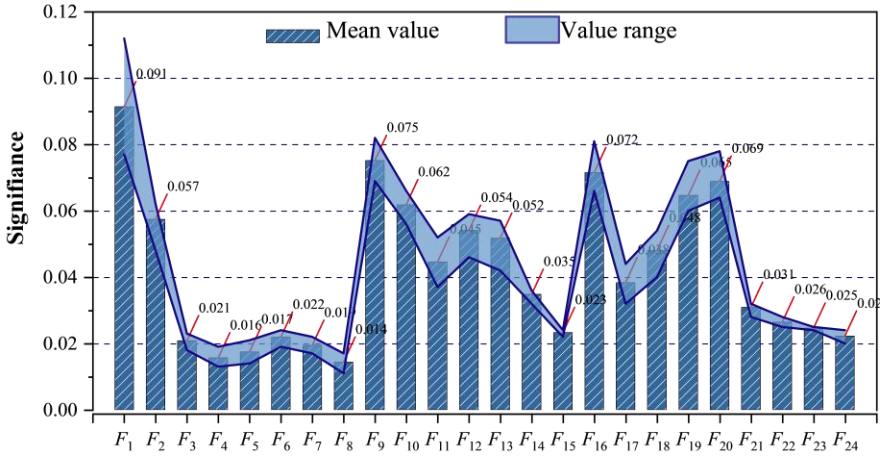

Figure 10. Secondary characteristics feature significance.

From **Figure 9**, it is evident that from a macroscopic perspective, the significance of CEC features reaches as high as 0.456, indicating their pivotal qualitative role in classifying ORS telecommunications fraud cases. Subsequently, with a value of 0.307, C emerges as another crucial feature implying its potential utility in warning models as an initial step for crime clue identification. By comprehensively considering both crime clues and crime evidence, law enforcement agencies can swiftly and effectively determine whether a case belongs to the category of ORS frauds. GCO features rank third with a value of 0.17 due to their discernible regional characteristics specific to ORS telecommunications fraud cases; this aspect becomes one of the key indicators distinguishing them from other types of telecom scams and facilitates prompt application of data investigation models by law enforcement agencies. Lastly, the influence factor percentage associated with the number of crimes stands at only 0.067, reflecting to some extent how social environment and law enforcement efforts impact

crime occurrence over time; however, its contribution in classification work remains relatively minor.

From a micro perspective, the ARFLGB-XGBoost model output reveals the significance of secondary characteristic features (**Figure 10**). These features can be categorized into three groups based on their average significance levels. The highly significant factors include F_1, F_9, F_{10}, F_{16}, F_{19}, and F_{20}. Factors with medium significance comprise F_2, F_{11}, F_{12}, F_{13}, F_{14}, F_{17}, F_{18}, and F_{21}. Lastly, factors with lower significance consist of F_3, F_4, F_5, F_6, F_7, F_8, F_{15}, F_{22}, F_{23}, and F_{24}. To reduce the incidence of ORS accidents, the following analysis and suggestions are put forward according to these three characteristics.

Highly significant factors, such as fake IP addresses and lower awareness of fraud prevention, play a pivotal role in early warning models. For instance, when d Replacement of communication equipment more than 5 times(F_1), they should promptly employ technical investigation methods and utilize the ARFLGB-XGBoost early warning model to prevent potential victims from becoming emotionally dependent on the manipulative tactics of criminal suspects before any crime occurs. Additionally, employing video surveillance techniques to capture incriminating evidence like deceptive links and characteristic messages can effectively reduce incidents of telecom fraud caused by criminals. Furthermore, it is imperative to enhance public awareness campaigns for fraud prevention.

