Forecasting Financial Crises with Public Macro-Demographic Indicators: A Comparison of Logistic, Tree-Based and LSTM Models

Dongsheng Bei[1], Pengwei Zhu[1,2,*]
[1]Bank of Beijing Postdoctoral Research Workstation, Bank of Beijing, Beijing 100033, PR China
[2]University of Texas Health Science Center at Houston, School of Public Health, Houston, Texas, 77030, USA
E-mail: bds323007@163.com, pengwei_zhu@163.com
[*]Corresponding author

Keywords: financial crisis, LSTM, random forest, logistic regression, machine learning, deep learning, macroeconomics, early warning indicator

Received: October 20, 2025

This paper examines how far financial crises can be anticipated using only publicly available macroeconomic, macro-financial and macro-demographic indicators. We construct a monthly US panel from OECD and Federal Reserve sources and transform standard aggregates into a rich feature set capturing volatility, momentum, higher-order moments, drawdowns and structural breaks. Crisis risk is modeled in a hazard-style early-warning framework: for each month in expansion, we define binary labels for crisis onsets within 6-, 12- and 18-month horizons, combined with lead-time weights that reward earlier, more operationally useful signals. Using this common information set, we compare three families of models: regularised logistic regressions, gradient-boosted decision trees (LightGBM and XGBoost) and a bidirectional LSTM with attention fed by fixed-length feature sequences. Models are evaluated with expanding-window cross-validation and strictly out-of-sample holdout tests. Across horizons, penalised logistic regressions deliver the most accurate and stable forecasts, achieving holdout ROC-AUCs up to 0.99 and F1 scores up to 0.86 at the 18-month horizon, while tree-based methods are competitive only at longer horizons and the Bi-LSTM substantially overfits, adding little incremental predictive power. These results suggest that, in small and highly imbalanced crisis datasets built from open macro-demographic indicators, well-regularised linear models can match or surpass more complex machine-learning and deep-learning approaches, and offer greater transparency for macroprudential policy use.

Povzetek: Članek pokaže, da je pri napovedovanju finančnih kriz iz javno dostopnih makroekonomskih kazalnikov najbolj zanesljiv dobro regulariziran logistični model, ki v primerjavi z drevesnimi metodami in Bi-LSTM dosega višjo stabilnost in natančnost ter je hkrati bolj transparenten za uporabo v politiki.

1 Introduction

Financial crises remain a recurring part of modern economic history and typically leave behind deep and long-lasting damage. Drawing on a large post-war cross-country sample, Greenwood et al. (2022) show that when both credit and asset prices expand rapidly over a three-year window, the likelihood of a financial crisis in the following three years is close to 40 percent, compared with about 7 percent when neither grows unusually fast. This pattern indicates that crises arise more often in specific macro-financial environments and that commonly used aggregate series embed information about future crisis risk.

Recent early-warning studies make systematic use of this information. Using data for 59 advanced and emerging economies, Chen and Svirydzenka (2021) find that equity prices, property prices, the credit gap and the output gap provide useful advance signals of banking crises when expressed as gaps from trend, with signals often emerging several years before the event. In a different setting, Allaj and Sanfelici (2023) design early-warning models for large asset-price declines based on logit regressions that use realised variance and a price–volatility feedback indicator as predictors. Liu et al. (2022) compare logistic regression with a range of machine-learning classifiers and report that random forest, gradient boosting and related methods deliver more accurate financial-crisis predictions in a multi-country panel.

Macroprudential practice has moved in a similar direction, relying heavily on publicly available indicators. Koponen (2024) builds a composite cyclical systemic-risk index for Finland by combining standard banking-crisis early-warning indicators based on credit, asset prices and external balances, and shows that this composite performs well according to conventional classification measures. A related index for Croatia, constructed by Škrinjarić (2023), is used to discuss the advantages and limitations of such composite indicators for tracking cyclical systemic risk. At the cross-country level, Mugrabi et al. (2025) analyse bank-level measures of systemic risk in 45 emerging and advanced economies and find that tighter macroprudential policies are associated with lower systemic risk, particularly in inflation-targeting regimes, highlighting the role of observable macro-financial indicators in the design and calibration of macroprudential tools.

What is less clear from the existing literature is how far financial crises can be anticipated when one restricts attention strictly to such public economic indicators. Existing contributions establish that public macro-financial variables carry predictive content and that more flexible models can extract this information more effectively, but there is little systematic evidence on the limits of crisis prediction under a public-data constraint, or on how different modelling traditions perform when they all use exactly the same set of public indicators. This paper addresses that gap by assessing the predictive power of publicly available macroeconomic, macro-financial and macro-demographic series and by comparing benchmark econometric models, tree-based classifiers and sequence models for financial-crisis prediction within a common, transparent information set.

2 Literature review

A number of recent surveys take stock of early-warning systems for financial crises as a whole. Namaki, Eyvazloo, and Ramtinnia (2023) conduct a bibliometric review of 616 studies on financial early-warning systems and document a shift from traditional econometric models towards machine-learning approaches, applied to banking, currency and sovereign crises and to corporate failure. Their review stresses that combining macroeconomic aggregates with sectoral and balance-sheet information has become the norm in crisis prediction. Firdaus and Santoso (2025) focus more narrowly on models used to forecast financial crises and emphasise that both econometric and machine-learning approaches typically rely on sets of standard macro-financial indicators, sometimes augmented with uncertainty indices.

A separate strand of work examines how crisis-warning tools should be eValuated for policy use. Candelon, Dumitrescu, and Hurlin (2012) propose a unified statistical framework for assessing early-warning systems that is based on receiver-operating-characteristic (ROC) curves and optimal cut-offs, making the trade-off between missed crises and false alarms explicit. Drehmann and Juselius (2014) and Yildirim and Sanyal (2022) extend this perspective by adding criteria such as signal horizon, stability and interpretability and show that credit aggregates and debt-service ratios often satisfy these policy-oriented requirements. Together, these contributions argue that crisis-prediction models must deliver not only accurate forecasts, but also simple and robust signals that can be communicated to macroprudential authorities.

