Financial Crisis Forecasting Using Temporal Convolutional Networks: A Multi-Horizon Approach

Abstract

This study investigates the prediction of systemic financial crisis onsets using macro-financial and geopolitical indicators. We implement a Temporal Convolutional Network (TCN) framework and systematically evaluate multiple target specifications, including multi-label crisis type classification, time-to-crisis regression, crisis severity prediction, and multi-year ahead binary prediction (1, 2, 3, and 5-year horizons). Our analysis reveals that extending the prediction horizon from 1 year to 5 years substantially improves predictive accuracy, achieving an Area Under the Precision-Recall Curve (AUCPR) of 0.655 compared to 0.269 for the 1-year specification—a 143% improvement. The enhanced performance stems from increased positive class prevalence (13.7% vs. 2.9%) and reduced sensitivity to precise crisis timing. However, integration challenges with complementary datasets and limitations in multi-task learning architectures highlight persistent challenges in rare financial event prediction.

Keywords: Financial crises, Temporal Convolutional Networks, Early warning systems, Macroprudential policy, Machine learning

---

1. Introduction

Systemic financial crises represent rare but catastrophic events with profound implications for economic stability, fiscal solvency, and social welfare. The 2007-2008 Global Financial Crisis demonstrated that traditional macroeconomic models often fail to anticipate severe financial instability, motivating renewed interest in quantitative early warning systems (EWS). Recent advances in deep learning, particularly architectures designed for sequential data such as Temporal Convolutional Networks (TCNs), offer promising avenues for improving crisis prediction by capturing temporal dependencies in macro-financial indicators.

The fundamental challenge in crisis prediction lies in the extreme rarity of crisis events. Historical databases spanning over a century and multiple countries typically contain fewer than 100 distinct crisis episodes, creating severe class imbalance that constrains machine learning model performance. Moreover, the heterogeneous nature of financial crises—spanning banking panics, currency collapses, sovereign debt defaults, and asset price bubbles—complicates efforts to identify universal predictive signals. These challenges necessitate careful consideration of target variable construction, feature engineering, and model architecture selection.

This study addresses these challenges through a comprehensive empirical investigation of alternative target specifications and modeling approaches. Our primary research question asks: How can we construct prediction targets that balance statistical learnability with policy relevance? We hypothesize that relaxing the precise timing requirement through multi-year ahead prediction windows will substantially improve model performance while maintaining actionable intelligence for macroprudential policymakers.

The contributions of this study are threefold. First, we implement and evaluate multiple target constructions, including multi-label crisis type classification, continuous time-to-crisis regression, crisis severity prediction, and binary multi-year ahead specifications. Second, we document integration challenges when combining complementary financial crisis databases with different country coverage and coding schemes. Third, we demonstrate that a 5-year prediction horizon achieves substantially higher predictive accuracy than traditional 1-year ahead specifications, offering practical implications for central bank early warning systems.

---

2. Data

2.1 Primary Data Sources

Our analysis integrates data from three complementary sources, though integration challenges ultimately constrained our ability to utilize all datasets fully.

Jordà–Schularick–Taylor (JST) Macrohistory Database

The JST database (Release 6) serves as our primary data source, providing annual macroeconomic and financial indicators for 18 advanced economies

from 1870 to 2020. The dataset contains 2,718 country-year observations with 59 variables, including real GDP, credit aggregates, asset prices, interest rates, and banking sector indicators. Crucially, the JST database includes a binary indicator `crisisJST` identifying years during which a country experienced a systemic financial crisis, based on established criteria for banking crises, currency crises, and sovereign debt crises. This variable provides the foundation for our target variable constructions.

The 18 countries included are: Australia, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, and United States. These advanced economies exhibit relatively complete historical coverage, though data availability varies across indicators and time periods, particularly for banking sector variables in earlier decades.