Medium significant factors in ORS play a significant role, including the presence of false information and video chat evidence. Taking Failed to pay more than 3 times in the "Transaction" stage(F_2) as an example, when multiple payment failures are detected by the police during technological investigations, it is crucial to promptly issue a transfer alert to the victim and clearly communicate the potential risks associated with the other party's account in order to prevent further transfer actions by the victim. Moreover, enhancing dissuasion strategies during case investigations can effectively mitigate challenges faced by law enforcement in combating fraudsters who employ technical means to evade detection features.

Although the lower significant factors in the early warning model, such as transaction characteristics and linguistic personality traits of fraudsters, may carry a relatively smaller weightage, it is imperative not to overlook their significance. Taking Characteristics of the current transaction (F_8) as an example, when investigating a Ponzi scheme, promptly analyzing its features and implementing fund tracking and control measures becomes crucial to minimize financial losses and prevent further escalation of damages. Moreover, enhancing the description of regional characteristics pertaining to criminal suspects and organizing police forces for arrest operations can effectively mitigate losses caused by prolonged multi-location movements of telecom fraud suspects.

5.2. Algorithm comparison

By contrasting the classification outcomes of the ARFLGB-XGBoost model with five traditional machine learning techniques across five key metrics such as AUC, F1, Recall, Precision, and Accuracy, we evaluate the optimal classification performance. The parameters of these traditional machine learning models have all been fine-tuned

through grid search parameter optimization. **Figure 11** displays the ultimate classification results, offering a comprehensive overview of the strengths and weaknesses of each model in terms of classification performance.

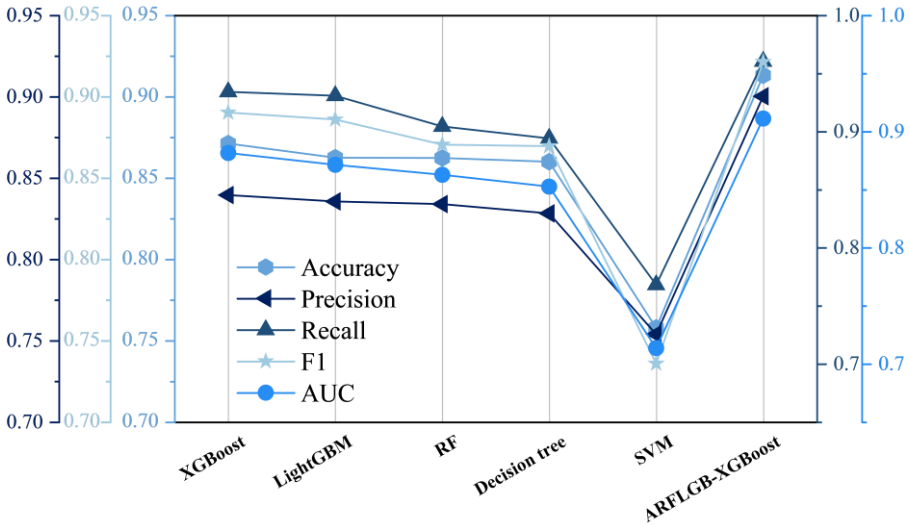

Figure 11. Comparison between ARFLGB-XGBoost model and traditional machine learning model.

Compared to the best-performing traditional machine learning models, the ARFLGB-XGBoost algorithm achieved a maximum accuracy of 0.939, an improvement of approximately 3.9%. It also reached a maximum precision of 0.926, demonstrating an increase of approximately 12.5%. In terms of recall, the model obtained a maximum value of 0.981, corresponding to an improvement of around 5.0%.

Furthermore, the ARFLGB-XGBoost algorithm achieved a maximum F1 score of 0.947 and an AUC value of 0.935, corresponding to improvements of approximately 5.6% and 5.2%, respectively. In addition to these metrics, the ARFLGB-XGBoost algorithm outperformed five other traditional machine learning models in various aspects, including output classification results, median, and quartiles. Therefore, it is evident that the ARFLGB-XGBoost algorithm significantly enhances classification performance and exhibits superior generalization ability and robustness in solving classification problems compared to other models available.

