Cross-Market Liquidity Stress Forecasting Using Explainable Temporal Fusion Networks Under Leakage-Safe Evaluation Protocols

Hugo C. White
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
white1980@oregonstate.edu

Giorgio Johnston
Department of Computer Science, University of Houston, Houston, TX, USA.
johnstongiorgio@uh.edu

Brendan L. Cohen
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
brendan.cohen@ku.edu

Abstract

Liquidity stress in financial markets propagates rapidly across asset classes and geographies, posing systemic risks that challenge both private risk management and regulatory oversight. While machine learning models offer considerable promise for early warning systems, their deployment in cross-market settings is hampered by two interrelated problems: the opacity of many deep learning architectures, which undermines trust and regulatory acceptance, and the prevalence of data leakage in evaluation protocols, which produces overly optimistic performance estimates that fail to replicate in real-time deployment. This paper develops an explainable temporal fusion network (TFT) framework specifically designed for cross-market liquidity stress forecasting. The TFT architecture combines multi-horizon attention mechanisms with built-in interpretability components, including variable selection networks and temporal self-attention, enabling analysts to identify the primary drivers of predicted stress events. To address the second challenge, we propose a leakage-safe evaluation protocol that enforces strict temporal consistency through purging, embargoing, and combinatorial purged cross-validation. The framework is tested on a multi-asset dataset spanning equity, bond, and foreign exchange markets over a fifteen-year period. Results demonstrate that the TFT model achieves statistically significant improvements in stress prediction accuracy compared to baseline approaches, while the leakage-safe evaluation reduces false positive rates by over forty percent relative to naive walk-forward methods. We further examine structural trade-offs involving model complexity, interpretability, and computational sustainability, and discuss governance implications for deploying such systems within regulatory stress-testing frameworks. The findings underscore the necessity of embedding both explanatory mechanisms and rigorous evaluation discipline into the design of financial early warning infrastructures.

Keywords

liquidity stress, temporal fusion networks, explainability, data leakage, cross-market, financial stability, evaluation protocols.

1. Introduction

Liquidity stress, defined as a sudden and severe deterioration in the ability to trade assets without affecting their prices, represents one of the most persistent vulnerabilities in modern financial systems. The interconnections among equity, fixed-income, and foreign exchange markets create channels through which localized liquidity shocks can amplify into systemic crises, as observed during the 2007–2008 global financial crisis and the COVID-19 induced turmoil of 2020 [3][8]. Quantitative early warning systems have therefore attracted sustained attention from both academic researchers and financial regulators, who seek models capable of anticipating cross-market liquidity dislocations with sufficient lead time to enable preemptive intervention. The adoption of deep learning techniques in this domain has been driven by their capacity to capture nonlinear dependencies and long-range temporal patterns that elude traditional econometric approaches [7][17]. However, the successful deployment of such models in practice hinges on two critical dimensions that are often underappreciated in the literature: the interpretability of the model's internal reasoning and the credibility of the evaluation framework used to assess its out-of-sample performance.

Deep learning architectures, particularly recurrent and transformer-based models, are frequently criticized for operating as black boxes, making it difficult for risk managers and regulators to understand why a specific stress prediction was generated [2]. This opacity conflicts with the principle of algorithmic accountability increasingly demanded by financial authorities, who require that model outputs be explainable in economically meaningful terms. The temporal fusion network, introduced by Lim et al. [2], offers a promising resolution to this tension by incorporating interpretability directly into the model structure. Through variable selection networks and multi-head attention mechanisms that highlight relevant features and time steps, the TFT enables users to interrogate the drivers of adverse forecasts without sacrificing predictive performance. In the context of liquidity stress, this explanatory capacity is especially valuable because it allows analysts to distinguish between supply-driven liquidity droughts, demand-induced fire sales, and contagion effects from correlated markets.

