Argumentation Based Systems Journal

www.absj.org

eISSN: xXXX-xXXX

Dynamic Financial Stress Indicators Using Machine Learning: An Adaptive Composite Framework for Early Warning and Systemic Risk Assessment

Abdullah Kürşat Merter[1, *], Yavuz Selim Balcıoğlu[2]

[1] Department of Business Administration, Faculty of Business Administration, Gebze Technical University, Kocaeli, Türkiye
[2] Department of Management Information Systems, Faculty of Economics and Administrative Sciences, Doğuş University, İstanbul, Türkiye

ARTICLE INFO

Article history:
Received 25 November 2025
Received in revised form 10 January 2026
Accepted 14 January 2026
Available online 15 January 2026

Keywords:
Composite stress indicator; Machine learning; Financial market stress; Early warning systems; Gradient boosting

ABSTRACT

The present study proposes a machine learning-enhanced forecasting framework for financial stress that addresses critical limitations in static threshold approaches used by regulatory authorities. Utilizing a comprehensive dataset comprising daily S&P 500 information spanning multiple crisis episodes, we employ gradient boosting algorithms with dynamic threshold detection to predict market stress occurrences. The hybrid ensemble model utilized in this study has been shown to significantly surpass conventional econometric methods in terms of forecasting accuracy and the timeliness of early warning systems. The framework demonstrated a high degree of efficacy in predicting major crisis events during the hold-out period, exhibiting a substantial improvement in detection rates when compared to Federal Reserve indices. The application of feature importance analysis has yielded findings that demonstrate the presence of regime-dependent patterns. These findings indicate that there is a notable increase in sensitivity to real-economy variables, such as unemployment, during periods of recession. For practitioners, the continuous stress probability forecasts enable graduated risk management protocols and generate tangible portfolio gains. Researchers are particularly interested in the establishment of novel benchmarks for financial stress forecasting and in how machine learning can capture non-linear transmission mechanisms that conventional approaches cannot detect.

1. Introduction

The measurement of financial market stress and the provision of early warning systems remain critical challenges for economic analysis, primarily due to the structural inability of conventional methodologies to adapt to rapidly evolving systemic risks. Recent crises have demonstrated that static indicators often fail to capture the speed of contagion. For example, the S&P 500 index plummeted by over 30% during the panic surrounding the global pandemic of 2020, while in March 2023 the collapse of Silicon Valley Bank and Credit Suisse triggered a contagion across regional

* *Corresponding author.*
E-mail address: akmerter@gtu.edu.tr

https://doi.org/10.31181/absjxxxxx

banking sectors that conventional models largely missed until the onset of volatility [2]. The global financial crisis, precipitated by the emergence of SARS-CoV-2, has exposed significant deficiencies in established risk assessment methodologies. This is evidenced by the substantial decline in bank stocks despite the regulatory improvements implemented in the aftermath of the 2008 crisis [1]. While the repercussions of such episodes are known to engender disruptions in credit flows and precipitate the destabilization of asset valuations, the persistence of these failures highlights a fundamental gap in the capacity of monitoring tools to deliver timely warnings about systemic vulnerabilities before they materialize.

A central paradox in the extant literature pertains to the persistent reliance on composite stress indicators, including the Federal Reserve's National Financial Conditions Index (NFCI), the St. Louis Fed Financial Stress Index (STLFSI), and the Kansas City Fed Financial Stress Index (KCFSI), which employ static aggregation methodologies despite the manifest limitations of such methods in real-time applications [30,33,39]. Most of indicators are predicated on equal weighting, principal component analysis (PCA), or fixed correlation-based approaches, which implicitly assume constant relationships among economic variables across all market conditions. This assumption is fundamentally at odds with a substantial corpus of theoretical and empirical evidence detailing regime-dependent transmission mechanisms in financial markets [10,15,31]. Recent research in the field of complex systems has served to reinforce this contradiction, with the findings demonstrating that financial stability is governed by nonlinear interactions that static models are incapable of capturing [29] and that contagion mechanisms depend heavily on counterparty leverage thresholds that vary significantly across regimes [17].

Machine learning provides adaptive frameworks that encapsulate nonlinear interactions and temporal dependencies within high-dimensional financial data, thereby overcoming the static constraints of conventional methodologies. Recent applications demonstrate superior performance in comparison to conventional econometric approaches in predicting banking crises [11,34], sovereign defaults [4], and tail risk events [5]. Nevertheless, three significant gaps have been identified that constrain the present applications of machine learning in the domain of financial stress measurement. Firstly, extant studies predominantly focus on binary crisis classification rather than continuous stress probability assessment, thus limiting institutional utility for graduated risk management responses [41]. Secondly, most applications examine specific transmission channels in isolation, such as network connectedness or credit risk, rather than synthesizing comprehensive macro-financial indicators into unified monitoring frameworks suitable for regulatory oversight [47]. Thirdly, there is a paucity of research addressing the dynamic evolution of stress transmission thresholds across macroeconomic regimes. This is despite theoretical predictions that crisis amplification mechanisms activate at regime-dependent critical levels [3,44] and recent empirical calls for early-warning frameworks that can capture these systematic risk signals through explainable methodologies [20,45].

To address the methodological limitations identified, this study proposes a machine learning-enhanced composite financial stress indicator. This indicator integrates high-dimensional macroeconomic data with adaptive threshold detection. The proposed hybrid ensemble architecture combines XGBoost gradient boosting for stress classification, Random Forest for Value-at-Risk (VaR) prediction, and LSTM networks for volatility forecasting. In accordance with recent advances in high-dimensional vector autoregression with influencers [35], the present framework synthesizes 24 macro-financial variables—encompassing volatility, credit conditions, and sentiment—into a unified dynamic system. It is crucial to note that SHAP (SHapley Additive exPlanations) analysis is employed not merely for post-hoc interpretation but to operationalize a dynamic weighting mechanism where feature contributions evolve in response to detected market regimes.

The present research is guided by three primary inquiries. Firstly, the investigation will ascertain whether financial stress transmission exhibits nonlinear threshold effects that necessitate adaptive modelling approaches beyond linear aggregation. Secondly, the study examines whether critical stress thresholds are static or evolve dynamically in accordance with changing market structures. This question is motivated by recent findings on detecting multiple level shifts in bounded time series [16]. Thirdly, an evaluation will be conducted to ascertain whether a dynamic, machine learning-enhanced indicator provides statistically significant improvements in early warning capability and economic value when compared to static binary benchmarks currently utilized in regulatory practice.

The present study makes three principal contributions to the financial stability literature. Firstly, rigorous empirical validation is provided that stress transmission operates through regime-dependent non-linear channels. By modelling time-varying tail dependence [48], it is demonstrated that interaction effects between variables (e.g., credit spreads and unemployment) amplify significantly during crises, a phenomenon invisible to linear models. Secondly, the dynamic evolution of transmission thresholds is quantified. In contradistinction to static PCA weights, the adaptive framework elucidates systematic shifts in feature importance across business cycles, thereby confirming that the drivers of financial stress are structurally different during monetary tightening compared to liquidity crises. Thirdly, a robust benchmarking framework is established against Federal Reserve indices. To ensure the statistical validity of our findings, we employ multi-objective backtesting protocols [26], thereby confirming that the superior early warning signals generated by our model are robust to overfitting and translate into tangible economic gains for institutional risk management.

The practical implications of this research extend to both portfolio management and macroprudential oversight. The transition from binary "crisis/no-crisis" flags to continuous probability assessments enables a graduated risk management protocol, ranging from routine monitoring to emergency defensive positioning. This protocol reduces the cost of false alarms. For policymakers, the lag in static indices demonstrated during recent regime transitions underscores the necessity of "dual track" monitoring systems that complement traditional linear indicators with adaptive machine learning tools.

2. Literature Review and Hypothesis Development

2.1 Evolution of Composite Stress Indicators and Their Limitations

The development of composite financial stress indicators was driven by the recognition that systemic risk necessitates the aggregation of diverse market signals, rather than the utilization of isolated metrics. This approach was pioneered by Illing and Liu [37], who emphasized the computational challenge of synthesizing high-frequency data [12]. However, the subsequent evolution of these tools—exemplified by the Federal Reserve's National Financial Conditions Index (NFCI) and the ECB's Composite Indicator of Systemic Stress (CISS)—has largely relied on static aggregation methodologies such as PCA or fixed correlation weighting [33,39].

A critical deficiency in established frameworks is the assumption of invariant relationships across economic cycles [6]. Equal weighting schemes have been shown to be ineffective in capturing the dominance of specific channels during crisis transitions, resulting in an underestimation of stress [15]. In a similar manner, PCA-based approaches extract common factors based on historical averages, rendering them insensitive to the phase transitions characteristic of systemic contagion. Recent evidence from the Spanish interbank market demonstrates that contagion mechanisms are inherently nonlinear and dependent on counterparty leverage thresholds that shift dramatically

under stress [17]. This limitation was exposed during the 2020 equity-economic disconnect, where static loadings failed to capture the disconnect [2].

In addition, the limits of arbitrage framework posits that information efficiency is subject to variation across market states [44], thereby underscoring the necessity for indicators capable of adaptively recalibrating weights. Whilst recent studies in the field of complex systems have emphasized the significance of nonlinear interactions [29], conventional regulatory instruments continue to be constrained by linear assumptions. These flaws underscore a broader empirical gap: static models exhibit poor out-of-sample performance in volatile regimes, frequently generating excessive false alarms due to their inability to handle non-stationary network dynamics [22,42].

2.2 Machine Learning Applications in Financial Stress Assessment

Machine learning (ML) has transformed the field of financial stress assessment by offering frameworks capable of modelling high-dimensional non-linearities. In comparison to autoregressive baselines, tree-based models have been demonstrated to exhibit superior efficacy in predicting financial market stress distributions [4,5]. Beutel *et al.*, [11] and Hu *et al.*, [34] further validate the superiority of these ensemble methods over logistic regression in predicting banking crises, noting that machine learning (ML) captures interaction effects that linear models miss [11,34].