6. Conclusion remarks

To mitigate the prevalence of telecommunications fraud cases and provide case warnings and classifications, this study presents a method for constructing an ORS warning model based on the ARFLGB-XGBoost framework. This model is derived from the integration of three traditional machine learning models, namely Random Forest, LightGBM, and XGBoost, coupled with an adaptive algorithm. In terms of feature engineering, the model employs Random Forest and LightGBM for output feature importance analysis and adopts an adaptive algorithm to construct the ARFLGB algorithm to obtain weighted features. These weighted features are adaptively adjusted based on the classification indicators of RF and LGB Output, and ultimately fed into the XGBoost model for classification. The dataset employed in this

research is derived from ORS cases, consisting of 24 independent variable features and one binary dependent variable feature. Among them, the independent variable features can be categorized into four dimensions: crime clues, crime evidence, crime regional characteristics, and factors influencing crime quantity.

To assess the efficacy of the proposed ARFLGB-XGBoost model, we conducted tests on a validation dataset. Based on its confusion matrix, the model demonstrates exceptional performance in binary classification accuracy, significantly surpassing the original XGBoost model. In comparison to the XGBoost model that solely incorporates RF and LGB fusion, all five major classification metrics exhibit varying degrees of enhancement. Relative to the RFLGB-XGBoost model that incorporates class ratio fusion, there is a nearly 3.9% increase in accuracy, a 4.2% increase in precision, a 0.8% increase in recall, a 1% increase in F1 score, and a 0.5% increase in AUC value. Moreover, in comparison to traditional machine learning models, such as SVM, decision trees, random forests, XGBoost, and LightGBM, our proposed ARFLGB-XGBoost model exhibits an approximately 3.9% increase in accuracy, along with a notable improvement of approximately 12.5% in precision, a 5.0% increase in recall rate, a 6% increase in F1 score, and a 5.2% increase in AUC value. These findings unequivocally demonstrate that the proposed ARFLGB-XGBoost model possesses outstanding classification performance.

Although this study offers enhanced precision and efficiency in decision support for fraud prevention and combating, several limitations should be noted: (1) The research may be limited by the availability of ORS data, particularly concerning timeliness issues associated with ORS cases. Furthermore, potential biases or noise in the data could affect the model's performance. (2) Micro factors, such as psychological influences on indicator characteristics from the perspective of victims, have not been comprehensively considered in the study of ORS indicators. (3) The analysis was exclusively based on ORS data, neglecting further exploration of other types of telecommunications fraud cases.

In future research, we can further accumulate ORS case data to gain deeper insights into the trends of criminal development. Real-time tracking of this data will enable us to accurately understand changes in criminal activities. Once the data accumulates to a certain scale, we can leverage advanced Conditional Generative Adversarial Network (CGAN) technology for feature recognition, thereby enhancing the precision of fraud detection alerts. Additionally, the ORS alert algorithm can be extended to other domains such as finance, healthcare, and government management, providing solutions for a wider range of societal issues.

Author contributions: Conceptualization, GC and WD; methodology, GC; software, YS; validation, GC, AW and LW; formal analysis, GC; investigation, GC; resources, LW; data curation, WD; writing—original draft preparation, GC; writing—review and editing, LW; visualization, GC; supervision, GC; project administration, GC; funding acquisition, WD. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the project Double First-Class Innovation Research Project for People's Public Security University of China (grant number: 2023SYL06).

Data availability statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Informed consent statement: Informed consent was obtained from all subjects involved in the study.

Conflict of interest: The authors declare no conflict of interest.