More recent studies explore model design and information sets. Gu (2022) develops hidden Markov–model based early-warning systems that generate short-term probabilities of financial instability and highlights the operational advantages of automated alerts for supervisors. Hidayat, Masyita, Nidar, Ahmad, and Syarif (2022) use a system-dynamics simulation framework to build an early-warning and early-action tool for bank solvency risk during the COVID-19 period, largely on the basis of publicly available banking data. These contributions show that flexible tools for crisis monitoring can be constructed from observable economic and financial indicators, but they do not provide a systematic comparison of different modelling traditions under a strict public-data constraint. The present paper addresses this gap by assessing how far financial crises can be anticipated using only public macroeconomic, macro-financial and macro-demographic series.

3 Related work

3.1 Econometric and machine-learning early-warning models

Recent work still uses logistic regression as a benchmark for binary crisis classification, but systematically shows that modern machine-learning methods perform better. Liu et al. (2022) compare logit with a range of classifiers on cross-country crisis data and find that tree-based ensembles such as random forests and gradient boosting improve both accuracy and variable selection in early-warning settings. Purnell et al. (2024) design an explainable early-warning system for financial networks, again finding that ensemble methods capture nonlinear contagion patterns better than logistic baselines while still allowing economically meaningful explanations.

3.2 Deep-learning models for crisis prediction

Deep learning has been used to move beyond static classifiers by modelling the time dimension of financial risk. Ouyang et al. (2021) use an attention-LSTM model to produce an early warning index for systemic financial risk in China and show systematic gains relative to ARIMA, support-vector regression and standard LSTM models. Yang et al. (2025) embed deep learning into a TVP-FAVAR system to produce a dynamic early-warning measure of systemic financial risk in China, while Elnaggar et al. (2025) develop a deep learning model trained directly on financial crisis datasets.

This emphasis on nonlinear and dynamic modelling is consistent with advances in nonlinear control, where feedback, fuzzy and neural methods are used to stabilise complex systems with uncertainty and unmeasured states. Relevant examples include output-feedback projective lag-synchronisation for chaotic systems with input nonlinearities (Boulkroune et al. 2017), adaptive fuzzy control achieving fixed-time synchronisation of fractional-order chaotic systems (Boulkroune et al. 2025), adaptive backstepping and robust neural adaptive control for uncertain nonlinear multi-input and single-input systems (Zouari et al. 2012; Zouari et al. 2013), nonlinear optimal control for gas-compressor–motor dynamics (Rigatos et al. 2023) and high-gain observer-based adaptive fuzzy control for multivariable nonlinear systems (Merazka, Zouari, and Boulkroune 2017). These works illustrate design principles for handling nonlinear dynamics, which motivates our use of Bi-LSTMs as flexible sequence models for crisis prediction.

3.3 Web3, DeFi and crypto-related instability

Recent work argues that decentralised finance, stablecoins and other crypto-assets can create additional channels for financial instability, for example through leveraged DeFi positions, liquidity mismatches in pools and run risk in stablecoins (European Central Bank, 2022; Financial Stability Board, 2023). Machine-learning studies using on-chain data show that tree-based and graph-based models can flag risky protocols and fraud in DeFi markets before large losses occur (Luo et al., 2024). However, these applications are mostly protocol- or asset-level and the available Web3 data do not yet form a long, consistent macro panel comparable to our macroeconomic and demographic series. For these reasons, we do not include Web3, DeFi or stablecoin indicators in our empirical crisis-prediction models. Instead, we treat them as a natural extension: once reliable, long-horizon measures of DeFi leverage, liquidations or stablecoin flows become available, they could be added alongside traditional macro-financial variables in the type of Bi-LSTM and random-forest framework developed here.

3.4 Macro-demographics, financial inclusion and crisis risk

Macro-demographic and financial-inclusion studies examine how population structure and access to finance affect stability, but typically without explicit crisis labels. Oanh et al. (2023) show that financial inclusion is positively related to financial stability, but that the relationship interacts in a nonlinear way with inflation and money growth in high- and low-development financial systems. Kebede et al. (2024) document a nonlinear “too little, too much” relationship between inclusion and stability, with threshold effects that matter for macroprudential policy. Schmieder and Imam (2024) analyse how demographic ageing changes bank balance sheets and the transmission of macro-financial shocks. These contributions underline the importance of macro-demographic and inclusion variables for financial stability, but they mainly use linear panel and VAR frameworks. Our study therefore combines a Bi-LSTM with macroeconomic and demographic indicators and benchmarks it against random forest and logistic regression in a financial-crisis prediction setting.

4 Materials and methods

4.1 Dataset

4.1.1 Dataset description

The dataset is a monthly US macro financial panel constructed from official OECD and Federal Reserve sources. Labour market indicators, including employment, unemployment, participation and working age population, are taken from the OECD infra-annual labour statistics, while business tendency balances and composite confidence measures for firms and households follow the OECD Composite Leading Indicators framework and related short term indicator collections (OECD, 2025a, OECD, 2025b). Price, trade and monetary variables draw on OECD consumer price indices for G20 economies, international merchandise trade statistics and short term monetary and real activity series that underpin the Key Short Term Economic Indicators (OECD, 2025c, OECD, 2025d).

Recession timing and financial conditions are added from composite indicators disseminated via the Federal Reserve Bank of St Louis. The crisis block combines the Real Time Sahm Rule Recession Indicator with the NBER based US recession dummy, while broader activity is summarised by the Chicago Fed National Activity Index, its diffusion index, moving average and expenditure and production subindexes (Federal Reserve Bank of St Louis, 2025a, Federal Reserve Bank of St Louis, 2025b, Federal Reserve Bank of St Louis, 2025c, Federal Reserve Bank of Chicago, 2025a). Financial stress is measured by the National Financial Conditions Index and its adjusted version, and overall real activity is cross checked using the Coincident Economic Activity Index for the United States from the Federal Reserve Bank of Philadelphia (Federal Reserve Bank of Chicago, 2025b, Federal Reserve Bank of Philadelphia, 2025).