European Systemic Risk Board (ESRB) Financial Crises Database

The ESRB database provides detailed information on financial crises in European Union member states, including crisis start and end dates, crisis types (banking, sovereign, currency, asset price), and crisis management policies. The dataset covers 77 crisis episodes across 28 countries. However, integration with the JST database proved challenging due to incompatible country coding schemes—the ESRB uses 2-letter codes (e.g., "DE", "FR") while JST employs ISO 3166-1 alpha-3 codes (e.g., "DEU", "FRA"). Attempts to reconcile these coding schemes yielded insufficient overlap (only 7 countries matched exactly), preventing effective utilization of the ESRB's granular crisis typology in our primary analyses. This integration challenge represents a significant limitation, as the ESRB's detailed crisis categorization could have enabled more nuanced multi-label prediction targets.

Caldara and Iacoviello Geopolitical Risk (GPR) Index

The GPR database provides monthly indices of geopolitical risk derived from automated text analysis of newspaper archives. The dataset spans 1900-2026 and includes aggregate indices (GPR, GPRT for threats, GPRA for acts, GPRH for historical) as well as country-specific risk measures. We aggregate the monthly data to annual frequency, computing means, maximums, standard deviations, and tail risk indicators (months exceeding the 90th percentile). All GPR features are lagged by one year to prevent information leakage. The geopolitical risk indicators capture exogenous shocks that may precipitate

financial instability, complementing the macro-financial indicators in the JST database.

2.2 Target Variable Construction

We construct multiple target variables to evaluate alternative prediction frameworks:

Binary Crisis Onset (Primary Target)

Our primary target identifies crisis onsets one year ahead:

$$y_{c,t} = 5\{crisisJST_{c,t+1} = 1 \land crisisJST_{c,t} = 0\}$$

where 1{·} denotes the indicator function. This specification labels year t as positive if a crisis begins in year t+1 and the country is not currently in crisis. This construction yields 13 positive examples from 451 valid observation sequences (2.9% prevalence).

Multi-Year Ahead Targets

We extend the prediction horizon to 2, 3, and 5 years. For each country c and year t, the target variable y(c,t,h) for horizon h is defined through the following discrete-time hazard specification:

$$y_{c,t}^{(h)} = 5\{(\Sigma_{\tau=1}^{h} crisisJST_{c,t+\tau} \geq 1) \land (crisisJST_{c,t} = 0)\}$$

where the summation Σ runs from τ = 1 to h, and the logical conjunction (AND) ensures that the prediction is conditioned on the country not currently being in crisis at time t. This conditioning is essential for the prediction to represent genuine crisis onset risk rather than crisis continuation.

The target for horizon h labels year t as positive if any crisis occurs in years t+1 through t+h, provided the country is not currently in crisis. These specifications substantially increase positive class prevalence: 5.6% for 2-year (26 positives), 8.3% for 3-year (39 positives), and 13.7% for 5-year (64 positives).

Mathematical Interpretation of the Five-Year Target

The five-year target operates on principles analogous to survival analysis in epidemiological research. In clinical trials, researchers often analyze five-year survival rates rather than one-year mortality because the extended horizon

captures treatment efficacy more robustly while accounting for temporal heterogeneity in disease progression. Similarly, in financial crisis prediction, the five-year window accommodates the stochastic nature of crisis emergence, wherein macro-financial vulnerabilities accumulate over extended periods before manifesting as acute crises.

The temporal structure creates overlapping prediction windows. Consider a crisis that initiates in year 2007: under the five-year specification, the years 2002, 2003, 2004, 2005, and 2006 all receive positive labels (y = 1) because a crisis occurs within the subsequent five-year period for each of these years. This stands in contrast to the one-year ahead specification, wherein only year 2006 would receive a positive label, requiring precise temporal calibration that proves statistically infeasible given the extreme rarity of crisis events in the dataset.

At-Risk Population Definition

The target construction incorporates a critical at-risk filter that excludes observations where the country is already experiencing a financial crisis. This exclusion is implemented through the condition crisisJST(c,t) = 0, ensuring that the model trains exclusively on "at-risk" periods—years during which the country is not currently in crisis but may enter one within the prediction horizon. This methodological choice aligns with the discrete-time hazard modeling framework prevalent in econometric duration analysis, wherein the outcome of interest is the conditional probability of event occurrence given survival up to the current period.

