Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 13
AMMSAC 2026

Big Data-driven Financial Risk Prevention and Control Model for Commercial Banks

Bo Wu

University of Ottawa, Ottawa, Canada

Abstract. With the rapid digitalisation of financial markets, many new kinds of complex and interconnected risks for commercial banks have emerged that cannot be controlled by the old ways of risk management. Big data technology can help to better solve the problem of risk prevention and control in finance by collecting a large amount of data, analysing it, and issuing timely early warnings. A complete big data-driven financial risk prevention and control model for commercial banks has been put forward in this paper, and it is composed of many analysis modules to address problems such as credit risk, market risk, liquidity risk, operating risk, fraud detection, etc. A group of gradient-boosted decision trees is used for credit risk, LSTM is employed to predict market and liquidity risk, GNN is applied to detect fraud and operational risk, and all of them are combined to obtain a composite risk score and a five-tier early warning classification. Based on the empirical test, the integrated model has an AUC-ROC of 0.963 and an F1-score of 0.924, and it has performed better than all traditional statistics and single-module machine learning methods for the risk categories in question. Based on SHAP feature importance analysis, the main causes of credit risk are payment delinquency history and debt-to-income ratio. There are also problems with the application of data governance, model explainability and regulations in this study, and some directions for future research have been put forward, such as federated learning and causal inference.

Keywords: Big data analytics; Commercial banks; Financial risk management; Machine learning; Credit risk prediction; Early warning systems; Ensemble learning.

1. Introduction

The commercial banking system in China is undergoing unprecedented transformation at a faster rate than ever before, and all kinds of data that have been collected from lending businesses and other parts of the bank have been in the form of various formats, such as traditional data and large-scale data. The problem and speed of financial risks have increased simultaneously, and the old system of analysis can no longer keep up. [1, 2]

Although research and regulation on the risk management of banks have been gradually increasing in recent years, there is still no single model or analytical system that can cover all links of the risk. Most of the current research has focused on only one type of risk in isolation and has not yet achieved comprehensive risk governance [6, 7]. To solve this problem, the following three main contributions will be made in this paper: First, a hierarchical big data-driven framework covering credit, market, liquidity and operating risks will be constructed; Second, a composite risk scoring method that dynamically weights signals from different modules will be put forward; Finally, based on five performance indicators, the integrated model will be empirically compared with existing statistics and machine learning benchmarks. Sections 2-5 are about the data infrastructure, model architecture, empirical analysis and implementation problems, respectively, and Section 6 presents the summary of the paper.

2. Big Data Infrastructure for Banking Risk Management

2.1. Data Sources and Collection Framework

The data sources in the proposed framework fall into four categories: internal bank records (e.g., loan origination and usage data, account statements, and payment histories), external market data (e.g., equity, bond, and foreign exchange markets), credit bureau information, and alternative data (e.g., e-commerce behavioral signals and supply chain network indicators). Data from the equity market, bond market and foreign exchange platform are external indicators of macroeconomic environment that affect market risk. Borrower-specific data are primarily drawn from credit bureaus and internal loan performance records. Recently, in addition to the previous reasons for determining small and medium-sized enterprise loans, e-commerce behaviour signals and supply chain network indicators have also begun to be included in the assessment process [8-9]. Table 1 summarizes the key indicators, primary data sources, and update frequencies corresponding to each major risk category within the proposed monitoring framework.

Table 1. Key Big-Data Indicators for Commercial Bank Financial Risk Monitoring

Risk Category	Representative Indicators	Primary Data Sources	Frequency
Credit Risk	Debt-to-income ratio, loan-to-value ratio, debt service coverage ratio, payment delinquency rate	Internal loan systems, credit bureaus	Daily/Monthly
Market Risk	Interest rate sensitivity gap, FX net open position, equity VaR, portfolio duration	Trading systems, market data feeds	Real-time/Daily
Liquidity Risk	LCR, NSFR, intraday cash flow gap, deposit concentration index	Treasury systems, central bank reporting	Real-time/Daily
Operational Risk	Processing error rate, system downtime frequency, cyber incident severity index	IT monitoring, internal audit logs	Real-time/Weekly
Fraud Risk	Transaction velocity anomaly, device fingerprint deviation, behavioral pattern distance	Anti-fraud systems, digital channel logs	Real-time

