ASSESSING MACHINE LEARNING STRATEGY APPROACHES FOR EARLY WARNING OF LIQUIDITY RISK

Thi Thu Ha Do [*], Thu Thuy Nguyen [**], Nguyen Phuc Duc Ho [*], Thi Van Khanh Truong [*], Minh Phuong Nguyen [*], Van Hieu Pham [***], Thi Hanh Duyen Nguyen [****]

* Faculty of Banking, Banking Academy of Vietnam, Hanoi, Vietnam
** *Corresponding author*, Thuongmai University, Hanoi, Vietnam
Contact details: Thuongmai University, 79 Ho Tung Mau, Tu Liem, Hanoi, Vietnam
*** Hanoi University of Business and Technology, Hanoi, Vietnam
**** Vinh University, Vinh, Vietnam

How to cite this paper: Do, T. T. H., Nguyen, T. T., Ho, N. P. D, Truong, T. V. K., Nguyen, M. P., Pham, V. H., & Nguyen, T. H. D. (2026). Assessing machine learning strategy approaches for early warning of liquidity risk. *Corporate and Business Strategy Review, 7*(2), 96–105.
https://doi.org/10.22495/cbsrv7i2art9

ISSN Online: 2708-4965
ISSN Print: 2708-9924

Received: 18.09.2025
Revised: 16.12.2025; 02.01.2026; 03.03.2026
Accepted: 10.03.2026

JEL Classification: C45, C53, G01, G21
DOI: 10.22495/cbsrv7i2art9

Abstract

Artificial intelligence (AI) and machine learning (ML) have been increasingly adopted in the banking sector due to their ability to analyze large-scale datasets, process complex variables, and uncover hidden patterns, especially in the context of liquidity risk, posing a significant challenge for commercial banks. This study contributes to the field by conducting a comprehensive evaluation of several widely used early warning models, such as least absolute shrinkage and selection operator (LASSO) regression, random forest (RF), and extreme gradient boosting (XGBoost), to identify the most suitable approach for forecasting liquidity risk in Vietnamese commercial banks (VCBs) based on VCBs data over the period of 2014–2023. By pinpointing key indicators associated with liquidity crises, these models can assist banks and regulatory authorities in implementing timely preventive measures and enhancing risk management strategies. As a result, the RF model outperforms other methods in identifying possible liquidity crises, according to the empirical results, with an accuracy rate of 99.8 percent. These findings provide bank managers and policymakers with a powerful tool for timely preventive measures, thereby enhancing the resilience and stability of the financial system.

Keywords: Banking Crisis, Early Forecasting Models, Liquidity Risk, Machine Learning

Authors' individual contribution: Conceptualization — T.T.H.D.; Methodology — N.P.D.H. and T.V.K.T.; Software — N.P.D.H.; Validation — T.T.N. and V.H.P.; Formal Analysis — N.P.D.H. and T.V.K.T.; Investigation — T.T.D.N.; Resources — T.T.D.N.; Data Curation — V.H.P.; Writing — Original Draft — T.T.H.D. and T.V.K.T.; Writing — Review & Editing — T.T.N. and M.P.N.; Visualization — N.P.D.H.; Supervision — M.P.N.; Project Administration — T.T.H.D.

Declaration of conflicting interests: The Authors declare that there is no conflict of interest.

Acknowledgements: The Authors gratefully acknowledge the financial support from the Banking Academy of Vietnam under Resolution No. 34/NQ-HDHV.

1. INTRODUCTION

In the field of finance, "liquidity" is a term that carries different meanings depending on the context. From an asset perspective, liquidity reflects the ease with which an asset can be converted into cash, or vice versa. An asset is considered highly liquid if it meets conditions such as the availability of trading volume, the existence of a market, the ability to transact in a short period of time, and a price that reflects its true value (Tavana et al., 2018). Rose (2004) argues that "an asset is regarded as highly liquid when it simultaneously meets two conditions: the existence of a trading market for conversion into cash, and the stability of its value without significant fluctuations due to transaction volume or timing" (p. 447). Therefore, the level of liquidity of an asset is usually assessed based on the time required and the costs incurred to complete the conversion into cash. Assets that can be quickly converted into cash at low cost are considered highly liquid. From a corporate perspective, liquidity refers to the total amount of cash and cash equivalents available. However, when applied to banking management, liquidity implies the bank's ability to promptly and adequately meet financial obligations arising in its operations, including deposit withdrawals, loan disbursements, settlements, and other financial transactions. According to the Basel Committee on Banking Supervision (2008), "Liquidity is the ability of a bank to fund increases in assets and meet obligations as they come due, without incurring unacceptable losses" (p. 1). Since liquidity is realized through cash, it only concerns cash flows. Failure to meet payment obligations leads to illiquidity. Liquidity is therefore not expressed by a specific figure or ratio, but rather reflects the extent to which a bank can meet its payment commitments. In contrast, "illiquidity" occurs when a bank is unable to fulfill its financial obligations at maturity. Under this understanding, liquidity is qualitative in nature, reflecting the financial strength and resilience of a bank in responding to payment demands.

Despite the critical importance of liquidity, traditional methods of risk management, which often rely on static financial ratios or standard statistical models, have shown limitations in capturing the complex, non-linear nature of modern financial data (Altman, 1968; Barongo & Mbelwa, 2024). Previous studies have largely focused on developed markets that may fail to predict liquidity crises accurately in volatile emerging markets like Vietnam (Nguyen, 2019). This represents a significant literature gap as the Vietnamese banking sector undergoes rapid digital transformation and integration into global markets, creating an urgent need for more advanced, predictive early warning systems.

To address this gap, this study applies the Diamond-Dybvig model and Asymmetric Information Theory as the theoretical framework to explain the underlying causes of liquidity hoarding and systemic risk (Diamond & Dybvig, 1983; Brunnermeier & Pedersen, 2009). The primary research aim of this paper is to evaluate and identify the most effective machine learning (ML) approach for forecasting liquidity risk. This investigation specifically addresses two central research questions:

RQ1: Which ML model — least absolute shrinkage and selection operator (LASSO) regression, random forest (RF), or XGBoost — exhibits the greatest predictive accuracy in assessing liquidity risk within Vietnamese commercial banks?