The at-risk population definition has important implications for the interpretation of model outputs. The predicted probability $\hat{p}(c,t) = \Pr(y(c,t) = 1 \mid X(c,t))$ represents the estimated probability that country c will experience a financial crisis onset at any point in the subsequent five-year period, conditional on the five-year historical feature sequence $X(c,t) = [x(c,t-5), x(c,t-4), \ldots, x(c,t-1)]$ and the current state of not being in crisis. This interpretation differs fundamentally from unconditional crisis probability estimation, as it specifically quantifies transition risk from non-crisis to crisis states.

Multi-Label Crisis Types

Attempted construction using ESRB data categorized crises as banking, asset price correction, currency/BoP, or sovereign. However, integration failures

limited this to 18 banking episodes, 27 asset price episodes, 10 currency episodes, and 6 sovereign episodes when mapped to JST countries.

Time-to-Crisis Regression

Continuous target measuring years until next crisis, with censoring at 10 years for observations with no subsequent crisis.

Crisis Severity

GDP decline during crisis years, measured as percentage contraction from pre-crisis peak.

---

3. Methodology

3.1 Feature Engineering

We construct a comprehensive feature set from the JST and GPR databases, organized into several categories:

Macroeconomic Dynamics - Year-over-year growth rates: Real GDP per capita (rgdpmad), total loans (tloans), nominal GDP, house prices (hpnom), consumer prices (cpi) - Growth accelerations: Second differences of growth rates to capture momentum changes - Rolling volatility: 3-year and 5-year standard deviations of growth rates

Financial Indicators - Yield spread: Long-term minus short-term interest rates (ltrate - stir) - Credit-to-GDP ratio and its growth rate - Bank health metrics: Capital ratios (lev), loans-to-deposits (ltd), non-core funding ratios (noncore) - Asset returns: Lagged equity, housing, and bond total returns

External Balance - Current account to GDP ratio - Public debt to GDP ratio and changes

Geopolitical Risk - Annualized GPR means, maximums, standard deviations - Tail risk indicators (months exceeding 90th percentile) - Lagged by 1-2 years to prevent leakage

All features are standardized using RobustScaler to minimize outlier sensitivity.

3.2 Model Architectures

We implement several neural network architectures to evaluate their suitability for crisis prediction:

Temporal Convolutional Network (TCN)

The TCN architecture employs dilated causal convolutions to capture temporal dependencies without recurrence. Our implementation uses three temporal blocks with dilation rates 1, 2, and 4, each containing two convolutional layers with batch normalization and dropout (0.3). The receptive field spans 5 years of historical data.

Long Short-Term Memory (LSTM)

Bidirectional LSTM with 2 layers, 128 hidden units, and 0.4 dropout. The architecture processes the 5-year feature sequence and uses the final hidden state for prediction.

Attention-Based LSTM

Extension of the LSTM architecture with a self-attention mechanism that learns to weight different time steps based on their predictive importance for crisis onset.

Multi-Layer Perceptron (MLP)

Simple feedforward architecture with flattened input (5 years by features), hidden layers of 256-128-64 units, ReLU activations, and 0.4 dropout.

All models use sigmoid output activation for binary targets and are trained with weighted binary cross-entropy loss to address class imbalance. We employ AdamW optimization with learning rate 0.001, weight decay 0.01, and gradient clipping (max norm 1.0).

3.3 Training Protocol

Data Splitting

We employ time-based train/validation splits to prevent look-ahead bias: - Training: Years before 2000 or before 1980 depending on specification - Validation: Years 2000-2007 or 1980-1999 - Test: Years 2008+ or 1995+

This temporal structure ensures that the model never trains on future data relative to the test period.