References

Abbasimehr, H., Paki, R., & Bahrini, A. (2023). A novel XGBoost-based featurization approach to forecast renewable energy consumption with deep learning models. Sustainable Computing: Informatics and Systems, 38. https://doi.org/10.1016/j.suscom.2023.100863

Afriyie, J. K., Tawiah, K., Pels, W. A., et al. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6. https://doi.org/10.1016/j.dajour.2023.100163

Ashraf, M. T., Dey, K., & Mishra, S. (2023). Identification of high-risk roadway segments for wrong-way driving crash using rare event modeling and data augmentation techniques. Accident Analysis & Prevention, 181. https://doi.org/10.1016/j.aap.2022.106933

Bahaghighat, M., Ghasemi, M., & Ozen, F. (2023). A high-accuracy phishing website detection method based on machine learning. Journal of Information Security and Applications, 77. https://doi.org/10.1016/j.jisa.2023.103553

Belly, G., Boeckelmann, L., Graciano, C. M. C., et al. (2023). Forecasting sovereign risk in the Euro area via machine learning. Journal of Forecasting, 42(3), 657–684. Portico. https://doi.org/10.1002/for.2938

Breiman, L. L. (2001). Random forest. Mach Learn. 45, 5-32. https://doi.org/10.1023/A:1010933404324

Cao, R., Wang, J., Mao, M., et al. (2023). Feature-wise attention based boosting ensemble method for fraud detection. Engineering Applications of Artificial Intelligence, 126. https://doi.org/10.1016/j.engappai.2023.106975

Coluccia, A., Pozza, A., Ferretti, F., et al. (2020). Online Romance Scams: Relational Dynamics and Psychological Characteristics of the Victims and Scammers. A Scoping Review. Clinical Practice & Epidemiology in Mental Health, 16(1), 24–35. https://doi.org/10.2174/1745017902016010024

Cross, C., & Holt, T. J. (2023). More than Money: Examining the Potential Exposure of Romance Fraud Victims to Identity Crime. Global Crime, 24(2), 107–121. https://doi.org/10.1080/17440572.2023.2185607

Domashova, J., & Kripak, E. (2022). Development of a generalized algorithm for identifying atypical bank transactions using machine learning methods. Procedia Computer Science, 213, 101–109. https://doi.org/10.1016/j.procs.2022.11.044

G.S., T., Dheeshjith, S., Iyengar, S. S., et al. (2021). A hybrid and effective learning approach for Click Fraud detection. Machine Learning with Applications, 3. https://doi.org/10.1016/j.mlwa.2020.100016

Gogineni, A., Panday, I. K., Kumar, P., et al. (2023). Predicting compressive strength of concrete with fly ash and admixture using XGBoost: a comparative study of machine learning algorithms. Asian Journal of Civil Engineering, 25(1), 685–698. https://doi.org/10.1007/s42107-023-00804-0

Izotova, A., & Valiullin, A. (2021). Comparison of Poisson process and machine learning algorithms approach for credit card fraud detection. Procedia Computer Science, 186, 721–726. https://doi.org/10.1016/j.procs.2021.04.214

Jiang, Z., Chen, K., Wen, H., et al. (2022). Applying blockchain-based method to smart contract classification for CPS applications. Digital Communications and Networks, 8(6), 964–975. https://doi.org/10.1016/j.dcan.2022.08.011

Kamboj, A., Kumar, P., Bairwa, A. K., et al. (2023). Detection of malware in downloaded files using various machine learning models. Egyptian Informatics Journal, 24(1), 81–94. https://doi.org/10.1016/j.eij.2022.12.002

Koc, K., Ekmekcioğlu, Ö., & Gurgun, A. P. (2021). Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. Automation in Construction, 131. https://doi.org/10.1016/j.autcon.2021.103896

Kolev, M. (2023). XGB-COF: A machine learning software in Python for predicting the friction coefficient of porous Al-based composites with Extreme Gradient Boosting. Software Impacts, 17. https://doi.org/10.1016/j.simpa.2023.100531

Kumar, M. (2023). Early detection of chronic kidney disease using recursive feature elimination and cross-validated XGBoost model. International Journal of Computational Materials Science and Engineering, 13(04). https://doi.org/10.1142/s2047684123500367