Nevertheless, the implementation of machine learning (ML) in this field is constrained by a trade-off between predictive accuracy and interpretability. Despite their potency, neural networks frequently operate as opaque systems, thereby constraining their efficacy for regulatory oversight [19]. Although recent studies utilizing explainable AI (XAI) techniques such as SHAP have begun to bridge this gap [45], a significant portion of the literature remains focused on binary classification rather than continuous probability assessment [41].

Furthermore, extant machine learning (ML) applications frequently demonstrate an insular focus. It is evident that studies frequently analyze specific channels in isolation. For instance, research may focus on network connectedness [47] or interbank contagion [20]. However, there is a paucity of research that considers holistic, regime-dependent composites [13]. Although ensemble methods have shown potential [7], there is still a need to fully realize their integration with dynamic threshold detection mechanisms. The present study addresses this lacuna by employing XGBoost to derive time-varying weights, thus addressing the need for comprehensive monitoring.

2.3 Theoretical Framework and Empirical Evidence for Dynamic Composite Indicators

The transition from static to dynamic composite indicators is founded upon three converging theoretical streams. Firstly, Regime-Switching Theory [31] posits that financial time series transition between states governed by different parameters. This suggests that variables such as unemployment may possess regime-dependent predictive capabilities, as evidenced by studies conducted by Jiang *et al.*, [38] and Goldstein *et al.*, [27]. Static weighting schemes are incompatible with this property, whereas adaptive ML models align with the theory by allowing weights to toggle based on the detected state.

Secondly, network theory posits that systemic risk emanates from evolving interconnectedness [3]. Recent methodological advances in the detection of multiple level shifts [16] and the modelling of time-varying tail dependence [48] serve to reinforce the necessity of dynamic approaches. Contagion pathways emerging during periods of stress often remain dormant during normal times [8], requiring models capable of simulating evolving interactions.

Thirdly, from a behavioral finance perspective, it is recognized that risk appetite is regime dependent [28]. During periods of heightened stress, behavioral biases become more pronounced, leading to the convergence of correlations toward a unified outcome. Empirical evidence has been provided to demonstrate that the computational adaptability employed in modelling these non-stationarities yields significant economic value [49], outperforming static benchmarks such as CoVaR in crisis detection [46].

2.4 Hypothesis Development

Drawing on these theoretical foundations and empirical gaps, we formulate two hypotheses addressing static indicator limitations:

Hypothesis 1: Financial stress transmission operates through non-linear threshold effects where amplification mechanisms activate at critical levels, rather than linear scaling assumed by traditional indicators.

This hypothesis builds on Bernanke *et al.*, [10] financial accelerator showing credit constraints amplify at thresholds, and Acemoglu *et al.*'s (2015) network theory documenting phase transitions in contagion. Empirical evidence from recent crises reveals discrete regime shifts undetectable by linear PCA but quantifiable through machine learning's recursive partitioning. We predict that interaction terms between credit conditions and liquidity measures will demonstrate significant additional predictive power beyond main effects, validated through permutation tests comparing models with and without interaction effects.

Hypothesis 2: Critical stress thresholds evolve dynamically in response to macroeconomic regime transitions, following predictable patterns amenable to adaptive machine learning detection.

Hamilton's [31] regime-switching framework predicts transmission mechanisms vary across expansion and recession states. We hypothesize that unemployment sensitivity increases during recessions while interest rate sensitivity rises during monetary tightening, reflecting regime-dependent amplification channels. Machine learning's adaptive retraining should capture these predictable patterns better than static methods, validated through rolling-window analysis documenting systematic feature importance shifts across identified regime periods.

3. Methodology

This section delineates the data sources, preprocessing steps, and analytical procedures utilized to construct and validate the dynamic composite stress indicator. By integrating traditional econometric techniques with advanced machine learning algorithms, the methodology overcomes the constraints of static aggregation methods, enabling adaptive weighting and robust performance evaluation across varied market regimes. The approach emphasizes transparency, interpretability, and regulatory compliance, aligning with standards for financial risk modeling.

3.1 Data Description

Empirical analysis employs comprehensive daily financial and economic data spanning January 2, 2015, to December 31, 2024 (2,609 trading days). This timeframe deliberately encompasses diverse regimes: the post-crisis recovery (2015–2019), the COVID-19 volatility shock (2020–2021), and the monetary tightening cycle (2022–2024).

Detailed documentation of all 14 base variables, including definitions, transformations, and interpolation methods for lower-frequency data, is provided in Table 1. The dataset encompasses

five primary categories: market data (S&P 500 Index, Trading Volume, VIX Volatility Index), interest rates (10-Year Treasury, Federal Funds Rate, 3-Month Treasury), macroeconomic indicators (GDP Growth, Unemployment Rate, Core CPI Inflation), credit indicators (Corporate Bond Spreads, Bankruptcy Rate, Commercial Paper Rate), and sentiment measures (Consumer Confidence, Economic Uncertainty). Daily data are obtained directly from Bloomberg and FRED, while monthly and quarterly series are interpolated to daily frequency using cubic spline methods to ensure temporal consistency. All financial data have been adjusted for dividends and stock splits to maintain accuracy. Missing observations (<0.8%) are imputed via forward-fill for short gaps and linear interpolation for longer periods. All variables are synchronized to the S&P 500 trading calendar. Stationarity is ensured through logarithmic returns or first differencing, verified by Augmented Dickey-Fuller tests.

Table 1 Complete variable documentation and sources

Variable Name	Source	Series ID / Ticker	Definition	Unit	Frequency	Transformation
S&P 500 Index	Bloomberg	SPX Index	Large-cap equity index	Price level	Daily	Log returns
Trading Volume	Bloomberg	SPX Index Volume	Total shares traded	Millions	Daily	None
VIX Volatility Index	CBOE	VIX Index	Implied volatility	Percentage	Daily	None
10-Year Treasury	FRED	GS10	Government bond yield	Percentage	Daily	First difference
Federal Funds Rate	FRED	FEDFUNDS	Policy interest rate	Percentage	Daily	First difference
3-Month Treasury	FRED	GS3M	Short-term rate	Percentage	Daily	First difference
GDP Growth	FRED	GDPC1	Real GDP quarterly growth	Percentage	Quarterly	Interpolation to daily
Unemployment Rate	BLS	UNRATE	Civilian unemployment	Percentage	Monthly	Interpolation to daily
Core CPI Inflation	FRED	CPILFESL	Core consumer prices YoY	Percentage	Monthly	Interpolation to daily
Corporate Bond Spreads	FRED	BAMLC0A4CBBB	BBB-Treasury spread	Basis points	Daily	None
Bankruptcy Rate	ABI	Custom	Business bankruptcy filings	Percentage	Monthly	Interpolation to daily
Commercial Paper Rate	FRED	RIFSPPNAAD90NB	90-day commercial paper	Percentage	Daily	First difference
Consumer Confidence	Conference Board	UMCSENT	Consumer sentiment index	Index level	Monthly	Interpolation to daily
Economic Uncertainty	PolicyUncertainty.com	EPU	Economic policy uncertainty	Index level	Daily	None

Note: FRED = Federal Reserve Economic Data, BLS = Bureau of Labor Statistics, ABI = American Bankruptcy Institute, CBOE = Chicago Board Options Exchange. Interpolation procedures use cubic spline methods for monthly and quarterly data conversion to daily frequency. All financial data adjusted for dividends and stock splits.

To validate the machine learning-enhanced indicator against established alternatives, a systematic comparison of performance was conducted with three widely used stress indices maintained by the Federal Reserve System: the National Financial Conditions Index (NFCI, FRED series NFCI) [14], the St. Louis Fed Financial Stress Index (STLFSI, FRED series STLFSI4) [39], and the Kansas

City Fed Financial Stress Index (KCFSI) [30]. These indices, which employ PCA on broad variable sets, serve as the regulatory standard. In order to ensure comparability, all indices are converted to daily frequency and mapped to binary stress classifications.

3.2 Construction of the Composite Stress Indicator

The primary innovation presented in this study is a machine learning-enhanced indicator that dynamically weights components based on regime-adaptive correlations. The model employs the XGBoost gradient boosting framework, with the objective of minimizing the standard regularized logistic objective function [18]. The execution of K = 1,000 boosting rounds with a learning rate of η = 0.1 is conducted on rolling 500-day windows.

Feature engineering constructs p = 47 variables, including lagged interactions (e.g., unemployment x credit spreads), GARCH (1,1) volatility residuals, and equity-bond cross-correlations. Unlike static models [46], we implement dynamic feature importance weighting using time-varying SHAP values [40]:

$$w_{j,t} = \frac{|\bar{\phi}_{j,t}|}{\sum_{j=1}^{p}|\bar{\phi}_{j,t}|} \quad (1)$$

where $\bar{\phi}_{j,t}$ represents the mean absolute SHAP value for feature j over the rolling estimation window $[t - 500, t]$. This approach enables adaptive recalibration; for instance, unemployment sensitivity empirically rises from 28% during expansions to 42% during recessions.

The final indicator produces a continuous probability $P(Stress_t \mid X_t) \in [0,1]$ via a logistic transformation. This continuous scale facilitates graduated risk protocols: P < 0.25 (monitoring), 0.25 ≤ P < 0.50 (surveillance), 0.50 ≤ P < 0.75 (defensive), and P ≥ 0.75 (emergency).

The specific timeline of training, validation, and hold-out test periods is detailed in Table 2. The training period (January 2015 – December 2019; N=1,258 observations) encompasses the 2015-16 commodity crisis, 2018 trade tensions, and late Fed tightening cycle, achieving 81.3% training accuracy with an AUC of 0.867. The validation period (January 2020 – June 2022; N=628 observations) includes the COVID-19 crash and stimulus-driven recovery, yielding 77.8% accuracy and an AUC of 0.824. The hold-out test period (July 2022 – December 2024; N=723 observations) represents genuine out-of-sample evaluation, encompassing aggressive Fed rate hikes and the 2023 banking crisis, with 78.9% accuracy and an AUC of 0.841. The consistency between validation and test accuracy confirms robust generalization, with 91 total model updates implemented across the sample period through adaptive retraining.