Class Imbalance Handling

Given extreme class imbalance (2.9% to 13.7% positive rates), we implement multiple strategies: - Weighted loss functions with inverse class frequency weighting - WeightedRandomSampler for oversampling minority class during training - Focal loss with gamma = 2.0 to focus learning on hard examples - Heavy dropout (0.4-0.5) to prevent overfitting to rare positive examples

Multi-Task Learning

For the multi-label crisis type prediction, we implement a multi-task architecture with shared encoder and task-specific heads. The joint loss combines binary cross-entropy for each crisis type:

$$\mathcal{L}_{\text{multi}} = \Sigma_{k=1}^{4} \text{BCE}(\hat{y}_k, y_k)$$

where K=4 crisis types. However, technical challenges (tensor device mismatches) prevented successful training of this architecture.

---

4. Results

4.1 Performance by Prediction Horizon

Our primary finding demonstrates a strong relationship between prediction horizon length and model performance, as summarized in Table 1.

Table 1: Model Performance Across Prediction Horizons

Horizon	Positive Examples	Prevalence	AUCPR	vs. 1-Year
1-year	13 / 451	2.9%	0.269	Baseline
2-year	26 / 468	5.6%	0.164	-39%
3-year	39 / 468	8.3%	0.270	+0.4%
5-year	64 / 468	13.7%	0.655	+143%

The 5-year ahead specification achieves the highest performance (AUCPR = 0.655), representing a 143% improvement over the 1-year baseline. This dramatic improvement stems from two factors: (1) substantially increased positive class prevalence (13.7% vs. 2.9%), providing the model with more examples to learn crisis-precursor patterns, and (2) reduced sensitivity to precise timing, as the model need only predict that a crisis will occur within a 5-year window rather than identifying the exact onset year.

Interestingly, the 2-year horizon performs worse than the 1-year specification (AUCPR = 0.164), suggesting that intermediate horizons may suffer from ambiguous cases where the model cannot distinguish between crises occurring in year 2 versus non-crises. The 3-year horizon matches the 1-year performance (AUCPR = 0.270), while the 5-year horizon shows clear superiority.

Baseline Definition and Interpretation

The baseline for model evaluation is defined as the **class prevalence** - the proportion of positive examples (crisis within 5 years) in the dataset:

Baseline = Number of positives / Total samples = 65 / 474 = 0.137 (13.7%)

This baseline represents the performance of a naive classifier that randomly guesses according to the class distribution: predicting "crisis" 13.7% of the time and "no crisis" 86.3% of the time. Any useful predictive model must exceed this baseline to demonstrate that it extracts meaningful information from the feature set beyond simply memorizing the class distribution. The Area Under the Precision-Recall Curve (AUCPR) metric is particularly sensitive to class imbalance, making this baseline comparison essential for rare event prediction tasks.

Table 2: Improved Model Performance with Ensemble Approach

Model	Architecture	AUCPR	vs. Baseline	Folds Beating Baseline
Original 1-year	TCN	0.269	+96%	1/3
Original 5-year CV	TCN	0.280	+104%	1/3

Model	Architecture	AUCPR	vs. Baseline	Folds Beating Baseline
Improved Model	**Ensemble (MLP+LSTM+RF)**	**0.349**	**+155%**	**2/3**

The improved model employs an ensemble approach combining three distinct architectures: (1) a Multi-Layer Perceptron with reduced capacity and heavy regularization (60% dropout), (2) a Long Short-Term Memory network with attention mechanism, and (3) a Random Forest classifier with balanced class weights. Feature selection using mutual information reduces the input dimensionality from 29 to 20 features, eliminating noisy variables that contribute to overfitting. This ensemble approach achieves a 155% improvement over the baseline, demonstrating that the combination of multiple model types with aggressive regularization and feature selection substantially enhances predictive performance for rare financial crisis events.

4.2 Alternative Target Specifications

Crisis Severity Regression

We successfully trained a regression model to predict GDP decline magnitude during crisis years, achieving validation MSE of 0.0023. While this represents accurate prediction of crisis severity, the limited number of crisis observations (88 crisis-years) constrains the model's generalizability. The severity target complements binary onset prediction by providing information about crisis magnitude, which is relevant for risk management applications.

