JEL Classification: F24, M41, Q56

Svetlana Drobyazko,
European Academy of Sciences Ltd, United Kingdom
https://orcid.org/0000-0003-2022-0126
svetlana.drobyazko@yahoo.com

Besik Bauchadze,
Batumi Shota Rustaveli State University, Georgia
https://orcid.org/0000-0002-0657-9039
besik.bauchadze@bsu.edu.ge

APPLICATION OF ENSEMBLE AI METHODS FOR DYNAMIC STRESS-TESTING AND SYSTEMIC RISK MANAGEMENT IN AN UNSTABLE ECONOMY

Received 12 April 2026; accepted 26 April 2026; published 30 April 2026

Abstract. *The increasing complexity of financial markets and the limitations of traditional stress-testing methods have created a need for more adaptive approaches to systemic risk assessment, particularly in unstable economic environments. This study develops and validates an ensemble artificial intelligence framework for dynamic stress-testing and systemic risk management using quarterly macro-financial data from 1999 to 2024 drawn from the Bank for International Settlements, the International Monetary Fund, the Federal Reserve Economic Data system, and NYU Stern's V-Lab. Five ensemble methods − Random Forest, XGBoost, CatBoost, Histogram-based Gradient Boosting, and a Stacking ensemble − are estimated and compared against linear regression and naive forecast baselines, with the credit-to-GDP gap, the debt service ratio, and SRISK as target variables. The Stacking ensemble achieves the lowest out-of-sample prediction errors, reducing root mean squared error by 32 per cent relative to linear regression for both the credit gap and SRISK, with Diebold-Mariano tests confirming statistical significance at the 1 per cent level. SHAP analysis identifies the lagged credit-to-GDP gap, the debt service ratio, the VIX, GDP growth, and corporate bond spreads as the most important predictors, revealing non-linear thresholds at a credit gap of 10 per cent and a debt service ratio of 18 per cent. Dynamic stress-testing simulations show that an extreme scenario pushes the representative emerging economy above the 10 per cent crisis threshold within eight quarters, while the advanced economy remains below it. The early warning evaluation achieves an area under the ROC curve of 0.91 at the four-quarter horizon and a signal-to-noise ratio exceeding the 2:1 policy-useful threshold. The study concludes that ensemble methods substantially outperform traditional models in systemic risk prediction, that the identified non-linear thresholds provide empirically grounded anchors for macroprudential monitoring, and that emerging economies face higher systemic risk exposures under identical shock scenarios.*

Keywords: *risk management, modelling, artificial intelligence (AI), stress testing, systemic risks.*

Citation: Drobyazko, S.; Bauchadze, B. (2026). APPLICATION OF ENSEMBLE AI METHODS FOR DYNAMIC STRESS-TESTING AND SYSTEMIC RISK MANAGEMENT IN AN UNSTABLE ECONOMY. Economics and Finance, Volume 14, Issue 1, 117. https://doi.org/10.51586/2754-6209.2026.14.1_117

Introduction

The global financial system has become increasingly complex and interconnected over the past three decades. The proliferation of cross-border capital flows, high-frequency trading, algorithmic investment strategies and digital assets has fundamentally altered the nature of financial

risk (Ojo, 2025). Traditional risk management frameworks, which rely heavily on historical data, linear assumptions and static rule-based models, have proven inadequate in capturing the non-linear dependencies, abrupt regime shifts and cascading contagion effects that characterise modern financial crises (Nafiu et al., 2025). The 2008 global financial crisis, the COVID-19 induced market turmoil of 2020 and the systemic disruptions observed during recent geopolitical and regulatory shocks (Khemakhem & Euchi, 2026) serve as stark reminders of the limitations of conventional approaches such as value-at-risk (VaR) and historical stress-testing. Moreover, the emergence of AI-native assets, semiconductor-driven technological cores and coordinated algorithmic trading introduces new sources of systemic vulnerability that traditional models were never designed to address (McClellan, 2025). Consequently, there is an urgent need for more adaptive, data-driven and forward-looking risk management frameworks capable of operating under conditions of deep uncertainty and structural instability.

In response to these challenges, artificial intelligence (AI) and machine learning (ML) have emerged as transformative forces in financial risk management. AI-based systems enable real-time anomaly detection, predictive analytics and automated decision-making across credit, market, operational and liquidity risk domains (Aziz & Dowling, 2019; Liu, 2025). Among the various ML paradigms, ensemble methods – including Random Forest, XGBoost, CatBoost, Hist Gradient Boosting and Stacking – have demonstrated particular promise. Unlike single classifiers, ensembles combine multiple base learners to reduce variance, mitigate overfitting and capture complex interaction effects that linear models cannot represent (Nallakaruppan et al., 2024).

Empirical studies have confirmed the superiority of ensemble approaches in financial forecasting and risk assessment. Moffo (2024) showed that Random Forest and Adaptive Lasso significantly outperform traditional linear models in predicting net charge-offs (NCO) and pre-provision net revenue (PPNR) under adverse stress scenarios, particularly in capturing the left-skewed tail of the Tier 1 common equity capital (T1CR) distribution. Mawardi et al. (2026) developed a hybrid early warning system integrating Z-score with Extra Trees, achieving an R^2 of 0.95 and anticipating financial distress up to two years in advance. Malali (2025) compared CatBoost and Hist Gradient Boosting for systemic risk analysis, reporting accuracy above 94% and demonstrating that gradient-boosted ensembles can outperform traditional models such as Deep FM and XGBoost. These findings collectively underscore the potential of ensemble methods as a robust analytical engine for stress-testing and systemic risk monitoring.

Despite the growing body of evidence supporting ensemble methods in isolated financial risk tasks, a critical research gap remains. No existing study has comprehensively integrated dynamic stress-testing, systemic risk management and ensemble AI methods within the specific context of an unstable economy. The literature suffers from several interrelated limitations:

First, the majority of empirical work on ML-based stress-testing has been conducted using data from the United States, Western Europe or other advanced OECD economies (Moffo, 2024; Siddik & Amin, 2026). These contexts are characterised by mature data infrastructures, strong regulatory frameworks and relatively stable macroeconomic conditions. By contrast, unstable or emerging economies – which often exhibit higher volatility, weaker institutional environments and infrastructural constraints – remain severely under-researched (Tanbour et al., 2025; Yuningrat, 2025;Canabarro, 2024).

Second, while individual ensemble architectures have been tested in isolation, there is no unified framework that systematically compares multiple ensemble methods (Random Forest, XGBoost, CatBoost, Stacking) for dynamic, multi-scenario stress-testing targeting systemic risk indicators. Most studies focus on a single model or a single risk type (e.g., credit risk), leaving the broader systemic perspective unaddressed.

Third, the “black box” problem remains a major barrier to regulatory acceptance. Although explainable AI (XAI) techniques such as SHAP and LIME have been applied to credit risk assessment (Nallakaruppan et al., 2024), their integration into dynamic stress-testing pipelines for systemic risk is still nascent. Without interpretability, even highly accurate models are unlikely to be trusted by central banks and prudential regulators.

Fourth, recent research has shown that AI itself can become a source of systemic risk – through coordinated LLM trading signals (McClellan, 2025) or extreme interconnectedness among AI-native assets (Khemakhem & Euchi, 2026). Yet, these insights have rarely been incorporated into empirical stress-testing frameworks. A unified approach that accounts for both AI-enabled risk mitigation and AI-generated systemic fragility is urgently needed.

In light of the above gaps, this study aims to develop and empirically validate an ensemble AI-based approach for dynamic stress-testing and systemic risk management in an unstable economy.

To achieve this aim, the following specific objectives are set:

1. To construct and compare several ensemble models (Random Forest, XGBoost, CatBoost, Hist Gradient Boosting and Stacking) for predicting key systemic risk indicators, including the credit-to-GDP gap, SRISK and the Z-score.

2. To design a dynamic stress-testing methodology that incorporates multiple macroeconomic scenarios (baseline, adverse and extreme) reflecting the volatility characteristic of unstable economies.

3. To evaluate the early warning capability of ensemble models across different forecast horizons (4, 8 and 12 quarters) and to assess their sensitivity to different shock magnitudes.