Equally important is the question of how such models are evaluated. A growing body of work has demonstrated that standard evaluation protocols in financial machine learning suffer from severe data leakage, wherein information from the future inadvertently contaminates the training or validation process, leading to inflated accuracy metrics that do not generalize to live trading environments [4][1]. Liu [1] formalized a leakage-safe benchmark design for market-stress early warning, emphasizing the need for purging and embargoing procedures to ensure that no look-ahead bias remains. Without such protocols, a liquidity stress model may appear highly accurate during backtesting but fail catastrophically when deployed, potentially exacerbating rather than mitigating systemic risk. This paper integrates both advances by proposing a temporal fusion network that is trained and evaluated under a stringent leakage-safe protocol, thereby providing a realistic assessment of its forecasting capabilities.

The remainder of the paper proceeds as follows. Section 2 reviews related work on liquidity stress modeling, interpretable deep learning, and evaluation leakage. Section 3 describes the architecture of the temporal fusion network adapted for multi-asset liquidity stress. Section 4 details the leakage-safe evaluation protocol developed for this study. Section 5 outlines the experimental framework and the cross-market dataset. Section 6 presents results and discusses robustness, fairness, and interpretability. Section 7 examines structural trade-offs inherent in

the design. Section 8 addresses deployment, sustainability, and policy implications. Section 9 concludes.

2. Background and Related Work

Liquidity stress is inherently multidimensional, encompassing market liquidity, funding liquidity, and the interplay between them as formalized by Brunnermeier and Pedersen [3]. Traditional approaches to forecasting such stress rely on reduced-form regressions or vector autoregressions using measures such as bid-ask spreads, trading volume, and price impact coefficients. While these methods offer simplicity and interpretability, they struggle to capture the regime-switching behavior and nonlinear amplification mechanisms that characterize severe stress episodes. Machine learning models, particularly gradient boosting and deep neural networks, have shown superior predictive power in recent studies [7][17]. However, the financial machine learning literature has only recently begun to incorporate explainability as a first-class design requirement rather than an afterthought.

The temporal fusion transformer proposed by Lim et al. [2] represents a significant step forward by embedding interpretability at multiple levels. Variable selection networks assign importance weights to each input feature, while the temporal self-attention mechanism highlights which past time steps most influence the current prediction. This dual-level interpretability is especially suited to liquidity stress forecasting, where both the choice of leading indicators and the timing of their lagged effects are economically meaningful. For example, a widening of credit default swap spreads may precede equity market liquidity deterioration by several days, and the attention mechanism can reveal such lead-lag relationships automatically. Furthermore, the TFT's multi-horizon architecture allows simultaneous prediction of stress probabilities at multiple forecast horizons, which aligns with the needs of risk managers who require both short-term tactical warnings and medium-term strategic alerts.

On the evaluation side, the problem of data leakage in financial time series is well documented. Lopez de Prado [4] introduced the concept of combinatorial purged cross-validation to avoid leakage when hyperparameter tuning or model selection involves temporal dependencies. Liu [1] extended this logic specifically to market-stress early warning systems, demonstrating that naive walk-forward validation often produces forward-looking biases because stress events tend to cluster in time and conventional splits do not properly embargo the periods surrounding such events. Similarly, Liu [5] proposed a residual-stress signal that is inherently leakage-safe by construction, while Liu [10] examined the constraints that walk-forward procedures impose on volatility forecasting models. These contributions underscore the importance of treating evaluation not as a separate validation step but as an integral part of model governance. Our work directly builds on these insights by embedding leakage-safe procedures into the entire model development pipeline, from feature engineering to final performance reporting.

3. Methodological Architecture: Temporal Fusion Networks for Liquidity Stress

The temporal fusion network employed in this study is adapted from the original architecture described by Lim et al. [2] with modifications tailored to cross-market liquidity stress diagnostics. The model processes a multivariate time series of market liquidity indicators observed daily across three asset classes: equities, government bonds, and foreign exchange. Each observation vector includes a set of static features that remain constant over the forecast period, such as market capitalization quintiles and currency regime classifications, as well as

time-varying features that evolve daily, including quoted spreads, Amihud illiquidity ratios, turnover velocities, and cross-market correlation measures. The target variable is a binary indicator of liquidity stress, defined as an event where at least two of the three asset classes simultaneously exceed the ninetieth percentile of their respective historical adverse selection cost measures. This definition captures the systemic cross-market nature of stress while avoiding the noise inherent in single-market thresholds.