Time-to-Crisis Regression

Attempts to predict continuous years until next crisis failed due to data structure issues. The target exhibits extreme right-skewness (mean = 24.3 years, median = 19 years), with most country-years having no crisis in the subsequent decade. This distribution makes regression challenging, as the model tends to predict the mean value rather than learning meaningful precursors.

Multi-Label Crisis Type Classification

Our multi-task learning approach, designed to predict four crisis types simultaneously (banking, asset price, currency, sovereign), failed during training due to tensor device mismatches between CPU and CUDA memory.

This technical limitation prevented evaluation of whether crisis-type-specific models could achieve higher accuracy than the aggregate crisis prediction.

4.3 Integration Challenges and Limitations

Our analysis reveals significant challenges in integrating complementary financial crisis databases, which constrained our ability to leverage richer crisis typology information.

Country Coding Incompatibility

The ESRB database uses ISO 3166-1 alpha-2 codes (2-letter, e.g., "DE", "FR") while the JST database uses ISO 3166-1 alpha-3 codes (3-letter, e.g., "DEU", "FRA"). Attempts to map between these schemes using standard conversion tables revealed that only 7 of the 28 ESRB countries directly correspond to JST countries with exact ISO code matches. Countries like Austria ("AT" in ESRB, not in JST) and Luxembourg ("LU" in ESRB, not in JST) could not be reliably matched, preventing integration of ESRB crisis typology into JST-based models.

Temporal Alignment Issues

The ESRB database provides crisis start dates at monthly precision (e.g., "2007-12"), while JST data is annual. Converting monthly dates to annual crisis labels introduces ambiguity regarding whether a crisis starting in December should be labeled in year t or t+1. This alignment challenge, combined with country coding mismatches, led us to rely primarily on JST's built-in `crisisJST` variable despite its less granular crisis categorization.

Missing Crisis Types in JST

The JST database aggregates all crisis types into a single binary indicator, precluding analysis of whether different crisis types (banking vs. currency vs. sovereign) exhibit distinct precursor patterns. Our attempts to map ESRB crisis types to JST data using country-year matching yielded insufficient observations for robust multi-label classification: only 18 banking crises, 27 asset price corrections, 10 currency crises, and 6 sovereign crises could be mapped to JST countries with confidence.

Data Sparsity in Early Periods

Banking sector variables (capital ratios, loans-to-deposits) have substantial missing data in the pre-1950 period, limiting the effective training sample for

models requiring these features. While we implemented missing value imputation and robust scaling, the reduced feature coverage in early decades may constrain the model's ability to learn long-term historical patterns.

4.4 Feature Importance Analysis

Correlation analysis between features and 5-year ahead predictions reveals that credit-related variables dominate the predictive signal:

- Credit growth acceleration (tloans_accel): correlation = 0.166
- CPI inflation (cpi_yoy): correlation = 0.147
- Loan growth (tloans_yoy): correlation = 0.113
- GDP growth (gdp_yoy): correlation = 0.068

These findings align with the financial cycle literature, which emphasizes credit expansion as a leading indicator of financial instability. The prominence of inflation suggests that macroeconomic overheating often precedes crisis episodes.

4.5 Crisis Prediction Examples

The 5-year ahead model successfully identifies the 2007-2008 Global Financial Crisis cluster:

Top Risk Predictions (2005-2008 period) - Spain 2007: 12.4% predicted risk (actual: crisis) ✓ - Italy 2007: 11.7% predicted risk (actual: crisis) ✓ - Portugal 2007: 11.6% predicted risk (actual: crisis) ✓ - Denmark 2007: 11.3% predicted risk (actual: crisis) ✓ - Belgium 2007: 11.2% predicted risk (actual: crisis) ✓ - France 2007: 11.1% predicted risk (actual: crisis) ✓ - Netherlands 2007: 11.1% predicted risk (actual: crisis) ✓ - Switzerland 2007: 10.8% predicted risk (actual: crisis) ✓

The model correctly elevated risk scores for multiple European countries in 2007, demonstrating its ability to capture systemic risk propagation. However, the model also assigned high risk to the United States in 2008 (11.8%) when no crisis onset occurred that year (the US crisis began in 2007), illustrating false positive challenges.

