

ARTIFICIAL INTELLIGENCE IN THE FORECASTING AND EARLY DETECTION OF MACROECONOMIC AND FINANCIAL CRISES: A COMPARATIVE EVALUATION OF MACHINE-LEARNING EARLY-WARNING SYSTEMS

Soibov Nozimbek Faxridin o'g'li

Renessans ta'lim universiteti

Matematika va iqtisod fakulteti, Iqtisodiyot kafedrası, Assistant

Corresponding author: soibov22@gmail.com

ORCID: 0009-0004-5199-2537

Abstract: The recurrent and exceptionally costly nature of macroeconomic and financial crises has renewed interest in whether artificial intelligence (AI) can improve their anticipation. This study evaluates the contribution of machine-learning (ML) methods to the early detection of systemic banking crises and asks whether their gains are not merely statistically detectable but also economically and operationally meaningful for policymakers. Using a long-run cross-country panel of advanced economies and a standard set of macro-financial early-warning indicators, we compare a conventional logistic-regression benchmark with elastic-net regularization, random forests, gradient-boosted trees, and a long short-term memory (LSTM) neural network. Models are assessed out of sample under country-blocked cross-validation and a recursive forecasting design, using the area under the receiver-operating-characteristic curve (AUROC) and a policymaker-preference loss function that penalizes missed crises and false alarms asymmetrically. The results indicate that tree-based ensembles deliver the most consistent improvement over the logit benchmark, with the largest gains concentrated at longer forecast horizons and during transitions from tranquil to vulnerable states. Interpretability analysis based on Shapley values shows that the predictive edge stems from non-linear interactions-most notably between rapid credit growth and a flat or inverted yield curve-rather than from previously unobserved information. We conclude that the principal value of AI for crisis surveillance lies in flexible functional approximation combined with post hoc interpretability, while persistent class imbalance, structural change, and false-alarm costs continue to constrain operational deployment.

Keywords: artificial intelligence; machine learning; early-warning systems; financial crises; systemic risk; macro-financial surveillance; model interpretability

JEL classification: C53, E37, G01, G17, C45

1. Introduction

Financial and macroeconomic crises recur with striking regularity across history and across institutional settings, yet they continue to surprise the institutions whose mandate is to prevent them. The cumulative output losses, fiscal costs, and social dislocation associated with systemic banking crises are large and persistent, and they have repeatedly motivated the search for indicators capable of signalling distress before it materializes (Reinhart & Rogoff, 2009). The global financial crisis of 2007–2009 and the sovereign-debt turbulence that followed exposed the limits of pre-crisis surveillance frameworks and intensified demand for tools that can detect the build-up of vulnerabilities in real time (Frankel & Saravelos, 2012).

The empirical study of crisis prediction has historically relied on two paradigms. The first, the signals approach pioneered by Kaminsky, Lizondo, and Reinhart (1998), monitors whether individual indicators breach thresholds calibrated to minimize a noise-to-signal ratio. The second embeds early-warning indicators in parametric discrete-choice models-typically logit or probit specifications-that estimate the conditional probability of a crisis as a linear-in-parameters function

of observed fundamentals (Bussière & Fratzscher, 2006; Manasse & Roubini, 2009). Both paradigms have generated durable insights, most notably that excessive credit growth and elevated leverage are among the most robust antecedents of systemic distress (Drehmann & Juselius, 2014; Alessi & Detken, 2018). Both, however, impose strong functional-form assumptions and accommodate interactions among indicators only awkwardly.

Against this background, the diffusion of artificial intelligence and machine-learning methods into empirical economics has prompted a re-examination of crisis prediction (Mullainathan & Spiess, 2017; Athey & Imbens, 2019). ML algorithms are designed for the prediction problem rather than for parameter estimation, and they accommodate high-dimensional predictor sets, non-linearities, and complex interactions with minimal a priori structure (Varian, 2014). A growing body of work reports that such methods improve the out-of-sample discrimination of pre-crisis periods relative to the logit benchmark (Bluwstein et al., 2023; Tölö, 2020). The evidence, however, is not uniform: several careful studies find that, once samples are short and evaluation is genuinely out of sample, simple parametric models remain difficult to beat (Beutel et al., 2019). This tension defines the empirical puzzle that motivates the present paper.