RQ2: What are the critical financial indicators that serve as precursors to a liquidity crisis?

Concerning the methodological approach, this research leverages a comprehensive dataset encompassing Vietnamese commercial banks (VCBs), spanning the years 2014 to 2023. We implement and subsequently compare three sophisticated ML algorithms: LASSO, RF, and XGBoost. The principal outcome of this study reveals that the RF model surpasses the performance of the other methodologies, attaining an outstanding accuracy rate of 99.8%. This study contributes to the existing body of literature by furnishing empirical support for the superiority of ML techniques over conventional approaches within the context of an emerging market, while also offering actionable policy recommendations for the improvement of risk management frameworks.

This paper's structure is as follows. Section 2 reviews the relevant literature and theoretical framework. Section 3 explains the context of commercial banks in Vietnam, and Section 4 presents methods, including how data was collected and how the models were set up. Section 5 presents the empirical results and discusses the findings in relation to existing research. Finally, Section 6 concludes the paper by discussing its implications and suggesting future research directions.

2. LITERATURE REVIEW

2.1. The concept of liquidity risk in banking

The Basel Committee on Banking Supervision (2008) defines liquidity risk as the risk that a financial institution is unable to generate sufficient funding to meet obligations as they fall due without adversely affecting its daily operations or financial condition. From a bank's perspective, liquidity reflects the capacity to promptly and adequately meet financial obligations such as deposit withdrawals, lending, settlements, and other transactions. Prolonged illiquidity leads to liquidity risk. Bonfim and Kim (2012) note that the complex financial intermediation role of banks inherently gives rise to liquidity risk. Banks employ their limited resources to extend loans to businesses and consumers to finance liquidity needs for investment and consumption. Moreover, most of these resources are liabilities, primarily deposits. In pursuit of profitability, banks typically engage in maturity transformation by using short-term liabilities (e.g., short-term deposits) to fund medium- and long-term loans, resulting in maturity mismatches and liquidity risk (Diamond & Dybvig, 1983). To mitigate the imbalance between asset and liability maturities — which undermines liquidity — banks may engage in liquidity risk management through designing an appropriate balance sheet structure, including maintaining a buffer of liquid assets. However, Bonfim and Kim (2012) emphasize that "holding highly liquid assets such as cash, short-term securities, or government bonds entails sacrificing potential

profitability, since these assets generally yield lower returns. The trade-off between liquidity safety and profitability makes maintaining adequate liquidity in banking practice a significant challenge" (p. 4).

Basel Committee on Banking Supervision (2008) further states that liquidity risk arises when a bank is unable to increase funding in assets or liabilities at the lowest cost. Brunnermeier and Pedersen (2009) stress that inadequate liquidity risk management inevitably exposes banks to liquidity shocks, forcing them to frequently liquidate assets and reduce lending to the economy. Liquidity risk at the level of individual banks, as well as systemic liquidity risk for the entire banking system, received little policy and managerial attention until the global financial crisis of 2007–2009. Thus, measuring and issuing warnings on potential systemic liquidity risk across commercial banks is critically important.

2.2. Early warning prediction models using artificial intelligence

Machine learning is increasingly being applied in various fields of finance, including liquidity risk prediction in commercial banks. Although ML is not the only tool for forecasting liquidity risk, ML models have started to demonstrate their superior value, particularly in handling large amounts of complex and unstructured data. One of the major challenges when applying ML in liquidity forecasting is the requirement for data quality. ML models demand a large and accurate dataset to achieve optimal performance. If the input data is incomplete or biased, the models may produce misleading results, thereby reducing forecasting reliability. Ekinci and Sen (2024) emphasize that data quality is a decisive factor in ensuring that ML models can deliver accurate and reliable predictions.

In supervised learning, models such as support vector machines (SVM), RF, and artificial neural networks (ANN) are powerful tools used to forecast financial events, including liquidity risk. These models learn from labeled data, meaning data with clear outcomes, to determine the relationship between independent variables and the probability of events occurring. For example, in liquidity prediction, SVM classifies banks into high- and low-liquidity-risk groups by finding an optimal hyperplane to separate the data. According to Ekinci and Sen (2024), SVM not only helps with classification but also optimizes the margin between classes, thus providing accurate predictions in unstable financial environments. A strong feature of SVM is its ability to handle classification problems with overlapping or ambiguous data, which traditional models often struggle with.

RF is another supervised ML model capable of handling large and complex datasets. This model builds an ensemble of decision trees, where each tree produces an individual prediction. The final outcome is the aggregate of all trees in the forest. This method helps reduce errors and improve forecasting accuracy, especially when working with complex financial data. Ekinci and Sen (2024) highlight that RF enhances liquidity risk prediction accuracy by combining multiple decision trees, thereby reducing the impact of noisy or unstable data.

ANN, with their complex structure consisting of hidden layers, can learn from nonlinear and complex financial data. ANNs are able to automatically optimize weights to detect hidden relationships among financial factors. According to Guerra et al. (2022), neural networks help model nonlinear relationships in financial indicators, which traditional models cannot achieve. ANNs can provide more accurate predictions of financial events, including liquidity risk, when input data is complex or unstructured. In addition to supervised learning, unsupervised learning also plays an important role in liquidity risk analysis. Unsupervised ML models, such as k-means clustering, help group data without prior knowledge of outcomes. Guerra et al. (2022) argue that k-means clustering can classify banks with similar liquidity risk, thereby revealing hidden patterns in data without requiring labels. This model can identify banks with high future liquidity risk even without pre-labeled group information.

