Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

# Innovative Applications of BP Neural Networks in Financial Fraud Recognition

Zibin Geng [1, *, #], Junchen Liu [1, #], Jiuzhou Wang [2]

[1] School of Information Technology, University of International Business and Economics, Beijing, China, 100029

[2] International business school, University of International Business and Economics, Beijing, China, 100029

* Corresponding Author Email: gzbww2024@163.com

[#]These authors contributed equally.

**Abstract.** In this study, through the use of machine learning models and BP neural network, a large amount of financial data is deeply mined and analysed. The research results show that the constructed model improves the financial fraud recognition accuracy to 0.94, which can achieve the discrimination of financial fraudulent account behaviours at a high accuracy rate, provide strong support for financial institutions, and effectively prevent financial fraud risks. This study aims to solve the problem of high variability and concealment of financial fraud, different from the direction of departure of the more significant results of fraud identification for financial transactions, but focuses on the portrayal and analysis of the characteristics of the financial fraud account, so as to achieve the efficient identification of the source of financial fraud decision-making.

**Keywords:** Financial fraud; XGBoost; BP neural network.

## 1. Introduction

As financial markets expand and innovate, the diversification and sophistication of financial fraud has become increasingly prominent. Globalisation has accelerated the growth of transnational financial fraud cases, challenging the international financial order. Technological advances have provided fraudsters with new modus operandi, and complex financial instruments have been exploited. The problem of information asymmetry has made it possible for fraudsters to take advantage of it. The problem of financial fraud has been exacerbated by factors such as lagging regulation and lack of ethics. Financial fraud has seriously affected the healthy operation of financial markets [1-2].

Previous research has focused on financial fraud detection techniques, machine learning algorithms, data analysis methods and risk assessment and early warning. Data mining reveals laws, machine learning builds models to achieve automated identification, such as anomaly detection, support vector machine and so on. Data analysis methods such as association rule mining are also applied. Risk assessment models predict the probability of fraud occurrence and provide early warning mechanisms [3].

In this study, by using machine learning models and BP neural network, a large amount of financial data is deeply mined and analysed, and BP neural network are applied to the field of financial fraud to construct an effective decision-making system for financial fraud judgement.

## 2. Selection of indicators and data processing

### 2.1. Data sources and processing

In this paper, 700,000 pieces of data were processed using the Kaggle Financial Fraud Accounts dataset. Anomalous data rows and rows with all zeros were removed, and the categorical data was transformed with unique hot coding. Three metrics covering credit card plan type, work status, and

living conditions were processed with inverted numerical correspondences. Indicators that were not quantitatively divided into ratings were processed with solo thermal coding. The data were normalised and covariance matrix was calculated for finding the correlation between features and fraud. PCA dimensionality reduction method was used to downscale the data to retain the main information and reduce the impact of complexity and outliers on the model. These steps help to improve the training effectiveness and predictive ability of the model and make it more adaptable [4-5].

### 2.2. Selection of indicators

Based on previous studies [6-10],the selection of indicators is shown in table 1 below.