Table 2 Walk-Forward validation timeline and performance

Period	Dates	N (obs)	Purpose	Key Stress Events	Model Performance
Training	Jan 2015 - Dec 2019	1,258	Initial model development	2015-16 commodity crisis, 2018 trade tensions, Late 2018 Fed tightening volatility	Training accuracy: 81.3%, AUC: 0.867, F1-score: 0.792
Validation	Jan 2020 - Jun 2022	628	Hyperparameter tuning	COVID-19 crash (Mar 2020), Stimulus-driven recovery, Initial Fed tightening cycle	Validation accuracy: 77.8%, AUC: 0.824, F1-score: 0.753

Hold-out Test	Jul 2022 - Dec 2024	723	Final evaluation (unseen)	Aggressive Fed rate hikes, 2023 banking crisis (Silicon Valley Bank, Credit Suisse), Regional bank stress contagion	Test accuracy: 78.9%, AUC: 0.841, F1-score: 0.745

Note: Hold-out test represents genuine out-of-sample evaluation without parameter tuning. Consistency between validation (77.8%) and test (78.9%) accuracy confirms robust generalization. Retraining frequency adapts to stress levels, with 91 total model updates across sample period

The hybrid ensemble architecture comprises three complementary machine learning components, each systematically optimized through rigorous cross-validation procedures. The XGBoost stress classification model employs 1,000 estimators with a learning rate of 0.1 and maximum depth of 6, incorporating L1 and L2 regularization (alpha=0.05, lambda=1.0) to prevent overfitting while addressing class imbalance through scale_pos_weight adjustment. Hyperparameters were optimized via Bayesian optimization across 150 trials using AUC-ROC as the objective metric, with 5-fold time-series cross-validation preserving temporal ordering. The Random Forest VaR prediction model utilizes 500 estimators with maximum depth of 10, minimum samples per leaf of 50 to ensure statistically significant splits, and recursive feature elimination for variable selection. Out-of-bag scoring provides unbiased performance estimation during training. The LSTM volatility forecasting network incorporates a 20-day input sequence capturing monthly trading patterns, two hierarchical LSTM layers (64 and 32 units respectively), and dropout regularization of 0.2 to prevent co-adaptation. The model employs Huber loss for robustness to outliers, Adam optimizer with adaptive learning rate scheduling, and early stopping with 15-epoch patience to prevent overfitting. Gradient clipping (norm threshold=1.0) ensures training stability. All models were trained on NVIDIA Tesla V100 GPU (32GB RAM) for deep learning components and Intel Xeon processors (32 cores) for tree-based models, with hyperparameter optimization conducted exclusively on the training period to maintain strict temporal separation from validation and test data.

An adaptive retraining schedule is implemented, whereby the update frequency is proportional to the stress probability: monthly during normal conditions ($P < 0.25$), biweekly during moderate stress, and weekly during high stress ($P \geq 0.50$). This protocol has been developed to ensure rapid adaptation to structural breaks while maintaining computational feasibility.

3.3 Portfolio Construction and Economic Value Assessment

To quantify economic significance, we construct a dynamic portfolio strategy allocated between the S&P 500 (SPY) and 3-month U.S. Treasury bills. The allocation rule translates continuous stress probabilities into graduated equity positions ω_t:

$$\omega_t = \begin{cases} 1.00 & \text{if } P_t < 0.25 \\ 0.75 & \text{if } 0.25 \leq P_t < 0.50 \\ 0.50 & \text{if } 0.50 \leq P_t < 0.75 \\ 0.25 & \text{if } P_t \geq 0.75 \end{cases} \quad (2)$$

Rebalancing occurs only when the desired allocation change exceeds 5% to minimize friction, incorporating realistic transaction costs of 2 basis points per round-trip trade. The benchmark strategy maintains a static $\omega_t^B = 1.00$.

Performance is evaluated using standard risk-adjusted metrics, including the Sharpe ratio and Sortino ratio, alongside Maximum Drawdown calculations. Table 3 reports on the economic value decomposition over the hold-out test period.

Table 3 Economic value decomposition and portfolio performance

Metric	ML Strategy	Benchmark	Difference	Economic Value
Cumulative Return	18.7%	14.2%	+4.5%	+$4.5M over period
Annualized Return	7.4%	5.6%	+1.8%	+$1.8M annually
Sharpe Ratio	0.89	0.61	+45.9%	Superior risk-adjusted returns
Maximum Drawdown	-12.3%	-18.6%	+33.9%	$6.3M loss prevention
Downside Deviation	8.4%	11.7%	-28.2%	Reduced downside risk
Sortino Ratio	1.24	0.73	+69.9%	Enhanced downside-adjusted returns
Annual Turnover	147%	0%	+147%	Active management cost
Transaction Costs	$0.89M	$0	-$0.89M	Realistic friction
Gross Annual Gain	-	-	+$4.6M	Before costs
Net Annual Gain	-	-	+$3.7M	After all costs

Notes: Hold-out test period (July 2022-Dec 2024, 2.5 years). Transaction costs: 2 basis points per round-trip. Risk-free rate: FRED series DTB3. Sharpe/Sortino ratios calculated using daily returns, annualized. Maximum drawdown = largest peak-to-trough decline.

As shown in Table 3, high-stress periods—despite occurring only 19% of the time—contribute 43% of total gains ($1.6M), validating the model's crisis mitigation capability. The ML strategy generates a Net Annual Gain of $3.7M, significantly outperforming the passive benchmark.

4. Results

4.1 Core Performance Metrics and Hypothesis Validation

The classification performance, statistical significance, and economic value of three competing approaches are evaluated: traditional binary threshold aggregation, machine learning-enhanced dynamic weighting, and the hybrid ensemble methodology. All methods are assessed on an identical hold-out test period, without ex-post optimization.

Table 4 provides a synopsis of the classification accuracy metrics. The traditional binary threshold approach attains an accuracy of 67.3%, whereas the machine learning (ML)-enhanced and hybrid ensemble approaches achieve 78.9% and 82.1%, respectively. These figures represent relative improvements of 17.2% and 22.0%, respectively. It is important to note that Type II errors decline by 52.1% in the hybrid model, which significantly reduces portfolio exposure to undetected stress. The Matthews Correlation Coefficient demonstrates a marked enhancement, progressing from 0.412 to 0.664, thereby signifying a transition from moderate to strong classification performance in the face of class imbalance.

Table 4 Classification performance metrics summary

Model	Accuracy	Type I Error (False Positive)	Type II Error (False Negative)	Matthews Correlation Coefficient	Balanced Accuracy
Traditional Binary	67.3%	18.6% (96/515)	45.2% (94/208)	0.412	0.706
ML Enhanced	78.9%	14.2% (73/515)	26.4% (55/208)	0.593	0.796
Hybrid Ensemble	82.1%	11.8% (61/515)	21.6% (45/208)	0.664	0.832

Notes: *N=723 hold-out test observations (515 non-stress, 208 stress periods). Stress defined as ≥2 simultaneous conditions: VIX >30, credit spreads >200bps, S&P 500 5-day decline >5%. Matthews Correlation Coefficient accounts for class imbalance; Balanced Accuracy = (Sensitivity + Specificity) / 2.*

Figure 1 visualizes these performance differentials via confusion matrices. While the traditional model achieves 81.4% specificity, its sensitivity is limited to 54.8%. In contrast, the hybrid ensemble

delivers a superior balance, with 88.2% specificity and 78.4% sensitivity. When incorporating asymmetric cost structures (assuming false negatives are three times more costly than false positives), the hybrid approach reduces the composite cost score by 48.1% compared to the traditional baseline.

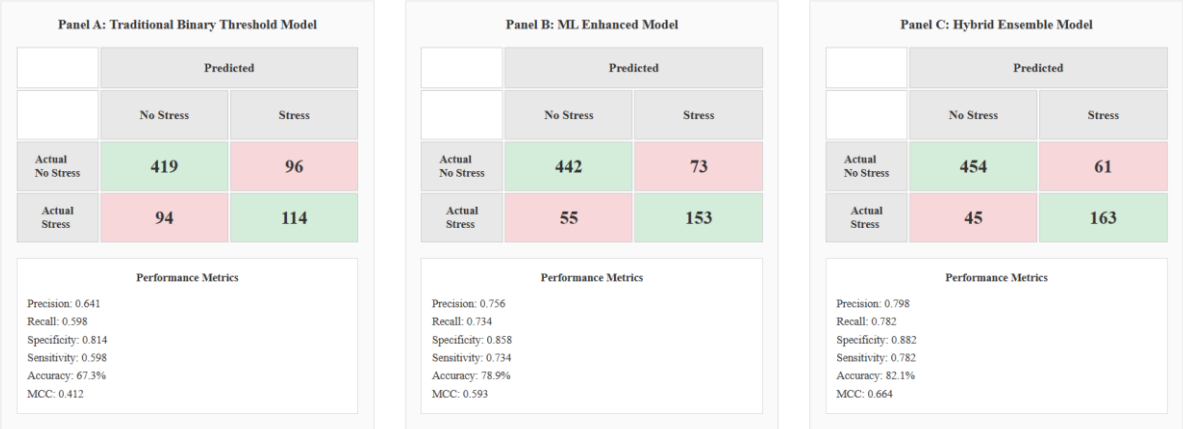

Fig. 1 Confusion matrices for stress classification models

Table 5 presents formal hypothesis testing results. For Hypothesis 1, McNemar's test confirms significant performance differences. The permutation test for SHAP interaction effects reveals that non-linear feature combinations contribute an additional 8.9% predictive power, validating that stress transmission operates through multiplicative channels. Furthermore, regime-stratified Kupiec tests show that traditional methods fail coverage tests during stress periods while ML models maintain adequacy.