One of the greatest advantages of ML models is their ability to learn from data without rigid assumptions about the relationships between financial variables. ML models can automatically improve and optimize predictions as new data becomes available, helping banks adjust liquidity management strategies flexibly and promptly. According to Drudi and Nobili (2021), ML models enable more accurate forecasting and can be applied in situations where financial data changes rapidly. ML can analyze complex datasets and identify major factors affecting liquidity that traditional models cannot capture. Recent literature also emphasizes that, besides liquidity risk, banking stability is increasingly influenced by diverse factors. For instance, Kou et al. (2021) highlighted the complexity of systemic risk analysis in modern financial sectors, while Barongo and Mbelwa (2024) demonstrated the efficacy of ML in specifically detecting liquidity anomalies. In the context of Vietnam, banks are also navigating challenges related to sustainable development awareness (Nguyen et al., 2024) and the adoption of international reporting standards (Nguyen et al., 2025).

3. THE CONTEXT OF COMMERCIAL BANKS IN VIETNAM

This study adopts an associative quantitative This section provides an overview of the banking sector in Vietnam with particular attention paid to business performance, the state of liquidity risk at the moment, and the early warning systems in place. Comprehending these contextual elements is essential to assessing the suitability and efficacy of the ML models suggested in this research.

3.1. Business performance of Vietnamese commercial banks

From 2019 to 2024, Vietnam's commercial banking system experienced significant growth and restructuring. Total operating income of commercial banks increased markedly, from approximately VND 400 trillion in 2019 to VND 600 trillion in 2023. Despite some fluctuations in the first quarter of 2024, the overall trend indicates stability and improved business performance. This growth was largely driven by banks' efforts to expand their operations, improve financial services, and leverage government support policies such as monetary

easing, lowered benchmark interest rates, and credit programs for small and medium-sized enterprises. According to Vietcombank Securities (VCBS, 2023), the average interest rate for term deposits decreased by 2% to 2.9% compared to the end of 2022, enhancing credit access and stimulating investment demand. In addition, profitability indicators of the banking system improved significantly between 2018 and 2021. Return on assets (ROA) and return on equity (ROE) showed steady upward trends, reflecting better asset and capital management. However, from 2022 onwards, these indicators began to decline, signaling potential underlying challenges, including increasing liquidity risk.

3.2. The reality of liquidity risk in Vietnamese commercial banks

Between 2017 and 2024, liquidity risk became a prominent concern for VCBs. During the 2017–2020 period, the interbank interest rate — a key indicator of liquidity pressure — fluctuated between 2% and 6%. In 2019, it spiked above 6%, indicating heightened end-of-year liquidity stress. However, in 2020, the COVID-19 pandemic drove rates down as the State Bank of Vietnam (SBV) adopted an accommodative monetary policy to inject liquidity into the system. From 2022 onward, liquidity pressure significantly increased amid global uncertainties, geopolitical tensions, and rising funding costs. By November 2024, SBV had injected approximately VND 110 trillion into the market via open market operations (OMO), underscoring the intensity of liquidity stress. Additionally, the system-wide loan to deposit ratio (LDR) remained high at 106%, surpassing the recommended safe threshold of 85-90%. Smaller and mid-sized banks such as BVB, ABBank, NamABank, and MSB were compelled to raise deposit interest rates and rely more heavily on short-term funding and interbank borrowing. This not only increased funding costs but also exposed banks to short-term liquidity shocks, raising the risk of mass withdrawals or funding imbalances.

3.3. The state of early warning models for liquidity risk in Vietnamese commercial banks

In the current financial landscape, VCBs have adopted a number of approaches to forecast and manage liquidity risk. These approaches are broadly categorized into asset-based, liability-based, and balanced liquidity management models. Under the asset-based approach, banks hold a substantial proportion of liquid assets, such as government securities or short-term interbank placements, which can be quickly converted into cash to meet urgent liquidity needs. This method is particularly important in emerging markets such as Vietnam, where liquidity shocks may arise suddenly due to changes in monetary policy or fluctuations in capital inflows. By contrast, the liability-based approach places emphasis on the mobilization of new funding from external sources, such as interbank borrowing, issuing negotiable instruments, or attracting wholesale deposits. Finally, the balanced approach integrates both strategies, seeking to provide a more comprehensive response to liquidity demands by simultaneously ensuring short-term stability and enhancing long-term capital efficiency.

The application of these models has led to several notable achievements in the Vietnamese banking sector. Regulatory reforms initiated by the SBV have played a crucial role, especially through Circular No. 26/2022/TT-NHNN, which stipulates prudential ratios and strengthens the regulatory framework for liquidity management. As a result, key indicators such as the LDR, short-term funding ratio, and other liquidity metrics have improved in recent years. These changes reflect a gradual alignment with global best practices, particularly the principles of Basel III, which emphasize liquidity coverage ratios (LCR) and net stable funding ratios (NSFR) as essential measures of resilience. Empirical evidence suggests that compliance with these standards has helped Vietnamese banks enhance their ability to withstand short-term shocks and reduce systemic vulnerabilities in the financial system.

Nevertheless, significant limitations remain in the implementation of these models. The asset-based approach, while effective in securing immediate repayment capacity, tends to constrain banks' profitability by increasing the opportunity cost of idle liquid assets. The liability-based approach, on the other hand, exposes banks to fluctuations in market funding costs and liquidity conditions, which can become critical in periods of financial stress. Although the balanced approach theoretically provides a more optimal solution, its successful implementation requires advanced risk measurement techniques, robust internal controls, and a sophisticated management infrastructure — conditions that many Vietnamese banks, especially small- and medium-sized ones, still lack. In practice, the reliance on traditional indicators such as LDR or the current liquidity ratio has constrained the predictive capacity of existing models, limiting their effectiveness as early warning tools.