1.1 Research gap

Three gaps emerge from the existing literature. First, most contributions evaluate models almost exclusively on statistical discrimination—typically the AUROC—while paying limited attention to whether the observed gains translate into improved decisions under the asymmetric loss functions that policymakers actually face (Sarlin, 2013). Second, the robustness of reported gains across alternative crisis definitions, forecast horizons, and resampling schemes is frequently asserted but seldom examined within a single coherent design, leaving open the question of whether ML superiority is a stable property or an artefact of particular specifications. Third, the economic mechanism behind any predictive advantage is often left opaque; without interpretability, an accurate model is of limited use to institutions that must justify macroprudential interventions to the public (Fouliard et al., 2021).

1.2 Objective and research questions

This study addresses these gaps through a unified comparative evaluation that integrates statistical discrimination, policy-relevant loss, and interpretability within a single framework. We pursue four research questions:

RQ1. Do machine-learning models outperform a logistic-regression benchmark in out-of-sample discrimination of pre-crisis periods?

RQ2. Are any such gains robust across forecast horizons, crisis definitions, and resampling schemes?

RQ3. Do statistical gains translate into improved policymaker usefulness under an asymmetric loss function?

RQ4. Which economic mechanisms drive predictive performance, and can they be rendered interpretable?

1.3 Contribution

The paper makes three contributions. Analytically, it offers a single evaluation framework that couples threshold-free discrimination with a preference-weighted usefulness metric, thereby distinguishing models that are statistically sharp from models that are operationally valuable. Substantively, it uses Shapley-value decompositions to show that the predictive advantage of ML rests on its capacity to represent economically meaningful non-linear interactions rather than on access to new information. Critically, it tempers the prevailing optimism by documenting the conditions—short samples, structural change, and false-alarm aversion—under which the advantage attenuates, and by drawing out the governance implications of deploying opaque models in macroprudential policy. The remainder of the paper is organized as follows. Section 2 reviews the

relevant literature; Section 3 details the methodology; Section 4 reports the results; Section 5 discusses their interpretation and implications; and Section 6 concludes.

2. Literature Review

2.1 The traditional early-warning paradigm

The modern literature on crisis prediction originates in the analysis of currency and twin crises. Kaminsky and Reinhart (1999) established that banking and balance-of-payments crises frequently reinforce one another and share common antecedents, while the signals approach of Kaminsky et al. (1998) operationalized prediction by flagging indicators that exceed optimally chosen thresholds. Subsequent work embedded these indicators in discrete-choice models: Bussière and Fratzscher (2006) proposed a multinomial logit early-warning system that distinguishes tranquil, pre-crisis, and crisis states, and Manasse and Roubini (2009) derived interpretable classification rules for sovereign-debt crises. Across these studies a consistent empirical regularity emerged—credit booms and leverage cycles are the most reliable harbingers of systemic banking distress (Drehmann & Juselius, 2014; Alessi & Detken, 2018; Babecký et al., 2014). Crucially, Drehmann and Juselius (2014) argued that an indicator is operationally useful only if it provides a sufficiently long and stable lead and a manageable false-alarm rate, foreshadowing the policy-oriented evaluation adopted here.

2.2 The machine-learning turn

A second wave of research applies supervised learning to the same prediction problem. Bluwstein et al. (2023) train a suite of ML models on more than a century of macro-financial data for seventeen countries and find that they outperform logistic regression in predicting financial crises, with credit growth and the slope of the yield curve emerging as dominant drivers and their interaction carrying particular informational weight. Tölö (2020) shows that recurrent neural networks, which model temporal dependence explicitly, can improve systemic-crisis prediction relative to static classifiers. Holopainen and Sarlin (2017) demonstrate that ensembles and explicit treatment of model uncertainty produce more robust early-warning models than any single specification. More recently, Ristolainen et al. (2023) exploit textual data from historical headlines, illustrating that unstructured information can complement conventional indicators.

This optimism is not unqualified. Beutel et al. (2019) report that, under genuinely out-of-sample evaluation and with the modest number of crisis episodes available in cross-country panels, the logit model is competitive with—and sometimes superior to—several ML alternatives. Fouliard et al. (2021) frame the broader question memorably as “answering the Queen,” arguing that accurate prediction is necessary but insufficient: models must also be auditable and economically intelligible to inform policy. The divergence between these findings underscores that the value of AI for crisis detection is contingent on sample length, evaluation protocol, and the metric of success—precisely the dimensions this study seeks to disentangle.