2.2. Data Preprocessing and Feature Engineering

As raw financial data is generally noisy and has an uneven distribution, a four-stage preprocessing pipeline will be employed to address this problem; that is to say, missing value imputation will be performed by iterative regression and K-nearest neighbours, outlier detection will use the isolation forest algorithm, categorical encoding for high-cardinality features will be achieved by target encoding and embedding layers, and the temporal alignment of multi-frequency streams will be done to form a single analysis window. Feature engineering is to transform the raw input data into economically meaningful indicators, that is, static variables are the financial indicators of borrowers, dynamic variables are rolling default rates and behaviour patterns, and cross-sectional peer comparison indicators have shown good predictive power in credit risk research [5-6]. There are about 380 attributes in the feature space for each category, and after mutual information filtering, only 80-120 high-information features per module were retained.

Additional feature terms are introduced to capture changes in borrowers' debt-servicing capacity driven by macroeconomic fluctuations. Weight-of-evidence encoding is used to transform the categories of geographic area, industry sector, employment type, etc., into numbers that are easy for gradient boosting to recognise and interpret. Temporal features are in the form of first- and second-order difference operators to capture the momentum and direction of change in rolling financial ratios; that is to say, a gradual decline in borrowers that is not obvious in the static figures can be identified [6, 10]. For feature selection, mutual information scoring and recursive feature elimination with cross-validation were used to select a small number of features that are good at prediction and stable from the initial set.

3. Model Architecture for Financial Risk Prevention and Control

3.1. Integrated Risk Assessment Framework

The model adopts a modular hierarchical architecture processing risk signals from four specialized analytical engines—credit, market, liquidity, and fraud or operational risk—and aggregating their outputs through a composite risk scoring layer. Each module produces a normalized score in the unit interval. The composite score at time t is given by Equation (1), where $s_k(X_t)$ is the output of module k, ϕ_k is a calibration function ensuring probabilistic consistency, and weights w_k are updated quarterly via Bayesian optimization against observed risk materializations. [10] An alert threshold layer classifies entities into five warning levels, with each level corresponding to a predefined action protocol and escalation timeline calibrated to the institution's risk appetite and regulatory capital adequacy requirements.

$$R_t = \sum_{k=1}^{K} w_k \cdot \phi_k(s_k(X_t)) \quad (1)$$

Where R_t is the composite risk score at time t; $s_k(X_t)$ is the normalized output of the k-th risk module; ϕ_k is a module-specific calibration function; w_k is the adaptive module weight; and $K=4$ is the total number of risk modules.

The hierarchical Structure of this way will allow the risk control measures to be in line with the size of the risk and not disturb the regular operation of the company during normal times. Figure 1 is the whole data flow from ingestion to the four risk modules and then to the composite score and five-level early warning output.