Lao, Z., He, D., Wei, Z., et al. (2023). Intelligent fault diagnosis for rail transit switch machine based on adaptive feature selection and improved LightGBM. Engineering Failure Analysis, 148. https://doi.org/10.1016/j.engfailanal.2023.107219

Lazarus, S., Button, M., & Kapend, R. (2022). Exploring the value of feminist theory in understanding digital crimes: Gender and cybercrime types. The Howard Journal of Crime and Justice, 61(3), 381–398. Portico. https://doi.org/10.1111/hojo.12485

Lazarus, S., Whittaker, J. M., McGuire, M. R., et al. (2023). What do we know about online romance fraud studies? A systematic review of the empirical literature (2000 to 2021). Journal of Economic Criminology, 2. https://doi.org/10.1016/j.jeconc.2023.100013

Liu, W., Fan, H., Xia, M., et al. (2022). Predicting and interpreting financial distress using a weighted boosted tree-based tree. Engineering Applications of Artificial Intelligence, 116. https://doi.org/10.1016/j.engappai.2022.105466

Liu, Z., Zhang, Z., Yang, H., et al. (2023). An innovative model fusion algorithm to improve the recall rate of peer-to-peer lending default customers. Intelligent Systems with Applications, 20. https://doi.org/10.1016/j.iswa.2023.200272

Ma, L., Zhou, C., Lee, D., et al. (2022). Prediction of axial compressive capacity of CFRP-confined concrete-filled steel tubular short columns based on XGBoost algorithm. Engineering Structures, 260. https://doi.org/10.1016/j.engstruct.2022.114239

Mohiuddin, G., Lin, Z., Zheng, J., et al. (2023). Intrusion Detection using hybridized Meta-heuristic techniques with Weighted XGBoost Classifier. Expert Systems with Applications, 232. https://doi.org/10.1016/j.eswa.2023.120596

Mokbal, F. M. M., Dan, W., Xiaoxi, W., et al. (2021). XGBXSS: An Extreme Gradient Boosting Detection Framework for Cross-Site Scripting Attacks Based on Hybrid Feature Selection Approach and Parameters Optimization. Journal of Information Security and Applications, 58. https://doi.org/10.1016/j.jisa.2021.102813

Murugan, M. S., & T, S. K. (2023). Large-scale data-driven financial risk management & analysis using machine learning strategies. Measurement: Sensors, 27. https://doi.org/10.1016/j.measen.2023.100756

Musbah, H., Ali, G., Aly, H. H., et al. (2022). Energy management using multi-criteria decision making and machine learning classification algorithms for intelligent system. Electric Power Systems Research, 203. https://doi.org/10.1016/j.epsr.2021.107645

Nanath, K., & Olney, L. (2023). An investigation of crowdsourcing methods in enhancing the machine learning approach for detecting online recruitment fraud. International Journal of Information Management Data Insights, 3(1). https://doi.org/10.1016/j.jjimei.2023.100167

Nti, I. K., & Somanathan, A. R. (2024). A Scalable RF-XGBoost Framework for Financial Fraud Mitigation. IEEE Transactions on Computational Social Systems, 11(2), 1556–1563. https://doi.org/10.1109/tcss.2022.3209827

Qian, Q., Sun, H., Wu, J., et al. (2020). AKI Prediction Models in ICU: A Comparative Study (Preprint). https://doi.org/10.2196/preprints.18257

Razavi, R., Gharipour, A., Fleury, M., et al. (2019). A practical feature-engineering framework for electricity theft detection in smart grids. Applied Energy, 238, 481–494. https://doi.org/10.1016/j.apenergy.2019.01.076

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Cornell University. https://doi.org/10.1145/2939672.2939778

Sha, Y., Faber, J., Gou, S., et al. (2022). An acoustic signal cavitation detection framework based on XGBoost with adaptive selection feature engineering. Measurement, 192. https://doi.org/10.1016/j.measurement.2022.110897

Singh, V. K. V., Rashwan, H. H., Akram, F. F., et al. (2018). Retinal Optic Disc Segmentation using Conditional Generative Adversarial Network. Cornell University.