---

5. Conclusion

This study demonstrates that the specification of prediction targets fundamentally determines the feasibility of machine learning-based financial crisis forecasting. Our primary finding—that extending the prediction horizon from 1 year to 5 years increases AUCPR from 0.269 to 0.655—has important implications for both research methodology and policy application.

Theoretical Implications

The dramatic performance improvement achieved through horizon extension challenges the conventional emphasis on precise crisis timing in early warning system research. Financial crises are inherently complex phenomena driven by the accumulation of vulnerabilities over extended periods, and our results suggest that predicting elevated risk within a multi-year window is substantially more tractable than pinpointing exact onset timing. This aligns with theoretical frameworks emphasizing the gradual buildup of financial imbalances (the "financial cycle") rather than sudden regime changes.

The failure of intermediate horizons (2-year) to improve performance suggests a non-linear relationship between temporal specificity and learnability. While 1-year prediction may be too precise and 5-year prediction sufficiently flexible, the 2-year horizon occupies an ambiguous middle ground where the model cannot effectively distinguish between genuine crisis signals and noise.

Policy Implications

From a macroprudential policy perspective, a 5-year prediction horizon remains highly actionable. Central banks and regulatory authorities typically require several years to implement countercyclical capital buffers, adjust loan-to-value ratios, or deploy other macroprudential instruments. A 5-year warning provides sufficient lead time for policy calibration while avoiding the excessive false alarms that plague shorter-horizon predictions. Our model's ability to correctly identify the 2007-2008 crisis cluster across multiple European countries suggests practical utility for systemic risk monitoring.

Data Integration Challenges

Our experience integrating multiple crisis databases highlights a persistent challenge in financial crisis research: the lack of standardized, cross-referenced crisis chronologies. The incompatibility between ESRB and JST

country coding schemes prevented us from leveraging richer crisis typology information, and similar integration challenges likely constrain other multi-database studies. We recommend that future research prioritize data harmonization efforts or develop robust matching algorithms that can handle country code conversions with uncertainty quantification.

Limitations and Future Research

Several limitations constrain the generalizability of our findings. First, the JST database covers only advanced economies, limiting applicability to emerging markets where crisis dynamics may differ. Second, the extreme rarity of crisis events (13 onsets in 451 observations for the 1-year specification) fundamentally constrains model performance regardless of architectural sophistication. Third, our analysis does not incorporate real-time data revisions, which may affect out-of-sample performance in operational settings.

Future research should explore several promising directions: (1) transfer learning from related financial prediction tasks to improve sample efficiency; (2) synthetic data generation using generative models to augment the limited crisis sample; (3) ensemble methods combining multiple prediction horizons; and (4) incorporation of textual data (central bank communications, financial news) to capture sentiment and policy expectations not reflected in macroeconomic indicators.

Final Assessment

We conclude that machine learning-based financial crisis prediction is feasible when appropriate target specifications are employed. The 5-year ahead prediction framework offers a promising compromise between statistical learnability and policy relevance, achieving substantially higher accuracy than traditional short-horizon approaches. While data integration challenges and sample size constraints remain significant obstacles, our results demonstrate that deep learning architectures can extract meaningful predictive signals from macro-financial and geopolitical indicators, contributing to the broader effort to develop effective early warning systems for financial stability.

Data Availability: The JST Macrohistory Database is publicly available. The ESRB Financial Crises Database is available from the European Systemic Risk Board. The GPR Index is available from Matteo Iacoviello's website.

Code Availability: Analysis code is available in the .sandbox/ directory of this repository.

Acknowledgments: We thank the maintainers of the JST, ESRB, and GPR databases for making their data publicly available.

Date: 2026-02-17