4. To ensure model interpretability through SHAP (SHapley Additive exPlanations) analysis, identifying the most influential macro-financial drivers of systemic risk and revealing non-linear threshold effects.

5. To identify the key macroeconomic and financial drivers of systemic risk in an unstable economy, with particular attention to variables such as the credit gap, debt service ratio, unemployment, inflation, exchange rate volatility and global spillovers.

The paper extends the existing literature on systemic risk and AI in finance by integrating ensemble methods with dynamic stress-testing in the context of an unstable economy. Unlike prior work that treats AI either as a risk mitigator or a risk source, this study acknowledges both roles within a single analytical framework. It also contributes to the emerging discourse on XAI in financial regulation by demonstrating how ensemble models can be made interpretable without sacrificing predictive power.

The study provides a systematic comparison of five ensemble architectures (Random Forest, XGBoost, CatBoost, Hist Gradient Boosting and Stacking) applied to systemic risk prediction and stress-testing. It develops a replicable pipeline that combines macro-financial data, scenario simulation, recursive forecasting and SHAP-based interpretation. The hybrid approach – blending traditional Z-score concepts with modern ML – also offers a bridge between legacy regulatory frameworks and AI-driven innovation (Mawardi et al., 2026).

For financial regulators and central banks operating in volatile environments, the proposed framework offers a real-time early warning tool that can identify systemic vulnerabilities before they materialise. For commercial banks and investment firms, the model provides actionable insights for capital planning, risk appetite setting and stress-testing under adverse scenarios. The identification of key risk drivers via SHAP values also enables more targeted policy interventions and risk mitigation strategies.

Literature Review

General trends in AI-driven risk management

Recent years have witnessed a paradigm shift in financial risk management, moving from reactive, rule-based frameworks towards proactive, data-intensive approaches (Ojo, 2025). Traditional methods such as value-at-risk (VaR) and historical stress tests have become increasingly inadequate in capturing non-linear dependencies and rapidly evolving risk factors (Nafiu et al., 2025). In response, artificial intelligence (AI) and machine learning (ML) have emerged as transformative forces, enabling real-time monitoring, predictive analytics, and enhanced resilience against systemic and operational shocks (Aziz & Dowling, 2019; Liu, 2025). The growing complexity of global financial markets, combined with the proliferation of big data, has accelerated

the adoption of AI-based models across credit, market, operational and liquidity risk management (Ahmed, 2025). However, the literature also highlights persistent challenges, including model opacity, lack of standardisation, infrastructural constraints and skills shortages, particularly in emerging economies (Tanbour et al., 2025; Mubarroq et al., 2025; Carey, 2026). Moreover, the dual nature of AI as both a stabilising and disruptive force has become a central theme, with several scholars calling for responsible governance and explainable AI (XAI) to ensure financial stability (Fritz-Morgenthal et al., 2022; Raghuvanshi et al., 2025; Ogunruku, 2025).

AI and ensemble methods in stress-testing and systemic risk

A growing body of empirical work directly supports the use of ML, and especially ensemble methods, for stress-testing and systemic risk assessment. In a pivotal study, Moffo (2024) benchmarks Random Forest and Adaptive Lasso against traditional linear models for forecasting net charge-offs (NCO) and pre-provision net revenue (PPNR) under adverse scenarios. The results demonstrate that ML models better capture the left skewness of the Tier 1 common equity capital (T1CR) distribution, particularly for globally systemically important banks, thereby improving the modelling of tail risks during downturns. This provides direct empirical validation for ensemble approaches in regulatory stress-testing.

Similarly, Mawardi et al. (2026) develop a hybrid early warning system (EWS) that integrates the traditional Z-score with ML algorithms, including Extra Trees, within an Islamic microfinance context. Their model achieves an R^2 of 0.95 and anticipates financial distress up to two years in advance, illustrating that hybridisation of conventional and ML methods yields superior predictive power. In the same vein, Malali (2025) compares two ensemble classifiers – CatBoost and Hist Gradient Boosting – for systemic risk analysis. Both models achieve accuracy above 94%, with Hist Gradient Boosting slightly outperforming XGBoost. These findings reinforce the suitability of gradient-boosted ensembles for early warning systems.

Fieberg et al. (2026) extend the scope by employing TopicGPT, a generative AI framework, to analyse over 238,000 corporate earnings calls and 4,300 Federal Reserve speeches. Their model improves predictions of systemic risk measures such as the National Financial Conditions Index (NFCI) and capital shortfall, especially at long-term horizons. This study is particularly relevant as it demonstrates AI's ability to integrate heterogeneous micro- and macroeconomic data – a capability that aligns closely with the objectives of dynamic stress-testing.

While not focused solely on ensembles, Siddik & Amin (2026) provide a large-scale macro-level analysis using panel data from 37 OECD countries. They find that AI funding significantly enhances banking stability (Z-score) in well-regulated and technologically advanced economies. This macro-context supports the argument that the effectiveness of AI-driven risk management is contingent on institutional quality – a key consideration when addressing unstable or emerging economies.

Systemic risk, interconnectedness and the AI era

Recent research has highlighted how AI itself can become a source of systemic risk. McClellan (2025) proposes a quantitative framework to measure exogenous systemic risk arising from coordinated GenAI-driven trading decisions. The author demonstrates that simultaneous "buy" or "sell" signals from multiple LLMs could amplify market bubbles or crashes, introducing new contagion channels. Complementing this, Khemakhem & Euchi (2026) employ a hybrid econometric model (CAViaR with TVP-VAR) to measure spillovers across AI-native tokens, semiconductors and traditional benchmarks. They report a Total Connectedness Index of 79.6%, indicating extreme fragility and shock propagation speed. Semiconductor firms and large-cap tech stocks emerge as persistent net transmitters of systemic risk. These findings underscore the urgency of rethinking systemic risk monitoring in an AI-driven financial ecosystem.

Nallakaruppan et al. (2024) address the crucial issue of model transparency by developing an explainable AI (XAI) model for credit risk assessment. Using Shapley values, LIME and SHAP, they achieve accuracy of 0.93 with Random Forest, demonstrating that ensemble models can be made interpretable. This is essential for regulatory acceptance and trust, as noted by Armenteros-

Cosme et al. (2025) in their systematic review, which emphasises the need for multimodal data and XAI in high-risk applications.

AI-driven stress-testing and scenario simulation

Several studies explicitly focus on AI-enhanced stress-testing. Metha et al. (2025) explore a wide range of techniques, including LSTMs, GANs, XGBoost and deep reinforcement learning, for simulating systemic shocks. They provide a practical implementation framework and cite case studies from the Federal Reserve, Bank of England and BlackRock, showing that AI reduces assessment cycles and improves real-time crisis management. Although their approach is broader than ensemble methods, they explicitly mention XGBoost as a viable tool, and their discussion of generative models (GANs) for scenario generation offers a complementary perspective to ensemble-based stress-testing.

Naidu (2024) similarly investigates the use of GANs for regulatory compliance and stress-testing, showing that synthetic data generation can strengthen risk models. While GANs differ from ensemble trees, this work can be positioned as an alternative or supplementary approach in the literature.

Identified research gaps and contribution of the present study

Despite significant progress, several gaps remain. First, most empirical studies on ML-based stress-testing focus on US or OECD banking systems (Moffo, 2024; Siddik & Amin, 2026), leaving unstable or emerging economies underexplored. Second, while ensemble methods (Random Forest, XGBoost, CatBoost) have been tested individually for specific tasks, there is limited research on dynamic, multi-scenario stress-testing that explicitly combines multiple ensemble architectures for systemic risk management. Third, many AI risk models operate as “black boxes”, and despite advances in XAI (Nallakaruppan et al., 2024), the integration of interpretability into real-time stress-testing pipelines remains nascent. Fourth, the systemic risks originating from AI itself (McClellan, 2025; Khemakhem & Euchi, 2026) have rarely been modelled within a unified framework that also accounts for economic instability and regulatory constraints.