Srivastava, A., Singh, V., & Drall, G. S. (2019). Sentiment Analysis of Twitter Data. International Journal of Healthcare Information Systems and Informatics, 14(2), 1–16. https://doi.org/10.4018/ijhisi.2019040101

Tan, B., Gan, Z., & Wu, Y. (2023). The measurement and early warning of daily financial stability index based on XGBoost and SHAP: Evidence from China. Expert Systems with Applications, 227. https://doi.org/10.1016/j.eswa.2023.120375

Tang, Z., Xiao, Y., Jiao, Y., et al. (2022). Research on Short-Term Low-Voltage Distribution Network Line Loss Prediction Based on Kmeans-LightGBM. Journal of Circuits, Systems and Computers, 31(13). https://doi.org/10.1142/s0218126622502280

Tianyu, B. B., Changbing, Z. Z., Chenlin, L. L., Economics, S.O.S. (2019). Design and application of pur-chase behavior recognition model in implicit feedback data based on Lightgbm algorithm. Wireless Internet Technology.

Wang, X., Gao, S., Guo, Y., et al. (2022). A Combined Prediction Model for Hog Futures Prices Based on WOA-LightGBM-CEEMDAN. Complexity, 2022(1). Portico. https://doi.org/10.1155/2022/3216036

Wang, X., Zhang, G., Lou, S., et al. (2022). Two-round feature selection combining with LightGBM classifier for disturbance event recognition in phase-sensitive OTDR system. Infrared Physics & Technology, 123. https://doi.org/10.1016/j.infrared.2022.104191

Wang, Z., & Thing, V. L. L. (2023). Feature mining for encrypted malicious traffic detection with deep learning and other machine learning algorithms. Computers & Security, 128. https://doi.org/10.1016/j.cose.2023.103143

Whitty, M. T., & Buchanan, T. (2016). The online dating romance scam: The psychological impact on victims – both financial and non-financial. Criminology & Criminal Justice, 16(2), 176–194. https://doi.org/10.1177/1748895815603773

Yan, Z., Chen, H., Dong, X., et al. (2022). Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost. Expert Systems with Applications, 207. https://doi.org/10.1016/j.eswa.2022.117943

Yang, L. L., Niu, X. X., Wu, J. J. (2021). RF-LighGBM: A probabilistic ensemble way to predict customer repurchase behaviour in community e-commerce. Cornell University.

Zhang, T., Zhu, W., Wu, Y., et al. (2023). An explainable financial risk early warning model based on the DS-XGBoost model. Finance Research Letters, 56. https://doi.org/10.1016/j.frl.2023.104045

Zhao, Z., Li, D., & Dai, W. (2023). Machine-learning-enabled intelligence computing for crisis management in small and medium-sized enterprises (SMEs). Technological Forecasting and Social Change, 191. https://doi.org/10.1016/j.techfore.2023.122492

Zheng, H.-L., An, S.-Y., Qiao, B.-J., et al. (2023). A data-driven interpretable ensemble framework based on tree models for forecasting the occurrence of COVID-19 in the USA. Environmental Science and Pollution Research, 30(5), 13648–13659. https://doi.org/10.1007/s11356-022-23132-3

Zhou, X., Zhao, C., & Bian, X. (2023). Prediction of maximum ground surface settlement induced by shield tunneling using XGBoost algorithm with golden-sine seagull optimization. Computers and Geotechnics, 154. https://doi.org/10.1016/j.compgeo.2022.105156

Zhu, C., Zhang, C., Wang, R., et al. (2023). Building of safer urban hubs: Insights from a comparative study on cyber telecom scams and early warning design. Urban Governance, 3(3), 200–210. https://doi.org/10.1016/j.ugj.2023.05.004