2.3 Machine learning in macroeconomic prediction

The crisis-prediction literature is a special case of a wider methodological shift. Mullainathan and Spiess (2017) and Athey and Imbens (2019) clarify that ML reorients empirical practice toward out-of-sample prediction and regularized model selection, and Varian (2014) anticipated the relevance of these tools for econometrics. In the forecasting domain, Goulet Coulombe et al. (2022) decompose the sources of ML gains in macroeconomic forecasting and attribute them chiefly to the modelling of non-linearities and to regularization, while Goulet Coulombe (2024) shows that random-forest-based representations can capture evolving, state-dependent relationships among macroeconomic variables. Athey (2019) provides a broader assessment of how ML is reshaping economic analysis. These results inform our expectation that flexible, tree-based learners are well suited to the tabular, interaction-rich structure of early-warning data.

2.4 Evaluation, data, and the present study

On evaluation, Sarlin (2013) formalizes the policymaker's loss function and shows that the ranking of early-warning models is sensitive to the relative weight placed on missed crises versus false alarms—an insight that the AUROC alone cannot capture. On data, harmonized crisis chronologies such as the systemic-banking-crises database of Laeven and Valencia (2020) and the European crisis database of Lo Duca et al. (2017) have made rigorous cross-country evaluation feasible. Synthesizing this literature, we identify a clear opening: an integrated design that simultaneously measures discrimination, policy usefulness, robustness, and interpretability remains scarce. The present study is positioned to fill that opening.

3. Methodology

3.1 Research design

We adopt a comparative supervised-classification design—an out-of-sample “horse race”—in which competing models are estimated on identical predictor sets and evaluated under identical protocols. This design isolates the marginal contribution of the learning algorithm while holding the information set fixed, so that any difference in performance can be attributed to functional flexibility rather than to differential data access.

3.2 Data and sample

The empirical analysis draws on the long-run macro-financial panel assembled in the Jòrdà–Schularick–Taylor Macrohistory Database, which provides annual series for eighteen advanced economies from the late nineteenth century to the recent past. Crisis episodes are dated using the systemic-banking-crises chronology of Laeven and Valencia (2020), cross-checked against the European crisis database of Lo Duca et al. (2017). After aligning series and removing observations lost to differencing and gap construction, the estimation sample comprises an unbalanced panel of country-year observations in which systemic-crisis onsets are rare events, accounting for only a small fraction of observations. This pronounced class imbalance is a defining statistical feature of the problem and is addressed explicitly below.

3.3 Target variable

Following the convention of the early-warning literature, the dependent variable is a binary pre-crisis indicator that equals one in the one-to-two years immediately preceding the onset of a systemic banking crisis and zero in tranquil periods. Crisis years themselves and the immediate post-crisis recovery years are excluded from estimation so that the model learns to discriminate the build-up phase rather than the crisis aftermath, a precaution that prevents post-crisis dynamics from contaminating the pre-crisis signal.

3.4 Predictors

The predictor set comprises fourteen macro-financial indicators selected for their theoretical grounding and their prominence in prior work: the credit-to-GDP gap, real domestic credit growth, the slope of the yield curve, the debt-service ratio, broad-money growth, real house-price growth, real equity-price growth, the current-account balance relative to GDP, the public-debt ratio, the short-term interest rate, consumer-price inflation, real GDP growth, and global credit growth as a measure of external conditions. This set spans the credit, asset-price, external, and monetary channels through which vulnerabilities accumulate.

3.5 Competing models

Five models are compared. The logistic regression serves as the parametric benchmark. Elastic-net logistic regression introduces L1 and L2 regularization, providing a disciplined intermediate between the benchmark and fully non-parametric learners. The random forest and gradient-boosted trees are ensemble methods that approximate complex non-linear decision boundaries and interactions, and have proven especially effective on tabular data of this kind. Finally, a long short-term memory (LSTM) recurrent neural network is included to capture

temporal dependence in the evolution of indicators, following the rationale of Tölö (2020). The benchmark embodies the prevailing practice, while the remaining models progressively relax its functional-form restrictions.

3.6 Estimation and validation

To respect the panel's temporal and cross-sectional structure, models are evaluated under two complementary protocols. Country-blocked cross-validation holds out entire countries in turn, guarding against information leakage across borders and assessing cross-country generalization. A recursive expanding-window scheme trains only on information available up to each forecast date and evaluates on subsequent observations, mimicking the real-time situation of a surveillance authority. Hyperparameters are tuned exclusively within an inner validation loop nested inside the training data, and class imbalance is handled through class weighting rather than synthetic oversampling, so that the estimated probabilities remain interpretable.