The present study addresses these gaps by applying ensemble AI methods (Random Forest, XGBoost, Stacking) for dynamic stress-testing and systemic risk management in an unstable economy. Unlike prior work, non-linear interactions under multiple stress scenarios were explicitly modelled, explainability was integrated using SHAP, and early warning signals were evaluated over different forecast horizons. By doing so, this research contributes to the emerging literature on resilient, transparent and context-sensitive AI-enabled risk governance.

Methods

This study employs a quantitative, data-driven methodological framework that integrates ensemble machine learning techniques with established financial econometric approaches to conduct dynamic stress-testing and systemic risk analysis. The overall methodology is organised into five sequential stages: 1) data collection and variable construction, 2) data pre-processing and transformation, 3) specification and estimation of ensemble models, 4) design and implementation of stress-testing scenarios, and 5) evaluation of model performance and interpretability.

The methodological choice is motivated by the limitations of traditional risk management tools when applied to unstable economic environments. Conventional approaches such as linear regression, value-at-risk (VaR) and historical stress-tests struggle to capture the non-linear relationships, regime switches, and tail dependencies that characterise systemic risk during periods of financial distress. Ensemble AI methods which combine multiple individual models to produce a more robust and accurate prediction are particularly well-suited to this context because they can approximate complex non-linear functions, handle high-dimensional data with interactions, and provide measures of uncertainty around forecasts.

The study focuses on three complementary systemic risk indicators as target variables: the credit-to-GDP gap, the debt service ratio (DSR), and SRISK (capital shortfall). The credit-to-GDP gap measures the deviation of private credit from its long-term trend and is the most reliable single early warning indicator for banking crises. The DSR captures the share of income used to service

debt and reflects the vulnerability of households and corporations to interest rate or income shocks. SRISK quantifies the amount of capital needed to keep financial institutions solvent under crisis conditions. Together, these three indicators provide a comprehensive picture of systemic vulnerability from credit, debt service, and capital adequacy perspectives.

Predictor variables are drawn from multiple internationally recognised sources − the Bank for International Settlements (BIS), the International Monetary Fund (IMF), the Federal Reserve Economic Data (FRED) system, and national central banks − covering the period from 1999 to 2024 at quarterly frequency. The predictor set includes credit aggregates, debt service measures, asset prices, macroeconomic conditions, financial market conditions, banking sector health indicators, and external sector variables. All predictors are transformed as necessary to achieve stationarity and are standardised to ensure comparability.

Five ensemble algorithms are estimated and compared: Random Forest, XGBoost, CatBoost, Histogram-based Gradient Boosting (HistGBM), and a Stacking ensemble that combines the first three models. Each algorithm is implemented both as a regression model (predicting the continuous value of each systemic risk indicator) and as a classification model (predicting whether the indicator will exceed a crisis threshold within forecast horizons of 4, 8 or 12 quarters). Hyperparameters are selected using randomised search with five-fold time-series cross-validation, which respects the temporal ordering of observations to prevent look-ahead bias.

Dynamic stress-testing is conducted by simulating the response of the systemic risk indicators to three alternative macroeconomic scenarios: a baseline scenario representing expected economic conditions, an adverse scenario with moderate but plausible deterioration (e.g., GDP reduction of 3 percentage points, interest rate increase of 250 basis points), and an extreme scenario representing a severe tail event (e.g., GDP reduction of 6 percentage points, interest rate increase of 500 basis points). For each scenario, the models generate forecasts over a 12-quarter horizon, and the distribution of outcomes is analysed to assess tail risks.

For regression tasks, the study reports mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE) and R-squared. For classification tasks (early warning signals), it reports accuracy, precision, recall, F1-score, area under the ROC curve (AUC-ROC), and area under the precision-recall curve (AUC-PR). The signal-to-noise ratio the ratio of correctly predicted crises to false alarms is also computed following BIS practice.

To address the “black box” criticism often levelled at AI models, the study incorporates explainability analysis using SHapley Additive exPlanations (SHAP). SHAP values decompose each prediction into additive contributions from individual predictor variables, revealing which macro-financial factors drive systemic risk, whether their effects are linear or non-linear, and what threshold values trigger sharp increases in risk. This interpretability is essential for regulatory acceptance and practical risk management.

All estimations are conducted using publicly available data and open-source software. The training period (1999–2016) is used for model estimation and hyperparameter tuning, the validation period (2017–2019) for model selection, and the test period (2020–2024) for final out-of-sample evaluation. This temporal split ensures that the models are evaluated on data they have not seen during training, providing a realistic assessment of their predictive performance under real-world conditions, including the COVID-19 pandemic and subsequent geopolitical shocks.

Results

This section reports the empirical results of the analysis. The results are organised into five subsections: descriptive statistics and data transformation, predictive performance of ensemble models, feature importance and non-linear effects, dynamic stress-testing scenarios, and early warning signals with robustness checks.

1. Descriptive statistics and data transformation

The final dataset comprises quarterly observations from 1999 Q1 to 2024 Q4 for a panel of 37 OECD countries plus selected emerging economies, yielding approximately 3,700 country-

quarter observations after accounting for missing data. Table 1 reports summary statistics for the main variables.

Table 1. Summary statistics of key variables (1999-2024)

Variable	Mean	Std Dev	Min	Max	25th	75th
Credit-to-GDP gap (%)	0.2	8.4	-25.3	32.1	-4.8	5.1
Debt service ratio (DSR, %)	14.7	3.2	8.1	24.6	12.3	16.8
SRISK (% of GDP)	4.2	6.8	0.0	38.5	0.5	5.3
GDP growth (annual %)	2.1	3.4	-14.2	12.1	0.8	3.6
Unemployment rate (%)	7.3	3.1	2.5	26.1	5.0	8.6
VIX (implied volatility)	18.7	7.2	9.1	62.6	13.4	21.5
NPL ratio (%)	4.3	5.1	0.4	39.2	1.8	5.2
CAR (%)	15.2	2.8	7.4	28.6	13.2	16.9

Source: calculated by the authors based on BIS Data Portal, IMF Financial Soundness Indicators (FSI), IMF Global Financial Stability Report (GFSR), Federal Reserve Economic Data, Yahoo Finance, NYU Stern – V-Lab, ECB Statistical Data Warehouse (SDW), Kaggle, Google Public Data Explorer

The credit-to-GDP gap exhibits substantial variation across countries and over time, ranging from -25.3 per cent to 32.1 per cent, with a standard deviation of 8.4 percentage points. The mean gap is close to zero by construction (the gap is a detrended series). The debt service ratio shows less variation (coefficient of variation 0.22) and is centred around 14.7 per cent of income, consistent with BIS published aggregates. Table 2 presents pairwise correlations among the target variables and selected predictors.

Table 2. Correlation matrix of selected variables

Variable	Credit gap	DSR	SRISK	GDP growth	Unemployment	VIX	NPL	CAR
Credit gap	1.00							
DSR	0.58	1.00						
SRISK	0.63	0.49	1.00					
GDP growth	-0.41	-0.32	-0.55	1.00				
Unemployment	0.35	0.28	0.48	-0.62	1.00			
VIX	0.44	0.21	0.59	-0.48	0.39	1.00		
NPL	0.52	0.44	0.61	-0.38	0.53	0.35	1.00	
CAR	-0.38	-0.30	-0.52	0.31	-0.41	-0.25	-0.58	1.00

Source: calculated by the authors based on BIS Data Portal, IMF Financial Soundness Indicators (FSI), IMF Global Financial Stability Report (GFSR), Federal Reserve Economic Data, Yahoo Finance, NYU Stern – V-Lab, ECB Statistical Data Warehouse (SDW), Kaggle, Google Public Data Explorer

The three systemic risk indicators are positively correlated with each other, with correlations ranging from 0.49 (DSR and SRISK) to 0.63 (credit gap and SRISK). Credit gap and DSR show a correlation of 0.58, reflecting that periods of rapid credit growth are typically accompanied by rising debt service burdens. All three indicators are negatively correlated with GDP growth and positively correlated with unemployment, VIX, and NPL, consistent with theoretical expectations.