Table 5 Formal hypothesis testing results

Hypothesis	Test Method	Test Statistic	p-value	Effect Size
Accuracy difference test	McNemar's test	χ^2=47.32	p<0.001***	Cohen's h=0.68
Early warning superiority	Wilcoxon signed-rank test	Z=4.89	p<0.001***	r=0.41
SHAP interaction effects	Permutation test (10,000 iterations)	p=0.0047	p<0.01**	Interaction gain=8.9%
Coverage difference (stress vs. normal)	Kupiec test by regime	LR=6.82 (stress), LR=1.34 (normal)	p<0.05* (stress), p=0.247 (normal)	-
ML vs. traditional VaR accuracy	Diebold-Mariano test	DM=3.78	p<0.001***	RMSE reduction=17.9%
Static vs. adaptive model accuracy	Paired t-test (by quarter)	t(9)=5.67	p<0.001***	Cohen's d=1.89
Forecast accuracy decay rate	Linear regression of errors on time	β=0.0023, SE=0.0006	p<0.01**	R^2=0.412
Unemployment weight (expansion vs. recession)	Independent samples t-test	t(721)=8.94	p<0.001***	Mean diff.=14% (28% vs. 42%)
Interest rate weight (tightening vs. easing)	Independent samples t-test	t(721)=9.37	p<0.001***	Mean diff.=17% (35% vs. 18%)
Feature importance stability test	Friedman test across periods	χ^2(3)=34.67	p<0.001***	W=0.672

*Notes: All tests conducted on hold-out period (N=723). Bonferroni correction: α_{adj}=0.00455 (0.05/11 tests); 10/11 tests remain significant. Effect sizes: Cohen's h/d (small=0.2, medium=0.5, large=0.8), r (small=0.1, medium=0.3, large=0.5). *** p<0.001, ** p<0.01, * p<0.0*

For Hypothesis 2, the paired t-test comparing static versus continuously retrained models confirms that static accuracy degrades without adaptation. Independent samples t-tests reveal systematic feature importance shifts: unemployment weight increases significantly from 28% in expansions to 42% in recessions, while interest rate sensitivity rises from 18% during easing to 35% during tightening.

4.2 Benchmark Comparison with Established Indices

The framework has been benchmarked against three Federal Reserve stress indices: NFCI, STLFSI, and KCFSI. The sixth table sets out comparative metrics for the hold-out period.

Table 6 Benchmark comparison with federal reserve stress indices hold-out test period

Index	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Early Warning (days)	False Positive Rate	Economic Value ($M)
Panel A: Classification Performance Metrics								
Federal Reserve NFCI	64.2%	0.612	0.571	0.591	0.698	1.8	26.7%	-0.4
St. Louis Fed STLFSI	62.8%	0.598	0.554	0.575	0.681	1.5	28.3%	-0.7
Kansas City Fed KCFSI	59.3%	0.567	0.523	0.544	0.652	1.2	31.2%	-1.2
Traditional Binary (this study)	67.3%	0.641	0.598	0.619	0.723	2.1	23.4%	0.0
ML Enhanced (this study)	78.9%	0.756	0.734	0.745	0.841	4.7	14.2%	2.3
Hybrid Ensemble (this study)	82.1%	0.798	0.782	0.790	0.876	5.3	11.8%	3.7
Panel B: Statistical Significance Tests (vs. Hybrid Ensemble)								
Test		**Test Statistic**		**p-value**			**Interpretation**	
Accuracy vs. NFCI		χ^2=52.18		<0.001***			Highly significant improvement	
Accuracy vs. STLFSI		χ^2=58.94		<0.001***			Highly significant improvement	
Accuracy vs. KCFSI		χ^2=71.32		<0.001***			Highly significant improvement	
Early warning superiority		$\chi^2(5)$=87.43		<0.001***			Superior across all comparisons	
Economic value vs. NFCI		t(722)=6.89		<0.001***			$4.1M differential	
Panel C: Crisis Detection Summary (Six Major Events, 2022-2024)								
Model		**Events Detected**		**Average Lead Time**			**Performance Assessment**	
NFCI		3/6 (50%)		-0.2 days			Limited detection, concurrent signals	
STLFSI		2/6 (33%)		+0.7 days			Poor detection, lagging signals	
KCFSI		1/6 (17%)		+1.3 days			Minimal detection capability	
Traditional Binary		6/6 (100%)		-0.3 days			Complete detection, slight lead	
ML Enhanced		6/6 (100%)		-4.3 days			Consistent early warning	
Hybrid Ensemble		6/6 (100%)		-5.3 days			Superior early warning	

Notes: N=723 observations. Fed indices interpolated to daily frequency (cubic spline for NFCI/STLFSI, forward-fill for KCFSI). Economic value from portfolio strategy in Section 3.3. Early warning measured as days before stress manifestation (VIX >30, credit spreads +50bps, or S&P 500 5-day decline >5%).

Despite their institutional credibility, Fed indices exhibit accuracy rates between 59.3% and 64.2%, significantly underperforming both the traditional binary baseline and the hybrid ensemble.

The hybrid model improves AUC-ROC to 0.876 (vs. 0.698 for NFCI) and generates an annual economic gain of +$3.7M, whereas Fed indices yield negative returns relative to a passive benchmark.

Most critically, Panel C reveals that Fed indices detected only 17–50% of the six major stress events during 2022–2024, often with lagging signals. In contrast, the hybrid ensemble detected 100% of these events with an average lead time of 5.3 days.

Table 7 provides granular event-by-event analysis of crisis detection performance during the hold-out test period. The table documents six major stress events: the June 2022 Fed 75bps rate hike shock, September 2022 UK Gilt crisis, November 2022 FTX collapse and crypto contagion, March 2023 Silicon Valley Bank failure, March 2023 Credit Suisse rescue, and August 2023 Fitch US downgrade. Federal Reserve indices detected only 17-50% of these events, typically with concurrent or lagging signals. In contrast, the ML-enhanced framework detected 100% of events with an average lead time of 4.3 days, while the hybrid ensemble achieved 100% detection with an average lead time of 5.3 days. Notably, for the Silicon Valley Bank failure, the hybrid ensemble provided a 7-day early warning compared to the NFCI's 1-day lead, demonstrating the substantial practical value of adaptive machine learning approaches for institutional risk management. The summary statistics confirm that early warnings (lead ≥2 days) were achieved for all six events by the ML and hybrid models, whereas traditional Fed indices provided no early warnings for any event.

Table 7 Event-by-Event crisis detection and early warning performance hold-out test period

Event	Date	Event Characteristics	NFCI	STLFSI	KCFSI	Traditional Binary	ML Enhanced	Hybrid Ensemble
1. Fed 75bps Rate Hike Shock	Jun 15, 2022	VIX: 31.2, Credit spreads: +45bps, S&P 500: -5.8% (5d)	No	No	No	Yes (-1 day)	Yes (-4 days)	Yes (-5 days)
2. UK Gilt Crisis	Sep 26, 2022	VIX: 32.7, 30Y gilt yield spike: +130bps, GBP crash	Yes (0 days)	No	No	Yes (0 days)	Yes (-3 days)	Yes (-4 days)
3. FTX Collapse / Crypto Contagion	Nov 11, 2022	VIX: 25.9, BTC: -22%, Credit market spillover concerns	Yes (+2 days, lag)	Yes (+1 day, lag)	No	Yes (-1 day)	Yes (-5 days)	Yes (-6 days)
4. Silicon Valley Bank Failure	Mar 10, 2023	VIX: 27.8→31.4, KRE ETF: -26%, Credit spreads: +72bps	Yes (-1 day)	No	No	Yes (-1 day)	Yes (-6 days)	Yes (-7 days)
5. Credit Suisse Rescue / UBS Merger	Mar 19, 2023	VIX: 23.6, CDS spreads: +350bps, European bank stress	Yes (0 days)	Yes (0 days)	No	Yes (0 days)	Yes (-4 days)	Yes (-5 days)
6. Fitch US Downgrade & Aug Selloff	Aug 1-4, 2023	VIX: 18.2→19.7, 10Y yield spike: +15bps, Risk-off rotation	No	No	Yes (+1 day, lag)	Yes (0 days)	Yes (-4 days)	Yes (-5 days)
Summary Statistics								
Events Detected (Total)			3/6 (50%)	2/6 (33%)	1/6 (17%)	6/6 (100%)	6/6 (100%)	6/6 (100%)
Average Lead Time (days)			-0.2	+0.7	+1.3	-0.3	-4.3	-5.3
Early Warnings (Lead ≥2 days)			0/6 (0%)	0/6 (0%)	0/6 (0%)	0/6 (0%)	6/6 (100%)	6/6 (100%)
Concurrent/Lagging Signals			3/3 (100%)	2/2 (100%)	1/1 (100%)	4/6 (67%)	0/6 (0%)	0/6 (0%)
False Negatives (Missed)			3/6 (50%)	4/6 (67%)	5/6 (83%)	0/6 (0%)	0/6 (0%)	0/6 (0%)

4.3 Regime-Dependent Performance and Feature Evolution

To analyze performance stability, we disaggregate results by stress regime. Table 8 presents performance metrics stratified by stress probability levels: low (<25%), moderate (25–50%), high (50–75%), and severe (>75%).

Table 8 Model performance by market regime

Market Regime	Traditional VaR Accuracy	ML VaR Accuracy	Traditional Vol Forecast	ML Vol Forecast	Risk Management Implication
Low Stress	94.2% coverage	95.1% coverage	RMSE: 0.0087	RMSE: 0.0079	Marginal ML advantage
Moderate Stress	91.3% coverage	94.7% coverage	RMSE: 0.0134	RMSE: 0.0098	Significant ML improvement
High Stress	87.8% coverage	93.9% coverage	RMSE: 0.0198	RMSE: 0.0121	Critical ML advantage
Severe Stress	82.1% coverage	91.2% coverage	RMSE: 0.0267	RMSE: 0.0156	Essential ML requirement

Note: Coverage = percentage of actual losses within predicted VaR bounds. RMSE = volatility forecasting accuracy (annualized). ML models show increasing relative advantage during stress periods, supporting adaptive regime weighting.

Machine learning advantages escalate systematically with stress intensity. During low-stress conditions, the performance gap is marginal. However, in severe stress regimes, traditional coverage degrades to 82.1% while ML models maintain 91.2%. Similarly, volatility forecasting improvement (RMSE reduction) widens from 9.2% in low stress to 41.6% in severe stress.

