Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Synergy Between Resilient Networks and Random Forests in Online Fraud Detection

Junxi Wang[1], Ningtao Sun[2], Yuhan Lv[3], Jiayi Zhou[4], Yue Xiao[5]*

[1]School of Intelligent Manufacturing and Aviation, Zhuhai College of Science and Technology, Zhuhai 519041, Guangdong, China
[2]School of Computer Science, Zhuhai College of Science and Technology, Zhuhai 519041, Guangdong, China
[3]School of Liberal Arts, Zhuhai College of Science and Technology, Zhuhai 519041, Guangdong, China
[4]School of Logistics Management and Engineering, Zhuhai College of Science and Technology, Zhuhai 519041, Guangdong, China
[5]School of Data Science, Zhuhai College of Science and Technology, Zhuhai 519041, Guangdong, China

***Corresponding author:** Yue Xiao, xiaoyue810@zcst.edu.cn

Abstract: This paper explores the synergistic effect of a model combining Elastic Net and Random Forest in online fraud detection. The study selects a public network dataset containing 1781 data records, divides the dataset by 70% for training and 30% for validation, and analyses the correlation between features using a correlation matrix. The experimental results show that the Elastic Net feature selection method generally outperforms PCA in all models, especially when combined with the Random Forest and XGBoost models, and the ElasticNet + Random Forest model achieves the highest accuracy of 0.968 and AUC value of 0.983, while the Kappa and MCC also reached 0.839 and 0.844 respectively, showing extremely high consistency and correlation. This indicates that combining Elastic Net feature selection and Random Forest model has significant performance advantages in online fraud detection.

Keywords: Fraudulent websites; Machine learning; Elastic Net; Random forests

Online publication: March 28, 2025

1. Introduction

With the rapid development of Internet technology, the Internet has become an indispensable part of daily life. It provides a convenient platform for information acquisition, communication, and business transactions. However, this convenience also brings a series of security challenges, especially the increasing number of network fraud. Network fraud not only harms the economic interests of users, but also threatens the security and trust of the entire network ecosystem. Therefore, how to effectively identify and prevent network fraud has become an important topic in the field of network security.

Among the many cybersecurity issues, the detection of fraudulent websites is particularly critical. Fraudulent

websites usually disguise themselves as legitimate websites and induce users to disclose sensitive information, such as login credentials and credit card information, which brings direct financial losses and privacy risks to users. Traditional rule-based detection methods appear to be incompetent in the face of ever-changing fraudulent tactics, so researchers have begun to seek more advanced techniques to improve the accuracy and efficiency of detection [1].

In recent years, machine learning techniques have received a lot of attention for their power in data analysis and pattern recognition. The ability of machine learning algorithms to learn and recognize complex patterns from large amounts of data has led to great potential in the field of fraud detection. In recent years Dou *et al.* proposed a new model called CAmouflage-REsistant GNN (CARE-GNN), which aims to strengthen the aggregation process of the GNN through three neural modules to prevent the fraudster's camouflage behaviors and demonstrated the effectiveness of the method on two real fraud datasets through experimental results [2]. Kim *et al.* proposed a fraud detection method based on graph neural network (GNN) to solve the fraud detection problem through a dynamic relational attention aggregation mechanism [3]. The method learns the node representation of each relation and uses a learnable attention function to aggregate the node representations, assigning different attention coefficients to each relation. Xu *et al.* proposed SEC-GFD, a semi-supervised GNN fraud detector, which improves detection through hybrid filtering and local environment constraints and experimentally proves that it outperforms other methods [4].

Based on this research, this paper will explore the application of models combining Elastic Net and Random Forest in fraudulent website detection. Elastic Net is an integrated regression technique that incorporates L1 and L2 regularization to enhance the generalization ability of the model by promoting sparsity in feature selection and enhancing model stability. Meanwhile, Random Forest, an integrated learning method, effectively improves the accuracy and robustness of the model by constructing multiple decision trees and integrating their predictions. This combination aims to optimize the performance of fraud detection to identify and prevent online fraud more effectively.

