Financial Market Crises Prediction using AI-Based Analytics for Risk Management Applications

A Hybrid Bi-Directional LSTM and Transformer Framework with Multi-Modal Data Integration

Knowledge Gaps, Research Design, and Experimental Methodology

Hemant Bhanawat[1], Otilia Manta[2]

[1] *Postdoctoral Researcher, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia*

[2] *Supervisor – Lincoln University College Program*

Romanian Academy, Bucharest, Romania

Romanian-American University, Bucharest, Romania

pdf.hemantbhanawat@lincoln.edu.my

Abstract

Building on the foundational literature synthesis presented at Conference 1, this paper shifts focus to the methodological core of the research programme. The study addresses a well-established inadequacy in the financial crisis prediction literature: the absence of a unified, real-time, multi-factor early warning framework capable of operating across multiple countries, asset classes, and data modalities simultaneously. Five critical knowledge gaps are formally defined and operationalised into testable experimental designs. The proposed methodology integrates Bi-Directional Long Short-Term Memory (Bi-LSTM) networks, Transformer-based temporal encoders, dynamic Graph Neural Networks, Hidden Markov Model regime-switching mechanisms, and Bayesian uncertainty quantification — applied to a curated multi-modal dataset spanning India, the United States, the European Union, and China from 2000 to 2024. Each experiment is mapped directly to a research hypothesis, supported by a defined dataset, specified tools, and measurable outcome criteria. The paper demonstrates methodological rigour by aligning experimental design with validation protocols drawn from both machine learning benchmarking standards and financial econometric practice.

Keywords: Knowledge gaps; Research methodology; Bi-LSTM; Transformer; Graph Neural Networks; Regime switching; Experimental design; Financial crisis prediction; Multi-modal data; Explainable AI; SHAP; Benchmarking

1. Introduction

Conference 1 established the research's theoretical and literature foundation: four thematic streams were synthesised, five hypotheses were articulated, and the broad architectural contours of the proposed hybrid early warning system were outlined. Conference 2 now demands a different kind of intellectual discipline — not the breadth of a literature review but the precision of a research design. The central task of this paper is to answer, rigorously, two questions: what exactly does the existing body of knowledge not yet know, and how, specifically, will this research find out?

These two questions structure the paper. Section 2 provides a formal definition of the five knowledge gaps identified from the literature, moving beyond descriptive summary to operationalise each gap as a researchable problem with specific measurable dimensions. Section 3 presents the full experimental methodology — dataset construction, model architecture, training protocol, validation approach, and benchmarking framework — in sufficient technical detail to be reproducible. Section 4 maps each hypothesis to its corresponding experiment, specifying data inputs, algorithmic procedures, and outcome metrics. Section 5 discusses the methodological limitations and quality assurance measures embedded in the research design. Section 6 concludes with a summary of the Conference 2 deliverables and a forward map to Conference 3.

2. Formal Definition of Knowledge Gaps

The following five knowledge gaps are not merely observations about what the literature has not done — they are formally bounded statements about the specific dimensions of the financial crisis prediction problem that remain empirically unresolved. Each gap is defined in terms of what currently exists, what is missing, and why the missing element matters for both theory and practice.

Gap	Label	What Currently Exists	What Is Missing	Why It Matters
G1	**Absence of Real-Time Multi-Factor EWS**	Monthly/quarterly macro EWS models (KLR, Logit, GARCH) operating on 1–3 indicator variables	No system continuously ingests high-frequency multi-asset data with AI-driven probabilistic alert generation	Crises develop over weeks; monthly data creates a structural detection lag of 4–8 weeks that makes pre-emptive intervention impossible
G2	**No Multi-Country Deep Learning Crisis Framework**	Bi-LSTM and Transformer studies exist but are all single-country, single-asset applications (Qiu et al., 2020; Fischer & Krauss, 2018)	No study models cross-border contagion dynamics among India, USA, EU, and China within a unified deep learning architecture	Cross-border crisis transmission accounts for 60–70% of crisis severity in emerging market episodes; ignoring it systematically underestimates systemic risk
G3	**No Hybrid Predictive-Forensic Architecture**	AI models serve either prediction (probability scoring) or forensics (post-event attribution) — never both simultaneously	No single deployable system combines forward-looking crisis probability with real-time SHAP-based causal attribution	Regulators require both early warning AND auditable explanation in the same tool; current literature cannot provide this dual output from one architecture
G4	**No Regulatory-Grade Explainability in Crisis Prediction**	SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) exist and are applied in other domains but not financial crisis prediction	No published crisis prediction study has delivered SHAP-attributed, variable-level alert justifications meeting central bank auditability requirements	Without interpretable outputs, AI EWS cannot be legally deployed by regulatory bodies under emerging algorithmic accountability frameworks (EU AI Act, 2024)
G5	**No Real-Time Dynamic Contagion Monitoring**	CoVaR (Adrian & Brunnermeier, 2016) and MES (Acharya et al.,	No AI framework monitors dynamic changes in financial	Financial network structure changes substantially in the weeks before a crisis; a system