4.1.2 Data processing

The raw dataset combines monthly macro-financial indicators with a binary recession flag. All series are coerced to a common monthly calendar keyed by a TIME_PERIOD index, sorted chronologically, and trimmed to the overlapping sample. When indicators are available at higher than monthly frequency, they are aggregated to monthly Values by taking simple averages within each calendar month before merging. This produces a single, time-ordered multivariate panel where each row corresponds to one month.

The binary recession flag is first converted into three auxiliary series start_onset, end_offset and state. The start_onset indicator marks months in which the economy switches from expansion to recession, end_offset marks the end of a recession spell, and state encodes the contemporaneous regime (expansion or recession). These derived variables are then used to identify contiguous recession segments and the corresponding non-recession intervals. A short post-recession buffer is appended after each recession end to represent a fragile recovery phase; months in this buffer are temporarily excluded from the set of candidates for new recession onsets.

On this basis, hazard-style labels and risk sets are constructed for each forecast horizon H. For every month t in expansion and outside the post-recession buffer, a horizon window (t, t+H] is scanned for the presence of a start_onset. If at least one start_onset occurs in that window, the hazard label is set to one; otherwise it is set to zero. Months that are in recession or in the post-recession buffer do not enter the estimation sample for that horizon. This yields, for each H, an expansion-only risk set with a corresponding binary hazard series.

Lead time and lead-time weights are precomputed together with the labels. For each horizon H and each month t in the risk set with $Y_t^H = 1$, the algorithm records the first onset month within (t, t+H] and defines the lead time $\Delta_t^H = u^*(t) - t$. On the basis of this lead time, a scalar weight is attached to each positive observation in order to favour earlier, more operationally useful warnings over very late signals that occur just before the onset. The positive weight is defined as $v(\Delta_t^H) = 1 + \alpha \frac{H-\Delta_t^H}{H}$, with a tuning parameter. Observations that are closer to the onset (small Δ_t^H) receive larger weights, while observations that are only marginally ahead of the onset receive weights close to one. Negative observations keep a weight of one. These weights are later combined with standard class weights in the loss function.

Feature processing is applied to the time-ordered macro-financial panel before sequence construction. Rolling means, standard deviations and simple linear trends are computed over medium- and long-horizon windows using only current and past Values. Structural-break indicators and similarity scores to selected historical prototype episodes are added, with changepoint models always fitted on training periods only in cross-Validation. Network-based features summarising conditional dependence and stability across indicators are derived from sparse graphical and vector autoregressive models estimated on the training slices of each fold; their outputs are aligned back to the monthly grid and carried forward when necessary. All feature blocks are merged into a common design matrix, sorted by time and variable name, and missing entries introduced by the joins are handled by forward filling along the time axis followed by zero filling if needed. This preserves temporal ordering and yields a numerically complete feature set.

For classical models, the hazard labels and features at each month in the risk set are combined into cross-sectional design matrices indexed by horizon. For the Bi-LSTM, the same standardised features are converted into sequences using fixed-length sliding windows. Given a chosen sequence length L, the data are reshaped into overlapping blocks [t-L+1,...,t] paired with the hazard label and lead-time weight at time t. Windows that would cross the boundary between training, Validation and holdout segments, or that extend beyond the available sample, are discarded. All preprocessing steps, including aggregation, label construction, feature engineering, scaling and sequence generation, are performed separately within each training partition in cross-Validation and then applied forward to the corresponding Validation and holdout segments, ensuring that no information from the future enters the inputs or labels used for model estimation.

4.2 Model specification and preparation

As part of data processing, we define and prepare baseline learners on structured tabular features. We configure three families of models: regularized logistic regressions to control variance and mitigate multicollinearity, gradient-boosted decision trees (LightGBM and XGBoost) to capture nonlinearities and interaction effects in tabular predictors, and a bidirectional LSTM to exploit short-horizon temporal structure created by rolling-window feature construction. This staged setup converts cleaned and scaled panels into model-ready matrices and sequences and keeps model specification aligned with the crisis labeling scheme and evaluation design.

In the regularized logistic specifications, we estimate L1, L2, and elastic-net variants, shrinking unstable coefficients and improving generalization in the presence of correlated macro-financial indicators; penalty type and strength are chosen by cross-validated hyperparameter search using the validation objective, in line with recent applications of penalized logit early-warning models for financial crises (Liu, Chen, and Wang, 2022). For the tree-based models, we train shallow, heavily regularized gradient-boosting ensembles on bootstrapped samples with feature subsampling, which allows the models to approximate nonlinear decision boundaries while remaining robust in small-sample, imbalanced crisis settings where similar methods have been shown to perform well relative to logit benchmarks (Liu, Chen, and Wang, 2022). For deep sequence modeling, we transform aligned panels into fixed-length sequences and feed them to a single-layer bidirectional LSTM; the forward and backward hidden states are concatenated and passed through a hazard-conditioned attention and output layer, an architecture choice motivated by evidence that recurrent networks can capture regime shifts and nonlinear dynamics in financial-stability applications when combined with careful regularization and early-warning evaluation procedures (Tölö, 2020).

5 Results

The models were eValuated on three prediction horizons of six, twelve and eighteen months using expanding-window cross Validation with separate holdout periods, and the results are summarised in the per-horizon tables and figures. Across all horizons, the regularised logistic regressions delivered the most stable and accurate out-of-sample performance, while tree-based machine learning models performed competitively at longer horizons and the Bi-LSTM showed much weaker generalisation despite high in-sample fit. For the six-month horizon, LOGISTIC (L1) reached a holdout ROC-AUC of 0.91 with precision-recall AUC 0.64, accuracy 0.83, F1 0.48 and Cohen’s kappa 0.41, outperforming LOGISTIC (L2) which achieved a holdout ROC-AUC of 0.90 and PR-AUC 0.51 with accuracy 0.80 and kappa 0.32. The elastic net variant obtained high holdout accuracy of 0.91 but did so with F1 equal to zero and negative kappa, indicating that it almost never predicted the crisis class in the holdout window. Among machine learning models at this horizon, LIGHTGBM and XGBOOST produced much lower holdout ROC-AUC Values of 0.61 and 0.79 respectively and F1 scores of 0.14 and 0.21, while the stacking ensemble achieved high training ROC-AUC above 0.92 but a holdout ROC-AUC of only 0.36. The Bi-LSTM displayed similar behaviour, with training ROC-AUC of 0.96 but a holdout ROC-AUC of 0.64, PR-AUC 0.11 and F1 0.08, even though Validation ROC-AUC remained above 0.91 during cross Validation.