The root causes of these challenges are both internal and external. Internally, many Vietnamese banks face structural barriers, including underdeveloped data management systems, limited integration of advanced risk management technologies, and shortages of qualified personnel with expertise in modern financial risk modeling. Externally, the regulatory framework, although progressively improving, continues to focus primarily on compliance with static prudential ratios rather than fostering the adoption of forward-looking, dynamic, and predictive models. This orientation leads to a gap between regulatory compliance and actual risk preparedness. Compared with international practices — where stress testing, scenario analysis, and artificial intelligence (AI)-driven models are increasingly employed — Vietnamese banks remain at an early stage in developing sophisticated early warning systems for liquidity risk.

From an international perspective, the gap is striking. In developed markets, banks have moved beyond ratio-based monitoring to integrate ML, AI, and big data analytics into their liquidity risk management frameworks. These tools allow banks not only to detect early warning signals but also to model complex interactions between macroeconomic shocks, market movements, and customer behaviors. In contrast, the majority of VCBs continue to rely heavily on retrospective data and conventional

ratios, which are insufficient in volatile financial environments characterized by uncertainty and rapid structural changes. This disparity underscores the urgent need for Vietnam to accelerate the adoption of advanced methodologies in line with global standards.

Taken together, the current situation of early warning models in VCBs reveals both progress and gaps. While compliance with SBV regulations and gradual convergence with Basel III standards have enhanced the overall resilience of the banking system, the prevailing reliance on traditional approaches constrains predictive accuracy and adaptability. These limitations highlight the necessity of adopting more advanced and dynamic models, particularly those based on AI, to develop early warning systems that are not only regulatory-compliant but also forward-looking, flexible, and capable of capturing the complexities of modern financial markets. Addressing these gaps is essential for VCBs to enhance their resilience, safeguard financial stability, and sustain long-term growth in an increasingly integrated and unpredictable global economy.

4. RESEARCH METHOD

In this section, we provide a comprehensive theoretical foundation for the application of AI models in predicting liquidity risk at commercial banks. Specifically, we focus on three advanced ML models: LASSO regression, RF, and Extreme Gradient Boosting (XGBoost). These models have gained increasing popularity in financial risk modeling due to their capacity to handle high-dimensional data, capture nonlinear patterns, and provide reliable classification accuracy.

Financial risk forecasting has historically made extensive use of statistical techniques like logit/probit regression and discriminant analysis (Altman, 1968). Nevertheless, these traditional models frequently rely on rigid linearity and normality assumptions, which restrict their ability to accurately represent the intricate, non-linear dynamics of contemporary financial markets. ML techniques, on the other hand, provide greater flexibility because they can handle high-dimensional data and find hidden non-linear patterns without strict presumptions (W. Liu et al., 2022). In order to improve the predictive accuracy of liquidity risk early warning systems, this study uses sophisticated ML techniques, namely LASSO, RF, and XGBoost, rather than conventional approaches.

4.1. Least absolute shrinkage and selection operator model

The least absolute shrinkage and selection operator regression model is a widely used statistical technique, particularly influential in the fields of econometrics, financial modeling, and ML. Initially introduced by Robert Tibshirani in 1996, the LASSO method was developed as a response to the growing challenge of high-dimensional datasets, where the number of explanatory variables can be large and often highly correlated. In many applied financial forecasting problems, the ability to efficiently identify and retain only the most relevant predictors — while eliminating irrelevant or redundant ones — is crucial for enhancing both model performance and interpretability.

LASSO operates on the principle of regularization, specifically L1 regularization, which modifies the standard regression model by introducing a penalty term into the loss function. This penalty is proportional to the sum of the absolute values of the regression coefficients. In contrast to ridge regression, which uses L2 regularization and shrinks coefficients toward zero without ever eliminating them entirely, LASSO has the unique property of forcing some coefficients to exactly zero. This results in automatic variable selection, whereby variables that do not contribute meaningfully to the predictive power of the model are excluded during the optimization process. Mathematically, the LASSO model aims to minimize the following objective function:

$$max_{\beta} l(\beta|y) - \lambda \sum_{i} ||\beta_i||_1$$

The penalty term in the LASSO regression is given by:

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

As λ increases, the model penalizes complexity more heavily. This encourages sparsity, i.e., setting many coefficients to zero, effectively excluding non-informative variables from the model. From an ML perspective, L1 regularization enhances the model's generalization ability by avoiding overfitting — a common issue when dealing with noisy, volatile financial data. As noted by Mahbobi et al. (2023), increasing the α (or λ) value in the regularization term leads to a simpler and more parsimonious model. The trade-off between model complexity and predictive accuracy is governed by this α value: smaller values allow more variables to remain in the model (thus increasing variance), while larger values reduce the model's variance at the cost of potentially omitting useful predictors.

In practical financial applications, such as liquidity risk prediction, multicollinearity among variables (e.g., different asset ratios, interest spreads, and credit quality indicators) is a frequent challenge. Traditional regression models may produce unstable coefficient estimates when predictors are highly correlated. LASSO helps mitigate this issue by selecting only a subset of these variables, thereby reducing model sensitivity and improving interpretability.

Liu and Yu (2022) emphasize that LASSO serves as a powerful technique for dimensionality reduction, especially in environments where the number of explanatory variables exceeds the number of observations, or when many variables have weak or no relationship with the outcome of interest. Moreover, it facilitates model transparency, allowing analysts and financial regulators to understand which variables are most influential in predicting outcomes such as default risk or liquidity shortfalls. Furthermore, Drudi and Nobili (2021) underscore the importance of regularization techniques like LASSO in enhancing model robustness. According to their findings, LASSO optimizes the log-likelihood of a binomial

distribution, subject to a penalty on the coefficient estimates, ensuring that the model remains flexible enough to detect early warning signs of financial distress while not becoming overly sensitive to random fluctuations in the data. In essence, the method estimates the conditional probability of a binary outcome (e.g., default vs. no default) by maximizing a penalized likelihood function, which can be expressed as:

$$l(\beta|y) = \frac{1}{N*T}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t} * \beta' x_{i,t} - log\left(1 + e^{\beta' x_{i,t}}\right) \quad (1)$$

where $p_i = \frac{1}{1+e^{-X_i\beta}}$ the predicted probability of default, and the second term imposes a constraint on the complexity of the coefficient vector β.