2. Model construction

2.1. Analysis of data sets

A publicly available web dataset was selected for this study, which contains 1781 data records covering 14 different feature dimensions. For model training and validation, the dataset was divided in a 7:3 ratio, where 70% of the data was used to construct a training set for model learning and the remaining 30% was used as a validation set to evaluate the model's performance and generalization ability. This division aims to ensure that the model can make full use of the data during the training process, and at the same time, the validation set is used to monitor the overfitting situation of the model, to achieve the optimization and refinement of the model. To assess the correlation between different features in the dataset, this study adopts the correlation matrix technique for the analysis. By constructing the correlation matrix, as shown in **Figure 1**, it is possible to visualize the correlation coefficients between the features.

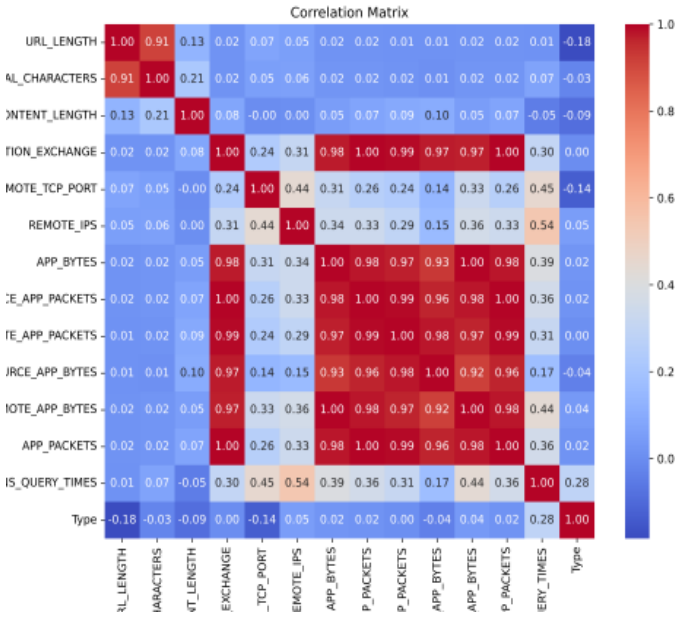


Figure 1. Correlation matrix diagram

2.2. Elastic networks

As shown in **Figure 1**, it can be observed that there is a high degree of positive correlation between features such as "APP_BYTES" and "APP_PACKETS", which may lead to the problem of multicollinearity during the modelling process, thus affecting the stability and prediction accuracy of the model [5, 6]. To solve this problem, this study adopts the Elastic Net regularization technique. The Elastic Net regularization technique combines the advantages of Lasso regression (L1 regularization) and Ridge regression (L2 regularization). By introducing L1 and L2 regularization at the same time, it is able to balance the feature selection and model smoothing, which effectively reduces the risk of overfitting and improves the prediction performance. The objective function of elastic network regression can be expressed as shown in Equation 1 below:

$$\mathcal{L}(\omega) = \frac{1}{2n}\|y - X_\omega\|_2^2 + \alpha\left(\rho\|w\|_1 + \frac{1-\rho}{2}\|w\|_2^2\right) \quad (1)$$

In this equation, X denotes the feature matrix, y is the target variable, the vector of regression coefficients to be estimated, n is the number of samples, is the number of features, and is the regularization strength parameter, which controls the degree of regularization of the model, and is the weighted ratio of L1 regularization (Lasso) to L2 regularization (Ridge). Specifically, when ρ = 1, the elasticity network degenerates into a Lasso regression model; when = 0, it degenerates into a Ridge regression model; and when 0 < ρ < 1, the elasticity network combines the advantages of L1 and L2 regularization, which is capable of feature selection as well as avoiding the problem of covariance between features.

In the objective function of the elastic network, the error term is $\frac{1}{2n}\|y - X_\omega\|_2^2$. The measure is the error of fit of the model, the regularization term is $\alpha\left(\rho\|w\|_1 + \frac{1-\rho}{2}\|w\|_2^2\right)$. It is used to control the complexity of the model and to prevent overfitting. The L1 regularization term (Lasso term) enables feature selection by summing the absolute values of the weights so that certain weight coefficients converge to zero, while the L2 regularization term (Ridge term) smoothes the model parameters and reduces instability due to covariance by summing the squares of

the weight coefficients.