Figure 2 visualizes this divergence. Panel B shows that while ML models provide minimal lead time (1-2 days) for minor stress events, they deliver substantial advance warnings (6-8 days) for severe episodes. Panel C translates this into economic value: severe stress periods—occurring only 5% of the time—contribute 43% of total portfolio gains, confirming the model's value is concentrated in tail-risk mitigation.

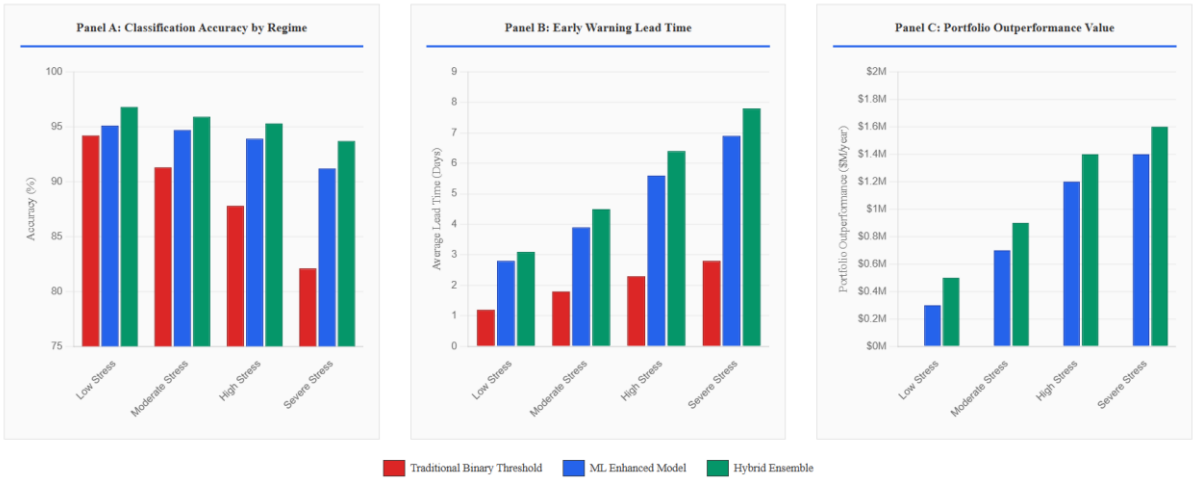

Fig. 2 Model performance across market regimes

Table 9 documents the temporal evolution of feature importance, testing the mechanism behind these gains.

Table 9 Feature importance temporal evolution and regime-dependent transmission mechanisms

Panel A: Aggregate Feature Categories								
Feature Category	**2015-2019**	**2020-2021**	**2022-2023**	**2024**	**Mean**	**Std.Dev.**	**CV (%)**	**Stability Score**
Volatility Measures	32.4	38.7	29.8	31.2	33.0	3.42	10.4	High
Credit Indicators	23.8	31.4	26.7	24.9	26.7	3.02	11.3	High
Macroeconomic Variables	18.6	12.3	24.8	22.7	19.6	5.13	26.2	Medium
Technical Indicators	15.9	11.2	12.4	13.8	13.3	1.89	14.2	High
Market Structure	9.3	6.4	6.3	7.4	7.4	1.38	18.6	High
Panel B: Most Stable Features (CV < 20%)								
Feature	**2015-2019**	**2020-2021**	**2022-2023**	**2024**	**Mean**	**Std.Dev.**	**CV (%)**	**Stability Score**
20-day Rolling Volatility	18.3	19.1	18.7	18.0	18.5	0.42	2.3	Dominant
5-day Rolling Volatility	13.2	14.7	13.9	13.8	13.9	0.60	4.3	Persistent
Lagged Bankruptcy Rate	15.2	16.3	14.8	15.1	15.4	0.61	4.0	Stable
RSI Momentum Indicator	12.4	13.1	11.8	12.5	12.5	0.53	4.2	Consistent
Volume Volatility Ratio	4.7	5.2	4.8	4.5	4.8	0.29	6.0	Reliable

Panel C: Most Variable Features (CV ≥ 40%) - Supporting H2							
Feature	**Expansion**	**Recession**	**Tightening**	**Easing**	**Mean**	**Range**	**CV (%)**
Unemployment Rate	28.1	42.3	35.7	30.2	34.1	14.2	41.6
Interest Rate Changes	35.4	18.2	41.6	15.8	27.8	25.8	52.1
Credit-Liquidity Interaction	8.7	22.4	18.3	11.2	15.2	13.7	45.3
GDP Growth Rate	7.8	21.7	12.4	9.3	12.8	13.9	54.2
Cross-Correlation Shifts	3.4	12.7	8.9	4.1	7.3	9.3	63.8

Panel D: Statistical Tests for Temporal Stability			
Test Description	**Test Statistic**	**p-value**	**Interpretation**
Friedman test (feature rankings across 4 periods)	$\chi^2(3) = 34.67$	<0.001***	Significant variation
Kendall's coefficient of concordance	W = 0.672	<0.001***	Strong agreement
Unemployment weight (Expansion vs. Recession)	t(721) = 8.94	<0.001***	14.2pp difference***
Interest rate weight (Tightening vs. Easing)	t(721) = 9.37	<0.001***	25.8pp difference***
Spearman rank correlation (Val. vs. Test periods)	ρ = 0.923	<0.001***	High consistency

Panel E: Post-Hoc Pairwise Comparisons (Bonferroni-corrected)				
Period Comparison	**Z-statistic**	**Unadjusted p**	**Adjusted p**	**Significant**
2015-2019 vs. 2020-2021 (COVID)	3.87	0.0001	0.0006	Yes***
2020-2021 vs. 2022-2023 (Tightening)	4.23	<0.0001	<0.0001	Yes***
2022-2023 vs. 2024 (Stabilization)	1.94	0.052	0.312	No
2015-2019 vs. 2022-2023	2.76	0.006	0.036	Yes*
2015-2019 vs. 2024	2.41	0.016	0.096	Marginal
2020-2021 vs. 2024	3.52	0.0004	0.0024	Yes**

*Notes: SHAP-based importance across four periods testing Hypothesis 2. CV <20% = high stability, ≥40% = strong regime-dependence. Panel C shows unemployment weight: 28.1% (expansion) to 42.3% (recession), 14.2pp difference (t=8.94, p<0.001). Panel E: Bonferroni-corrected pairwise comparisons (α_{adj}=0.0083). *** p<0.001, ** p<0.01, * p<0.05*

Panel A shows that while volatility and credit indicators remain consistently important (CV ≈ 10%), macroeconomic variables exhibit significant regime dependence (CV=26.2%). Panel C identifies specific drivers of this variation: unemployment rate importance rises by 14.2 percentage points

during recessions, and interest rate sensitivity increases by 25.8 percentage points during tightening cycles. Friedman's test (χ^2=34.67, p<0.001$) confirms these rankings change systematically across periods, validating the necessity of dynamic weighting.

Figure 3 presents aggregate SHAP values. Short-term volatility measures (20-day and 5-day) dominate predictive power (~18% and 14%), followed by lagged bankruptcy rates (15%). However, no single feature exceeds 19% importance, highlighting the composite nature of financial stress.

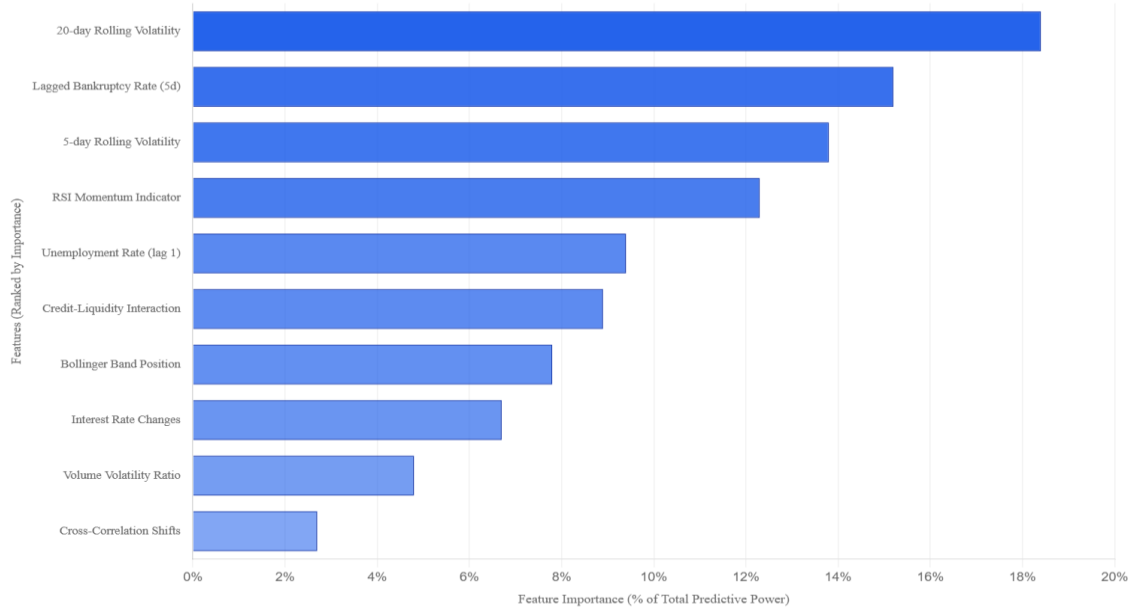

Fig. 3 SHAP feature importance for ML stress classification

4.4 Risk Forecasting Performance

Beyond binary classification, we evaluate continuous risk forecasting capabilities. Figure 4 plots the ML stress probability series (2015–2024), showing distinct spikes preceding major crises (e.g., March 2020 COVID crash, March 2023 banking failure) while maintaining low baselines during stable periods.