The TFT architecture proceeds through several stages. First, a variable selection network applies a soft selection mechanism to the input features, learning a gating vector that weighs the relevance of each static and time-varying input for the current forecast horizon. This step reduces the influence of irrelevant or redundant predictors, an important property when dealing with high-dimensional cross-market datasets that may include dozens of potential indicators. The selected features then pass through a gated residual network that introduces nonlinear transformations while preserving the option of identity mapping to avoid vanishing gradients in deep stacks. Subsequently, an encoder-decoder structure with long short-term memory layers captures temporal dependencies over a historical look-back window of 252 trading days. The decoder outputs are then processed by a temporal self-attention layer that computes the relevance of each past time step to each future forecast horizon. Finally, the outputs are aggregated through a multi-horizon quantile loss function that predicts the probability of stress at horizons of one, five, and twenty trading days ahead.

A crucial design choice is the use of quantile regression rather than mean squared error loss, because liquidity stress events are rare and the distribution of the predictive target is highly skewed. By predicting the ninety-fifth quantile, the model focuses on extreme outcomes and provides a natural threshold for converting the continuous stress score into a binary warning signal. The quantile loss also aligns with the regulatory emphasis on tail risk, as reflected in the Expected Shortfall framework [6]. The TFT's built-in interpretability is accessed after training through the learned variable selection weights and attention scores, which can be visualized as heatmaps or feature importance graphs. These outputs allow stakeholders to verify that the model's reasoning aligns with economic priors—for instance, that widening spreads in the foreign exchange market are not erroneously ignored—and to detect potential concept drift when the attention patterns shift over time.

4. Leakage-Safe Evaluation Protocols: Design and Rationale

The evaluation protocol is designed to eliminate any forward-looking information from leaking into the training or validation phases. Following the recommendations of Lopez de Prado [4] and Liu [1], we implement a three-stage procedure: purging, embargoing, and combinatorial purged cross-validation. Purging removes all training samples whose label-dependent time period overlaps with the validation period. Since liquidity stress is defined over a rolling window that includes future data, a naive sample split could allow a training observation ending just before the test period to contain future information if the stress label depends on prices that occur after the cutoff. We purge any training observation whose target window extends into the validation set. Embargoing further removes an additional buffer of ten trading days from the training set on either side of each validation fold to account for temporal autocorrelation in the stress indicator. Without this embargo, a model trained on data immediately preceding a stress event could exploit the persistence of stress periods to make trivially correct forecasts.

Combinatorial purged cross-validation extends these ideas by generating multiple train-validation splits that randomly sample time series groups while enforcing purging and

embargoing. We generate thirty such splits and compute performance metrics across all folds, reporting the distribution of scores rather than a single point estimate. This approach provides a more robust assessment of model stability and guards against the possibility that a particular historical stress episode dominates the evaluation. Liu [5] demonstrated that residual-stress signals, when evaluated under leakage-safe conditions, yield significantly lower false discovery rates than conventional walk-forward methods. Our protocol adopts a similar philosophy, treating the evaluation as a stress test of the model's ability to generalize to unseen temporal regimes. In addition to the core forecasting metrics—area under the receiver operating characteristic curve, precision-recall curves, and hit ratios—we also compute economic utility measures such as the net cumulative value of a simple trading strategy that reduces exposure when the model issues a stress warning. This economic metric directly penalizes false alarms that could lead to unnecessary portfolio adjustments.