Optimization of the model can be achieved by adjusting the parameters α and ρ, which govern the strength and balance of the regularization terms. The parameter α controls the overall regularization strength; larger values of α increase the penalty, leading to a more simplified model that can prevent overfitting. Meanwhile, ρ determines the relative contributions of L1 and L2 regularizations within the elastic net framework, allowing the model to flexibly select relevant features while suppressing irrelevant ones. Through the appropriate tuning of these parameters, the model's generalization ability is enhanced, ultimately improving its detection accuracy.

2.3. Random forest

In recent years, significant progress has been made in Random forest techniques, making them a prominent choice for ransomware detection due to their robustness and ability to handle high-dimensional data [7]. Similarly, Pathak *et al.* used various machine learning models for phishing website detection and after comparison found that random forest model was able to show better performance in this task [8]. The valid data used for detecting fraudulent websites in this study contains multiple dimensions and the data in each dimension is discrete, so the random forest technique can also be used for fraudulent website detection. Random forest, as an integrated learning model proposed based on bagging strategy (Bagging), consists of multiple decision trees. In random forest, each decision tree is an independent and uncorrelated classifier and each tree will only learn on a certain subset of the whole data. Because of this, the random forest can capture the nonlinear structure in the data through different decision trees and comprehensively acquire important features in the data while effectively reducing the risk of overfitting during the training process. The principle of the training and testing steps of the random forest model in this study is shown in **Figure 2**.

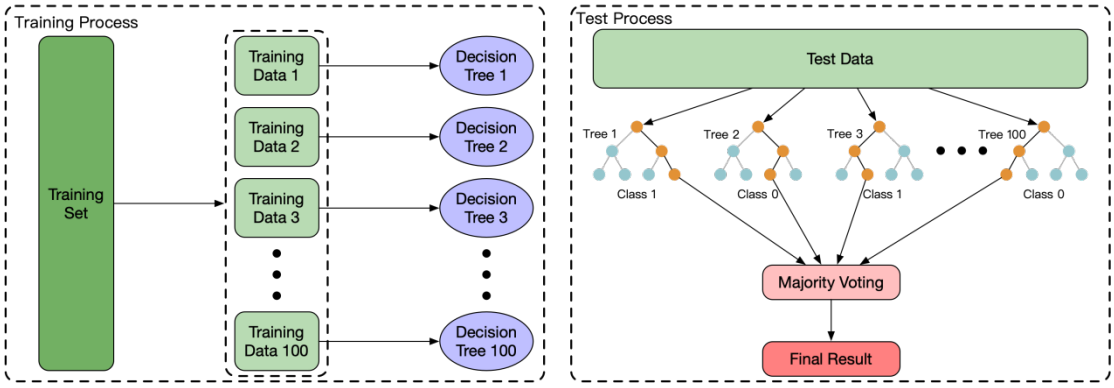


Figure 2. Random forest structure diagram

To ensure that the learning model used in this study has sufficient predictive power, the random forest model designed in this study consists of 100 decision trees. By using the fraud detection website dataset, the data was divided into 100 parts to train each decision tree, and the majority voting result of multiple decision trees was selected as the final result in the test. Ultimately, the depth of each decision tree in the random forest model trained in this study was moderate, with an average depth of 14.04. The training data indicated that the model was trained in a way that ensured that the model was able to capture the data features effectively while avoiding the risk of overfitting. With the bagging strategy feature, the random forest model applied to fraudulent website detection uses the majority voting method to predict the results of all decision trees with the prediction formula shown in Equation 2:

$$\hat{y} = arg \max_{C \in \{0,1\}} \sum_{t=1}^{T} 1(h_t(x) = C) \quad (2)$$

3. Results of the experiment

3.1. Model evaluation indicators

(1) Accuracy: Accuracy is one of the most intuitive metrics to evaluate in machine learning, measuring the number of samples the model predicts correctly as a percentage of the total number of samples.

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

(2) Specificity: Specificity refers to the rate of true negative cases, i.e., the proportion of cases correctly predicted as negative by the model, and it reflects the model's ability to classify negative cases.

$$Specificity = \frac{TN}{TN+FP}$$

(3) Log Loss: In binary classification problems, log-loss is a metric for assessing the performance of a classification model, which measures the difference between the model's predicted probability and the actual occurrence of a label.