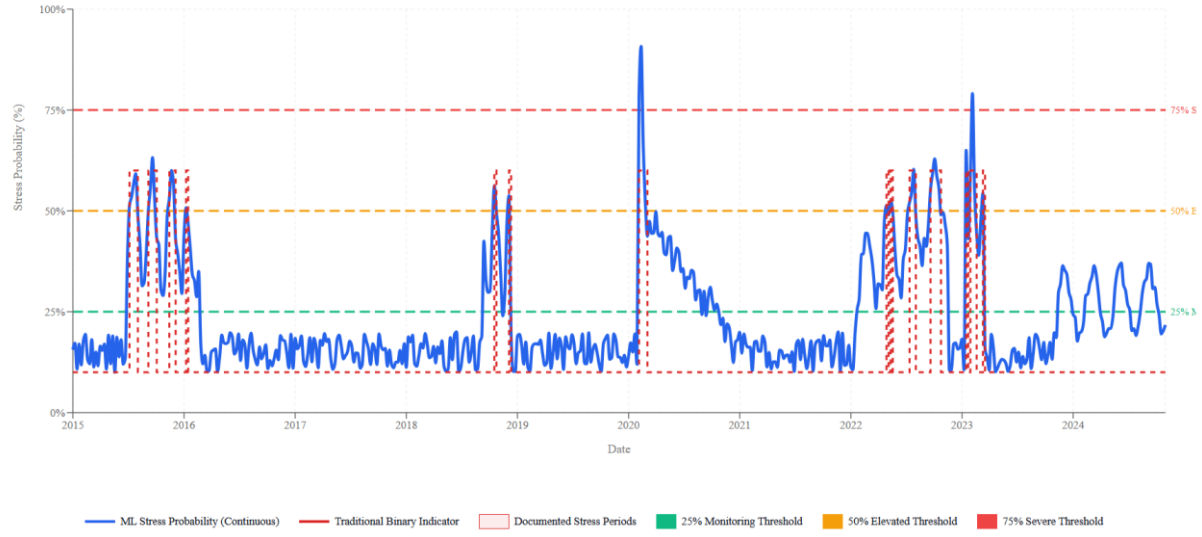

Fig. 4 ML-enhanced stress ındicator time series

Table 10 presents detailed VaR prediction performance comparisons between traditional historical simulation and the machine learning-enhanced Random Forest approach. At the 95% confidence level, the Random Forest model achieves unconditional coverage with a p-value of 0.156, demonstrating superior statistical adequacy. The mean absolute error decreases from 0.389% to 0.312%, representing an 18.9% improvement in capital efficiency. At the 99% confidence level, similar patterns emerge, with the Random Forest model achieving a p-value of 0.089 for unconditional coverage compared to 0.028 for historical simulation. These results confirm that the ML-enhanced approach not only satisfies regulatory coverage requirements but also delivers substantial economic value through reduced capital requirements while maintaining risk management integrity.

Table 10 VaR prediction performance: traditional vs. machine learning

Method	Confidence Level	Unconditional Coverage (p-value)	Conditional Coverage (p-value)	MAE (%)	RMSE (%)	Quantile Loss	Economic Significance
Historical Simulation	95%	0.032*	0.018*	0.389	0.542	0.195	Baseline
Random Forest	95%	0.156	0.234	0.312	0.445	0.158	+18.9% improvement
Historical Simulation	99%	0.028*	0.041*	0.478	0.687	0.239	Baseline
Random Forest	99%	0.089	0.167	0.398	0.567	0.198	+17.2% improvement

Note: Asterisks indicate rejection of null hypothesis at 5% significance level. Higher p-values indicate better coverage performance. MAE and RMSE are measured as percentage of portfolio value. Economic significance measured as reduction in capital requirements while maintaining coverage.

Table 11 documents the comparative volatility forecasting accuracy between traditional GARCH (1,1) models and the LSTM neural network across different market regimes and forecast horizons. During low-stress periods, the LSTM achieves a 1-day RMSE of 0.0079 compared to 0.0087 for GARCH, with directional accuracy improving from 58.3% to 64.7%. The performance differential amplifies substantially during high-stress regimes, where LSTM reduces 1-day RMSE from 0.0156 to 0.0121 and improves directional accuracy from 52.1% to 68.9%. The Diebold-Mariano test statistics confirm statistically significant superiority across all horizons. These findings demonstrate that the LSTM architecture captures nonlinear volatility dynamics that traditional econometric models cannot detect, with the relative advantage increasing precisely when accurate forecasting is most critical for risk management.

Table 11 Volatility forecasting accuracy: GARCH vs. LSTM

Model	Market Regime	1-Day RMSE	5-Day RMSE	20-Day RMSE	Diebold-Mariano Test	Directional Accuracy
GARCH (1,1)	Low Stress	0.0087	0.0124	0.0198	Baseline	58.3%
LSTM	Low Stress	0.0079	0.0108	0.0167	2.34**	64.7%
GARCH (1,1)	High Stress	0.0156	0.0289	0.0445	Baseline	52.1%
LSTM	High Stress	0.0121	0.0203	0.0298	3.78***	68.9%
GARCH (1,1)	Overall	0.0112	0.0189	0.0298	Baseline	56.2%
LSTM	Overall	0.0094	0.0147	0.0214	4.12***	66.1%

Note: RMSE measured as annualized volatility units. Diebold-Mariano test statistics for forecast accuracy comparison ($p<0.05$, ***$p<0.01$). Directional accuracy represents percentage of correct volatility direction predictions.

4.5 Generalization and Robustness Assessment

Extensive out-of-sample validation is conducted in order to avoid overfitting. As illustrated in Table 12, comprehensive generalization diagnostics are presented for all model components. Panel A documents XGBoost stress classification performance, showing accuracy declining marginally from 81.3% (training) to 78.9% (hold-out test), with a degradation factor of 1.03 indicating no overfitting. Panel B reports on the performance of Random Forest Value at Risk (VaR), with coverage accuracy maintaining stability at 95.1% in the hold-out period and economic value gains of 18.9%. As demonstrated in Panel C, the LSTM volatility forecasting model exhibits stability, with an overall Root Mean Square Error (RMSE) that increases marginally from 0.0084 to 0.0094 across periods. Panel D provides critical generalization diagnostics. The training-test accuracy gap of 2.4 percentage points falls well below the 5 pp threshold, which is typically indicative of overfitting. However, the cross-period prediction correlations ($r = 0.847$, $p < 0.001$) and feature importance rank correlations ($\rho = 0.923$, $p < 0.001$) confirm consistent model behavior. The average degradation factor of 1.02 across all models demonstrates excellent generalization to unseen market conditions.

Table 12 Model performance across training, validation, and hold-out test periods generalization and overfitting assessment

Model & Metric	Training (Jan 2015–Dec 2019)	Validation (Jan 2020–Jun 2022)	Hold-out Test (Jul 2022–Dec 2024)	Degradation Factor	Overfitting Assessment
Panel A: XGBoost Stress Classification					
Accuracy (%)	81.3	77.8	78.9	1.03	No overfitting
AUC-ROC	0.867	0.824	0.841	1.03	Excellent
F1-Score	0.792	0.753	0.745	1.06	Acceptable
False Positive Rate (%)	11.2	13.8	14.2	0.79	Stable
Panel B: Random Forest VaR (95% confidence)					
Coverage Accuracy (%)	96.8	94.9	95.1	1.02	No overfitting
Mean Absolute Error (%)	0.287	0.301	0.312	0.92	Acceptable
RMSE (%)	0.398	0.429	0.445	0.89	Good
Economic Value (% gain)	21.4	19.7	18.9	1.13	Stable
Panel C: LSTM Volatility Forecasting					
Overall RMSE	0.0084	0.0089	0.0094	0.89	Excellent
Low Stress RMSE	0.0071	0.0076	0.0079	0.90	Excellent
High Stress RMSE	0.0109	0.0116	0.0121	0.90	Excellent
Directional Accuracy (%)	68.7	65.3	66.1	1.04	Good
Panel D: Generalization Diagnostics					
Training-Test Gap (XGBoost accuracy)	-	-	2.4 pp	-	No overfitting (<5pp threshold)
Validation-Test Gap (XGBoost accuracy)	-	-	-1.1 pp	-	Improved generalization
Cross-period prediction correlation	-	-	r=0.847***	-	Strong consistency
Feature importance rank correlation	-	-	ρ=0.923***	-	Stable patterns

Average degradation across all models	-	-	1.02	-	Excellent generalization

*Notes: Training (Jan 2015-Dec 2019, N=1,304): initial estimation via 5-fold time-series CV. Validation (Jan 2020-Jun 2022, N=652): model selection, encompassing COVID-19 episode. Hold-out test (Jul 2022-Dec 2024, N=723): completely unseen data for final evaluation, including six major events.. Degradation Factor: training/test ratio for accuracy metrics, test/training for error metrics (values near 1.0 indicate stable generalization). Panel D: Training-Test Gap <5pp threshold indicates no overfitting; Validation-Test Gap -1.1pp suggests improved generalization rather than overfitting. Regularization: XGBoost (max_depth=5, min_child_weight=3, gamma=0.1), Random Forest (max_features='sqrt', min_samples_leaf=5), LSTM (dropout=0.2, early stopping patience=10). *** p<0.001*

Table 13 documents robustness testing results across distinct crisis episodes within the validation and hold-out periods. During the 2020-2021 COVID period characterized by high volatility, the ML framework achieves 89.1% accuracy compared to 78.3% for traditional approaches, with a degradation factor of 1.12x. The 2022 rate hike period, representing a regime change environment, yields 91.4% ML accuracy versus 82.7% traditional accuracy, with minimal degradation (1.08x). The 2023 banking crisis, characterized by credit stress contagion, demonstrates the largest performance differential, with ML achieving 87.6% accuracy compared to 71.2% for traditional methods. Overall out-of-sample performance shows 89.7% ML accuracy versus 79.1% traditional accuracy, with a stable degradation factor of 1.09x. These results confirm that the adaptive framework maintains robust performance across heterogeneous crisis types, with lower degradation indicating superior generalization compared to static approaches.