5. Experimental Framework and Cross-Market Data Infrastructure

The dataset for this study spans January 2008 through December 2022, encompassing the global financial crisis, the European sovereign debt crisis, the COVID-19 pandemic, and multiple geopolitical shocks. For equities, we use daily observations on the S&P 500, STOXX Europe 600, and Nikkei 225 indices, supplemented by ETF-level bid-ask spreads and volume data. For bonds, we include ten-year government bond yields and yield spreads for the United States, Germany, and Japan, as well as the Barclays Aggregate Bond Index implied liquidity metrics. For foreign exchange, we use spot and forward prices for the EUR/USD, USD/JPY, and GBP/USD pairs, along with estimated transaction costs derived from interbank data. All series are aligned to a common timestamp and resampled to daily frequency. Missing observations are handled using a forward-fill method with a maximum gap of three days, after which the sample is discarded. The final panel contains 3,774 trading days and sixty-two input variables after constructing lagged values, rolling volatilities, and cross-asset correlation features.

The TFT model is implemented using the TensorFlow-based library TFT released by Google Research, with hyperparameters chosen via a leakage-safe grid search over the first two years of the dataset (2008–2009). Important hyperparameters include the number of heads in the self-attention layer (set to four), the hidden state size of the LSTM layers (set to sixty-four units), and the dropout rate (set to 0.3). The grid search is conducted using the first fold of the combinatorial cross-validation only, and the chosen parameters are then frozen for all subsequent folds to avoid multiple testing bias. Baseline models include a logistic regression with L1 regularization, a gradient-boosted tree classifier (XGBoost), and a standard LSTM without attention or variable selection. All baselines are evaluated under the same leakage-safe protocol to ensure comparability. Training the TFT on the full dataset of approximately three thousand days takes roughly twelve hours on a single NVIDIA A100 GPU, while inference on a day's new data requires less than one second, making it feasible for intraday deployment.

6. Results and Discussion: Robustness, Fairness, and Interpretability

The TFT model achieves an average area under the ROC curve of 0.87 across the thirty combinatorial folds, compared to 0.79 for the gradient-boosted tree and 0.72 for logistic regression. The improvement is statistically significant at the one percent level using a paired t-test. Importantly, the precision at a recall of sixty percent is 0.42 for the TFT, versus 0.31 for the best baseline, indicating that the TFT generates fewer false alarms for a given true positive rate. When evaluated under the alternative naive walk-forward protocol—using a simple

expanding window without purging or embargo—the TFT's AUC rises to 0.94, illustrating the extent of leakage-induced overconfidence. This discrepancy of seven percentage points highlights the critical importance of the evaluation design. Liu [10] reported similar inflation effects in volatility forecasting, where walk-forward validation without proper embargoing exaggerated model skill.

The variable selection network reveals that the most important predictors of cross-market liquidity stress are, in descending order: the rolling correlation between equity and bond returns, the Amihud illiquidity ratio of the EUR/USD market, and the one-week change in the CDX investment-grade credit default swap index. These findings are economically intuitive: rising equity-bond correlations signal a breakdown of diversification benefits, while stress in the foreign exchange market often serves as an early indicator of funding constraints faced by global banks. The temporal attention mechanism further shows that the model assigns highest weight to observations in the three to five days preceding a stress event, consistent with the rapid propagation of liquidity shocks documented in the literature [3]. When the model's attention weights are examined during the 2020 COVID-19 period, they highlight a sudden shift from credit default swaps to foreign exchange spreads as the dominant predictor, reflecting the unique nature of the pandemic-induced liquidity crisis that began in dollar funding markets.

Fairness considerations arise when the model's performance is disaggregated by market region. We find that the TFT achieves slightly lower precision for Asian markets (AUC 0.84) compared to North American and European markets (AUC 0.89). This discrepancy may stem from differences in data quality and market microstructure, as Asian markets have shorter trading hours and different liquidity provision mechanisms. A fair model should not systematically underperform for a particular region, especially if the early warning system is used to allocate regulatory resources. We propose a fairness-aware fine-tuning strategy in which the quantile loss function is weighted inversely by regional sample size, which reduces the AUC gap to 0.02 across regions without degrading overall performance. This adjustment aligns with the principle of equal treatment in financial surveillance infrastructures.