The optimization formula in LASSO is as follows:

$$minimize[\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|]$$

Explanation of terms:

• y_i: the real value of the target variable for observation i,

• x_{ij}: the value of the j^{th} independent variable for the i^{th} observation,

• β_j: the regression coefficient corresponding to the independent variable x_j,

• λ: a regularization parameter that adjusts the level of penalty.

Meaning of parameter λ:

• When $\lambda = 0$: Without constraints, LASSO becomes ordinary linear regression (ordinary least squares (OLS)).

• As λ increases: The regression coefficients β_j begin to shrink. When λ is large enough, some coefficients will be pulled toward zero, leading to variable selection.

In summary, the LASSO regression model offers a mathematically rigorous and practically useful framework for building early warning systems in the financial sector. By enhancing interpretability, controlling model complexity, and improving predictive accuracy, LASSO has become an indispensable tool for analysts, policymakers, and risk managers seeking to monitor and manage systemic risks in banking and financial markets.

4.2. Random forest

"Random Forest (RF) is a machine learning algorithm belonging to the ensemble learning group", proposed and developed by Breiman (2001, p. 5), "to improve the predictive performance of the traditional decision tree model (Decision Tree) by combining many independent decision trees together" (p. 6). This algorithm works based on the Bootstrap Aggregation (Bagging) principle, which is to create many subsets of data sampled from the original data set, train individual decision trees in these subsets, and then aggregate the prediction results of all trees to make the final decision (Breiman, 2001). Using multiple decision trees instead of a single tree helps RF reduce overfitting and improve the generalization ability of the model (W. Liu et al., 2022). One of the most important characteristics of RF is the ability to handle highly non-linear and noisy data. According to T. Liu and Yu (2022), RF can detect complex relationships between input variables that linear models, such as OLS regression, cannot identify. This makes RF an effective tool in financial problems where data are often highly volatile, and many factors are acting simultaneously. Basic equations: for a classification problem, suppose that there are M trees in the forest, each decision tree $h_m(x)$ produces a prediction. The final result of the prediction y of the RF is decided by majority vote:

$$y = mode(h_1(x), h_2(x), \ldots, h_m(x)) \quad (2)$$

For a regression problem, the final output y of the RF is calculated as the average of the predictions from the trees:

$$y = \frac{1}{M}\sum_{m=1}^{M} h_m(x) \quad (3)$$

Meaning of important parameters:

• Number of trees in the forest (*n estimators*): This parameter determines the number of decision trees constructed in the RF model. Increasing the number of trees generally improves the accuracy of the prediction, but also increases training time.

• Number of features randomly selected at each split node (*max features*): Specifies how many features are randomly selected from the total set of features when splitting a node. This randomness reduces correlation among trees and contributes to the overall robustness of the model.

• Maximum depth of each tree (*max_depth*): Limits the depth to which each individual decision tree can grow. Without this constraint, trees can grow too complex and overfit the training data by perfectly classifying all samples.

• Sample size constraints (*min samples split*, *min_samples).* These parameters control the minimum number of samples required to split a node (*min samples split*) and the minimum number of samples required to be at a leaf node (*min samples leaf*). They help regulate the complexity of the model and prevent overfitting.

• Bootstrap sampling method (*bootstrap*): Determines whether bootstrap sampling (sampling with replacement) is used when building each tree. Using bootstrap increases diversity among trees, improving generalization.

• Class weights (*class weight*): Assign different weights to classes in classification problems. This is particularly useful in unbalanced datasets where some classes are underrepresented.

RF is a powerful ensemble learning method proposed by Breiman (2001), widely used in classification and regression tasks. It constructs multiple decision trees during training and aggregates their outputs to improve predictive performance and reduce overfitting. The method relies on two main techniques: bootstrap aggregation (bagging) and random feature selection.

Let the dataset be $D = \{(X_i y_i)\}_{i=1}^{n}$, where $X_i \in R^p$. The algorithm proceeds as follows:

1. For $b = 1$ to B:

• Draw a bootstrap sample D_b.

• Train a decision tree h_b on D_b, selecting m features randomly at each split.

2. Aggregate predictions using majority vote:

$$\hat{y} = mode\{h_B(x)\}_{b=1}^{B} \quad (4)$$

RF is non-parametric and robust to noise and outliers. It can rank feature importance, handle high-dimensional data, and model nonlinear relationships, making it ideal for financial risk prediction.

4.3. Extreme gradient boosting

"Extreme Gradient Boosting (XGBoost) is a machine learning algorithm belonging to the ensemble learning group", proposed by Peng et al. (2022, p. 4). XGBoost works on the principle of boosting, in which decision trees are built sequentially, each new tree trying to correct the error of the previous tree by optimizing the prediction error based on the second derivative of the loss function. This process consists of three main steps. First, the model is initialized with a simple decision tree to predict the input data. Then, each new tree added learns from the remaining errors of the previous tree, with adjusted weights to reduce the prediction error. Finally, the model uses a weighted aggregation method to combine the decision trees, which improves the prediction ability compared to bagging methods such as RF (Peng et al., 2022). Thanks to the boosting mechanism, XGBoost can leverage the information of each tree to improve the forecasting quality, instead of training independent trees like RF. This helps XGBoost achieve higher accuracy, but at the same time, it also has the risk of overfitting if the parameters are not adjusted properly (Drudi & Nobili, 2021). In financial studies, XGBoost is one of the most effective algorithms for predicting liquidity risk. When applied to financial risk forecasting, XGBoost has been shown to yield superior predictive performance compared to traditional statistical models and other ensemble approaches like random forest (W. Liu et al., 2022).

Mathematically: Suppose our model is a set of trees f k, and the model after t iterations is:

XGBoost model formulation. The predicted value is given by:

$$\hat{y} = \sum_{k=1}^{t} f_k(x) \quad (5)$$

where: $f_k(x)$ are the new trees added in round k.