For the twelve-month horizon, performance again favoured the regularised logistic models, with the elastic net specification emerging as the strongest overall. LOGISTIC (Elastic Net) reached holdout ROC-AUC of 0.94 and PR-AUC of 0.59 with accuracy 0.89, F1 0.74 and kappa 0.68, while LOGISTIC (L2) attained holdout ROC-AUC of 0.84, PR-AUC 0.37, accuracy 0.80 and F1 0.52. LIGHTGBM matched these results closely with holdout ROC-AUC 0.87, PR-AUC 0.55, accuracy 0.87 and F1 0.58, whereas XGBOOST yielded a similar accuracy of 0.89 but lower discrimination with ROC-AUC 0.69. The stacking ensemble again showed a large gap between training and holdout performance, with training ROC-AUC 0.92 but holdout ROC-AUC 0.51 and PR-AUC 0.15. The Bi-LSTM exhibited the sharpest divergence between in-sample and out-of-sample performance at this horizon: training ROC-AUC reached 0.92 and PR-AUC 0.87, while the holdout ROC-AUC fell to 0.32, PR-AUC to 0.12, and F1 to zero with negative kappa.

At the eighteen-month horizon, predictive accuracy increased further for the best traditional models. LOGISTIC (L2) delivered the strongest overall results with holdout ROC-AUC 0.99, PR-AUC 0.98, accuracy 0.93, F1 0.86 and kappa 0.82. LIGHTGBM achieved very similar performance with holdout ROC-AUC 0.95, PR-AUC 0.74, accuracy 0.93, F1 0.87 and kappa 0.82, while LOGISTIC (Elastic Net) also performed well with holdout ROC-AUC 0.93 and PR-AUC 0.73. LOGISTIC (L1) produced a holdout ROC-AUC of 0.91 and PR-AUC 0.79, but with lower F1 of 0.40 due to more conservative classification of crises. XGBOOST lagged behind with holdout ROC-AUC only 0.71 and F1 0.24, and the stacking ensemble again underperformed with holdout ROC-AUC 0.97 but accuracy 0.24 and kappa effectively zero, reflecting extreme imbalance in predicted labels. The Bi-LSTM improved relative to shorter horizons with holdout ROC-AUC 0.66, PR-AUC 0.64, accuracy 0.87, F1 0.62 and kappa 0.55, but still did not match the best logistic or tree-based models, even though its training ROC-AUC and PR-AUC were both around 0.84 and higher.

Model CV and holdout performance across horizons

Horizon=6		**Accuracy**	**PR-AUC**	**ROC-AUC**	**F1**	**Kappa**
LOGISTIC (L1)	Train	0.54	0.36	0.69	0.48	0.26
	Val	0.26	0.34	0.68	0.22	0.06
	Holdout	0.83	**0.64**	**0.91**	**0.48**	**0.41**
LOGISTIC (L2)	Train	0.92	0.77	0.90	0.73	0.69
	Val	0.59	**0.93**	**0.99**	0.31	0.21
	Holdout	0.80	0.51	0.90	0.40	0.32
LOGISTIC (Elastic Net)	Train	0.90	0.75	0.90	0.71	0.65
	Val	**0.83**	0.68	0.90	0.35	**0.28**
	Holdout	**0.91**	0.19	0.78	0.00	-0.01
LIGHTGBM	Train	0.35	0.22	0.58	0.34	0.08
	Val	0.33	0.26	0.62	**0.36**	0.20
	Holdout	0.67	0.12	0.61	0.14	0.02
XGBOOST	Train	**0.97**	**0.87**	**0.98**	**0.89**	**0.87**
	Val	0.53	0.55	0.85	0.34	0.21
	Holdout	0.69	0.35	0.79	0.21	0.09
STACKING	Train	0.92	0.80	0.93	0.77	0.72
	Val	0.60	0.55	0.92	0.28	0.18
	Holdout	0.87	0.07	0.36	0.00	-0.06
Bi-LSTM	Train	0.93	0.82	0.96	0.86	0.82
	Val	0.64	0.71	0.91	0.34	0.25
	Holdout	0.69	0.11	0.64	0.08	-0.05

Horizon=12		**Accuracy**	**PR-AUC**	**ROC-AUC**	**F1**	**Kappa**
LOGISTIC (L1)	Train	0.67	0.50	0.73	0.65	0.37
	Val	0.19	0.52	0.64	0.31	0.00
	Holdout	0.83	0.36	0.84	0.00	-0.01
LOGISTIC (L2)	Train	0.87	0.79	0.90	0.75	0.66
	Val	0.40	0.65	0.80	0.38	0.12
	Holdout	0.80	0.37	0.84	0.52	0.40
LOGISTIC (Elastic Net)	Train	0.84	0.77	0.86	0.72	0.61
	Val	0.30	**0.78**	**0.94**	0.34	0.04
	Holdout	**0.89**	**0.59**	**0.94**	**0.74**	**0.68**
LIGHTGBM	Train	0.43	0.33	0.55	0.49	0.09
	Val	0.40	0.33	0.59	0.45	0.18
	Holdout	0.87	0.55	0.87	0.58	0.50
XGBOOST	Train	**0.92**	**0.87**	**0.95**	**0.88**	**0.82**
	Val	**0.73**	0.72	0.92	**0.64**	**0.48**
	Holdout	**0.89**	0.53	0.69	0.45	0.41
STACKING	Train	0.90	0.83	0.92	0.82	0.75
	Val	0.38	0.49	0.80	0.39	0.12
	Holdout	0.65	0.15	0.51	0.16	-0.05
Bi-LSTM	Train	0.90	**0.87**	0.92	0.84	0.77
	Val	0.37	0.73	0.88	0.38	0.13
	Holdout	0.58	0.12	0.32	0.00	-0.25