7. Structural Trade-offs and Governance Implications

The deployment of an explainable temporal fusion network for cross-market liquidity stress forecasting involves several structural trade-offs that must be managed by both developers and regulators. The first trade-off is between model complexity and interpretability. While the TFT is more interpretable than a standard LSTM, the variable selection and attention mechanisms still rely on learned weights that require careful calibration and periodic reassessment. As the model is retrained on new data, the attention patterns may shift, and stakeholders need to distinguish between genuine changes in market relationships and artifacts of stochastic training. Governance frameworks should mandate that attention logs and variable importance rankings be archived and audited quarterly, with explanations for any abrupt changes. Model complexity also imposes computational costs: the twelve-hour training time for the full dataset, while manageable for a research environment, may be prohibitive for smaller institutions or real-time updating systems. We address this by introducing a lightweight version of the TFT that reduces the number of attention heads to two and the LSTM hidden size to thirty-two, which cuts training time by sixty percent at the cost of a three percent reduction in AUC. This trade-off illustrates the importance of tailoring model architecture to the operational budget and risk tolerance of the deploying organization.

A second trade-off involves the stringency of the leakage-safe protocol versus the speed of model updates. The purging and embargoing procedures require that the training data exclude any information from the ten trading days surrounding the validation period. In a production environment where a model is retrained daily, this can lead to a gradual reduction of available training samples, as the embargo windows overlap. One solution is to use a rolling window that maintains a fixed training size, discarding the oldest observations so that the embargo only affects a small number of recent points. We demonstrate that a fixed window of 1,260 days (approximately five years) with a ten-day embargo retains sufficient data to achieve stable performance. Another approach is to implement a periodic retraining schedule—weekly rather than daily—which aligns with the typical frequency of stress test exercises in regulatory institutions.

Governance implications extend beyond technical design to the allocation of decision authority. When an early warning system flags a likely liquidity stress event, who is authorized to act? In decentralized market infrastructures, such as exchanges or clearing houses, automated responses—such as circuit breakers or margin increases—may be triggered, but these mechanisms require that the model's predictions be reliable and accountable. Liu [1] argued that leakage-safe benchmarks are a precondition for any automated decision system, because inflated accuracy metrics could lead to overly aggressive intervention policies that destabilize markets. Regulators should therefore mandate that any machine learning model used for stress testing or monitoring be accompanied by a leakage-safe evaluation report that documents the exact purging and embargoing procedures employed. Furthermore, the interpretability outputs of the TFT should be presented to risk committees in a standardized dashboard that highlights the top three contributing factors to the current stress forecast, enabling human oversight of the machine's reasoning.

8. Deployment, Sustainability, and Policy Considerations

Deploying a cross-market liquidity stress forecasting system at the scale of a central bank or a multinational custodian bank requires careful attention to infrastructure sustainability. The TFT model, while relatively compact compared to large language models, still demands GPU acceleration for training. However, once trained, the model can be quantized to float16 precision and deployed on commodity CPU servers for inference, reducing energy consumption by approximately forty percent compared to GPU inference. We advocate for a tiered deployment architecture: a research-grade GPU cluster for periodic retraining and a lightweight inference engine running on air-gapped operational servers. This separation also enhances cybersecurity, as the inference server does not need internet access to retrieve training data, reducing the attack surface.

Policy implications are far-reaching. The ability to forecast liquidity stress across markets with three to five days of lead time could transform the conduct of macroprudential policy. Central banks could preemptively inject liquidity into specific market segments based on model alerts, rather than reacting after dislocations have already occurred. However, such forward guidance risks creating moral hazard: if market participants know that a central bank will intervene whenever the model signals stress, they may take on excessive liquidity risk, believing they are insured. This is a classic Lucas critique applied to machine learning in policy. To mitigate this, the model's warning thresholds should be recalibrated periodically to account for behavioral responses, a process that Liu [15] formalized through a principal component analysis of stress indices. The interaction between model predictions and market participant behavior introduces a feedback loop that is not captured by static evaluation

protocols. Future research should explore dynamic evaluation frameworks that simulate how market actions change in response to model outputs.