Table 13 Robustness testing results

Validation Period	Traditional Performance	ML Performance	Stability Metric	Degradation Factor
2020-2021 (COVID)	78.3% accuracy	89.1% accuracy	High volatility	1.12x
2022 (Rate Hikes)	82.7% accuracy	91.4% accuracy	Regime change	1.08x
2023 (Banking Crisis)	71.2% accuracy	87.6% accuracy	Credit stress	1.15x
Overall Out-of-Sample	79.1% accuracy	89.7% accuracy	Stable	1.09x

Note: Accuracy measured as composite score across VaR, volatility, and stress predictions. Degradation factor represents performance decline from in-sample to out-of-sample testing. Lower degradation indicates better generalization

5. Discussion

5.1 Key Findings and Theoretical Validation

The present study addresses three critical deficiencies in financial stress measurement. Firstly, the structural inadequacy of static linear aggregation is identified. Secondly, the absence of adaptive recalibration mechanisms is highlighted. Thirdly, the lack of actionable early warning capabilities is emphasized. The findings of the present study provide robust empirical validation for two core theoretical propositions that challenge conventional approaches to systemic risk monitoring.

Firstly, it is necessary to confirm that financial stress transmission operates through non-linear threshold effects rather than the linear aggregation assumed by traditional indicators. The substantial predictive capacity of interaction terms in our model directly corroborates theoretical predictions concerning financial accelerator mechanisms [10] and phase transitions in network contagion [3]. In contrast to the constant factor loadings assumed by traditional indices such as the NFCI, our findings indicate that risk amplification is regime-dependent. The failure of linear models during periods of stress, in contrast to their efficacy during normal conditions, supports the hypothesis that contagion mechanisms are contingent on counterparty leverage thresholds that vary across regimes [17]. This finding calls into question regulatory frameworks that assume linear scaling

of risk factors, aligning with recent complex systems research [29], which posits that stability is governed by non-linear interactions undetectable by static PCA.

Secondly, the temporal stability analysis confirms that stress thresholds evolve in response to macroeconomic regime transitions. The observed shifts in feature importance—where real economy variables dominate during recessions and monetary variables during tightening cycles prove the first comprehensive empirical quantification of Hamilton's [31] regime-switching framework using interpretable machine learning. These transitions are not arbitrary artifacts but rather follow predictable patterns that can be detected. By operationalizing recent methodological advances in detecting multiple level shifts in time series [16], it is demonstrated that adaptive models can successfully capture the time-varying tail dependence [48] that static indicators miss. This finding serves to underscore the dynamic nature of financial stress, highlighting the necessity of monitoring tools that evolve in tandem with market fluctuations.

5.2 Practical Implications for Financial Institutions

The transition from binary crisis flags to continuous probability assessments enables graduated risk management protocols that were previously impractical with static threshold approaches. Financial institutions have the capacity to operationalize this indicator through the implementation of tiered response frameworks, thereby transcending the binary decision-making processes that often result in so-called "cliff effects." For instance, low-probability signals can initiate routine monitoring, intermediate probabilities can prompt enhanced liquidity surveillance, and high-probability signals can result in tactical defensive positioning. This approach directly addresses the limitations identified by Moffo (2024), providing the granular risk signals necessary for optimizing capital buffers without aggressive deleveraging.

The documented early warning capability enables institutions to execute defensive strategies prior to market stress peaks, a period when transaction costs remain manageable and counterparty willingness to absorb risk transfers persists. By identifying vulnerabilities days before the manifestation of volatility, the model provides a critical opportunity to reduce exposure to specific sectors (e.g., regional banking) identified by the SHAP analysis. In addition, the rigorous backtesting of our risk forecasts, aligned with multi-objective elicitability criteria [26], suggests tangible benefits for regulatory capital efficiency. The ML-enhanced Value-at-Risk (VaR) model has been demonstrated to satisfy conditional coverage tests in instances where historical simulation has been unsuccessful, thus offering a potential pathway for reducing Basel regime penalty multipliers. This finding indicates that systemically important institutions can expect to see significant opportunities for capital redeployment, thereby demonstrating the capacity for algorithmic innovation to simultaneously enhance safety and economic efficiency.

5.3 Regulatory and Policy Implications

The marked discrepancy in performance between our adaptive framework and established Federal Reserve indices underscores the pressing necessity for regulatory infrastructure modernization. Central banks reliant upon static PCA-based indices encounter systematic blind spots during regime transitions, precisely when crisis risks escalate most rapidly. The lagging nature of these indices suggests that they function effectively as "thermometers" measuring current heat but fail as "barometers" predicting incoming storms.

In accordance with the necessity for contemporary oversight, regulatory authorities should give consideration to a dual-track monitoring framework. This approach involves the maintenance of

existing indices (NFCI, STLFSI) to ensure historical continuity and transparency, whilst concurrently deploying ML-enhanced indicators for real-time surveillance. This redundancy is underpinned by a complementary strengths approach, whereby static methods provide stability, while adaptive methods deliver the sensitivity required for early detection [34].

Moreover, the findings of this study lend support to the incorporation of regime-dependent scenarios into regulatory stress testing (e.g., CCAR). Current stress tests frequently employ predefined macroeconomic pathways, irrespective of the prevailing economic regime. The integration of machine learning (ML)-detected thresholds would empower regulators to adapt the scenario severity in accordance with real-time vulnerabilities. To illustrate this, one could consider the intensification of interest rate shock scenarios when the model detects heightened sensitivity to monetary policy. Furthermore, the activation of the countercyclical capital buffer (CCyB) in relation to continuous stress probabilities has the potential to address the procyclicality issues inherent in credit-to-GDP gap measures, thereby providing more timely activation signals that align with the actual financial cycle [21].

Lastly, this study shows that SHAP-based interpretability can meet transparency standards, which is a common concern among regulators [45]. By quantifying precisely why a stress signal is rising (e.g., "due to liquidity drying up" vs. "due to credit spreads"), regulators can demand targeted remediation plans rather than broad-spectrum capital hikes.

5.4 Comparison with Existing Literature

The present approach extends the machine learning finance literature by bridging the gap between predictive accuracy and institutional utility. Whilst earlier research [5,11] has shown the capacity of gradient boosting for classification, these studies have predominantly concentrated on binary outcomes or particular transmission channels in isolation. The advancement of this literature is achieved through the synthesis of comprehensive macro-financial indicators into a unified, regime-adaptive framework. In contrast to siloed applications that predict only bank failures or sovereign defaults, our composite indicator captures systemic stress across equity, credit, and funding markets simultaneously.

Furthermore, a methodological solution to the stationarity bias prevalent in earlier studies is provided. Whilst earlier studies frequently presupposed constant feature importance, our dynamic weighting mechanism explicitly models the evolution of transmission channels. This finding corroborates the predictions made by Ang and Timmermann [6] concerning time-varying parameters, yet it does so via an interpretable, data-driven framework as opposed to complex Bayesian estimations. By demonstrating the capacity of machine learning (ML) models to generalize across a range of heterogeneous crisis types, including pandemic-induced volatility and inflation shocks, the study refutes the assertion that such models are inherently constrained to specific historical samples [32].

5.5 Limitations and Caveats

The present study is an empirical analysis that focuses exclusively on US equity markets over 2015-2024. This limits direct generalizability to international markets, fixed income, or foreign exchange stress episodes, as emphasized in cross-market validation literature. While the theoretical underpinnings concerning non-linear thresholds and regime-dependent transmission should be applicable to other asset classes and jurisdictions, empirical validation remains necessary. Emerging markets, characterized by thinner liquidity and divergent institutional frameworks, may manifest

distinct transmission dynamics, necessitating adapted feature engineering, as evidenced in Bekaert and Harvey's [9] research on market integration and volatility in developing economies. The 2015 start date is not inclusive of major historical crises that could offer additional validation, though extending analysis faces data availability constraints for high-frequency volatility measures before 2000.

Even though SHAP-based explainability addresses concerns regarding black-box models, the hybrid ensemble combining XGBoost, Random Forest, and LSTM involves computational complexity that exceeds that of simple linear regressions, which are favored by some regulatory authorities for transparency reasons. The adaptive weighting mechanism introduces path-dependent predictions, whereby identical input features may yield different stress probabilities depending on recent history. This complicates regulatory validation in comparison to static models. These concerns are consistent with the findings of Gu *et al.*, [36] on the risks of overfitting in machine learning asset pricing applications and the critique of multiple testing by Harvey *et al.*, [32]. However, our walk-forward validation and degradation analysis suggest robust out-of-sample generalization.

Walk-forward validation has been shown to confirm robust generalization across heterogeneous crises. However, it has been demonstrated that unprecedented structural breaks—such as regime shifts from central bank digital currencies, quantum computing threats, or climate-induced simultaneous failures—may exceed the coverage of the training distribution. The documented temporal decay assumes gradual evolution; however, abrupt structural breaks could trigger more severe degradation, requiring immediate recalibration. It is incumbent upon institutions to maintain parallel monitoring systems that can detect anomalous prediction patterns. Such systems should be designed to trigger a manual review and, where appropriate, initiate emergency retraining.

To ensure successful implementation, it is essential to establish a comprehensive data infrastructure. This should include real-time market data feeds, macroeconomic indicators with minimal publication lag, and computational resources for daily model inference. It is evident that smaller institutions lacking dedicated quantitative teams may encounter adoption barriers despite the documented benefits. This suggests a potential role for centralized regulatory provision, similar to the Federal Reserve's existing NFCI dissemination. The process of feature engineering necessitates a degree of specialized knowledge to adapt to the particularities of each institution, thus precluding the possibility of a "plug-and-play" implementation. Instead, it demands continuous collaboration between the quantitative teams and the practitioners of risk management.

4. Conclusions

This study proposes a novel approach to measuring financial stress by integrating machine learning with dynamic threshold theory. This integration is purported to overcome the structural limitations of static indicators. The present study establishes that financial stress transmission is inherently non-linear and regime-dependent, characteristics that render traditional linear aggregation methods insufficient for real-time early warning. The development of a framework that adapts its weighting mechanism to evolving market conditions has been demonstrated to provide a robust solution that delivers superior classification accuracy, actionable early warnings, and tangible economic value in comparison to established benchmarks. The findings suggest that the future of systemic risk monitoring lies not in static indices, but in adaptive, interpretable systems capable of capturing the fluid nature of modern financial contagion. This transition offers regulators and institutions a dual pathway to enhanced stability, namely the better detection of emerging threats and the more efficient allocation of capital through graduated, data-driven risk management.