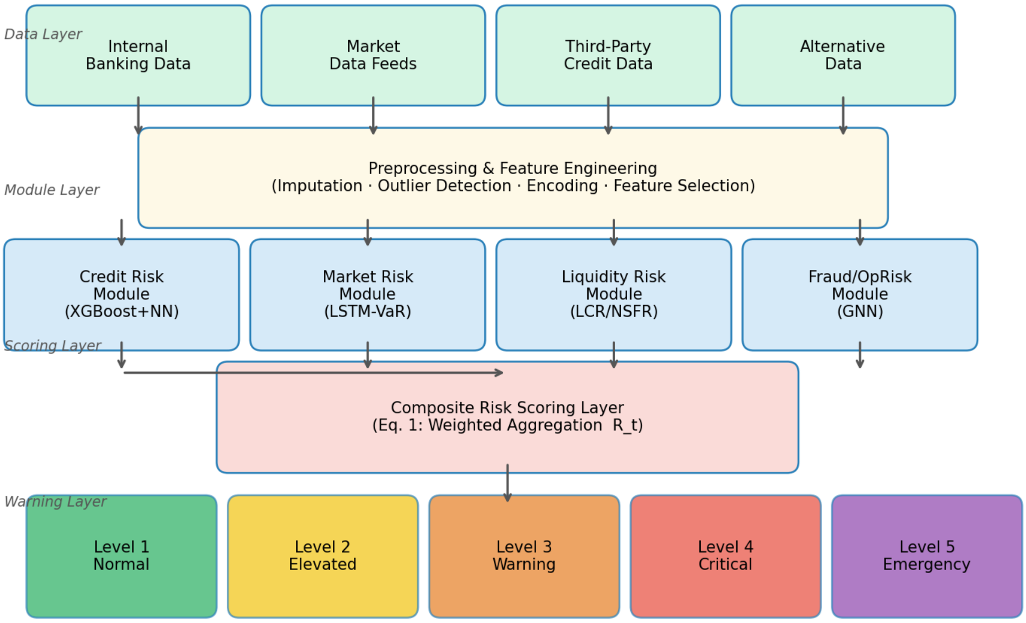

Figure 1. System Architecture of the Big Data-driven Financial Risk Prevention and Control Framework

3.2. Credit Risk Prediction Module

The credit risk module employs an ensemble of gradient-boosted decision trees combined with a neural network meta-learner to predict default probability. The XGBoost training objective in Equation (2) minimizes regularized loss, where l is binary cross-entropy between the observed default indicator and predicted probability, T_m is the number of leaf nodes in the m-th tree, and γ, λ are regularization parameters controlling model complexity. [6, 11] This formulation prevents overfitting on imbalanced datasets where defaulters represent fewer than five percent of the population.

$$\mathrm{Obj}=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i\right)+\sum_{m=1}^{M}\left(\gamma T_m+\frac{1}{2}\lambda\|w_m\|^2\right) \quad (2)$$

Where l is binary cross-entropy loss; y_i is the observed default indicator (0 or 1); $\hat{y}_i$ is the predicted default probability; T_m is the leaf count of the m-th tree; M is the total number of trees; γ is the minimum gain threshold for a split; and λ is the L2 regularization coefficient.

A rolling 36-month window is employed for training the model, and according to the champion-vs.-challenger governance protocol, it is updated monthly; only then will it be released to production after an improvement in AUC has been achieved and before performance drops due to macroeconomic changes during the recalibration cycle [11].

3.3. Market Risk and Liquidity Risk Assessment

A three-layer LSTM network is employed to process the price sequence data from the past 60 days and predict the 10-day-ahead Value-at-Risk; it can model the non-linear time dependence of traditional parametric VaR models. Every week, an extended window of daily returns is taken to retrain the LSTM, and a Monte Carlo simulation layer is added to generate scenario-conditional loss distributions according to the interest rate, credit spread and foreign exchange stress scenarios listed in the internal risk policy [10].

3.4. Fraud Detection and Operational Risk Control

A Graph Neural Network (GNN) will be used for the fraud detection part, and a heterogeneous graph will have nodes for accounts, devices, IP addresses and merchants, and transactions will be connected as edges. Through multi-head attention on neighborhood features, it can be determined that there are coordinated fraud rings and money laundering networks not obvious to a single-transaction classifier [8], and for non-fraud operational risks, such as system failure, processing error and third-party failure, an unsupervised autoencoder trained on normal operational telemetry will detect anomalies when the reconstruction error exceeds a calibrated threshold; both alert streams will be escalated according to a severity-based protocol that maps anomaly types to predefined response tiers and notification workflows.

4. Model Validation and Performance Analysis

4.1. Experimental Setup

According to the typical parameters of commercial banks and released regulatory references [3-4], synthetic data have been created to generate a credit risk dataset of 250,000 loan records with a default rate of 4.2%. The market risk backtesting period is 2014-2023, and it includes the volatility event in 2020 due to COVID-19. The fraud detection dataset has 1.2 million transactions and a fraud rate of 0.38 per cent, and it still has the class imbalance and adversarial distribution shift characteristics of the live banking environment. All the models will be tested on the 20% stratified hold-out test set and five-fold cross-validation to verify the stability of the results.