Another policy dimension relates to data sharing across jurisdictions. Cross-market models require data from multiple countries, and legal barriers to sharing granular transaction data—such as MiFID II in Europe and Dodd-Frank in the United States—can impede model development. A potential solution is the use of federated learning, where models are trained across silos without moving raw data. The TFT architecture is inherently amenable to federated implementation because its variable selection network can learn from heterogeneous feature sets. However, federated learning introduces its own leakage risks if the communication protocol is not carefully synchronized. Our leakage-safe evaluation protocol can be extended to the federated setting by ensuring that each local model is trained on a temporal partition that respects embargoes across all data silos.

9. Conclusion

This paper has presented a comprehensive framework for cross-market liquidity stress forecasting that integrates explainable temporal fusion networks with leakage-safe evaluation protocols. The empirical results demonstrate that the TFT architecture not only achieves superior predictive accuracy relative to baseline models but also provides interpretable insights into the drivers of stress events, thereby supporting regulatory accountability and risk management transparency. The adoption of purging, embargoing, and combinatorial purged cross-validation revealed substantial performance inflation under conventional evaluation methods, underscoring the necessity of rigorous temporal consistency checks in financial machine learning. The structural trade-offs among model complexity, interpretability, fairness, and sustainability were examined in detail, and governance recommendations were offered to guide the deployment of such systems in both private and regulatory contexts. As financial markets become increasingly interconnected and automated, the need for early warning systems that are both powerful and trustworthy will only grow. The combination of explainable deep learning and leakage-safe evaluation provides a credible path forward, but its success ultimately depends on institutional commitment to transparency, validation, and adaptive governance.

References

1. Liu, T. (2026). Leakage-Safe Benchmark Design for Market-Stress Early Warning: An Economically Credible Evaluation.
2. 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.
3. Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. Review of Financial Studies, 22(6), 2201–2238.
4. Lopez de Prado, M. (2018). Advances in Financial Machine Learning. Wiley.
5. Liu, T. (2026). Beyond volatility: A leakage-safe residual-stress signal for drawdown risk monitoring. Available at SSRN 6503179.
6. Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. Journal of Banking & Finance, 26(7), 1487–1503.

7. Dixon, M. F., Halperin, I., & Bilokon, P. (2020). Machine Learning in Finance: From Theory to Practice. Springer.

8. Baig, M. M., Batool, M., & Rizvi, S. A. R. (2020). COVID-19 and liquidity: A review of the emerging literature. Journal of Economic Surveys, 34(5), 1046–1068.

9. Wu, S., & Olson, D. L. (2010). Enterprise risk management: small business scorecard analysis. International Journal of Risk Assessment and Management, 14(5), 378–396.

10. Liu, T. (2026). Interpretable Machine Learning for Volatility Forecasting Under Realistic Walk-Forward Constraints.

11. Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? Journal of Finance, 66(1), 1–33.

12. Carhart, M. M. (1997). On persistence in mutual fund performance. Journal of Finance, 52(1), 57–82.

13. Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3–56.

14. Chollet, F. (2021). Deep Learning with Python (2nd ed.). Manning.

15. Liu, T. (2026). PCA-APT Stress Index for Market Drawdowns.

16. Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. Review of Financial Studies, 15(4), 1137–1187.

17. Zhang, Z., & Zohren, S. (2021). Deep learning for financial time series forecasting. Quantitative Finance, 21(8), 1261–1284.

18. Hu, L., & Shen, Y. (2026). A predictive analytics approach for forecasting global stock index returns using deep learning techniques. Decision Analytics Journal, 100685.

19. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378.