		2010) measure static historical network risk — adopted by ECB and Fed for stress testing	network topology in real time and generates contagion probability alerts	blind to this topology shift cannot detect the precursor signal it carries

Table 1. Formal Definition of Five Knowledge Gaps: Existing Evidence, Missing Elements, and Research Significance

2.1 Gap Operationalisation: From Descriptive to Measurable

Each gap above translates into a specific measurable research deficiency. G1 is operationalised as: no published EWS generates crisis probability outputs with a lead time exceeding four weeks using daily-frequency multi-asset data. G2 is operationalised as: no Bi-LSTM or Transformer model has been trained and validated on a four-country, four-asset-class crisis dataset with cross-country correlation features. G3 is operationalised as: no published model produces both a probability score and a variable-level attribution vector from the same forward pass. G4 is operationalised as: no crisis prediction study reports SHAP values for individual prediction outputs alongside model performance metrics. G5 is operationalised as: no AI framework updates and reads financial network adjacency matrices at daily or sub-daily frequency and incorporates the resulting topology features as crisis predictors.

These operationalised definitions are not merely semantic refinements. They determine, directly, what data must be collected, what model components must be built, what experiments must be run, and what metrics must be reported to claim that each gap has been addressed. The experimental design in Section 3 is structured around these requirements.

3. Research Methodology and Experimental Design

3.1 Overall Research Design

The study employs a quantitative, longitudinal, multi-method experimental design. The overarching goal is to construct, train, validate, and compare a hybrid AI early warning system against eight benchmark models across four crisis episodes and four national financial markets spanning 2000 to 2024. The design is explicitly hypothesis-driven: each of the five hypotheses stated at Conference 1 is assigned a corresponding experiment with pre-specified inputs, procedures, and outcome criteria. Walk-forward cross-validation — the methodological gold standard for time-series model evaluation — is used throughout to prevent data leakage and ensure that performance estimates reflect genuine out-of-sample generalisation.

3.2 Dataset Construction

3.2.1 Data Sources and Variable Categories

The dataset integrates four variable categories across four economies (India, USA, EU, China) at daily frequency from January 2000 to December 2024. Table 2 presents the complete data architecture.

Category	Variables	Source	Frequency	Gaps Addressed
Market Data	Equity indices (Nifty 50, S&P 500, DAX 40, Hang Seng, FTSE 100); OHLCV; 10Y-2Y yield spread; IG/HY	Bloomberg Terminal, Refinitiv, Yahoo Finance	Daily	**G1, G2, G5**

	CDS spread; EUR/USD, USD/INR, DXY; VIX, India VIX			
Macroeconomic	Real GDP growth (QoQ); CPI inflation; current account/GDP; M2 growth; credit-to-GDP gap (BIS method); FII net flows; sovereign debt/GDP	FRED, IMF DataMapper, World Bank, RBI Statistical Databases	Monthly / Quarterly (interpolated to daily)	**G1, G2**
Network Topology	Rolling 30-day cross-market correlation matrices (10 markets); BIS bilateral bank exposure data; dynamic network density, clustering coefficient, average path length	BIS, Refinitiv, computed from market data	Daily (computed)	**G2, G5**
Multi-Modal Sentiment	Bloomberg News Sentiment Score; Reuters NLP Index; FinBERT-classified news (positive/negative/neutral probability); Google Trends: 'financial crisis', 'bank run', 'stock crash'; FOMC/ECB/RBI communication tone (Transformer NLP)	Bloomberg, Reuters, FinBERT model, Google Trends API, Fed/ECB/RBI archives	Daily (news); Monthly (central bank)	**G1, G3**
Crisis Labels	Binary crisis indicator (0/1) per country-day, labelled using Reinhart-Rogoff (2009) crisis dates; 90-day pre-crisis window labelled as 'stress' (class 2) for three-class specification	Reinhart-Rogoff (2009), BIS Crisis Database, IMF Crisis Dates	Daily	**All Gaps**