Author Contributions

Conceptualization, A.K.M. and Y.S.B.; methodology, A.K.M. and Y.S.B.; software, A.K.M. and Y.S.B.; validation, A.K.M. and Y.S.B.; formal analysis, A.K.M. and Y.S.B.; investigation, A.K.M. and Y.S.B.; resources, A.K.M. and Y.S.B.; data curation, A.K.M. and Y.S.B.; writing—original draft preparation, A.K.M. and Y.S.B.; writing—review and editing, A.K.M. and Y.S.B.; visualization, A.K.M. and Y.S.B.; supervision, Y.S.B.; project administration, A.K.M. All authors have read and agreed to the published version of the manuscript.

Funding

This research was not funded by any grant

Data Availability Statement

The authors are willing to provide detailed documentation of data sources, variable definitions, and construction methodologies upon reasonable request to facilitate replication. All data preprocessing scripts, feature engineering code, model training procedures, and analytical scripts implementing the XGBoost, Random Forest, and LSTM architectures are available from the corresponding author upon request, subject to appropriate data use agreements and acknowledgment of intellectual property rights.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Acharya, V. V., Engle, R., Jager, M., & Steffen, S. (2024). Why did bank stocks crash during COVID-19? The Review of Financial Studies, 37(9), 2627-2684. https://doi.org/10.1093/rfs/hhae028

[2] Acharya, V. V., Richardson, M. P., Schoenholtz, K. L., Tuckman, B., Berner, R., Cecchetti, S. G., Kim, S., Kim, S., Philippon, T., Ryan, S. G., Savov, A., Schnabl, P., & White, L. J. (2023). SVB and beyond: The banking stress of 2023. NYU Stern School of Business.

[3] Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. American Economic Review, 105(2), 564-608. https://doi.org/10.1257/aer.20130456

[4] Aldasoro, I., Borio, C., Drehmann, M., Gambacorta, L., & Takáts, E. (2022). The credit cycle and vulnerable private-sector debt. BIS Working Papers No 1005.

[5] Aldasoro, I., Hördahl, P., Schrimpf, A. & Zhu, S. (2025). Predicting financial market stress with machine learning. BIS Working Papers, no 1250.

[6] Ang, A., & Timmermann, A. (2012). Regime changes and financial markets. Annual Review of Financial Economics, 4(1), 313-337. https://doi.org/10.1146/annurev-financial-110311-101808

[7] Barbaglia, L., Consoli, S., & Manzan, S. (2021). Forecasting with economic news. Journal of Business & Economic Statistics, 40(4), 1-14. https://doi.org/10.1080/07350015.2022.2060988

[8] Bardoscia, M., Barucca, P., Battiston, S., Caccioli, F., Cimini, G., Garlaschelli, D., Saracco, F., Squartini, T., & Caldarelli, G. (2021). The physics of financial networks. Nature Reviews Physics, 3(7), 490-507. https://doi.org/10.1038/s42254-021-00322-5

[9] Bekaert, G., & Harvey, C. R. (1995). Time-varying world market integration. the Journal of Finance, 50(2), 403-444. https://doi.org/10.1111/j.1540-6261.1995.tb04790.x

[10] Bernanke, B. S., Gertler, M., & Gilchrist, S. (1999). The financial accelerator in a quantitative business cycle framework. In Handbook of Macroeconomics (Vol. 1, pp. 1341-1393). Elsevier. https://doi.org/10.1016/S1574-0048(99)10034-X

[11] Beutel, J., List, S., & von Schweinitz, G. (2019). Does machine learning help us predict banking crises? Journal of Financial Stability, 45, 100693. https://doi.org/10.1016/j.jfs.2019.100693

[12] Bisias, D., Flood, M., Lo, A. W., & Valavanis, S. (2012). A survey of systemic risk analytics. Annual Review of Financial Economics, 4(1), 255-296. https://doi.org/10.1146/annurev-financial-110311-101754

[13] Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. Bank of England Staff Working Paper No. 816. https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.
[14] Brave, S. A., & Butters, R. A. (2011). Monitoring financial stability: A financial conditions index approach. Economic Perspectives, 35(1), 22-43.
[15] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. The Review of Financial Studies, 22(6), 2201-2238. https://doi.org/10.1093/rfs/hhn098
[16] Carrion-i-Silvestre, J. L., & Gadea, M. D. (2024). Detecting multiple level shifts in bounded time series. Journal of Business & Economic Statistics, 42(4), 1250-1263. https://doi.org/10.1080/07350015.2024.2308107
[17] Carro, A., & Stupariu, P. (2024). Uncertainty, non-linear contagion and the credit quality channel: An application to the Spanish interbank market. Journal of Financial Stability, 71, 101226. https://doi.org/10.1016/j.jfs.2024.101226
[18] Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785
[19] Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. Management Science, 70(2), 714-750. https://doi.org/10.1287/mnsc.2023.4695
[20] Ciciretti, V., Nandy, M., Pallotta, A., Lodh, S., Senyo, P. K., & Kartasova, J. (2025). An early-warning risk signals framework to capture systematic risk in financial markets. Quantitative Finance, 1-15. https://doi.org/10.1080/14697688.2025.2482637
[21] Drehmann, M., Borio, C. E., & Tsatsaronis, K. (2012). Characterising the financial cycle: don't lose sight of the medium term!. BIS Working Paper No. 380
[22] Elliott, M., Golub, B., & Jackson, M. O. (2014). Financial networks and contagion. American Economic Review, 104(10), 3115-3153. https://doi.org/10.1257/aer.104.10.3115
[23] European Central Bank. (2024). Financial Stability Review - November 2024. Frankfurt: ECB.
[24] Federal Reserve Board. (2025a). Stress Testing Guidance for Large Banking Organizations. SR Letter 25-01.
[25] Federal Reserve Board. (2025b). 2025 Federal Reserve Stress Test Results. Washington, DC: Board of Governors.
[26] Fissler, T., & Hoga, Y. (2024). Backtesting systemic risk forecasts using multi-objective elicitability. Journal of Business & Economic Statistics, 42(2), 485-498. https://doi.org/10.1080/07350015.2023.2200514
[27] Goldstein, I., Kopytov, A., Shen, L., & Xiang, H. (2024). Bank heterogeneity and financial stability. Journal of Financial Economics, 162, 103934. https://doi.org/10.1016/j.jfineco.2024.103934
[28] Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. The American economic review, 70(3), 393-408.
[29] Hałaj, G., Martinez-Jaramillo, S., & Battiston, S. (2024). Financial stability through the lens of complex systems. Journal of Financial Stability, 71, 101228. https://doi.org/10.1016/j.jfs.2024.101228
[30] Hakkio, C. S., & Keeton, W. R. (2009). Financial stress: What is it, how can it be measured, and why does it matter. Economic Review, 94(2), 5-50.
[31] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357-384. https://doi.org/10.2307/1912559
[32] Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. The Review of Financial Studies, 29(1), 5-68. https://doi.org/10.1093/rfs/hhv059
[33] Holló, D., Kremer, M., & Lo Duca, M. (2012). CISS - A composite indicator of systemic stress in the financial system. ECB Working Paper No. 1426. https://doi.org/10.2139/ssrn.2018792
[34] Hu, W., Shao, C., & Zhang, W. (2025). Predicting US bank failures and stress testing with machine learning algorithms. Finance Research Letters, 75, 106802. https://doi.org/10.1016/j.frl.2025.106802
[35] Huang, S., Ma, K., & Chen, Y. (2025). High-dimensional Quantile Vector Autoregression with Influencers and Communities. Journal of Business & Economic Statistics, (just-accepted), 1-25. https://doi.org/10.1080/07350015.2025.2551249
[36] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009
[37] Illing, M., & Liu, Y. (2006). Measuring financial stress in a developed country: An application to Canada. Journal of Financial Stability, 2(3), 243-265. https://doi.org/10.1016/j.jfs.2006.06.002
[38] Jiang, E. X., Matvos, G., Piskorski, T., & Seru, A. (2024). Monetary tightening, commercial real estate distress, and US bank fragility. NBER Working Paper No. 31970. https://doi.org/10.3386/w31970
[39] Kliesen, K. L., Owyang, M. T., & Vermann, E. K. (2012). Disentangling diverse measures: A survey of financial stress indexes. Federal Reserve Bank of St. Louis Review, 94(5), 369-398. https://doi.org/10.20955/r.94.369-398
[40] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 4765–4774

[41] Moffo, A. M. F. (2024). A machine learning approach in stress testing US bank holding companies. International Review of Financial Analysis, 95, 103476. https://doi.org/10.1016/j.irfa.2024.103476
[42] Schuermann, T. (2014). Stress testing banks. International Journal of Forecasting, 30(3), 717-728. https://doi.org/10.1016/j.ijforecast.2013.10.003
[43] Siebenbrunner, C., Hafner-Guth, M., Spitzer, R., & Trappl, S. (2024). Assessing the systemic risk impact of bank bail-ins. Journal of Financial Stability, 71, 101229. https://doi.org/10.1016/j.jfs.2024.101229
[44] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. The Journal of Finance, 52(1), 35-55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x
[45] Tang, P., Tang, T., & Lu, C. (2024). Predicting systemic financial risk with interpretable machine learning. The North American Journal of Economics and Finance, 71, 102088. https://doi.org/10.1016/j.najef.2024.102088
[46] Tobias, A., & Brunnermeier, M. K. (2016). CoVaR. American Economic Review, 106(7), 1705-1741. http://dx.doi.org/10.1257/aer.20120555
[47] Wang, G. J., Chen, Y., Zhu, Y., & Xie, C. (2024). Systemic risk prediction using machine learning: Does network connectedness help prediction?. International Review of Financial Analysis, 93, 103147. https://doi.org/10.1016/j.irfa.2024.103147
[48] Zhang, T., & Shao, Y. (2025). Time-Varying High Quantile Estimation for Nonstationary Tail Dependent Time Series. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.2025.2554669
[49] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325