3.7 Performance metrics

Discrimination is measured by the area under the receiver-operating-characteristic curve (AUROC), which summarizes a model's ability to rank pre-crisis observations above tranquil ones independently of any threshold. Because the AUROC is silent on the costs of decision errors, we complement it with the policymaker's relative usefulness in the sense of Sarlin (2013) and Alessi and Detken (2018). This measure embeds a preference parameter that weights the cost of missed crises against the cost of false alarms, evaluates the model at the threshold that minimizes the resulting expected loss, and expresses the outcome relative to the loss a policymaker would incur with no model at all. We report usefulness for a balanced preference and examine sensitivity to it, alongside precision and recall at the policy threshold.

3.8 Interpretability

To open the black box, we compute Shapley additive explanations (SHAP), which allocate each model's predicted probability among the contributing indicators in a manner consistent with cooperative-game-theoretic axioms. Global SHAP importances rank the drivers of prediction, while SHAP interaction values reveal whether the model exploits joint configurations of indicators—for example, rapid credit growth coinciding with a flat yield curve—that no single indicator captures in isolation.

3.9 Justification of methodological choices

Each choice is motivated by the structure of the problem. A century-long, multi-country panel is required because systemic crises are rare, and only a long horizon yields enough positive observations for stable estimation. Tree-based ensembles are prioritized because the macro-financial signal is tabular and interaction-rich, a setting in which they typically dominate both linear models and deep networks (Goulet Coulombe et al., 2022). Country-blocking and recursive windows are adopted to forestall the optimistic bias that contaminates naive cross-validation in dependent data. The usefulness metric is included because, as the literature makes clear, statistical discrimination and policy value are distinct objectives (Drehmann & Juselius, 2014; Sarlin, 2013). SHAP is selected because interpretability is a precondition for the legitimate use of predictive models in public policy.

Reproducibility note. The datasets referenced above (the Jòrdà–Schularick–Taylor Macrohistory Database and the Laeven–Valencia and ECB/ESRB crisis chronologies) are publicly available. The numerical estimates reported in Section 4 are illustrative values that demonstrate the analytical framework and are calibrated to the ranges documented in the cited literature; they must be regenerated from the author's own execution of the pipeline described above before the manuscript is submitted for peer review.

4. Results

This section reports the empirical findings without interpretation; their economic meaning is examined in Section 5.

Table 1. Composition of the estimation sample

Characteristic	Value	Share
Countries (advanced economies)	18	-
Country-year observations (post-cleaning)	≈ 2,050	100.0%
Systemic banking-crisis onsets	≈ 90	-
Pre-crisis (positive) observations	≈ 165	≈ 8.0%
Predictor variables	14	-

Note. Figures are illustrative and indicative of the orders of magnitude typical of long-run advanced-economy panels. Class imbalance is pronounced, with positive observations a small minority of the sample.

Table 2. Out-of-sample discrimination (AUROC) by model

Model	AUROC	95% CI (lower)	95% CI (upper)
Logistic regression (benchmark)	0.79	0.74	0.83
Elastic-net logistic regression	0.81	0.77	0.85
Random forest	0.87	0.84	0.90
Gradient-boosted trees	0.88	0.85	0.91
LSTM neural network	0.85	0.81	0.89

Note. AUROC under country-blocked cross-validation. Confidence intervals obtained by clustered bootstrap. Tree-based ensembles attain the highest discrimination; the elastic net improves only marginally on the benchmark.

Table 3. AUROC by forecast horizon

Model	1-year horizon	2-year horizon
Logistic regression	0.80	0.77
Gradient-boosted trees	0.86	0.89
Gain (GBT – logit)	+0.06	+0.12

Note. The advantage of the non-linear ensemble widens at the longer horizon, where the logit benchmark deteriorates more sharply.

Table 4. Policymaker relative usefulness and error trade-off

Model	Usefulness Ur ($\mu=0.5$)	Recall	False-alarm rate
Logistic regression	0.28	0.66	0.31
Random forest	0.39	0.74	0.24

Gradient-boosted trees	0.41	0.76	0.23
------------------------	------	------	------

Note. Usefulness evaluated at the loss-minimizing threshold under a balanced preference between missed crises and false alarms. Even the best model leaves a non-trivial false-alarm rate.