XGBoost minimizes the following objective function:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k) \quad (6)$$

With: $L(y_i, \hat{y}_i)$: the loss function between the actual value y_i and the predicted value $\hat{y}_i$; $\Omega(f_k)$: the regularization component, which helps prevent overfitting, defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j{}^2 \quad (7)$$

where:

• T is the number of leaves in the tree,

• ω_j is the weight of the j-th leaf,

• γ and λ are the tuning hyperparameters.

Gradient boosting updates the model based on the derivative of the loss function. If $L(y_i, \hat{y}_i)$ is the loss function, then at step $t + 1$, the model is updated by adding a new tree $\hat{y}_t + 1$ to minimize the error:

$$\hat{y}_t + 1 = \hat{y}_t - n\nabla_{\hat{y}_t} L \quad (8)$$

where:

• n is the learning rate,

• $\nabla_{\hat{y}_t} L$ is the gradient of the loss function at iteration t. XGBoost is widely adopted due to its fast-processing speed, strong ability to prevent overfitting, and high predictive accuracy.

4.4. Data and variables

This study uses a data set of 239 observations, collected from financial statements and public data sources of 24 commercial banks operating in Vietnam in the period from December 31, 2014, to December 31, 2023, with an annual frequency, reflecting the fluctuations of the banking system in a challenging economic context. To ensure security and consistency in analysis, banks are encrypted by stock code (CK Code). In addition, macroeconomic data is collected from the World Bank, ensuring transparency and high reliability in analysis. The dependent variable (default) is used to measure the liquidity risk of commercial banks in Vietnam and is constructed using three key liquidity ratios: the loan-to-deposit ratio (LDR), the liquidity coverage ratio (LCR), and the net stable funding ratio (NSFR). These ratios are calculated from the financial statements of commercial banks in accordance with Basel III standards. From there, the study proposes two hypotheses as follows:

H1: Default = 0: The bank does not have liquidity risk.

H2: Default = 1: The bank has liquidity risk.

Independent variables: include 11 variables used to early forecast liquidity risks of commercial banks, including: capital adequacy ratio (*CAR*), *LCR,* gross domestic product (GDP) growth (*GGDP*), inflation (*INF*), *ROA,* unemployment rate (*UR*), bad debt ratio (*NPL*), *LDR, SIZE, NSFR,* net interest margin (*NIM*). The variables are described specifically in Table 1. These variables are popular micro and macro financial indicators that have been widely used in previous research.

Given prediction $\hat{y}_i = \hat{y}_i + f_t(x_i)$, the objective function is:

$$\mathcal{L}(f) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)} + \sum_{k=1}^{t} \Omega(f_k) \quad (9)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$.

Using Taylor expansion:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (10)$$

With $g_i = \partial_{\hat{y}_i} L(y_i, \hat{y}_i)$ and $h_i = \partial_{\hat{y}_i}{}^2 l(y_i, \hat{y}_i)$.
The optimal split is determined by gain:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_{L+} H_R + \lambda} \right] - \gamma \quad (11)$$

XGBoost offers robust performance for imbalanced data, supports regularization, and captures complex feature interactions, making it suitable for banking applications such as liquidity risk forecasting.

5. RESULTS AND DISCUSSIONS

Data from Table 1 shows that the average *CAR* reached 10.64%, while the *LDR* was 75.22%. *NPL* is low, averaging 1.77%, and real *GGDP* reached 4.79%. The average *INF* rate is 3.46%, and the *UR* remains low at 73.19%. In terms of operating efficiency, the average *ROA* reaches 1%. The default variable (*default*) shows that 68.5% of cases belong to the default group. The data shows relative stability, but there are some outliers to watch out for.

Table 1. Descriptive statistics of variables in the model

Variables	*Average*	*Std. dev.*	*Min*	*Max*	*Obs.*
CAR	0.1064	0.0032	0	0.2453	238
LDR	0.7522	0.0070	0.3181	1.0571	238
LCR	$2.10 * 10^{11}$	$9.32 * 10^{10}$	0	$1 * 10^{13}$	238
NSFR	126.5592	1.2740	95.3638	236.2279	238
ROA	0.0100	0.0005	0	0.0365	238
NIM	0.0318	0.0009	0	0.0937	238
SIZE	32.8681	0.0725	30.3925	35.3721	238
NPL	0.0177	0.0007	0	0.0691	238
GGDP	6.0479	0.1195	2.5537	8.1235	238
INF	3.4366	0.0727	1.8347	6.3120	238
UR	1.7319	0.0227	1.1610	2.3850	238
Default	0.6849	0.0302	0	1	238

Source: Authors' calculation.

The results of the multicollinearity test through the 'correlation matrix' between independent variables in the model indicate no clear signs of severe multicollinearity. Specifically, according to common conventions, multicollinearity is generally considered serious when the correlation coefficient between two variables exceeds the threshold of 0.8 or 0.9 (Gujarati & Porter, 2009). In the obtained correlation matrix (Table 2), no pair of variables reaches this level, suggesting that the likelihood of high multicollinearity is negligible.

However, some variable pairs exhibit significant correlations, such as the relationship between *ROA* and *NIM*, with a correlation coefficient of 0.74. This demonstrates the practical reality that a bank's profitability is often closely related to operational efficiency. In addition, the *NSFR* shows a relatively high negative correlation with the *LDR* (-0.69), indicating that the more robust a bank's liquidity strategy, the lower its *LDR*.

Moreover, the relationships among macroeconomic variables are also noteworthy. The *UR* has a strong negative correlation with *GGDP* (-0.70), which aligns with macroeconomic theory, as higher economic growth generally leads to lower *URs* (Okun, 1962). At the same time, *INF* shows a positive correlation with *GGDP* (0.38), reflecting the tendency for economic growth to be accompanied by inflationary pressures. Some other variables, such as the *CAR*, have very low correlations with most other variables, indicating that this variable's impact on other factors in the model is negligible.