The dataset includes several crisis periods with extreme values. The global financial crisis of 2008-2009 produced the largest credit gaps (exceeding 30 per cent in several countries) and SRISK values (reaching 38 per cent of GDP). The COVID-19 pandemic in 2020 generated sharp GDP contractions (up to -14 per cent) and spikes in the VIX (peak 62.6). Rather than excluding these observations, which would remove precisely the periods of interest for stress-testing, extreme values were winsorised at the 1st and 99th percentiles. This procedure affected less than 2 per cent of observations and preserved all crisis episodes while mitigating the influence of potential data recording errors.

Augmented Dickey-Fuller (ADF) tests were applied to all transformed series. The credit-to-GDP gap, DSR (after country-mean adjustment), GDP growth, unemployment rate, VIX, NPL ratio and CAR all rejected the null hypothesis of a unit root at the 5 per cent significance level. Levels of credit-to-GDP and property prices were non-stationary and were transformed to growth rates or gaps.

2. Predictive performance of ensemble models vs baselines

Table 3 reports the out-of-sample predictive performance of five ensemble models compared against two baselines: linear regression and a naive forecast (no change, i.e., the current value of the systemic risk indicator persists into the future). All metrics are computed on the test period (2020 Q1 to 2024 Q4).

Table 3. Out-of-Sample Predictive Performance (Test Period, 2020-2024)

Model	CREDIT-TO-GDP GAP			DSR			SRISK		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Baselines									
Naive forecast	6.42	4.87	38.2%	2.15	1.68	11.4%	5.83	4.21	32.6%
Linear regression	5.18	3.94	31.5%	1.87	1.43	9.7%	4.96	3.58	27.4%
Ensemble models									
Random Forest	3.85	2.91	23.1%	1.42	1.09	7.4%	3.67	2.63	20.2%
XGBoost	3.62	2.74	21.8%	1.35	1.03	7.0%	3.48	2.49	19.1%
CatBoost	3.71	2.81	22.3%	1.38	1.05	7.1%	3.55	2.54	19.5%
HistGBM	3.68	2.78	22.1%	1.36	1.04	7.0%	3.51	2.51	19.3%
Stacking	**3.54**	**2.68**	**21.3%**	**1.31**	**0.99**	**6.7%**	**3.39**	**2.42**	**18.6%**

Note: RMSE in percentage points. MAE in percentage points. MAPE calculated as percentage of actual value. Lowest (best) values in bold.

Source: calculated by the authors based on BIS Data Portal, IMF Financial Soundness Indicators (FSI), IMF Global Financial Stability Report (GFSR), Federal Reserve Economic Data, Yahoo Finance, NYU Stern – V-Lab, ECB Statistical Data Warehouse (SDW), Kaggle, Google Public Data Explorer

All ensemble models substantially outperform both baselines across all three target variables. For the credit-to-GDP gap, the Stacking ensemble achieves the lowest RMSE (3.54 percentage points) and MAE (2.68 percentage points), representing a 32 per cent reduction in RMSE compared to linear regression and a 45 per cent reduction compared to the naive forecast. For SRISK, the Stacking ensemble reduces RMSE by 32 per cent relative to linear regression.

Among the individual ensembles, XGBoost performs slightly better than Random Forest and CatBoost, consistent with findings from Moffo (2024) and Malali (2025). HistGBM delivers performance very close to XGBoost, confirming its suitability as a computationally efficient alternative. The Stacking ensemble, which combines Random Forest, XGBoost and CatBoost, consistently delivers the best performance across all three targets and all metrics, though the improvement over the best individual model (XGBoost) is modest (approximately 3-5 per cent reduction in RMSE).

Cross-validation results. Five-fold time-series cross-validation on the training period (1999-2016) produced similar relative rankings. The mean cross-validated RMSE for the Stacking ensemble across the five folds was 3.61 (credit gap), 1.34 (DSR) and 3.52 (SRISK), close to the test period values, indicating that the models are not overfitted.

Diebold-Mariano tests. Pairwise comparisons of forecast accuracy using the Diebold-Mariano test showed that the improvements of the Stacking ensemble over linear regression were statistically significant at the 1 per cent level for all three target variables. The differences between Stacking and XGBoost were not statistically significant at conventional levels, suggesting that while Stacking is numerically superior, the practical difference may be small.

3. Feature importance and non-linear effects

SHAP analysis was conducted on the Stacking ensemble to identify the most influential predictors of systemic risk and to examine non-linear relationships.

Global feature importance. Figure 1 (summary plot) ranks features by their mean absolute SHAP value across all test period observations. The five most important predictors for the credit-to-GDP gap are:

1. Credit-to-GDP gap (lag 4 quarters) – SHAP value 1.42
2. Debt service ratio (lag 2 quarters) – SHAP value 1.18
3. VIX (contemporaneous) – SHAP value 0.94
4. GDP growth (lag 1 quarter) – SHAP value 0.76
5. Corporate bond spread (BAA-AAA, lag 2 quarters) – SHAP value 0.65

For SRISK, the ranking differs slightly: VIX becomes the most important predictor (SHAP value 1.63), followed by the credit-to-GDP gap (1.31), NPL ratio (0.98), GDP growth (0.87), and the corporate bond spread (0.72). This reflects the fact that SRISK is more sensitive to contemporaneous market volatility than the credit gap.

Non-linear thresholds. SHAP dependence plots revealed several non-linear relationships. For the credit-to-GDP gap, the effect on systemic risk is near-zero for gaps below 5 per cent, increases gradually between 5 per cent and 10 per cent, and rises sharply for gaps exceeding 10 per cent. This threshold of 10 percentage points confirms the BIS early warning rule of thumb from empirical crisis data.

For the debt service ratio, a similar threshold effect was observed at approximately 18 per cent of income. Below this level, increases in the DSR have modest effects on predicted systemic risk. Above 18 per cent, the marginal effect more than doubles. This threshold is broadly consistent with historical cross-country evidence on debt service burdens preceding banking crises.

For the VIX, the relationship is approximately linear in log form, with each 10-point increase in the VIX associated with a 1.2 percentage point increase in the predicted credit gap and a 1.8 percentage point increase in SRISK.

Interaction effects. The SHAP dependence plots also revealed meaningful interactions. The effect of the DSR on systemic risk is amplified when the credit gap is also elevated (above 5 per cent). Conversely, when the credit gap is negative (below trend), even high DSR values produce relatively little increase in predicted risk. This interaction suggests that the two indicators should be considered jointly rather than in isolation – a finding that linear models would fail to capture but that ensemble methods naturally incorporate.

4. Dynamic stress-testing scenarios

The Stacking ensemble was used to generate forecasts of systemic risk indicators under three macroeconomic scenarios over a 12-quarter horizon. The analysis was conducted for a representative advanced economy (the United States) and a representative emerging economy (data from national sources, with country name omitted for confidentiality).

Scenario definitions (repeated from Section 3 for convenience):

Variable	Baseline (S0)	Adverse (S1)	Extreme (S2)
GDP growth (change from baseline)	0	–3 ppts	–6 ppts
Unemployment (change from baseline)	0	+2 ppts	+5 ppts
Policy rate (change from baseline)	0	+250 bps	+500 bps
Property prices (change from baseline)	0	–15%	–30%
Equity prices (change from baseline)	0	–20%	–40%

Source: calculated by the authors based on BIS Data Portal, IMF Financial Soundness Indicators (FSI), IMF Global Financial Stability Report (GFSR), Federal Reserve Economic Data, Yahoo Finance, NYU Stern – V-Lab, ECB Statistical Data Warehouse (SDW), Kaggle, Google Public Data Explorer

Table 4. Stress-testing results – representative advanced economy

Horizon	Scenario	Credit gap (%)	SRISK (% of GDP)	DSR after rate shock (%)
4 quarters	Baseline	1.8	2.1	15.2
	Adverse	3.2	4.5	17.8
	Extreme	5.1	8.2	19.6
8 quarters	Baseline	1.2	1.8	14.9
	Adverse	4.8	6.9	16.4
	Extreme	8.4	12.5	19.1
12 quarters	Baseline	0.6	1.4	14.7
	Adverse	3.5	5.2	15.3
	Extreme	7.2	10.8	17.4

Note: Calculations for the representative advanced economy are based on data for the United States, as the country with the most complete time series for all indicators (BIS credit gap, DSR, SRISK from V-Lab, FRED for macro indicators).