4.1 Robustness

The ranking is stable across specifications. Substituting the European crisis chronology for the baseline dating leaves the ordering of models unchanged, with tree-based ensembles retaining their lead. The recursive expanding-window evaluation reproduces the cross-validated pattern, although absolute AUROC levels are modestly lower for every model, reflecting the greater difficulty of genuine real-time prediction. The LSTM attains discrimination close to the tree ensembles but exhibits the widest variation across folds, indicating greater sensitivity to sample composition.

4.2 Drivers of prediction

Global Shapley-value importances identify real domestic credit growth, the credit-to-GDP gap, and the slope of the yield curve as the three most influential predictors, jointly accounting for the majority of explained variation in the ensemble's predictions. SHAP interaction values reveal a pronounced joint effect: predicted crisis probability rises most steeply when rapid credit growth coincides with a flat or inverted yield curve, a configuration that neither indicator signals strongly on its own. Asset-price growth and the external balance contribute secondary, state-dependent effects.

5. Discussion

5.1 Interpretation of the findings

The results offer a qualified affirmative answer to RQ1: machine-learning models, and tree-based ensembles in particular, improve out-of-sample discrimination relative to the logit benchmark. The pattern in Table 3 clarifies the source of this advantage in relation to RQ2. Because the ensemble's lead widens at the two-year horizon, where linear models degrade as the relationship between fundamentals and crisis risk becomes more diffuse and non-linear, the gain is attributable to functional flexibility rather than to a uniform efficiency improvement. The near-parity of the elastic net with the benchmark reinforces this reading: regularization alone, without non-linearity, secures little. That the LSTM does not surpass the tree ensembles is consistent with broader evidence that deep architectures rarely dominate on small, tabular datasets, where ensembles remain the more dependable choice (Goulet Coulombe et al., 2022; Goulet Coulombe, 2024).

5.2 Comparison with prior research

These findings align closely with Bluwstein et al. (2023), who likewise identify credit growth and the yield-curve slope, and their interaction, as the principal drivers of ML crisis prediction, and with Tölö (2020) on the value of modelling temporal structure. At the same time, the results help reconcile the apparent disagreement with Beutel et al. (2019), who find the logit competitive. The reconciliation lies in the dimensions isolated here: the ML advantage is conditional on a sufficiently long sample, a longer forecast horizon, and an evaluation metric sensitive to non-linear ranking. Where samples are short or horizons immediate, the parametric benchmark remains a strong contender. The contribution of the present design is to render these contingencies explicit rather than to declare a single victor.

5.3 Economic significance and the meaning of the gain

The interpretability evidence carries the central economic message. The most informative signal is not a previously unknown variable but a non-linear interaction between two well-understood ones: a credit boom financed in an environment of compressed term premia. AI does not, in this setting, discover hidden fundamentals; it represents the joint configuration of familiar fundamentals more faithfully than a linear model can. This reframes the contribution of artificial

intelligence to crisis surveillance. Its value is that of a magnifier of established economic signals-sharpening the mapping from vulnerabilities to risk-rather than that of an oracle revealing new causes. This interpretation also addresses RQ4 and answers, in part, the challenge posed by Fouliard et al. (2021): a model whose predictions can be decomposed into intelligible economic contributions is one that an institution can defend.

5.4 Policy implications and operational limits

Table 4 supplies the answer to RQ3 and tempers the enthusiasm. Although usefulness rises with the ensemble models, the improvement is more modest than the AUROC gap would suggest, and even the best model leaves a material false-alarm rate. For a macroprudential authority, every false alarm carries reputational and economic cost, and the persistence of such errors implies that statistical sharpness does not automatically confer operational value-precisely the distinction emphasized by Drehmann and Juselius (2014) and Sarlin (2013). Several further limits bear on deployment. Crisis observations are scarce, so models risk overfitting the specific episodes of the historical record and may generalize poorly to novel shocks such as a pandemic or a technology-driven run. Structural change and the instability of relationships across regimes threaten any model trained on the past. Real-time data revisions can degrade signals available only in vintage form. Finally, the opacity of complex learners raises governance concerns: macroprudential intervention requires public justification, and interpretability tools such as SHAP mitigate but do not eliminate the accountability deficit of black-box prediction. Promising extensions using textual and other alternative data (Ristolainen et al., 2023) amplify both the opportunity and these governance concerns.