Table 2. Multi-Modal Dataset Architecture: Variable Categories, Sources, Frequency, and Gap Linkage

3.2.2 Pre-Processing Pipeline

The pre-processing pipeline applies six sequential steps: (1) missing data imputation using forward-fill for market data and cubic spline interpolation for macroeconomic series; (2) stationarity testing using the Augmented Dickey-Fuller (ADF) test and KPSS test, with first-differencing applied to non-stationary series; (3) feature scaling using min-max normalisation within each rolling 252-day training window to prevent data leakage; (4) lag feature engineering creating T-1 through T-20 lagged versions of all core variables; (5) rolling statistical features including 10-day, 30-day, and 90-day moving averages, standard deviations, and cross-asset Pearson correlations; and (6) class imbalance

handling using SMOTE (Synthetic Minority Over-sampling Technique) oversampling on the training set, given that crisis observations represent approximately 8–12% of total observations across the four-country dataset.

3.3 Hybrid Model Architecture

3.3.1 Temporal Encoding Module: Bi-LSTM and Transformer

The temporal encoding module employs two parallel sub-architectures whose outputs are subsequently fused. The first is a three-layer Bi-Directional LSTM network. Each LSTM layer consists of 128 hidden units in each direction (256 total per layer), with a dropout rate of 0.30 between layers. The input sequence length is set to 90 trading days, reflecting the empirical observation that systemic crisis precursors typically develop across a three-to-six month horizon. Bidirectionality enables the model to process both the forward temporal trajectory of crisis build-up and the backward temporal pattern from confirmed crisis outcomes — a property of particular value during walk-forward backtesting on historical episodes.

The second sub-architecture is a Temporal Fusion Transformer (Lim et al., 2021) configured with eight attention heads, a model dimension of 128, and two encoder layers. The TFT's multi-head self-attention mechanism dynamically weights the importance of each historical time step for the current prediction — complementing the LSTM's sequential gating with a global attention view over the full input window. The two pathways are fused through a learned attention-weighted concatenation layer that produces a 512-dimensional state vector passed to the prediction head.

3.3.2 Contagion Module: Dynamic Graph Neural Network

The contagion module represents the financial system as a dynamic graph $G(t) = (V, E(t))$, where nodes V correspond to the ten financial markets in the dataset and edge weights $E(t)$ are the pairwise Pearson correlations computed on a rolling 30-day window, updated daily. A Graph Convolutional Network (GCN — Kipf & Welling, 2017) with two convolutional layers and a graph-level pooling layer learns representations of network topology that are passed as a 64-dimensional contagion embedding to the final prediction layer. The GNN is specifically trained to detect the topology signature that Forbes and Rigobon (2002) identify with genuine financial contagion: a structural break in cross-market correlations that exceeds two standard deviations of the historical correlation distribution.

3.3.3 Regime-Switching Module: Hidden Markov Model

A three-state Hidden Markov Model (expansion, stress, crisis) is fitted to the joint distribution of realised volatility, yield spreads, and credit spreads using the Baum-Welch algorithm. The HMM operates as a wrapper layer: its regime posterior probability vector is concatenated to the fused temporal-contagion representation and passed through regime-specific output heads. This ensures that the model's implicit weighting of predictor variables shifts appropriately across market phases — a design motivated by the well-documented empirical finding that variable-outcome relationships in financial data are highly non-stationary across regimes.

3.3.4 Uncertainty Quantification: Monte Carlo Dropout and Bayesian Output Layer

Probabilistic uncertainty quantification is implemented through two mechanisms. During training, standard dropout (rate = 0.25) is applied in all LSTM and Transformer layers. During inference, dropout is kept active and 100 forward passes are executed per input, producing a distribution of crisis probability estimates from which the mean (point estimate) and 95% credible interval are extracted. A Bayesian Neural Network output layer using variational inference with a reparameterisation trick (Gal & Ghahramani, 2016) further calibrates output uncertainty by placing prior distributions over network weights. This combination produces crisis probability outputs of the form $P(\text{crisis} \mid X_t) = \mu \pm \sigma$, where both the central estimate and confidence interval are reported to institutional users.