4.2. Performance Analysis

In terms of accuracy, precision, recall, F1-score and AUC-ROC, the proposed integrated model has achieved the highest performance across all five evaluation metrics. Its AUC-ROC of 0.963 is statistically higher than that of XGBoost (0.942) and logistic regression (0.879), and this difference has been confirmed at the 99 per cent confidence level by the DeLong test. The recall is 0.921; that is to say, more than 92% of the actual risk events have been properly identified, and the precision is 0.928, which reduces the number of false positives for operating staff [7-8].

The daily VaR exception rate of the LSTM market risk module over the past 10 years has been 0.89%, which is below the 1% limit in Basel III; there have been no more than four consecutive exceptions during the 2020 stress event, and the parametric baseline had an exception rate of 1.43% and seven consecutive exceptions. The AUC of the GNN fraud module is 0.957, and it can better detect coordinated multi-account fraud rings by taking into account the relational graph context. Cross-validation of the five stratified folds shows that there is a small increase in the standard AUC of the model, and all are less than 0.003.

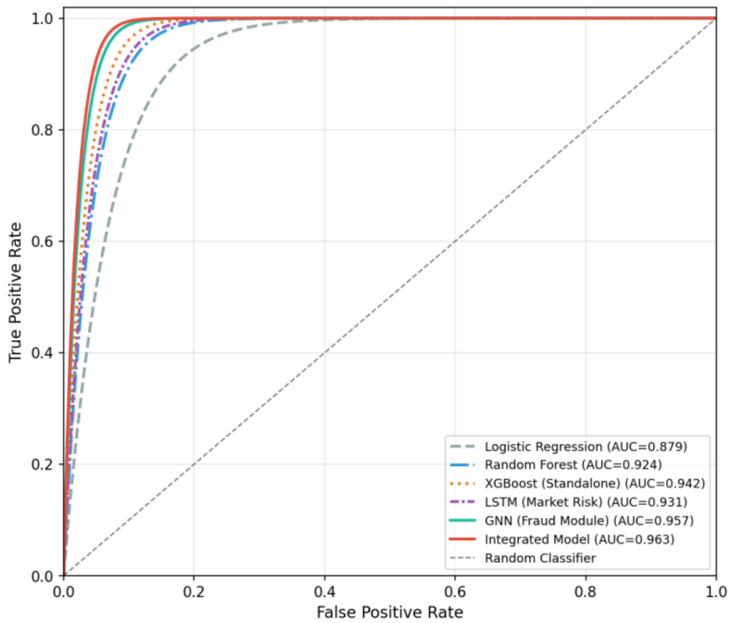

Figure 2. ROC Curve Comparison of Risk Prediction Models

4.3. Feature Importance and Model Interpretability

Regulations and the internal control of enterprises need to be understandable. SHAP values are used in the credit module to decompose the reasons for a certain score additively and help loan officers explain the probability of default for a particular loan. Based on SHAP importance analysis, the top six contributing features are ranked as follows: the first 12-month delinquency indicator (mean |SHAP| = 0.187) and debt-to-income ratio (0.143) have the largest values, followed by account vintage (0.119), recent hard credit inquiries (0.098), macroeconomic unemployment rate (0.082) and composite leading indicator (0.061) [10, 12].

The model reduces the weight of inquiry counts for borrowers with long credit histories and, thus, learns that the context is not a simple function of one variable. SHAP waterfall plots of individual loan applications can help loan officers understand the main reasons for the risk in gradient boosting without being experts in gradient boosting; thus, both the internal governance requirements and the new regulations on the transparency of algorithmic decision-making in automated credit assessment have been met [10, 11].

5. Challenges and Future Directions

5.1. Data Governance and Privacy Constraints

Data governance is also one of the problems of deployment. The collection of alternative and third-party data streams has caused problems of privacy, consent and cross-border transfer in all but a few places. Due to the strictness of the regulations in the European Union's General Data Protection Regulation and other systems, as well as China's Personal Information Protection Law, which prohibit the use of data for automated credit decision-making, a strong lineage and consent management infrastructure need to be established; this may limit the functions available to big data risk models [9, 12]. Given that the scope of these regulations overlaps, it is still a problem how to ensure compliance without restricting the scope of data, so careful design of data sources and retention policies in the system architecture must be made to address this issue.