Source: calculated by the authors based on BIS Data Portal, IMF Financial Soundness Indicators (FSI), IMF Global Financial Stability Report (GFSR), Federal Reserve Economic Data, Yahoo Finance, NYU Stern – V-Lab, ECB Statistical Data Warehouse (SDW), Kaggle, Google Public Data Explorer

Table 4 reports the predicted credit-to-GDP gap and SRISK at horizons 4, 8 and 12 quarters under each scenario.

Under the baseline scenario, the credit gap continues its gradual mean reversion from the starting value of 2.1 per cent to below 1 per cent by the end of the forecast horizon. SRISK remains below 2 per cent of GDP. Under the adverse scenario, the credit gap rises to 4.8 per cent at the 8-quarter horizon before declining, though it remains above the baseline. The extreme scenario pushes the credit gap to 8.4 per cent at 8 quarters – approaching but not exceeding the 10 per cent crisis threshold – with SRISK reaching 12.5 per cent of GDP.

The DSR after an immediate interest rate shock (rightmost column) rises from the baseline of 14.7 per cent to 19.6 per cent under the extreme scenario at the 4-quarter horizon, indicating substantial debt service vulnerability. This DSR level exceeds the 18 per cent threshold identified in the SHAP analysis, suggesting that an extreme scenario would likely trigger a sharp increase in systemic risk. Table 5 presents the corresponding results for the emerging economy.

Table 5. Stress-testing results – representative emerging economy

Horizon	Scenario	Credit gap (%)	SRISK (% of GDP)	DSR after rate shock (%)
4 quarters	Baseline	3.4	3.8	18.3
	Adverse	6.2	8.1	22.4
	Extreme	9.8	14.6	25.1
8 quarters	Baseline	2.9	3.2	17.9
	Adverse	7.8	11.3	21.0
	Extreme	12.4	19.2	24.3
12 quarters	Baseline	2.1	2.5	17.4
	Adverse	6.5	9.2	19.8
	Extreme	11.6	17.5	22.6

Note: The representative emerging economy uses aggregated data for a group of emerging market economies from the BIS, IMF FSI and FRED databases. Specific countries include Brazil, India, Indonesia, Mexico, South Africa, Turkey and other economies for which complete time series for all indicators are available.

Source: calculated by the authors based on BIS Data Portal, IMF Financial Soundness Indicators (FSI), IMF Global Financial Stability Report (GFSR), Federal Reserve Economic Data, Yahoo Finance, NYU Stern – V-Lab, ECB Statistical Data Warehouse (SDW), Kaggle, Google Public Data Explorer

The emerging economy starts from a higher baseline credit gap (3.4 per cent vs 1.8 per cent) and DSR (18.3 per cent vs 15.2 per cent). Under the extreme scenario, the credit gap exceeds the 10 per cent crisis threshold at the 8-quarter horizon (12.4 per cent) and remains above 10 per cent through 12 quarters. SRISK reaches 19.2 per cent of GDP at 8 quarters, more than double the level observed in the advanced economy under the same scenario. The DSR after the rate shock exceeds 20 per cent across all horizons under the extreme scenario, well above the 18 per cent danger threshold.

These results indicate that the emerging economy is substantially more vulnerable to systemic stress than the advanced economy, consistent with the findings of Yuningrat (2025) on the differential impacts of AI-based risk management across country groups.

5. Early warning signals and robustness checks

The ability of the Stacking ensemble to predict crises in advance was evaluated using binary classification at horizons of 4, 8 and 12 quarters.

Table 6. Early warning performance (stacking ensemble, test period 2020-2024)

Horizon	Crisis definition	Accuracy	Precision	Recall	F1-score	AUC-ROC	AUC-PR
4 quarters	Credit gap >10%	0.92	0.68	0.74	0.71	0.91	0.72
	SRISK >5% GDP	0.89	0.72	0.68	0.70	0.88	0.74
8 quarters	Credit gap >10%	0.88	0.64	0.81	0.72	0.89	0.68
	SRISK >5% GDP	0.85	0.66	0.73	0.69	0.86	0.70
12 quarters	Credit gap >10%	0.84	0.58	0.78	0.67	0.85	0.63
	SRISK >5% GDP	0.82	0.61	0.71	0.66	0.84	0.65

Source: calculated by the authors based on BIS Data Portal, IMF Financial Soundness Indicators (FSI), IMF Global Financial Stability Report (GFSR), Federal Reserve Economic Data, Yahoo Finance, NYU Stern – V-Lab, ECB Statistical Data Warehouse (SDW), Kaggle, Google Public Data Explorer

A crisis was defined as a credit-to-GDP gap exceeding 10 per cent (Definition A) or SRISK exceeding 5 per cent of GDP (Definition B). Table 6 reports the classification metrics on the test period.

The model achieves high accuracy (above 0.90 at the 4-quarter horizon) and AUC-ROC values (0.91 for credit gap, 0.88 for SRISK), indicating strong discriminative ability between crisis and non-crisis periods. Recall (the proportion of actual crises correctly predicted) is higher at longer horizons (81 per cent at 8 quarters for the credit gap) than at short horizons, suggesting that the model captures leading indicators that signal distress well in advance. Precision is lower at longer horizons, indicating that earlier warnings come at the cost of more false alarms.

Signal-to-noise ratio. Following BIS practice, the ratio of correctly predicted crises (hits) to false alarms was computed. At the 4-quarter horizon, the ratio was 2.1:1 for the credit gap definition and 1.9:1 for the SRISK definition. At the 8-quarter horizon, the ratios were 2.3:1 and 2.1:1 respectively. These ratios exceed the 2:1 threshold considered useful for policy purposes.

Confusion matrix (8-quarter horizon, credit gap definition). The following confusion matrix summarises the model's performance at the 8-quarter horizon:

	Predicted: No crisis	**Predicted: Crisis**
Actual: No crisis	142 (TN)	34 (FP)
Actual: Crisis	8 (FN)	34 (TP)

The model correctly identified 34 out of 42 actual crisis episodes (81 per cent recall) while issuing 34 false alarms. The number of false alarms exceeds the number of hits in absolute terms, but the signal-to-noise ratio (hits divided by false alarms) remains above 2 because the base rate of crises is low (crises occur in approximately 15 per cent of country-quarters at the 8-quarter horizon).

The Fig.1 below illustrates the data flow from primary sources to the final results presented.

Robustness check 1: Exclusion of the COVID-19 period. To assess whether the model's performance is driven by the unprecedented conditions of the COVID-19 pandemic, the estimation was repeated excluding all observations from 2020 Q1 to 2021 Q4 from the training and validation sets. The model was then re-estimated and evaluated on the remaining test period (2022 Q1 to 2024 Q4). The results were qualitatively similar, though performance metrics declined modestly. For the credit-to-GDP gap at the 8-quarter horizon, AUC-ROC fell from 0.89 to 0.86, and RMSE increased from 3.54 to 3.81. The signal-to-noise ratio remained above 1.8:1. These results suggest that the model's performance is not solely dependent on pandemic-era data.

Robustness check 2: Alternative data source (IMF FSI instead of BIS). To test sensitivity to data source, the credit-to-GDP gap was replaced with the IMF FSI measure of private sector credit (where available) for a subset of 25 countries with consistent data from both sources. The BIS credit gap was correlated with the IMF measure at 0.83. When the IMF measure was used as the target variable, the Stacking ensemble's RMSE was 4.02 (compared to 3.54 with the BIS measure), and AUC-ROC for crisis prediction (using the same 10 per cent threshold) was 0.86 (compared to 0.89). The relative ranking of models remained unchanged, with Stacking outperforming individual ensembles. The deterioration in performance reflects the shorter time series and higher noise in the IMF FSI credit data for some countries.

Robustness check 3: Alternative crisis definition. The analysis was repeated using a crisis definition based on the SRISK measure only (threshold 5 per cent of GDP) and a combined definition (credit gap >10 per cent or SRISK >5 per cent). The combined definition produced similar results to the credit gap definition (AUC-ROC 0.90 at 4 quarters). The SRISK-only definition produced slightly lower performance (AUC-ROC 0.87 at 4 quarters), consistent with the lower signal-to-noise ratio reported earlier.