3.3.5 Explainability Layer: SHAP and LIME

Post-prediction explainability is implemented using two complementary methods. SHAP GradientExplainer computes Shapley additive explanation values for every prediction output, decomposing the crisis probability into the marginal contributions of each input feature at each time step. LIME generates locally faithful linear surrogate models around individual prediction instances to provide a second, model-agnostic attribution. Both outputs are integrated into the system's alert dashboard, which displays, alongside every crisis probability estimate, the top five contributing variables ranked by absolute SHAP value, enabling regulatory users to audit the model's reasoning for any specific alert.

3.4 Training Protocol

3.4.1 Walk-Forward Cross-Validation

The study employs an expanding-window walk-forward validation scheme, the methodological standard for time-series model evaluation. The training set begins at January 2000 and expands in annual increments; the test window is fixed at 252 trading days (one calendar year). This produces twelve validation folds. Three of the twelve folds contain a major crisis episode (2008, 2010–12, 2020); the remaining nine constitute non-crisis validation. This deliberate asymmetry is intentional: the model must demonstrate that it maintains low false-positive rates during non-crisis periods as well as high sensitivity during crisis periods — a balance that point-in-time evaluation on crisis windows alone cannot assess.

3.4.2 Hyperparameter Optimisation

Bayesian hyperparameter optimisation using the Optuna framework is applied to the following parameters: LSTM hidden units {64, 128, 256}, dropout rate {0.1, 0.2, 0.3, 0.4}, learning rate {1×10^{-4} to 1×10^{-2}}, batch size {32, 64, 128}, Transformer attention heads {4, 8, 16}, and GNN convolutional layers {1, 2, 3}. A Tree-structured Parzen Estimator (TPE) acquisition function is used with 150 trials per fold. The objective function is PR-AUC on the validation set, selected in preference to standard AUC-ROC because crisis prediction is a severely imbalanced classification problem in which AUC-ROC is known to overestimate performance (Davis & Goadrich, 2006).

3.4.3 Implementation Environment

The model is implemented in Python 3.11 using TensorFlow 2.14 and Keras for the Bi-LSTM and Transformer components, PyTorch Geometric 2.4 for the GNN component, and the hmmlearn library for the HMM regime-switching layer. SHAP 0.44 and LIME 0.2 are used for explainability. Training is conducted on NVIDIA A100 GPU instances. Reproducibility is ensured through fixed random seeds (42 for all stochastic components), version-locked dependency management via conda environment files, and complete code archiving in a private GitHub repository to be made public upon publication.

3.5 Benchmark Models and Evaluation Metrics

3.5.1 Benchmark Models

Eight benchmark models are specified, covering the full methodological spectrum from classical EWS to modern deep learning, ensuring that performance gains can be attributed to specific architectural choices rather than data advantages. Table 3 presents the benchmark specification.

#	Model	Type / Rationale	Key Parameters	Gap(s) Tested

B1	**GARCH-BEKK**	Volatility baseline — establishes the performance floor for traditional EWS	Bivariate GARCH-BEKK; p=1, q=1; MLE estimation; 252-day rolling window	**G1**
B2	**VAR (p=4)**	Macroeconomic multivariate baseline	4-lag VAR on GDP, CPI, credit gap, yield spread; Cholesky identification	**G1**
B3	**Binary Logit**	Classical probabilistic EWS baseline (Demirguc-Kunt & Detragiache, 1998)	Stepwise variable selection; clustered standard errors; AUROC threshold optimisation	**G1, G3**
B4	**KLR Signal Extraction**	First-generation EWS benchmark (Kaminsky et al., 1999)	15 indicators; percentile threshold = 80th; signal window = 24 months	**G1**
B5	**Random Forest**	Ensemble ML baseline	500 trees; max depth 10; min samples leaf 5; Gini criterion; OOB error	**G1, G2**
B6	**XGBoost**	Gradient boosting baseline (best-performing ML in ECB study)	η=0.05; max depth 6; subsample 0.8; colsample 0.8; early stopping	**G1, G2**
B7	**CNN-LSTM Hybrid**	Deep learning sequence baseline	1D CNN (32 filters, kernel 3) → 2-layer LSTM (128 units); same data as proposed model	**G1, G2**
B8	**TFT (standalone)**	Architectural ablation — isolates Transformer contribution	Same TFT as in proposed model but without Bi-LSTM, GNN, or HMM modules	**G1, G2, G3**

Table 3. Eight Benchmark Models: Specification, Rationale, and Knowledge Gap Linkage

3.5.2 Evaluation Metrics

Table 4 specifies the full evaluation metric framework. Metrics are selected to address the specific statistical challenges of crisis classification: severe class imbalance, probabilistic output, and the operational primacy of lead-time over raw accuracy.