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i\ log\ (p_i) + (1-y_i)\ log\ (1-p_i)]$$

(4) AUC: AUC indicates the ability of the model to identify positive classes; the higher the AUC value, the better the classification ability of the model.

(5) Kappa: The Kappa coefficient is a measure of classification accuracy that evaluates the agreement between two evaluators (or classifiers) by taking into account both observed agreement and chance agreement.

$$Kappa = \frac{P_A - P_e}{1-P_e}$$

(6) MCC: MCC is a metric for assessing the performance of classification models that takes into account the values of all four confusion matrices and is not affected by category imbalance.

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

3.2. Model results

Table 1 shows that two feature selection methods, PCA and ElasticNet, were used to compare model performance in online fraud detection.

Table 1. Table of evaluation indicators for different outcome model analyses

Method	Model	Accuracy	Specificity	Log Loss	AUC	Kappa	MCC
PCA	Logistic Regression	0.867	0985	0.308	0.809	0.0007	0.001
PCA	SVM	0.878	1.000	0.349	0.739	0.0000	0.000
PCA	XGBoost	0.933	0.970	0.237	0.893	0.667	0.669
PCA	Random Forest	0.943	0.989	0.288	0.920	0.697	0.711
ElasticNet	Logistic Regression	0.908	0.983	0.227	0.918	0.450	0.485
ElasticNet	SVM	0.886	1.000	0.239	0.933	0.103	0.233
ElasticNet	XGBoost	0.959	0.974	0.094	0.989	0.809	0.810
ElasticNet	Random Forest	0.968	0.993	0.102	0.983	0.839	0.844

Results indicate that ElasticNet outperformed PCA across all models, especially in Random Forest and XGBoost. The ElasticNet + Random Forest model achieved the highest accuracy (0.968), AUC (0.983), Kappa (0.839), and MCC (0.844). ElasticNet + XGBoost also performed well with an accuracy of 0.959 and AUC of 0.989. Under PCA, Random Forest had the best accuracy (0.943) and AUC (0.920). SVM achieved perfect specificity (1.000) under both methods, while Logistic Regression had lower specificity (0.983 under ElasticNet). In terms of log loss, ElasticNet + XGBoost and ElasticNet + Random Forest had the lowest values (0.094 and 0.102, respectively). Overall, ElasticNet + Random Forest provided the best performance for web fraud detection.

The performance of ElasticNet regularization combined with four machine learning models (Logistic Regression, SVM, XGBoost, and Random Forest) for classification is shown in **Figure 3** and **Figure 4**. The confusion matrices indicate that ElasticNet + Random Forest had the best performance for positive classes (only 3 misclassifications), while ElasticNet + Logistic Regression was best for negative classes (462 correct predictions, 8 misclassifications). The scatterplots show that ElasticNet + SVM had good linear separability, while ElasticNet + XGBoost and ElasticNet + Random Forest had some misclassifications but still performed well overall. Overall, ElasticNet + Random Forest appears to be the best model for this classification task.