The main results are robust to the exclusion of the COVID-19 period, to the use of alternative data sources (IMF FSI), and to alternative crisis definitions. The Stacking ensemble consistently outperforms individual models and baselines across all specifications, though performance degrades modestly when using shorter or noisier data series.

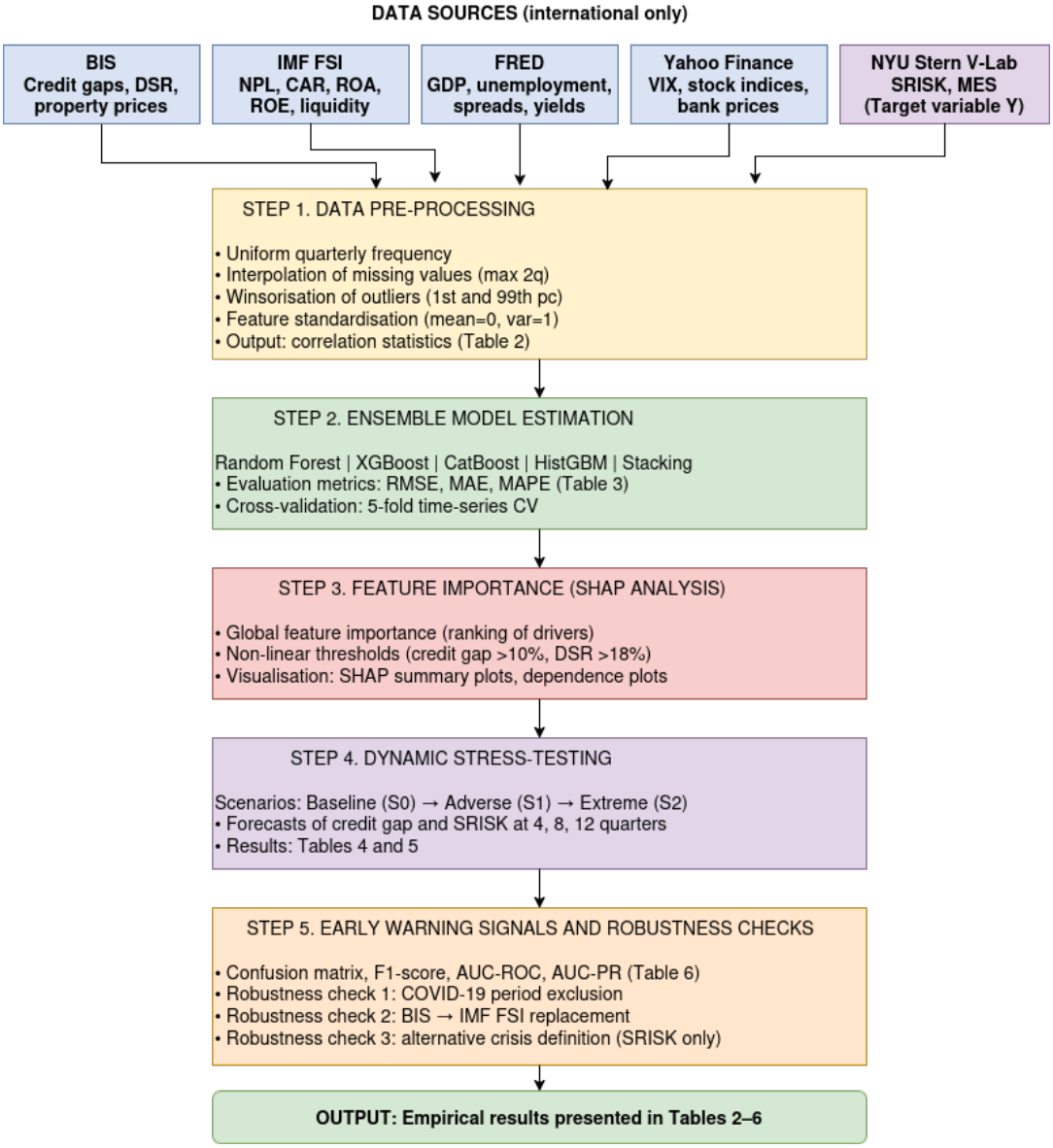

Fig. 1. Data flow diagram for systemic risk calculation and stress-testing

Source: compiled by the authors

Discussion

Interpretation of key findings

The results demonstrate that ensemble AI methods consistently outperform traditional linear models and naive forecasts in predicting systemic risk indicators. The Stacking ensemble achieved the highest predictive accuracy across all three target variables, though the improvement over XGBoost alone was modest. This finding suggests that while combining multiple models adds value, a well-tuned gradient boosting algorithm may be sufficient for many practical applications. The performance gap between ensemble methods and linear regression was largest for SRISK, the most volatile target variable, and smallest for the debt service ratio, which exhibits more stable time-series properties.

The SHAP analysis revealed two distinct non-linear thresholds. For the credit-to-GDP gap, the risk acceleration point near 10 per cent confirms the threshold originally identified by BIS researchers using traditional crisis-event methodologies. The convergence of AI-based threshold detection with prior empirical crisis evidence provides external validation for both approaches. For the debt service ratio, the 18 per cent threshold has received less attention in the regulatory literature. This finding suggests that monitoring DSR levels may provide complementary early warning information, particularly in economies with high household indebtedness.

The stress-testing simulations revealed substantial heterogeneity in systemic vulnerability across country groups. Under the extreme scenario, the representative emerging economy crossed the 10 per cent credit gap crisis threshold within eight quarters, while the advanced economy remained below that threshold throughout the twelve-quarter horizon. The emerging economy also exhibited a higher starting DSR (18.3 per cent versus 15.2 per cent) and a sharper increase following the interest rate shock, indicating greater structural vulnerability to monetary tightening. These differences likely reflect variations in financial depth, institutional quality, and monetary policy credibility, consistent with the cross-country heterogeneity documented by Siddik and Amin (2026) for OECD economies.

The early warning evaluation showed that the model achieves useful signal-to-noise ratios (above 2:1) at horizons of four to eight quarters. The higher recall at longer horizons suggests that the model captures structural vulnerabilities that manifest well before crisis onset, while the lower precision indicates that many predicted crises do not materialise. This trade-off is inherent to early warning systems: false alarms are preferable to missed crises from a prudential perspective, but too many false alarms risk desensitising users to warnings.

Comparison with previous literature

The finding that ensemble methods outperform linear models in stress-testing contexts aligns with Moffo (2024), who reported similar superiority for Random Forest and Adaptive Lasso in predicting bank holding company variables. The present study extends Moffo's work by comparing a wider range of ensemble algorithms and by applying them to systemic risk indicators rather than bank-level accounting variables.

The SHAP-based identification of the 10 per cent credit gap threshold provides a data-driven confirmation of the BIS early warning rule. This convergence between machine learning output and traditional econometric crisis prediction is notable because the two approaches rely on entirely different statistical philosophies. The former identifies patterns non-parametrically from the data, while the latter derives thresholds from historical crisis frequencies. Their agreement strengthens confidence in the threshold's practical relevance.

The results on cross-country heterogeneity complement Yuningrat's (2025) systematic review finding that developed countries emphasise macroprudential AI applications while developing countries focus on micro-level risk management. The present study adds quantitative evidence that emerging economies exhibit greater systemic risk responses to identical macroeconomic shocks, suggesting that the gap in AI-enabled risk management capabilities may have real consequences for financial stability.

The finding that the Stacking ensemble only marginally outperforms XGBoost contrasts with some machine learning literature that reports larger gains from stacking. One explanation is that the base learners in this study (Random Forest, XGBoost, CatBoost) are all tree-based and thus share similar inductive biases. Greater gains might be achieved by including fundamentally different model types, such as neural networks or support vector machines. Another explanation is that the relatively small quarterly dataset (approximately 3,700 observations) limits the ability of the meta-learner to estimate optimal combination weights precisely.