Metric	Formula / Definition	Rationale	Target (Proposed Model)
F1-Score	*2×(P×R)/(P+R) where P=Precision, R=Recall*	Harmonic mean balancing false positives and false negatives — appropriate for imbalanced classes	**> 0.83 across ≥3 of 4 crisis episodes**

PR-AUC	*Area under Precision-Recall curve*	Primary metric for imbalanced classification; preferred over AUC-ROC (Davis & Goadrich, 2006)	**> 0.80 on all validation folds**
Brier Score	$(1/N)\Sigma(\hat{p}_i - y_i)^2$	Proper scoring rule for probability calibration; measures how well confidence matches outcomes	**< 0.10 (lower = better)**
Lead-Time Gain	*Weeks of advance warning before Reinhart-Rogoff onset date*	Primary operational effectiveness metric for an EWS — captures actionable advantage	**6–8 weeks ahead of crisis onset**
Accuracy	$(TP+TN)/(TP+TN+FP+FN)$	Overall classification correctness — reported for comparability with published benchmarks	**> 87%**
Precision	$TP/(TP+FP)$	Rate at which crisis alerts are genuine — low precision means regulatory alert fatigue	**> 85%**
Recall	$TP/(TP+FN)$	Rate at which actual crises are detected — low recall means dangerous false security	**> 83%**
MCC	$(TP\times TN-FP\times FN)/\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}$	Matthews Correlation Coefficient — robust to class imbalance; single balanced quality measure	**> 0.75**

Table 4. Evaluation Metric Framework: Definitions, Rationale, and Performance Targets

4. Hypothesis-to-Experiment Mapping

Table 5 presents the complete mapping of each hypothesis to its corresponding experiment, specifying input data, methodology, evaluation criteria, and the knowledge gap it addresses. This mapping constitutes the core experimental programme.

H	Hypothesis	Experiment	Input Data	Key Method / Tool	Gap

H1	Hybrid model outperforms all 8 benchmarks on F1, PR-AUC, Brier Score	E1: Full benchmark comparison across 12 walk-forward folds	All 4 variable categories; 4 countries; 2000–2024	Walk-forward CV; Wilcoxon signed-rank test; Friedman rank test across folds	**G1, G2**
H2	Multi-modal fusion yields significantly higher PR-AUC than single-category models	E2: Ablation study — 4 data-category combinations tested	Market-only; Macro-only; Sentiment-only; Full multi-modal	One-way ANOVA with Tukey HSD post-hoc; PR-AUC as dependent variable	**G1, G3**
H3	GNN detects Forbes-Rigobon contagion breaks ≥4 weeks before crisis onset	E3: GNN topology analysis across all 4 crisis episodes; Forbes-Rigobon structural break test	Rolling correlation matrices; BIS exposure data; crisis onset dates	Chow test for structural break; GNN edge weight trajectory analysis; lead-time measurement	**G2, G5**
H4	HMM regime classifications are statistically consistent with ex-post expert labels	E4: HMM state sequence vs. NBER/CEPR expert recession/crisis dates; ablation vs. non-switching baseline	Volatility, spreads, credit indicators; NBER/CEPR dates	Cohen's κ agreement; McNemar test; classification error comparison (crisis onset subwindow)	**G1**
H5	SHAP rankings are consistent across all 4 crisis episodes (top-3 variables stable)	E5: SHAP attribution computed per episode; Spearman rank correlation across episode pairs	Full feature set; model outputs per crisis episode	SHAP GradientExplainer; Spearman ρ between SHAP rank vectors; bootstrapped 95% CI	**G3, G4**

Table 5. Hypothesis-to-Experiment Mapping: Inputs, Methods, and Knowledge Gap Linkage

5. Methodological Rigour and Quality Assurance

5.1 Threats to Internal Validity and Mitigation

Three threats to internal validity are identified and addressed. The first is data leakage — the inadvertent use of future information in model training, which inflates performance estimates in time-series contexts. Mitigation: strict expanding-window walk-forward validation ensures that each test fold contains only observations strictly after the training window closes; all scaling parameters are computed within the training window and applied to the test window without re-fitting. The second threat is class imbalance bias, in which a model achieves high accuracy by predominantly predicting the majority (non-crisis) class. Mitigation: SMOTE oversampling on the training set; PR-AUC and MCC as primary metrics; threshold optimisation on the validation fold rather than the test fold. The third threat is multiple comparison inflation, arising from the large number of model comparisons and metrics reported.