**Table 1.** Selection of indicators.

| Indicator fields | Indicator Description |
|---|---|
| is_fraud | A binary indicator that indicates whether a transaction has been flagged as fraudulent. |
| annual_income_quantile | A quartile of annual income that may be used to indicate income level. |
| name_email_match_score | A score for matching name and email address, which may be used to detect false identities. |
| previous_address_months | Number of months lived at previous address, which may be associated with stability. |
| current_address_months | Number of months lived at current address, again may be correlated with stability. |
| age_decade | Age range, which may be used to assess credit risk. |
| days_post_request | Number of days since application, which may be used to assess the urgency of the applicant. |
| initial_transfer_amount | Initial transfer amount, which may be related to credit needs. |
| credit_card_plan | Type of credit card plan, which may affect credit scores. |
| zipcode_applications_4w | Number of applications in the same postcode in the last 4 weeks, may be used to detect gang fraud. |
| hourly_apps_velocity_6h | Application rate within the past 6 hours. |
| hourly_apps_velocity_24h | Application rate within the past 24 hours. |
| hourly_apps_velocity_4w | Application rate in the last 4 weeks. |
| bank_branch_apps_8w | Number of applications from bank branches in the last 8 weeks. |
| unique_emails_dob_4w | Number of unique emails based on date of birth within the last 4 weeks. |
| job_status | Job status, which may be used to assess financial stability. |
| internal_credit_score | Internal credit score. |
| is_free_email | Whether or not you use a free email service, which may be associated with credit risk. |
| living_condition | Living conditions, which may affect credit scores. |
| is_home_phone_valid | Whether home phone number is valid. |
| is_mobile_phone_valid | Whether mobile phone number is valid. |
| previous_bank_account_months | Months in previous bank accounts. |
| has_multiple_cards | Whether multiple credit cards are held. |
| credit_limit_proposal | The amount of credit proposed. |
| is_foreign_application | Whether it is a foreign application. |
| application_channel | Application channel. |
| online_session_duration | Duration of online session. |
| device_operating_system | Device operating system. |
| is_session_persistent | Whether the session is ongoing. |
| unique_device_emails_8w | Number of unique device-based emails in the last 8 weeks. |
| device_fraud_cases | Number of device fraud cases. |
| application_month | Month of application. |

## 3. Modelling and solving

### 3.1. Model Introduction

XGBoost is a powerful integrated machine learning algorithm known for its efficient training speed and excellent generalisation ability. It cleverly combines gradient boosting techniques to efficiently handle complex data structures and provide accurate solutions. XGBoost performs well on large-scale datasets, mining data for underlying patterns and enabling accurate prediction and analysis. For large data sets, XGBoost is able to efficiently process and identify complex non-linear relationships, using various features such as income, residence stability, age, etc. to identify fraudulent behaviours, ensuring excellent performance in the training and generalisation phases [5-6].

The structure of the neural network, the number of neurons, the activation function and other parameters are adjusted for different tasks and data to adapt flexibly. When dealing with complex nonlinear problems, BP neural networks may capture data features more accurately. BP neural networks can use multi-GPUs or distributed computing frameworks to improve training efficiency and are suitable for applications with high real-time requirements. In contrast, XGBoost has limited parallelisation ability and is less efficient under large-scale data. BP neural networks are tolerant to abnormal data and have strong generalisation ability, while XGBoost is sensitive to anomalies and has a weak generalisation ability.BP neural networks have a strong ability to adapt to changes in dynamic environments and can detect new fraud patterns in time to adapt to changes in user behaviour, whereas XGBoost is slower to update and adapt and is difficult to capture such changes.

BPFi model takes BP neural network model as the basic model, BP neural network, i.e., back-propagation neural network (Back-propagation), is a kind of artificial neural network adapted to nonlinear pattern recognition and classification pre-problems. BP network has the characteristics of self-learning, self-adaptation, highly nonlinear and strong generalisation ability, with the ability to approximate the nonlinear relationship with arbitrary accuracy, which is widely used in the Neural networks are widely used. It is through the training of the sample data, and constantly amend the network weights and thresholds [7-8], so that the error function declines along the negative gradient direction, approaching the desired output.