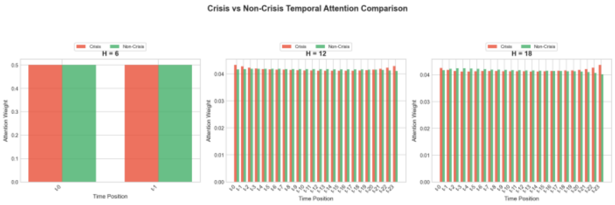
Horizon=18		**Accuracy**	**PR-AUC**	**ROC-AUC**	**F1**	**Kappa**
LOGISTIC (L1)	Train	0.55	0.53	0.58	0.66	0.17
	Val	0.38	0.46	0.60	0.45	0.05
	Holdout	0.82	0.79	0.91	0.40	0.34
LOGISTIC (L2)	Train	0.84	0.80	0.88	0.83	0.68
	Val	0.33	0.45	0.68	0.36	-0.07
	Holdout	0.93	**0.98**	**0.99**	0.86	**0.82**
LOGISTIC (Elastic Net)	Train	0.77	0.76	0.81	0.77	0.56
	Val	0.30	0.48	0.63	0.36	-0.09

	Holdout	0.87	0.73	0.93	0.73	0.64
LIGHTGBM	Train	0.64	0.59	0.67	0.70	0.30
	Val	0.37	0.39	0.66	0.45	0.05
	Holdout	**0.93**	0.74	0.95	**0.87**	0.82
XGBOOST	Train	0.82	0.83	0.85	0.85	0.61
	Val	**0.46**	0.62	0.78	0.35	0.01
	Holdout	0.79	0.65	0.71	0.24	0.20
STACKING	Train	**0.87**	**0.90**	**0.92**	**0.85**	**0.75**
	Val	0.44	0.30	0.57	0.31	0.09
	Holdout	0.24	0.88	0.97	0.39	0.00
Bi-LSTM	Train	0.83	0.83	0.84	0.84	0.61
	Val	0.42	**0.70**	**0.87**	**0.47**	**0.09**
	Holdout	0.87	0.64	0.66	0.62	0.55

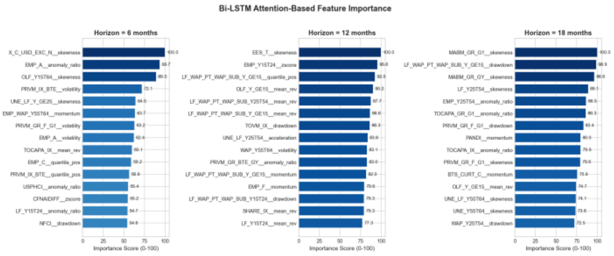
The Bi-LSTM attention-based feature importance analysis offers a complementary view of which indicators matter most when the deep learning model is used. At the six month horizon, the network placed its highest scaled weights on X_C_USD_EXC_N__skewness, EMP_A__anomaly_ratio, OLF_Y15T64__skewness and PRVM_IX_BTE__volatility, together with UNE_LF_Y_GE25__skewness, EMP_WAP_Y55T64__momentum, EMP_A__volatility, TOCAPA_IX__mean_rev and PRVM_GR_F_G1__volatility, and also on higher order moments of labour market variables and financial stress indices such as NFCI__drawdown and USPHCI__anomaly_ratio. At the twelve-month horizon, attention concentrated on EES_T__skewness, EMP_Y15T24__zscore, LF_WAP_PT_WAP_SUB_Y_GE15__quantile_pos, OLF_Y_GE15__mean_rev and LF_WAP_PT_WAP_SUB_Y25T54__mean_rev, as well as drawdowns and momentum of order book and wage indicators, accelerations in labour force participation and volatility of wage and equity indices. For the eighteen-month horizon, the most important features were MABM_GR_G1__skewness, LF_WAP_PT_WAP_SUB_Y_GE15__drawdown, MABM_GR_GY__skewness, LF_Y25T54__skewness, EMP_Y25T54__anomaly_ratio and TOCAPA_GR_G1__anomaly_ratio, along with PRVM_GR_F_G1__drawdown, PANDI__momentum, TOCAPA_IX__anomaly_ratio and several additional skewness and drawdown measures. These rankings indicate that the Bi-LSTM primarily exploits higher order moments, anomaly ratios, quantile positions and drawdowns of exchange rate, money, labour market and order book indicators rather than the volatility and momentum measures emphasised by the best logistic and tree models.

Temporal attention patterns from the Bi-LSTM further clarify how the deep learning model weights information across the two year input window. For the six month horizon, the learned attention vector is extremely concentrated on the two most recent months, with weights of approximately 0.50 at positions t-1 and t-0 for both crisis and non crisis sequences and differences across regimes of only about three ten thousandths. For the twelve- and eighteen-month horizons, the attention weights are nearly uniform across all twenty-four lags, ranging from about 0.041 to 0.043 for each time step. The crisis versus non crisis comparison shows that differences in attention between the two regimes are very small at every lag, mostly on the order of a few thousandths.

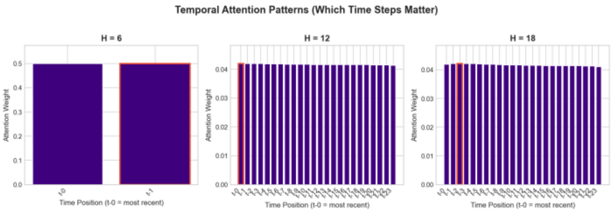
Crisis vs Non-Crisis Temporal Attention Comparison



Bi-LSTM Attention-Based Feature Importance



Temporal Attention Patterns



6 Discussion

The surprisingly strong performance of regularized logistic regression relative to complex models in predicting financial crises can be attributed to several factors. First, the dataset for crisis prediction is relatively small and imbalanced: crises are rare events, and even with many indicators, the number of positive instances is very limited. In such settings, high-capacity models like gradient-boosting trees or Bi-LSTM networks are prone to overfitting, learning spurious patterns that do not generalize. The logistic models, especially with L1/L2 penalties, effectively constrained the feature set and model complexity, focusing on a few salient indicators that consistently signal crises. This pattern stands in contrast to the early-warning results of Holopainen and Sarlin (2017), who show that conventional statistical models are generally outperformed by more advanced machine-learning and ensemble methods in a European setting, suggesting that the relative advantage of complex models is highly context dependent and may disappear when the sample is small and crises are rare. Our findings are, however, consistent with evidence from the forecasting literature that greater model complexity does not systematically improve accuracy and can even increase forecast error, so that simple, evidence-based procedures frequently dominate more elaborate approaches (Green and Armstrong, 2015).