Mitigation: Bonferroni correction applied to all pairwise statistical significance tests; primary hypotheses are pre-registered with the LGPR Programme Office before data analysis begins.

5.2 External Validity and Generalisability

The study deliberately includes four structurally distinct crisis episodes and four economies representing different financial system architectures (market-based — USA; bank-based — EU; hybrid transitional — China; emerging — India). This diversity is designed to ensure that findings are not specific to any single crisis typology or financial system configuration. Stress testing under three extreme out-of-distribution scenarios (40% single-day equity drawdown; sudden stop in capital flows; multi-country simultaneous sovereign downgrade) further probes generalisability beyond the historical training data.

5.3 Reproducibility and Open Science

Full reproducibility is ensured through: fixed random seeds (42) for all stochastic components; conda environment lock files capturing all package versions; complete pre-processing pipeline code in a structured Python package; and archiving of the curated dataset (excluding proprietary Bloomberg data) in a public repository upon journal submission. The LGPR Programme Office receives a pre-registration document specifying all hypotheses, experimental procedures, and primary outcome metrics before the analysis phase begins, preventing post-hoc hypothesis adjustment.

6. Conference 2 Deliverables and Forward Map to Conference 3

6.1 Conference 2 Deliverables

The deliverables from this conference milestone are: (1) formal operationalisation of all five knowledge gaps with measurable research deficiency statements; (2) complete dataset specification with source documentation, variable definitions, and pre-processing pipeline; (3) full model architecture specification with layer configurations, hyperparameter search space, and training protocol; (4) eight benchmark model specifications with implementation parameters; (5) evaluation metric framework with pre-specified targets; (6) hypothesis-to-experiment mapping with statistical testing procedures; and (7) methodological quality assurance protocol covering internal validity, external validity, and reproducibility.

6.2 Forward Map: Conference 3 Scope

Conference 3 will present the empirical results of the five experiments defined here. The specific deliverables anticipated for Conference 3 are: trained and validated Bi-LSTM-Transformer-GNN model with Bayesian uncertainty quantification; full benchmark comparison table with statistical significance tests; SHAP attribution maps for each crisis episode; HMM regime sequence analysis; and the first complete draft of the journal article targeted at a SCI/WoS Q1 journal. Data collection and feature engineering are currently underway. Model training will commence following dataset finalisation and pre-registration submission.

References

Econometric Models and Early Warning Systems

- Adrian, T., & Brunnermeier, M. K. (2016). CoVaR. American Economic Review, 106(7), 1705–1741.
- Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. (2010). Measuring systemic risk. Review of Financial Studies, 30(1), 2–47.

- Allen, F., & Gale, D. (2000). Financial contagion. Journal of Political Economy, 108(1), 1–33.
- Bussiere, M., & Fratzscher, M. (2006). Towards a new early warning system of financial crises. Journal of International Money and Finance, 25(6), 953–973.
- Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of ICML 2006, 233–240.
- Demirguc-Kunt, A., & Detragiache, E. (1998). The determinants of banking crises in developing and developed countries. IMF Staff Papers, 45(1), 81–109.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987–1007.
- Forbes, K. J., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. Journal of Finance, 57(5), 2223–2261.
- 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.
- Kaminsky, G., Lizondo, S., & Reinhart, C. (1999). Leading indicators of currency crises. IMF Staff Papers, 45(1), 1–48.
- Reinhart, C. M., & Rogoff, K. S. (2009). This time is different: Eight centuries of financial folly. Princeton University Press.

Machine Learning and Ensemble Methods

- Alessi, L., & Detken, C. (2018). Identifying excessive credit growth and leverage. Journal of Financial Stability, 35, 215–225.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD 2016, 785–794.
- Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. Computers and Operations Research, 32(10), 2513–2522.
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168.

Deep Learning, Transformers, and Graph Neural Networks

- Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of ICML 2016, 48, 1050–1059.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
- Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. Proceedings of ICLR 2017.

- Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in NeurIPS, 30, 4765–4774.
- Qiu, J., Wang, B., & Zhou, C. (2020). Forecasting stock prices with long-short term memory neural network based on attention mechanism. PLOS ONE, 15(1), e0227222.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. Proceedings of KDD 2016, 1135–1144.
- Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. Proceedings of IEEE BigData 2019, 3285–3292.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
- Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv:2303.17564.
- Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. arXiv:2006.08097.