### 3.2. Modelling applications and optimisation

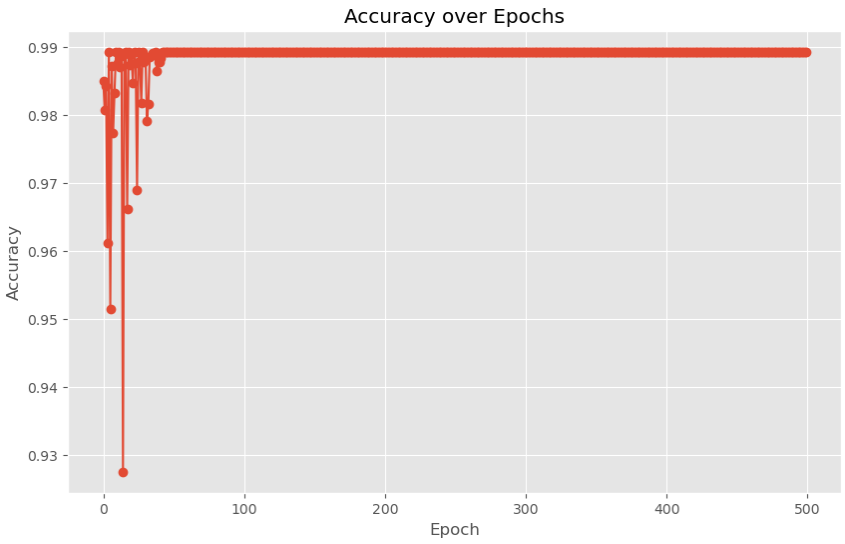


**Figure 1.** Projected results.

The Figure 1 shows that the model has a prediction accuracy of 99%, there is still an overfitting problem and the model needs to be optimised.

Even though financial fraud occurs repeatedly, the proportion of behavioural accounts involved in financial fraud as a percentage of the total is small - from the dataset we chose, which contains

700,000 pieces of data, there are just over 7,000 of these behavioural accounts involved in financial fraud. It is as if there are 700,000 people in a large community, but only 7,000 of them have a particular type of bad behaviour. If we simply assume that everyone is normal, then it seems reasonable in terms of proportions, but in reality it does not accurately reflect the true picture. Considering similar extremes predicting all outcomes as non-fraudulent would also still make for a good fit, but in reality the model predicts poorly. Assuming we go to the extreme of predicting all outcomes as non-fraudulent, it may also result in a better fit in terms of data scaling. However, the actual predictive effect of the model in this case does not truly reflect the actual situation and is unsatisfactory. Therefore, in order to avoid the insensitivity of the model to financial fraud judgement, we need to use the up-sampling method for optimisation to prevent overfitting or false accuracy of the prediction results due to data imbalance [9-10].

Taking credit card theft as an example, credit card theft generally occurs when the cardholder's information is stolen by criminals and then copy the card for consumption or when the credit card is activated for consumption after being fraudulently claimed by others. Once credit card fraud occurs, both the cardholder and the bank will suffer certain economic losses. Therefore, it is particularly important for banks to build financial anti-fraud models through big data. In practice, even though credit card theft is frequent, it is still a small number compared to the total number of credit card transactions. If the model is trained only based on a few cases of theft, it may lead to misjudgement of normal transactions, and predicting all transactions as non-fraud can also get a better fitting result, but this is obviously not in line with the actual situation. Therefore, in order to avoid this situation, an up-sampling method is used to increase the number of samples of stolen transactions, so that the model can better learn the characteristics of fraudulent behaviours and thus improve the accuracy of prediction. At the same time, other technical means, such as deep learning and reinforcement learning, can be combined to further improve the performance of the model.

In addition, the model's insensitivity to financial fraud can easily lead to misjudgment of fraudulent behaviours. As far as the model itself is concerned, the training data of the model cannot completely cover all types of financial fraud, and the variables selected by this group are judged through the anomalous nature of financial accounts as a starting point, which leads to insensitivity to the unconventional and abnormal effects of some emerging or rare fraud patterns. As far as the financial fraud concealment attribute is concerned, financial fraudsters tend to use various means to hide their behaviours so that they are not easily detected, such as falsifying transaction records and exploiting complex transaction structures. As for the external financial environment, there is information asymmetry between financial institutions and fraudsters, which makes the information obtained by the model incompletely accurate, and there may be noise and errors in the data, interfering with the model's judgement. And all these factors will increase the possibility of financial fraud misjudgement, making it an easy force majeure factor.

For these reasons, this group uses up-sampling method to process the data.