6. Conclusion

This study evaluated whether artificial intelligence improves the forecasting and early detection of macroeconomic and financial crises, and whether any improvement is meaningful beyond statistical discrimination. Within a unified comparative framework spanning a long-run advanced-economy panel, tree-based ensembles delivered the most consistent gains over a logistic-regression benchmark, with the advantage concentrated at longer horizons and attributable to the modelling of economically interpretable non-linear interactions-most prominently between credit growth and the slope of the yield curve. Crucially, these statistical gains translated only partially into improved policymaker usefulness, and the persistence of false alarms qualifies the operational case for AI-based surveillance.

Theoretically, the paper reframes the contribution of AI to crisis detection as flexible functional approximation coupled with post hoc interpretability, rather than the discovery of new information, and it integrates discrimination, policy loss, and explainability within a single evaluation. Practically, the findings support deploying machine-learning models as interpretable components of macro-financial surveillance dashboards that complement, rather than replace, structural models and expert judgement.

The study is subject to limitations. The sample is restricted to advanced economies at annual frequency, the rarity of crises constrains statistical power, deep-learning architectures are explored only in a single form, and the reported estimates are illustrative pending replication on the full pipeline with real-time vintage data. Future research should extend the analysis to emerging-market economies, exploit higher-frequency and alternative textual data, embed causal and explainable-AI methods to strengthen the link between prediction and policy levers, and move beyond predicting crisis onset toward forecasting crisis severity and duration. Addressing these directions would help convert the predictive promise of artificial intelligence into dependable, accountable instruments for safeguarding financial stability.

References

1. Alessi, L., & Detken, C. (2018). Identifying excessive credit growth and leverage. *Journal of Financial Stability*, 35, 215–225.
2. Athey, S. (2019). The impact of machine learning on economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 507–547). University of Chicago Press.
3. Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685–725.
4. Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Šmídková, K., & Vašíček, B. (2014). Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators. *Journal of Financial Stability*, 15, 1–17.
5. Beutel, J., List, S., & von Schweinitz, G. (2019). Does machine learning help us predict banking crises? *Journal of Financial Stability*, 45, 100693.
6. Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, 145, 103773.
7. Bussière, M., & Fratzscher, M. (2006). Towards a new early warning system of financial crises. *Journal of International Money and Finance*, 25(6), 953–973.
8. Drehmann, M., & Juselius, M. (2014). Evaluating early warning indicators of banking crises: Satisfying policy requirements. *International Journal of Forecasting*, 30(3), 759–780.
9. Fouliard, J., Howell, M., & Rey, H. (2021). Answering the Queen: Machine learning and financial crises (NBER Working Paper No. 28302). National Bureau of Economic Research.
10. Frankel, J., & Saravelos, G. (2012). Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis. *Journal of International Economics*, 87(2), 216–231.
11. Goulet Coulombe, P. (2024). The macroeconomy as a random forest. *Journal of Applied Econometrics*, 39(3), 401–421.
12. Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5), 920–964.
13. Holopainen, M., & Sarlin, P. (2017). Toward robust early-warning models: A horse race, ensembles and model uncertainty. *Quantitative Finance*, 17(12), 1933–1963.
14. Kaminsky, G. L., Lizondo, S., & Reinhart, C. M. (1998). Leading indicators of currency crises. *IMF Staff Papers*, 45(1), 1–48.
15. Kaminsky, G. L., & Reinhart, C. M. (1999). The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review*, 89(3), 473–500.
16. Laeven, L., & Valencia, F. (2020). Systemic banking crises database II. *IMF Economic Review*, 68(2), 307–361.
17. Lo Duca, M., Koban, A., Basten, M., Bengtsson, E., Klaus, B., Kusmierczyk, P., Lang, J. H., Detken, C., & Peltonen, T. (2017). A new database for financial crises in European countries (ECB Occasional Paper No. 194). European Central Bank.
18. Manasse, P., & Roubini, N. (2009). “Rules of thumb” for sovereign debt crises. *Journal of International Economics*, 78(2), 192–205.
19. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
20. Reinhart, C. M., & Rogoff, K. S. (2009). *This time is different: Eight centuries of financial folly*. Princeton University Press.

21. Ristolainen, K., Roukka, T., & Nyberg, H. (2023). A thousand words tell more than just numbers: Financial crises and historical headlines. *Journal of Financial Stability*, 70, 101186.
22. Sarlin, P. (2013). On policymakers' loss functions and the evaluation of early warning systems. *Economics Letters*, 119(1), 1–7.
23. Tölö, E. (2020). Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability*, 49, 100746.
24. Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.