Feature engineering and domain knowledge further help explain why traditional models were so competitive. The inputs were derived from macro-financial indicators that were transformed into volatility, momentum, skewness, drawdowns, and related statistics in order to highlight stress dynamics. A regularized linear model can exploit these engineered features effectively. In principle, a deep learning model could discover additional nonlinear interactions or subtle temporal patterns across these same features, yet in practice the Bi-LSTM did not add predictive value beyond what the engineered features already provide. The failure of the Bi-LSTM to outperform the logit suggests that either nonlinearities and temporal dependencies are not very strong in these data or that they cannot be reliably learned given the limited number of crisis episodes. Traditional early-warning models often assume mostly linear or threshold effects, for example that credit growth beyond a critical threshold sharply raises crisis risk, and our findings indicate that the relationship between the transformed indicators and crisis probabilities is largely monotonic and additive, which a logistic regression can capture. By contrast, using a multi-country dataset and a broader set of algorithms, Liu, Chen, and Wang (2022) find that machine-learning models—particularly random forests, gradient boosting decision trees, and ensembles—deliver substantially better out-of-sample crisis forecasts than a logistic model and argue that such methods should play a central role in early-warning system design.

The performance gap between traditional and deep models also reflects differences in stability and generalization. Logistic regression coefficients were comparatively stable across cross-validation folds, while the Bi-LSTM and ensemble methods showed high variance, with small changes in the training set sometimes leading to markedly different predictions. For policy applications, an early-warning system needs not only good accuracy but also consistent behavior and clear interpretation (Tölö 2020). The black-box nature of the Bi-LSTM, together with its instability, makes it difficult for policymakers to understand why a given warning is issued or to trust that the signal will be reliable in new environments. By contrast, logistic regression provides straightforward marginal effects, such as the effect of a one-unit increase in unemployment volatility on crisis odds, which supports communication and helps connect the model's behavior to economic mechanisms. Interpretability has become a central concern in applying machine learning to financial stability analysis, and even when complex models improve raw predictive power regulators increasingly emphasize the need to understand why a model flags elevated risk (Bluwstein et al. 2021). In our setting, the top features in the best logistic specification correspond to intuitive economic phenomena such as labor-market disruptions, credit boom–bust dynamics, and financial-market stress, whereas the deep model's inferred interactions did not translate into clearer or more accurate warnings. While some studies report that neural networks and ensembles can outperform logistic early-warning models under cross-validation, for example Tölö (2020) and Liu, Chen, and Wang (2022), those gains can disappear in truly out-of-sample tests and often come at the cost of interpretability. Our results underline that robustness and transparency are critical design criteria in crisis prediction and in many cases may be more important than squeezing out marginal in-sample performance gains.

The feature rankings also provide economic insight into the precursors of financial crises in an open-data framework. High importance scores for unemployment and labor force indicators, particularly for youth and prime-age groups, indicate that labor market deterioration tends to precede and accompany financial instability. Rising unemployment and falling participation can trigger loan defaults and weaken household demand, which in turn strains bank balance sheets. Similarly, the frequent appearance of credit and leverage indicators, especially measures related to private credit growth, volatility, and mean-reversion, is consistent with the view that rapid credit expansion followed by unstable credit dynamics signals a boom–bust cycle that often culminates in crisis. Price dynamics and term-structure variables also play a central role: inflation momentum and interest-rate spreads enter prominently in the feature rankings, and a sharp pickup in inflation or inflation volatility can trigger tighter monetary policy, higher short-term rates, and a flat or inverted yield curve, conditions that frequently precede recessions and banking-sector stress. The yield curve is a particularly informative recession indicator, and an inversion where short-term rates exceed long-term rates often signals tightening financial conditions and an impending contraction of credit. That these core predictors all come from public sources such as central bank statistics, FRED, and OECD releases underscores that an effective crisis-warning system can be built without proprietary data and that transparency and reproducibility can be maintained alongside strong predictive performance. These interpretations are consistent with recent early-warning evidence based on large macro-financial panels, which finds that credit conditions, asset-price dynamics, and broader macro indicators jointly deliver useful advance signals of financial crises when modeled in flexible forecasting frameworks (Tölö 2020; Liu, Chen, and Wang 2022).

The results also reveal limitations and point to avenues for further work. Class imbalance implies that even high AUC values may coexist with missed crises or false alarms, and at the 6-month horizon the best logistic model attains a holdout F1 of about 0.48, which leaves scope to refine the trade-off between precision and recall. Deep learning was introduced to capture potential nonlinear precursors, for example combinations of rapid credit growth, an increasingly flat yield curve, and rising unemployment that might jointly matter even if each component is only mildly abnormal. That strategy did not improve forecasting accuracy in this dataset, yet it suggests a direction for future research. With longer samples, richer cross-country panels, or more stringent regularization, such as Bayesian neural networks or transfer learning from simulated macro-financial scenarios, deep architectures might still uncover interactions that are not well captured by linear models. Horizon-specific behavior is another consideration. At the 18-month horizon even the Bi-LSTM performs more competitively, which may reflect that the buildup of vulnerabilities typically becomes clearer a year or more before crises, when credit and asset-price booms peak and macro-financial imbalances are most visible. Shorter horizons, such as 6 months, may be dominated by transitory disturbances that are harder to distinguish from noise. While the regularized logistic model delivers robust performance across horizons, different lead times may benefit from horizon-specific models or ensembles that emphasize indicators whose predictive content is strongest for that horizon.