Table 2. Correlation matrix of independent variables

Variables	*CAR*	*LDR*	*LCR*	*NFSR*	*ROA*	*NIM*	*SIZE*	*NPL*	*GGDP*	*INF*	*UR*	*Default*
CAR	1											
LDR	0.020481	1										
LCR	0.042929	0.034139	1									
NFSR	0.176169	-0.69172	-0.09934	1								
ROA	0.043965	0.390838	0.16268	-0.35274	1							
NIM	0.172751	0.41634	0.440596	-0.28093	0.737405	1						
SIZE	0.018381	0.451794	0.026742	-0.30481	0.485429	0.279153	1					
NPL	0.056269	0.034645	0.176132	0.008439	-0.06094	0.217882	-0.04564	1				
GGDP	0.042524	-0.01311	0.068214	0.108901	-0.12983	-0.01401	-0.1346	-0.01942	1			
INF	0.004912	-0.15692	0.077126	0.3248	-0.27036	-0.09317	-0.20875	-0.02749	0.377331	1		
UR	0.025265	0.047794	-0.05604	-0.11461	0.064434	0.037283	0.082838	0.004357	-0.70128	-0.28595	1	
Default	-0.0789	-0.5219	-0.21616	0.305103	-0.28247	-0.30642	-0.36789	0.075988	-0.00996	0.064023	0.064891	1

Source: Authors' calculation.

Overall, although there are no clear signs of "severe multicollinearity", further examination using the variance inflation factor (VIF) is necessary to reach a more precise conclusion about the level of multicollinearity in the model. If significant multicollinearity is detected, remedial measures such as removing highly correlated variables, "using ridge regression" (Hoerl & Kennard, 1970), or applying principal component analysis (PCA) can be employed to minimize the impact of multicollinearity and ensure the stability of the regression model.

Table 3. Results of machine learning models

Model	R^2
LASSO	0.8723
XGBoost	0.9787
Random fores	0.9980

Source: Authors' calculation.

The article emphasizes that RF is the most accurate model for predicting liquidity risks in commercial banks, achieving an accuracy of up to 99.8% at Table 3. This observation aligns with prior research concerning financial distress prediction. Specifically, W. Liu et al. (2022) and Barongo and Mbelwa (2024) arrived at comparable conclusions, demonstrating that ML methodologies, especially ensemble approaches, substantially surpass conventional statistical models, including logit and probit, in the early detection of banking crises. This model significantly outperforms traditional methods, providing better predictions of a bank's solvency and contributing to financial stability. RF's key advantage is its ability to handle complex relationships between financial variables while avoiding overfitting, emphasized by Breiman (2001). Moreover, although Ekinci and Sen (2024) highlighted the utility of cost-sensitive methodologies in predicting bank failures within the United States, our findings indicate that the RF's robustness offers a more dependable structure for safeguarding financial stability in an emerging market such as Vietnam. By accurately identifying critical liquidity indicators, the model empowers bank managers to institute timely preventative actions. This paper efficiently processes large and complex datasets, enabling banks to gain a clearer view of liquidity risks and implement more effective risk management strategies. This helps minimize potential financial instability and ensures long-term resilience. Conversely, it is essential to recognize that although this study incorporates both bank-specific ratios and significant domestic macroeconomic indicators, including GDP and inflation, etc., its foundation rests predominantly on structured quantitative data. Therefore, we propose that future research should enhance predictive models by integrating global financial volatility indicators, such as oil prices or the exchange rate fluctuations of major currencies, alongside unstructured data, including news sentiment or textual analysis. This broadened scope would enable banks to gain a more thorough understanding of external disruptions impacting liquidity risks, thereby improving the dependability and adaptability of their risk management strategies over time.

6. CONCLUSION

In the era of digital transformation, the development of big data along with powerful computing capabilities has facilitated the deployment of ML models to analyze and forecast risks with higher accuracy. Many large financial institutions have begun to integrate advanced algorithms into early warning systems to detect abnormalities in cash flow and liquidity. Therefore, it is necessary for commercial banks in Vietnam to apply ML models to forecast risks in the near future. The study has contributed to that future by comprehensively analyzing liquidity risks in commercial banks in Vietnam and proposing a method of applying AI to improve the efficiency of early forecasting of liquidity risks. The research results show that liquidity risks not only affect individual banks but also impact the entire financial system, along with the accuracy of ML models in early forecasting of liquidity risks in commercial banks. Traditional methods such as LCR, NSFR, and other liquidity indicators, although important, still have many limitations, especially in the context of increasingly complex and volatile financial markets. These methods mainly rely on historical data and are unable to predict abnormal liquidity risks early. The development of AI and ML has opened up a new approach to liquidity risk management. Algorithms such as RF and XGBoost have demonstrated their ability to analyze big data, detect abnormal trends, and provide more accurate forecasts. Our results show that algorithms such as RF and XGBoost have demonstrated superior ability to analyze big data and provide more accurate forecasts, with RF achieving an accuracy of up to 99.8%. According to international studies, the application of AI not only helps improve forecast accuracy but also supports banks in optimizing capital resources, limiting losses due to liquidity loss, and improving operational performance.

Theoretically, this study validates that ensemble learning techniques (RF, XGBoost) are more appropriate for the Vietnamese market than conventional linear models. This implies that in order to capture the intricate, non-linear dynamics of financial data, researchers should concentrate more on AI models. Practically, data quality and asynchronous technology infrastructure continue to be major obstacles to the application of AI models in liquidity risk management in Vietnam. Therefore, for AI to be as effective as possible, the government and the SBV must implement policies that support digital transformation. In order to identify risks earlier, commercial banks should give priority to investing in data systems and incorporating these early warning models.

This study still has some limitations. First, we only used quantitative data from financial statements. We did not use qualitative information, such as news or management reports. Second, the sample size is 239 observations, which is relatively small compared to global datasets. In the future, researchers can expand the data to include other Association of Southeast Asian Nations (ASEAN) countries or use monthly data to improve the timeliness of the predictions.