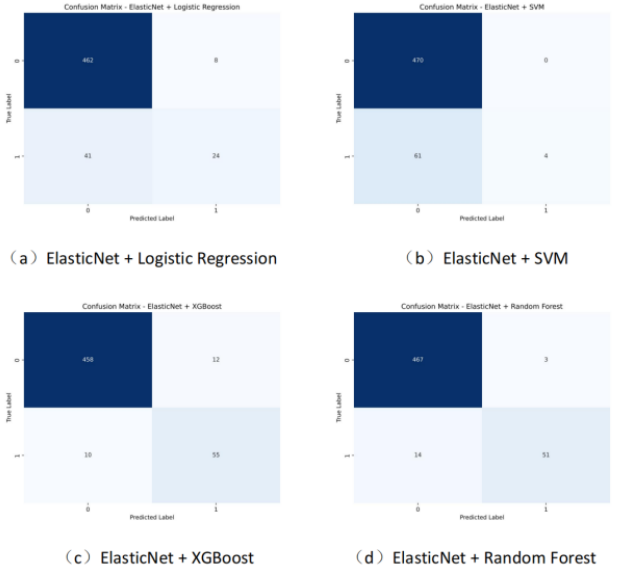
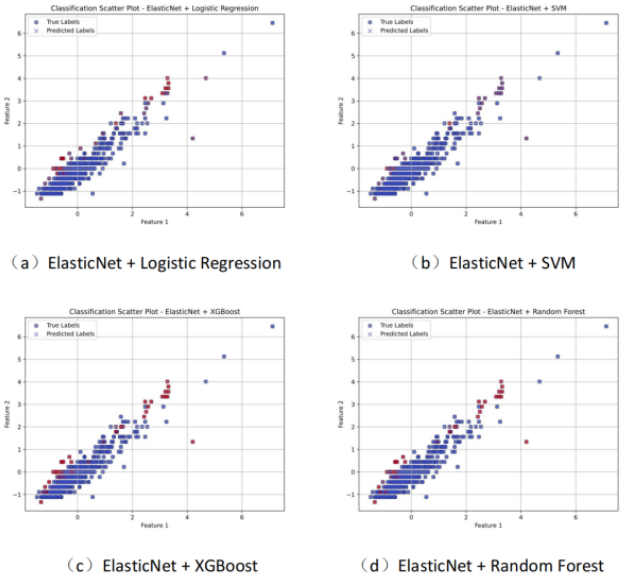


Figure 3. Confusion matrix for each combined model



（a）ElasticNet + Logistic Regression　（b）ElasticNet + SVM

（c）ElasticNet + XGBoost　（d）ElasticNet + Random Forest

Figure 4. Calibration curves for each combination

4. Conclusion

In this study, we validate the effectiveness of the model by combining ElasticNet with Random Forest and in online fraud detection. By comparing different feature selection methods and machine learning models, it is found that the combination of ElasticNet and Random Forest performs the best in terms of evaluation metrics such as accuracy, AUC value, Kappa, and MCC. The ElasticNet regularization technique not only reduces the feature complexity and the risk of overfitting but also achieves the feature selection, while the Random Forest effectively captures the nonlinear structure and important features in the data through the majority voting method. This combination not only improves the generalization ability of the model but also enhances the robustness and accuracy of the model, providing an efficient technical solution in the field of online fraud detection.

Funding

Guangdong Innovation and Entrepreneurship Training Programme for Undergraduates “Automatic Classification and Identification of Fraudulent Websites Based on Machine Learning” (Project No.: DC2023125)

Disclosure statement

The authors declare no conflict of interest.

Author contributions

Model construction and data analysis: Junxi Wang, Ningtao Sun
Writing: Yuhan Lv, Jiayi Zhou
Project guidance: Yue Xiao

References

[1] Pourhabibi T, Ong KL, Kam BH, et al., 2020, Fraud Detection: A Systematic Literature Review of Graph-Based Anomaly Detection Approaches. Decision Support Systems, 133: 113303.

[2] Dou Y, Liu Z, Sun L, et al., 2020, Enhancing Graph Neural Network-Based Fraud Detectors Against Camouflaged Fraudsters. Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020: 315–324.

[3] Kim H, Choi J, Whang JJ, 2023, Dynamic Relation-Attentive Graph Neural Networks for Fraud Detection. 2023 IEEE International Conference on Data Mining Workshops (ICDMW), 2023: 1092–1096.

[4] Xu F, Wang N, Wu H, et al., 2024, Revisiting Graph-Based Fraud Detection in Sight of Heterophily and Spectrum. Proceedings of the AAAI Conference on Artificial Intelligence, 38(8): 9214–9222.

[5] Zhang Y, Ye W, Zhang J, 2017, A Generalized Elastic Net Regularization with Smoothed ℓq Penalty for Sparse Vector Recovery. Computational Optimization and Applications, 2017, 68: 437–454.

[6] Zou H, Hastie T, 2005, Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2): 301–320.

[7] Rafapa J, Konokix A, 2024, Ransomware Detection Using Aggregated Random Forest Technique with Recent Variants. Authorea, 2024: 1–8.

[8] Pathak P, Shrivas AK, 2024, Development of Proposed Model Using Random Forest with Optimization Technique for Classification of Phishing Website. SN Computer Science, 5: 1059.