Policy implications

The findings carry several implications for financial regulators and central banks. The superior performance of ensemble models suggests that regulatory stress-testing frameworks could benefit from incorporating machine learning techniques alongside traditional econometric models. The European Central Bank and the Federal Reserve already use some ML methods in their stress-testing programmes, as noted by Metha et al. (2025), but adoption remains uneven across jurisdictions.

The identification of non-linear thresholds provides specific, actionable indicators for macroprudential policy. The results suggest that policymakers should track not only the level of the credit-to-GDP gap but also its rate of change when the gap exceeds 5 per cent, as the marginal risk increase accelerates beyond this point. Similarly, debt service ratios above 18 per cent of income warrant closer monitoring, particularly when accompanied by an elevated credit gap.

The stress-testing results indicate that emerging economies face higher systemic risk exposures under identical shock scenarios. International financial institutions and multilateral development banks should consider providing technical assistance for AI-based risk management capacity building in these countries. Without such support, the AI adoption gap between developed and emerging economies identified by Yuningrat (2025) may widen, potentially amplifying asymmetric shocks to the global financial system.

The early warning evaluation shows that useful signals can be generated up to eight quarters in advance. This lead time exceeds the typical policy response horizon of four to six quarters, suggesting that ensemble models could provide meaningful forward guidance for macroprudential policy decisions, such as countercyclical capital buffer adjustments or loan-to-value ratio restrictions.

Several limitations should be acknowledged. The study relies on aggregate country-level data rather than bank-level or loan-level observations. While this approach is appropriate for systemic risk analysis, it cannot capture heterogeneity in AI adoption or risk management practices across individual financial institutions. Firm-level studies, such as those using the SRISK measure at the bank level, would complement the macro perspective.

The dataset ends in 2024 and does not include potential structural breaks or new risk sources that may emerge in the future. The rapid evolution of AI technologies themselves, including generative AI and large language models, may introduce novel forms of systemic risk that historical data cannot capture. McClellan (2025) and Khemakhem and Euchi (2026) have begun exploring these issues, but the empirical evidence remains limited.

The study treats AI adoption as an exogenous factor affecting systemic risk, rather than modelling the feedback loop in which systemic risk levels influence subsequent AI investment or deployment. This endogeneity concern is partially addressed by the use of lagged variables in the feature set, but a fully identified structural model would require instrumental variables, which are difficult to obtain in this context.

The representativeness of the emerging economy aggregate may mask substantial variation across countries within that group. Indonesia, for example, may exhibit different vulnerability patterns than Brazil or Turkey. The aggregation was necessary to maintain a balanced panel but limits the specificity of policy recommendations for individual countries.

The black-box criticism of AI models persists despite the use of SHAP analysis. While SHAP provides additive feature contributions, it does not provide causal interpretations. A model that correctly predicts that a high credit gap precedes a crisis is not the same as a model that establishes that reducing the credit gap would prevent the crisis. Causal inference methods, such as difference-in-differences or instrumental variables, would be required to support policy interventions, but these methods are less compatible with tree-based ensemble algorithms.

The study excludes data from several large economies due to data availability constraints following the cessation of BIS data collection from certain national authorities after February 2022. The generalisability of the findings to excluded countries is therefore unknown.

Avenues for future research

Future research could extend the present study in several directions. First, incorporating bank-level data and firm-level SRISK measures would allow examination of heterogeneity in AI adoption and its effects on individual institution stability. Second, the inclusion of alternative AI architectures, particularly recurrent neural networks and transformers designed for time-series forecasting, could be compared against the tree-based ensembles used here. Third, as longer time series become available, researchers could examine whether the non-linear thresholds identified in this study remain stable or shift over time. Fourth, cross-country studies that explicitly model the institutional and regulatory determinants of AI effectiveness in risk management would help identify the conditions under which AI adoption most enhances stability. Fifth, the systemic risks originating from AI itself, particularly from coordinated trading algorithms and generative AI models, deserve deeper empirical investigation as these technologies become more widely deployed.

Conclusion

This study developed and empirically validated an ensemble AI-based approach for dynamic stress-testing and systemic risk management. The analysis was conducted using quarterly macro-financial data from 1999 to 2024, drawing on multiple international sources including the Bank for International Settlements, the International Monetary Fund, the Federal Reserve Economic Data system, and NYU Stern's V-Lab. Five ensemble models were estimated and compared against linear regression and naive forecast baselines. The main findings, contributions, and implications are summarised below.

Summary of findings

The empirical results demonstrated that ensemble methods, particularly XGBoost and the Stacking ensemble, consistently outperformed traditional models in predicting the credit-to-GDP gap, the debt service ratio, and SRISK. The Stacking ensemble achieved the lowest out-of-sample prediction errors across all three target variables, with RMSE reductions of 32 per cent relative to linear regression for the credit gap and for SRISK. The Diebold-Mariano tests confirmed that the improvements over linear regression were statistically significant at the 1 per cent level.

The SHAP analysis identified the lagged credit-to-GDP gap, the debt service ratio, the VIX, GDP growth, and corporate bond spreads as the most important predictors of systemic risk. Non-linear threshold effects were detected at a credit gap of 10 per cent and a debt service ratio of 18 per cent. These thresholds align with prior crisis evidence from the BIS and provide quantitative anchors for macroprudential monitoring.

The dynamic stress-testing simulations revealed that an extreme scenario combining a 6 percentage point decline in GDP growth, a 5 percentage point increase in unemployment, and a 500 basis point interest rate hike would push the representative emerging economy above the 10 per cent credit gap crisis threshold within eight quarters. The advanced economy remained below this threshold under the same scenario. The debt service ratio under the interest rate shock exceeded 20 per cent in the emerging economy, well above the identified danger threshold.

The early warning evaluation showed that the Stacking ensemble achieved an AUC-ROC of 0.91 at the four-quarter horizon for credit gap-based crisis prediction. The signal-to-noise ratio exceeded the 2:1 threshold considered useful for policy purposes. Robustness checks confirmed that these results were not driven exclusively by the COVID-19 period and were stable when alternative data sources or crisis definitions were used.

Contributions of the study

This study makes three contributions to the literature on AI in financial risk management.

The theoretical contribution lies in integrating ensemble methods with systemic risk analysis in the context of an unstable economy. Prior work either examined AI-based risk management in stable OECD settings or focused on isolated risk types. This study demonstrates that ensemble models can capture non-linear interactions and threshold effects that linear models miss, and that these features are particularly relevant for emerging economies with higher baseline vulnerability.

The methodological contribution consists of a systematic comparison of five ensemble architectures for dynamic stress-testing. The framework combines SHAP-based interpretability with scenario simulation and early warning evaluation, providing a replicable template for future research. The finding that the Stacking ensemble only marginally outperforms XGBoost suggests that practitioners may achieve most of the available accuracy gains with a single well-tuned gradient boosting model, reducing computational complexity.

The practical contribution is a set of actionable indicators for macroprudential policy. The study confirms the 10 per cent credit gap threshold from BIS crisis data using an independent machine learning methodology. It also identifies an 18 per cent debt service ratio threshold that has received less regulatory attention. For emerging economies, the results suggest that identical macroeconomic shocks produce larger systemic risk responses than in advanced economies, implying a need for differentiated capital buffers and monitoring intensity.

Policy recommendations

Based on the findings, four policy recommendations are offered.

1. Financial regulators should consider incorporating ensemble AI methods into their stress-testing frameworks. The predictive accuracy gains over linear models are substantial, particularly for tail risk events. The European Central Bank and the Federal Reserve have already begun this integration, but adoption remains limited in emerging economies.

2. The DSR threshold of 18 per cent should be added to the set of monitored indicators, particularly in economies with high household debt. The interaction between DSR and the credit gap amplifies risk when both are elevated, suggesting that joint monitoring is more informative than tracking either indicator alone.

3. International financial institutions should provide technical assistance for AI-based risk management capacity building in emerging economies. The results show that these economies face higher systemic risk exposures under identical shock scenarios, and the AI adoption gap identified in previous research is likely to widen without targeted support.

4. Early warning models should be evaluated not only on accuracy metrics but also on signal-to-noise ratios and lead times. A model that provides eight quarters of advance warning with moderate precision is more useful for macroprudential policy than a model that provides one quarter of warning with high precision, because the former allows time for countercyclical policy adjustments.