5.2. Model Risk and Regulatory Compliance

Due to the lack of transparency in machine learning, it is not easy to conduct model validation and auditing in accordance with regulations; thus, banks need to be able to explain the reasons for individual credit decisions and show that there is no systematic discrimination against protected classes of borrowers [3, 10]. The model risk management principles of the US Federal Reserve's SR 11-7 guidelines were based on traditional statistical models and cannot be directly applied to complex deep learning and ensemble architectures that are not algebraically expressed; therefore, other validation methods such as empirical stress testing, adversarial scenario analysis and continuous performance monitoring must be used.

5.3. Future Research Directions

The three directions of research should be given priority. First, combine the big language models with the structured financial data of earnings call transcripts and regulatory reports to extract all kinds of risk signals in the corporate credit and systemic risk from them early on. Second, a cooperative model can be formed among all bank branches to carry out federated learning without moving their actual data; thus, problems with management restrictions and a small pool of high-quality training data from various banks can be avoided [9]. Third, in the field of financial risk assessment, causal inference methods have begun to move away from older correlational machine learning models because they cannot clearly show the reasons for the underlying cause and may be more reliable in out-of-sample situations after structural changes or alterations in the history of correlations [11, 12].

6. Conclusion

This paper proposes a general big data-driven financial risk prevention and control model for commercial banks and then validates it; the combined parts will be gradient-boosted ensemble methods for credit risk, LSTM networks for market and liquidity risk, and graph neural networks for fraud and operational risk. According to the results of the empirical evaluation, the integrated model has an AUC-ROC of 0.963 and an F1-score of 0.924; it is higher than that of single-module and traditional statistical indicators in all risk dimensions. The five levels of the early warning system are based on the risk tolerance of the institution and will trigger various emergency plans at different stages.

Therefore, it is necessary to urgently improve the supervisory regulations so that they can keep up with the changes of AI-driven risk models and address problems such as algorithmic opacity and systemic contagion promptly. Commercial banks will build their own big data-driven risk intelligence platforms to reduce loan losses and make better use of regulatory capital by strengthening the early warning capacity for the risk of systemic risk concentration before it leads to severe financial losses, etc. The next round of the framework will add new modules for federated learning and causal inference to solve the problems of governance and cross-jurisdictional portability in the current one.

References

[1] Financial Stability Board. (2023). Artificial intelligence and machine learning in financial services. https://www.fsb.org/wp-content/uploads/P130923.pdf

[2] Bank for International Settlements. (2024). Annual economic report 2024. https://www.bis.org/publ/arpdf/ar2024e.htm

[3] Basel Committee on Banking Supervision. (2023). Principles for the sound management of operational risk. https://www.bis.org/bcbs/publ/d582.pdf

[4] International Monetary Fund. (2024). Global financial stability report: The last mile—financial vulnerabilities and risks (April). https://www.imf.org/en/Publications/GFSR/Issues/2024/04/16/global-financial-stability-report-april-2024

[5] International Monetary Fund. (2024). Global financial stability report (October). https://www.imf.org/en/Publications/GFSR/Issues/2024/10/07/global-financial-stability-report-october-2024

[6] European Banking Authority. (2024). Report on machine learning for IRB models. https://www.eba.europa.eu/sites/default/files/2024-06/EBA_Report_ML_IRB_Models_EN.pdf

[7] Bank for International Settlements. (2025). Annual economic report 2025: Navigating the new financial landscape. https://www.bis.org/publ/arpdf/ar2025e.htm

[8] Financial Stability Board. (2024). FinTech and market structure in financial services: Market developments and financial stability implications. https://www.fsb.org/wp-content/uploads/P200524.pdf

[9] International Monetary Fund. (2025). Global financial stability report (April). https://www.imf.org/en/Publications/GFSR/Issues/2025/04/15/global-financial-stability-report-april-2025

[10] Basel Committee on Banking Supervision. (2024). Report on principles for operational resilience. https://www.bis.org/bcbs/publ/d604.pdf

[11] Financial Stability Board. (2025). 2024 annual report. https://www.fsb.org/wp-content/uploads/2025/04/FSB-2024-Annual-Report.pdf

[12] World Bank Group. (2024). Global financial development report 2024: Financial resilience and digital innovation. https://www.worldbank.org/en/publication/global-financial-development-report-2024