7 Conclusion

Overall, the evidence from this study points to a central role for simple, interpretable models in financial-crisis forecasting, with machine-learning and deep-learning methods used selectively and only after careful assessment of their stability and incremental value. Regularized logistic regression provides a clear benchmark in this setting, delivering strong and transparent out-of-sample performance despite the small number of crisis observations, while more complex models struggle to convert their in-sample flexibility into reliable forecasts. At the same time, recent multi-country work shows that tree-based and ensemble methods can outperform logistic specifications when richer data are available (Liu, Chen, and Wang 2022), and applications of recurrent networks to financial stability emphasize the need for extensive robustness checks before using such models for policy (Tölö 2020). Future research can build on these results by extending the panel across countries and time, refining the set of open macro-financial indicators, and applying flexible algorithms within a framework that preserves regularization and interpretability, so that any gains from complexity are grounded in data and can be scrutinized by policymakers and practitioners.

Author statement

Dongsheng Bei: Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Pengwei Zhu: Methodology, Software, Validation, Resources, Visualization, Writing – review & editing.

Equal Contribution Statement: Dongsheng Bei and Pengwei Zhu contributed equally to this work and share co-first authorship.

References

[1] Allaj, Erindi, and Simona Sanfelici. 2023. "Early Warning Systems for Identifying Financial Instability." International Journal of Forecasting 39 (4): 1777-1803.

[2] Mugrabi, Farah, Mohamed Belkhir, Sami Ben Naceur, Bertrand Candelon, and Woon Gyu Choi. 2025. “Macroprudential Policy and Bank Systemic Risk: Does Inflation Targeting Matter?” Emerging Markets Review, article 101397.

[3] Chen, Sally, and Katsiaryna Svirydzenka. 2021. "Financial Cycles-Early Warning Indicators of Banking Crises?" IMF Working Paper 2021/116. Washington, DC: International Monetary Fund.

[4] Greenwood, Robin, Samuel G. Hanson, Andrei Shleifer, and Jakob Ahm Sørensen. 2022. "Predictable Financial Crises." Journal of Finance 77 (2): 863-921.

[5] Koponen, Heidi. 2024. "Constructing a Composite Indicator to Assess Cyclical Systemic Risks: An Early Warning Approach." BoF Economics Review 3/2024. Helsinki: Bank of Finland.

[6] Liu, Lanbiao, Chen Chen, and Bo Wang. 2022. "Predicting Financial Crises with Machine Learning Methods." Journal of Forecasting 41 (5): 871-910.

[7] Škrinjarić, Tihana. 2023. "Introducing a Composite Indicator of Cyclical Systemic Risk in Croatia: Possibilities and Limitations." Public Sector Economics 47 (1): 1-39.

[8] Purnell, Daren, Jr., Amir Etemadi, and John Kamp. 2024. “Developing an Early Warning System for Financial Networks: An Explainable Machine Learning Approach.” Entropy 26 (9): 796.

[9] Ouyang, Zi-sheng, Xi-te Yang, and Yongzeng Lai. 2021. "Systemic Financial Risk Early Warning of Financial Market in China Using Attention-LSTM Model." The North American Journal of Economics and Finance 56: 101383.

[10] Candelon, Bertrand, Elena-Ivona Dumitrescu, and Christophe Hurlin. 2012. “How to EValuate an Early-Warning System: Toward a Unified Statistical Framework for Assessing Financial Crises Forecasting Methods.” IMF Economic Review 60 (1): 75–113.

[11] Drehmann, Mathias, and Mikael Juselius. 2014. “EValuating Early Warning Indicators of Banking Crises: Satisfying Policy Requirements.” International Journal of Forecasting 30 (3): 759–780.

[12] Firdaus, N. T., and N. Santoso. 2025. “Early Warning Systems for Financial Crisis Prediction: A Systematic Literature Review of Econometrics, Machine Learning and Uncertainty Indices.” MALCOM: Indonesian Journal of Machine Learning and Computer Science 5 (4): 1415–1422.

[13] Gu, Xing. 2022. Early-Warning Alert Systems for Financial-Instability Detection: An HMM-Driven Approach. PhD dissertation, University of Western Ontario.

[14] Hidayat, Taufiq, Dian Masyita, Sulaeman Rahman Nidar, Fauzan Ahmad, and Muhammad Adrissa Nur Syarif. 2022. “Early Warning Early Action for the Banking Solvency Risk in the COVID-19 Pandemic Era: A Case Study of Indonesia.” Economies 10 (1): 6.

[15] Namaki, Ali, Reza Eyvazloo, and Shahin Ramtinnia. 2023. “A Systematic Review of Early Warning

Systems in Finance." arXiv preprint arXiv:2310.00490.

[16] Yildirim, Yusuf, and Anirban Sanyal. 2022. "EValuating the Effectiveness of Early Warning Indicators: An Application of Receiver Operating Characteristic Curve Approach to Panel Data." Scientific Annals of Economics and Business 69 (4): 557–597.

[17] Yang, Hufang, Luyi Liu, Jieyang Cui, Wenbin Wu, and Yuyang Gao. 2025. "Research on Dynamic Measurement and Early Warning of Systemic Financial Risk in China Based on TVP-FAVAR and Deep Learning Model." Systems 13 (8): 720.

[18] Elnaggar, Hoda A., Marwa Elsherif, and Mohamed I. Marie. 2025. "A Deep Learning-Based Model for Financial Crisis Prediction." Research Square preprint, version 1.

[19] Boulkroune, Abdesselem, Sarah Hamel, Farouk Zouari, Abdelkrim Boukabou, and Asier Ibeas. 2017. "Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities." Mathematical Problems in Engineering 2017: 8045803.

[20] Boulkroune, Abdesselem, Farouk Zouari, and Amina Boubellouta. 2025. "Adaptive Fuzzy Control for Practical Fixed-Time Synchronization of Fractional-Order Chaotic Systems." Journal of Vibration and Control.