The upsampling fit_resample method is a technique used to upsample data. Its basic principle is to increase the number of samples of a few classes by resampling the original data to balance the proportion of samples of different classes in the data set. In financial fraud judgement problems, there may be an imbalance in the data, i.e., there are relatively few fraud samples. By using the up-sampling method, the number of fraud samples can be increased so that the model can better learn fraud patterns and improve the detection of fraud. In this paper, we use the RandomOverSampler class in imblearn to upsample the training set by calling the fit_resample method of the RandomOverSampler object.

Let the original fraud data be $\{x_i\}_{i=1}^{7753}$ ,To perform up-sampling to M points (M>7753), the SMOTE algorithm chosen in this paper is an optimal up-sampling method to change the data distribution of an unbalanced dataset by analysing the minority class samples and synthesising new samples to be added to the dataset by employing linear interpolation to synthesise new samples between the two minority class samples, and the simple expression of linear interpolation is given as:

$$L(x) = \frac{x - x_i}{x_{i+1} - x_i} y_{i+1} + \frac{x_{i+1} - x_i}{x_{i+1} - x_i} y_i \tag{1}$$

In defining financial fraud, judgement is made through the binary variable is_fraud that indicates whether a transaction is fraudulent or not, and the metric is essentially judged by comparing it to 0.5. By upsampling, the number of points in the dataset judged to be financial fraud transactions increased, and the BPFi prediction analysis was re-run to obtain the following prediction accuracy results:Figure 2 is Prediction accuracy.

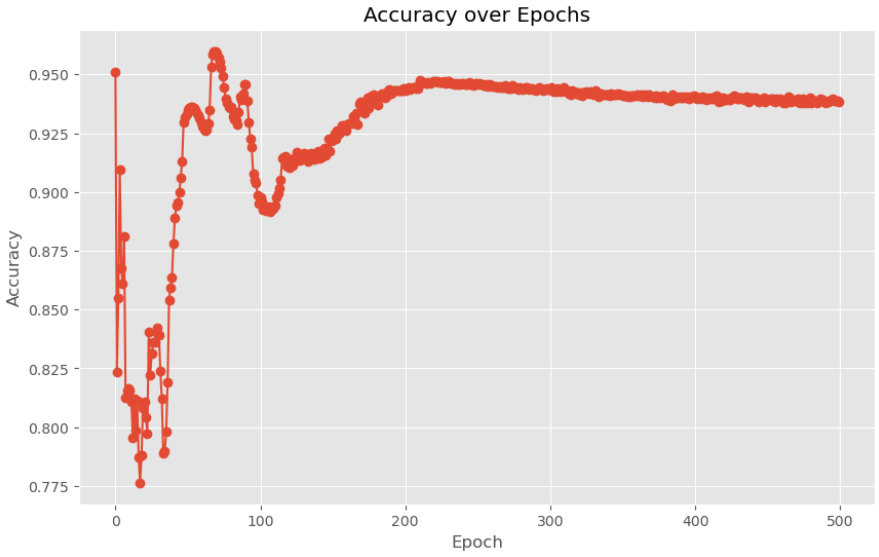


**Figure 2.** Prediction accuracy.

The results show that the up-sampled model is also more reasonable compared to the model with simple dataset partitioning due to the removal of the factor of fitting falsity in extreme cases.

At 500 rounds of model training, the prediction accuracy is close to 94%, when the model tends to be stable. However, the number of model training rounds is close to the stability inflection point. In order to examine the stability of the model, the group continues to train the model, and when the number of training rounds reaches 2000, the prediction accuracy is shown in figure 3.