REFERENCES

Akhter, N. (2023). Determinants of commercial bank's non-performing loans in Bangladesh: An empirical evidence. *Cogent Economics & Finance*, *11*(1), Article 2194128. https://doi.org/10.1080/23322039.2023.2194128

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.2307/2978933

Alves, C., Kretschmar, B., & Kim, K. S. (2023, February 8). *Liquidity management — A comparison of approaches to modeling cash flows for variable bank products*. BankingHub. https://www.bankinghub.eu/finance-risk/liquidity-management-cash-flows-variable-bank-products

Aspachs, O., Nier, E. W., & Tiesset, M. (2005). *Liquidity, banking regulation and the macroeconomy*. https://doi.org/10.2139/ssrn.673883

Barongo, S. I., & Mbelwa, J. T. (2024). Using machine learning for detecting liquidity risk in banks. *Machine Learning with Applications*, *15*, Article 100511. https://doi.org/10.1016/j.mlwa.2023.100511

Basel Committee on Banking Supervision. (2008, September). *Principles for sound liquidity risk management and supervision*. Bank for International Settlements. https://www.bis.org/publ/bcbs144.htm

Berger, A. N., & DeYoung, R. (1997). Problem loans and cost efficiency in commercial banks. *Journal of Banking & Finance*, *21*(6), 849–870. https://doi.org/10.1016/S0378-4266(97)00003-4

Bonfim, D., & Kim, M. (2012, July). *Liquidity risk in banking: Is there herding?* European Banking Authority. https://surli.cc/nbhxry

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, *91*(3), 401–419. https://doi.org/10.1086/261155

Drudi, M. L., & Nobili, S. (2021). *A liquidity risk early warning indicator for Italian banks: A machine learning approach* (Working Paper No. 1337). Bank of Italy Temi di Discussione. https://doi.org/10.2139/ssrn.3891566

Ekinci, A., & Sen, S. (2024). Forecasting bank failure in the U.S.: A cost-sensitive approach. *Computational Economics*, *64*, 3161–3179. https://doi.org/10.1007/s10614-023-10537-6

Gregova, E., Valaskova, K., Adamko, P., Tumpach, M., & Jaros, J. (2020). Predicting financial distress of Slovak enterprises: Comparison of selected traditional and learning algorithms methods. *Sustainability*, *12*(10), Article 3954. https://doi.org/10.3390/su12103954

Guerra, P., Castelli, M., & Côrte-Real, N. (2022). Machine learning for liquidity risk modelling: A supervisory perspective. *Economic Analysis and Policy*, *74*, 175–187. https://doi.org/10.1016/j.eap.2022.02.001

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill/Irwin.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., & Herrera-Viedma, E. (2021). Machine learning methods for systemic risk analysis in financial sectors. *Technological and Economic Development of Economy, 25*(5), 716–742. https://doi.org/10.3846/tede.2019.8740

Liu, T., & Yu, Z. (2022). The analysis of financial market risk based on machine learning and particle swarm optimization algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2022*, Article 31. https://doi.org/10.1186/s13638-022-02117-3

Liu, W., Fan, H., Xia, M., & Pang, C. (2022). Predicting and interpreting financial distress using a weighted boosted tree-based tree. *Engineering Applications of Artificial Intelligence, 116*, Article 105466. https://doi.org/10.1016/j.engappai.2022.105466

Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2023). Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*, *330*, 609–637. https://doi.org/10.1007/s10479-021-04114-z

Nguyen, M. P., Nguyen, T. T. H., & Phan, A. (2024). Enhancing sustainable development awareness in Vietnamese commercial banks: Analyzing environmental, social, and governance factors. *Pakistan Journal of Life and Social Sciences*, *22*(2), 17555–17567. https://doi.org/10.22495/cbsrv5i3art13

Nguyen, M. P., Phan, A., & La, T. T. T. (2025). Determinants of international financial reporting standards adoption at commercial banks: Evidence from an emerging country. *Humanities and Social Sciences Letters*, *13*(2), 396–408. https://doi.org/10.18488/73.v13i2.4112

Nguyen, V. T. (2019). The impact of artificial intelligence on banking operations. *Journal of Banking Science and Training*, *214*, 1–8. https://shorturl.at/EU6GM

Okun, A. M. (1962). Potential GNP: Its measurement and significance. In *Proceedings of the business and economic statistics section of the American Statistical Association* (pp. 98–104). American Statistical Association. https://www.scribd.com/document/347717593/Potential-GNP-Its-Measurement-and-Significance

Peng, H., Lin, Y., & Wu, M. (2022). Bank financial risk prediction model based on big data. *Scientific Programming*, *2022*, Article 3398545. https://doi.org/10.1155/2022/3398545

Rose, P. S. (2004). *Commercial bank management* (6th ed.). McGraw-Hill/Irwin.

State Bank of Vietnam. (2019, November 15). *Circular No. 22/2019/TT-NHNN. Limits and prudential ratios of banks and foreign bank branches*. Apolat Legal. https://apolatlegal.com/wp-content/uploads/2025/07/Circular-22_2019_TT-NHNN.pdf

State Bank of Vietnam. (2022, December 31). *Circular No. 26/2022/TT-NHNN. Amending and supplementing a number of articles of Circular No. 22/2019/TT-NHNN*. LuatVietnam. https://surli.cc/ddctjb

Tavana, M., Abtahi, A.-R., Di Caprio, D., & Poortarigh, M. (2018). An artificial neural network and Bayesian network model for liquidity risk assessment in banking. *Neurocomputing*, *275*, 2525–2554. https://doi.org/10.1016/j.neucom.2017.11.034

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vietcombank Securities (VCBS). (2023). *Banking sector report 2023*. https://api.vcbs.com.vn/storage/upload/media/toan-van-bao-cao-thuong-nien-2023.pdf

Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, *58*, 93–101. https://doi.org/10.1016/j.eswa.2016.04.001