[21] Zouari, Farouk, Kamel Ben Saad, and Mohamed Benrejeb. 2012. "Robust Neural Adaptive Control for a Class of Uncertain Nonlinear Complex Dynamical Multivariable Systems." International Review on Modelling and Simulations 5 (5): 2075–2103.

[22] Zouari, Farouk, Kamel Ben Saad, and Mohamed Benrejeb. 2013. "Adaptive Backstepping Control for a Class of Uncertain Single Input Single Output Nonlinear Systems." In 2013 10th International Multi-Conference on Systems, Signals and Devices (SSD).

[23] Rigatos, Gerasimos, Masoud Abbaszadeh, Bilal Sari, Pierluigi Siano, Gennaro Cuccurullo, and Farouk Zouari. 2023. "Nonlinear Optimal Control for a Gas Compressor Driven by an Induction Motor." Results in Control and Optimization 11: 100226.

[24] Merazka, Loubna, Farouk Zouari, and Abdesselem Boulkroune. 2017. "High-Gain Observer-Based Adaptive Fuzzy Control for a Class of Multivariable Nonlinear Systems." In 2017 6th International Conference on Systems and Control (ICoSC), 96–102.

[25] European Central Bank. 2022. "Decrypting Financial Stability Risks in Crypto-Asset Markets." Financial Stability Review, May.

[26] Financial Stability Board. 2023. The Financial Stability Risks of Decentralised Finance. Basel: Financial Stability Board.

[27] Luo, Bingqiao, Zhen Zhang, Qian Wang, Anli Ke, Shengliang Lu, and Bingsheng He. 2024. "AI-Powered Fraud Detection in Decentralized Finance: A Project Life Cycle Perspective." arXiv preprint 2308.15992.

[28] Oanh, Tran Thi Kim, Le Thi Thuy Van, and Le Quoc Dinh. 2023. "Relationship between Financial Inclusion, Monetary Policy and Financial Stability: An Analysis in High Financial Development and Low Financial Development Countries." Heliyon 9 (6): e16647.

[29] Kebede, Jeleta Gezahegne, Saroja Selvanathan, and Athula Naranpanawa. 2024. "Financial Stability and Financial Inclusion: A Non-Linear Nexus." Journal of Economic Studies 52 (4): 742–761.

[30] Schmieder, Christian, and Patrick A. Imam. 2024. Aging Gracefully: Steering the Banking Sector through Demographic Shifts. BIS Working Papers No. 1193, Bank for International Settlements, 12 June.

[31] Organisation for Economic Co-operation and Development (OECD). 2025a. "Infra-Annual Labour Statistics (IALFS)." OECD Data Explorer. Accessed November 19, 2025. https://data-explorer.oecd.org/.

[32] Organisation for Economic Co-operation and Development (OECD). 2025b. "Composite Leading Indicators (CLI)." OECD Data and Datasets. Accessed November 19, 2025. https://www.oecd.org/en/data/datasets/oecd-composite-leading-indicators-clis.html.

[33] Organisation for Economic Co-operation and Development (OECD). 2025c. "G20 – Consumer Price Indices, All Items." OECD Data Explorer. Accessed November 19, 2025. https://data-explorer.oecd.org/.

[34] Organisation for Economic Co-operation and Development (OECD). 2025d. "International Merchandise Trade Statistics." OECD Data Explorer. Accessed November 19, 2025. https://data-explorer.oecd.org/.

[35] Federal Reserve Bank of St. Louis. 2025a. "Real-Time Sahm Rule Recession Indicator (SAHMREALTIME)." FRED, Federal Reserve Bank of St. Louis. Accessed November 19, 2025. https://fred.stlouisfed.org/series/SAHMREALTIME.

[36] Federal Reserve Bank of St. Louis. 2025b. "NBER Based Recession Indicators for the United States from the Period following the Peak through the Trough (USREC)." FRED, Federal Reserve Bank of St. Louis. Accessed November 19, 2025. https://fred.stlouisfed.org/series/USREC.

[37] Federal Reserve Bank of St. Louis. 2025c. "Chicago Fed National Activity Index (CFNAI) and Related Series." FRED, Federal Reserve Bank of St. Louis. Accessed November 19, 2025. https://fred.stlouisfed.org/series/CFNAI.

[38] Federal Reserve Bank of Chicago. 2025a. "Chicago Fed National Activity Index: Current Data." Federal Reserve Bank of Chicago. Accessed November 19, 2025. https://www.chicagofed.org/research/data/cfnai/current-data.

[39] Federal Reserve Bank of Chicago. 2025b. “National Financial Conditions Index: About the NFCI and Current Data.” Federal Reserve Bank of Chicago. Accessed November 19, 2025. https://www.chicagofed.org/research/data/nfci/about.

[40] Federal Reserve Bank of Philadelphia. 2025. “State Coincident Indexes: Coincident Economic Activity Index for the United States (USPHCI).” Federal Reserve Bank of Philadelphia. Accessed November 19, 2025. https://fred.stlouisfed.org/series/USPHCI.

[41] Holopainen, Markus, and Peter Sarlin. 2017. “Toward Robust Early-Warning Models: A Horse Race, Ensembles and Model Uncertainty.” Quantitative Finance 17 (12): 1933–1963.

[42] Green, Kesten C., and J. Scott Armstrong. 2015. “Simple versus Complex Forecasting: The Evidence.” Journal of Business Research 68 (8): 1678–1685.

[43] Liu, Lanbiao, Chen Chen, and Bo Wang. 2022. “Predicting Financial Crises with Machine Learning Methods.” Journal of Forecasting 41 (5): 871–910.

[44] Tölö, Eero. 2020. “Predicting Systemic Financial Crises with Recurrent Neural Networks.” Journal of Financial Stability 49: 100746. https://doi.org/10.1016/j.jfs.2020.100746

[45] Bluwstein, Kristina, Marcus Buckmann, Andreas Joseph, Sujit Kapadia, and Özgür Şimşek. 2021. “Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach.” ECB Working Paper 2614. Frankfurt am Main: European Central Bank.