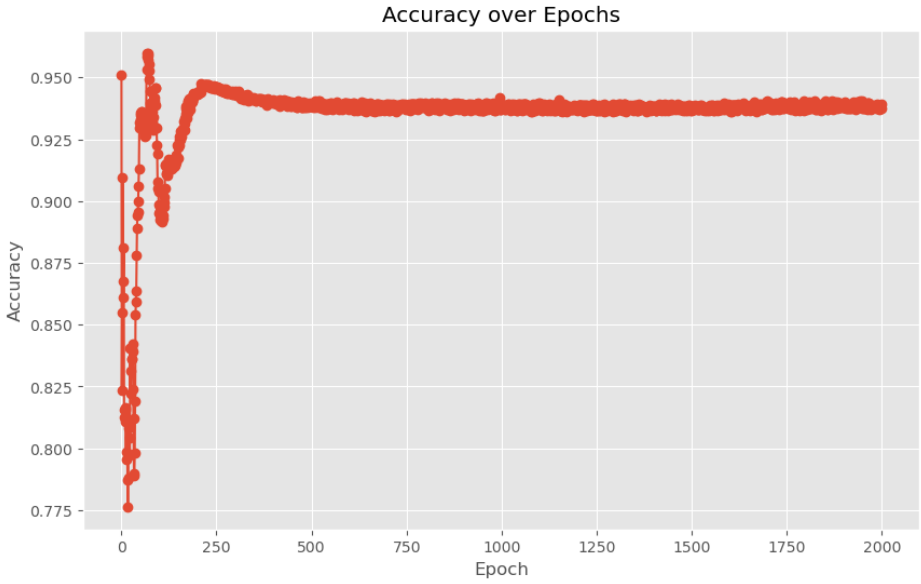


**Figure 3.** Prediction accuracy.

## 4. Conclusions

In this paper, XGBoost machine learning model with integrated function as well as BPFi model are established, the differences and advantages and disadvantages of different models are compared, and finally the better BPFi model is selected to train the financial fraud dataset, and the evaluation model that can provide financial fraud decision-making is obtained. Among them, BPFi successfully improved the prediction and accuracy to 0.94, which is highly exploitable. Using this model, financial accounts with financial fraud risk can be analysed and judged initially.

In addition, in the process of model analysis and optimisation, this paper analyses the unavoidable problems and mechanisms that exist at this stage of financial fraud judgement, providing innovative ideas and optimisation directions for research in different directions from fraudulent behaviour as well as user characteristics.

For the financial fraud judgment decision model, this paper is from the source of fraud, that is, the existence of fraudulent accounts to start judging, this model has the advantage of being able to effectively reduce the problem of lack of timeliness of the model due to the variability of fraudulent behaviour. Because of this, the model is only able to judge whether the subject of the transaction involved is risky by using the account credit score in the face of different types of fraudulent transactions. In the face of new accounts, hidden fraudulent accounts cannot be accurately identified and pinpointed to each transaction itself.

Therefore, the group suggests combining the fraud source identification model established in this paper with the more significant research results of today's academic judgment model of transaction behaviour, to further build both real-time, efficient and accurate, while including multi-dimensional monitoring and evaluation of the transaction of the effective model, to become a powerful weapon to jointly resist financial fraud.

## References


[1] Lv Jun. Research on fraud risk management of fintech platform based on artificial intelligence [D]. Beijing University of Posts and Telecommunications, 2023.

[2] P.C. Ning. Research on financial fraud detection method based on graph neural network [D]. Donghua University, 2023.

[3] Xinzhou Ruan. Research on financial service transaction data analysis and mining technology [D]. University of Electronic Science and Technology, 2023.

[4] Xia, Pingfan. Research on intelligent risk prediction method for digital financial fraud [D]. Hefei University of Technology, 2022.

[5] Liu Li-Ming, Meng Fang. Analysis of the use of neural network model in bank internet financial anti-fraud [J]. Digital Communication World, 2020, (12):100-101.

[6] Shiqi Jiang. Research on credit card anti-fraud model [D]. Chongqing University, 2021.

[7] Ding Shuangs. Research on the identification of internet financial fraud behaviour based on big data [D]. Capital University of Economics and Business, 2017.

[8] Zhang Yuanyuan. Fraud identification based on feature engineering and mean uncertain logistic regression in advertising and banking [D]. Shandong University, 2024.

[9] Gui Qin. Research on Identification and Prevention and Control of Fraudulent Users in Internet Financial Platform [D]. Harbin Institute of Technology, 2021.

[10] Liu Xiaofei. Research and application of knowledge graph in the field of anti-fraud in financial system [D]. University of International Business and Economics, 2024.

[11]