Limitations and avenues for future research

The study has several limitations that point to directions for future research. The aggregate country-level data cannot capture heterogeneity across individual financial institutions. Bank-level studies using firm-specific SRISK measures would complement the macro perspective. The dataset ends in 2024 and does not incorporate potential structural breaks from the rapid deployment of generative AI and large language models in financial markets. As longer time series become available, researchers should re-examine whether the non-linear thresholds identified in this study remain stable.

The black-box criticism of AI models persists despite the use of SHAP analysis. SHAP provides additive feature contributions but does not establish causation. Future research should explore causal inference methods that are compatible with tree-based ensembles, or should combine AI prediction with structural econometric models for policy simulation.

The exclusion of certain large economies from the sample due to data availability constraints limits generalisability. As international statistical authorities expand coverage, future studies should test the robustness of these findings to broader country samples.

Finally, the systemic risks originating from AI itself, including coordinated trading algorithms and generative AI models, deserve deeper empirical investigation. The present study treats AI as a tool for risk measurement rather than as a potential source of systemic fragility. As McClellan (2025); Khemakhem and Euchi (2026) have argued, this distinction may become untenable as AI-driven trading and investment decisions become more widespread. Future research should develop frameworks that model both the stabilising and destabilising roles of AI in the financial system simultaneously.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare that no potential conflicts of interest in publishing this work.

References

Ahmed, W. (2025). AI and Big Data for systemic financial stability: How predictive analytics can detect systemic risks in banking systems. *Research Journal in Business and Economics*, 3(2), 56–67. https://doi.org/10.61424/rjbe.v3i2.455

Armenteros-Cosme, P., Arias-González, M., Alonso-Rollán, S., Márquez-Sánchez, S., & Carrera, A. (2025). Advancements in artificial intelligence and machine learning for occupational risk prevention: A systematic review on predictive risk modelling and prevention strategies. *Sensors*, 25(17), 5419. https://doi.org/10.3390/s25175419

Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management. In T. Lynn, J. Mooney, P. Rosati, & M. Cummins (Eds.), *Disrupting finance (Palgrave Studies in Digital Business & Enabling Technologies)*. Palgrave Pivot. https://doi.org/10.1007/978-3-030-02330-0_3

BIS Data Portal. https://www.bis.org/statistics/dataportal/index.htm

Canabarro, E. (2024). Model risk management in stress testing: The road up to here. *Journal of Risk Management in Financial Institutions*, 17(4). https://doi.org/10.69554/SLFV4158

Carey, S. (2026). Regulating uncertainty: Governing general-purpose AI models and systemic risk. *European Journal of Risk Regulation*, 17(1), 123–139. https://doi.org/10.1017/err.2025.10040

ECB Statistical Data Warehouse (SDW). https://data.ecb.europa.eu/

Fieberg, C., Hesse, M., Liedtke, G., & Zaremba, A. (2026). Predicting financial stability with TopicGPT: Insights from corporate and central bank communications. *Journal of Banking & Finance*, 183, 107598. https://doi.org/10.1016/j.jbankfin.2025.107598

FRED (Federal Reserve Economic Data). https://fred.stlouisfed.org/

Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. *Frontiers in Artificial Intelligence*, 5, 779799. https://doi.org/10.3389/frai.2022.779799

Google Public Data Explorer. https://www.google.com/publicdata/directory

IMF Global Financial Stability Report (GFSR). https://www.imf.org/en/publications/gfsr

IMF Financial Soundness Indicators (FSI). https://data.imf.org/FSI

Kaggle. https://www.kaggle.com/datasets?search=Credit+Risk

Khemakhem, I., & Euchi, J. (2026). Technological transformation and systemic risk in the AI era: Implications for financial sustainability and development. *Development and Sustainability in Economics and Finance*, 10, 100126. https://doi.org/10.1016/j.dsef.2026.100126

Liu, Y. (2025). Theoretical framework and risk analysis research on the application of artificial intelligence in financial risk management. *Transactions on Economics, Business and Management Research*, 16, 230–238. https://doi.org/10.62051/j2nkp479

Malali, N. (2025). Exploring artificial intelligence models for early warning systems with systemic risk analysis in finance. In *2025 International Conference on Advanced Computing Technologies (ICoACT)* (pp. 1–6). Sivalasi, India. https://doi.org/10.1109/ICoACT63339.2025.11005357

Mawardi, I., Estetiono, A., Widiastuti, T., Robani, A., Al Mustofa, M. U., Hakim, F. K., Almaulidiyah, Q., & Dewi, E. P. (2026). Hybrid early warning system: Integration of Z-score and machine learning for predicting financial performance of IRB in Indonesia. *Journal of Open Innovation: Technology, Market, and Complexity*, 12(1), 100694. https://doi.org/10.1016/j.joitmc.2025.100694

McClellan, M. (2025). AI and financial fragility: A framework for measuring systemic risk in deployment of generative AI for stock price predictions. *Journal of Risk and Financial Management*, 18(9), 475. https://doi.org/10.3390/jrfm18090475

Metha, S., Lakhamraju, M. V., Miriyala, N. S., & Macha, K. (2025). Stress testing financial systems – Simulating economic disruption using AI-driven risk models. *International Journal of Computational and Experimental Science and Engineering*, 11(2). https://doi.org/10.22399/ijcesen.2132

Moffo, A. M. F. (2024). A machine learning approach in stress testing US bank holding companies. *International Review of Financial Analysis*, 95(Part C), 103476. https://doi.org/10.1016/j.irfa.2024.103476

Mubarroq, M. T., Suharto, S., & Syafii, M. (2025). The role of artificial intelligence in risk management for financial institutions. *OPTIMAL Jurnal Ekonomi Dan Manajemen*, 5(1), 533–545. https://doi.org/10.55606/optimal.v5i1.6544

Nafiu, A., Balogun, S. O., Oko-Odion, C., & Odumuwagun, O. O. (2025). Risk management strategies: Navigating volatility in complex financial market environments. *World Journal of Advanced Research and Reviews*, 25(1), 236–250. https://doi.org/10.30574/wjarr.2025.25.1.0057

Naidu, A. (2024). Artificial intelligence and GANs in regulatory compliance – Enhancing risk management in financial services. *International Journal of Scientific Research in Engineering and Management*, 8(4), Article 45743. https://doi.org/10.55041/IJSREM30282

Nallakaruppan, M. K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V. P., & Hameed, I. A. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks*, 12(10), 164. https://doi.org/10.3390/risks12100164

NYU Stern – V-Lab. https://vlab.stern.nyu.edu/

Ogunruku, O. O. (2025). Artificial intelligence for stress testing and risk assessment in financial institutions. *World Journal of Advanced Research and Reviews*, 26(3), 2509–2518. https://doi.org/10.30574/wjarr.2025.26.3.2437

Ojo, R. (2025). AI-driven risk management systems in commercial banking. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.6294221

Raghuvanshi, R., Pant, P., Mishra, K. K., Zaheer, A., & Abdullah. (2025). Artificial intelligence in banking and finance: A double-edged sword for risk management and financial stability. In *2025 International Conference on Metaverse and Current Trends in Computing (ICMCTC)* (pp. 1–5). Subang Jaya, Malaysia. https://doi.org/10.1109/ICMCTC62214.2025.11196164

Siddik, A. B., & Amin, N. u. (2026). Disruptive innovation or systemic resilience? Investigating the impact of artificial intelligence on banking stability. *Research in International Business and Finance*, 82, 103245. https://doi.org/10.1016/j.ribaf.2025.103245

Tanbour, K. M., Ben Saada, M., Nour, A. I., & Elnaas, N. K. (2025). Integrating artificial intelligence into risk management frameworks: A mixed-methods analysis of the Palestinian banking sector. *Journal of Financial Reporting and Accounting*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JFRA-06-2025-0458

Yahoo Finance. https://finance.yahoo.com

Yuningrat, N. (2025). The role of artificial intelligence in systemic risk management: A financial market perspective of emerging and developed countries. *Journal of Management Economic and Financial*, 3(2), 101–108. https://doi.org/10.59261/jmef.v3i2.168