Explainable Macro-Financial Fragility Detection Using Residual Stress Factors and Temporal Attention Models

Dan Dai
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
dand@ucf.edu

Jeffrey Zimmerman
Department of Computer Science, University of North Texas, Denton, TX, USA.
jeffrey.zimmerman@unt.edu

Abstract

The increasing complexity and interconnectivity of global financial systems have rendered traditional risk metrics inadequate for anticipating systemic fragility. This paper proposes a novel framework that integrates residual stress factors—latent signals extracted from high-frequency market data—with temporal attention mechanisms drawn from deep learning architectures to detect macro-financial vulnerabilities in a more interpretable and forward-looking manner. The residual stress factor construct, grounded in statistical arbitrage and stress-testing theory, captures nonlinear, regime-dependent deviations from equilibrium that conventional volatility-based measures often miss. Temporal attention models, specifically transformer-style architectures, provide the capacity to learn long-range dependencies and identify the most influential time segments preceding crisis events, thereby offering a degree of explainability that is essential for regulatory acceptance and policy formulation. We examine the structural trade-offs inherent in deploying such a system at the systemic level, including the balance between model complexity and computational sustainability, the tension between predictive accuracy and interpretability, and the challenges of data governance across heterogeneous jurisdictions. The architecture is discussed as a socio-technical infrastructure requiring robust feedback loops between machine learning outputs and human judgment. We further consider the implications for financial stability monitoring, macroprudential policy, and equitable risk allocation across institutions and economies. The proposed framework is not intended as a standalone predictor but as a complementary layer within a broader surveillance ecosystem. By coupling residual stress signals with attention-based explanations, the system aims to support regulators in identifying nascent fragility before it cascades into systemic crises, while also addressing fairness concerns related to model bias and transparency. The paper concludes with a forward-looking perspective on the governance and institutional design necessary to operationalize such explainable fragility detection at scale.

Keywords

macro-financial fragility, residual stress factors, temporal attention, explainable AI, systemic risk, financial stability, deep learning, socio-technical infrastructure.

1. Introduction

Macro-financial systems are characterized by deep nonlinearities, feedback loops, and sudden phase transitions that defy traditional linear risk models. The global financial crisis of 2008

and the more recent liquidity disruptions of 2020 have amply demonstrated that commonly used risk measures, such as value-at-risk and volatility indices, often fail to provide early warning signals for systemic collapse [1]. In response, a growing body of research has turned to machine learning methods to capture the high-dimensional, time-varying dependencies that underpin fragility. However, the opacity of many deep learning models has created a tension between predictive power and the explainability that regulators and policymakers require [2]. This paper addresses that tension by proposing a hybrid framework that combines a theoretically motivated residual stress factor—drawn from the literature on statistical arbitrage and drawdown risk monitoring—with a temporal attention mechanism that surfaces the most salient historical periods for each fragility signal.

The concept of residual stress factors originates from the insight that market equilibria are rarely fully efficient; persistent deviations from cointegrating relationships across asset classes carry information about accumulating imbalances [3]. Unlike volatility, which is symmetric and often misleading during calm periods, residual stress signals measure the degree of mispricing relative to a long-term stable relationship, making them particularly sensitive to the kind of structural strain that precedes crises. Recent work has formalized this idea through a leakage-safe residual stress signal designed for drawdown risk monitoring, demonstrating that such signals can anticipate tail events well before volatility spikes [15]. However, extracting actionable signals from these factors at the macro level requires models capable of processing long sequences of interdependent financial time series and identifying the temporal pattern that precedes fragility.

Temporal attention models, particularly the transformer architecture, offer a promising solution. Unlike recurrent neural networks, transformers employ self-attention mechanisms that can weigh the importance of every time step in a sequence without suffering from vanishing gradients, enabling the detection of dependencies spanning months or even years [4]. Moreover, attention weights provide a natural form of explainability: by visualizing which historical periods the model attends to when making a fragility prediction, analysts can trace the model's reasoning and validate its logic against known economic events. This paper integrates residual stress factors as input features to a temporal attention network, thereby combining a theoretically grounded fragility measure with a transparent decision-making process.

We adopt a system-level perspective that emphasizes the architectural choices and trade-offs involved in deploying such a model for macro-financial surveillance. The discussion covers the selection of input data, the handling of non-stationarity, the computational costs of training very deep attention models on high-frequency data, and the governance implications of relying on an attention-based explanation in a regulatory context. The paper is organized as follows. Section 2 reviews related work on systemic risk measurement, deep learning for finance, and explainable AI. Section 3 develops the conceptual framework linking residual stress factors and temporal attention. Section 4 details the system architecture and the structural trade-offs inherent in its design. Section 5 examines deployment and infrastructure considerations, including data pipelines and real-time monitoring. Section 6 addresses governance, fairness, and policy implications. Section 7 concludes with future directions.

2. Background and Related Work

The detection of macro-financial fragility has long been a central concern of financial stability research. Traditional approaches rely on aggregate indicators such as credit-to-GDP gaps, asset price volatilities, and interbank network exposures [5]. While these measures capture

system-wide imbalances, they often suffer from a lack of timeliness: by the time credit gaps widen or volatility surges, the crisis may already be underway. More recent work has explored machine learning models capable of extracting early warning signals from high-dimensional datasets, including random forests, gradient boosting, and neural networks [6]. These models have shown improved predictive performance, but their black-box nature has limited their adoption by central banks and financial regulators, who require interpretable justifications for policy actions.

The concept of residual stress in financial markets arises from the cointegration framework, where long-run equilibrium relationships among asset prices are defined. Deviations from these relationships represent temporary mispricings that can be exploited via statistical arbitrage. However, these deviations also encode information about structural stress: when many such relationships simultaneously break down or widen, it may indicate that the underlying financial architecture is under strain [7]. The residual stress factor formalizes this idea by constructing a portfolio that is market-neutral but leveraged to the cross-sectional residuals of a cointegrating system. A surge in the variance of this factor—or a sudden collapse in its mean reversion—can serve as a leading indicator of drawdown risk [15]. This signal is particularly attractive because it is derived from observable prices and does not require arbitrary calibration to tail probabilities.

Temporal attention models emerged from natural language processing but have found increasing application in time-series forecasting. The transformer architecture, introduced by Vaswani et al., uses multi-head self-attention to allow each element of a sequence to attend to every other element, capturing long-range dependencies without the sequential processing constraints of recurrent networks [4]. In financial contexts, transformers have been applied to stock price prediction, volatility forecasting, and portfolio optimization, often outperforming LSTM-based models [8]. More relevantly, attention weights have been used to identify the time steps that most influence a model's prediction, offering a form of interpretability that is particularly valuable for fragility detection [9]. For instance, a model trained to predict systemic crisis events might assign high attention to periods of quantitative tightening or geopolitical shocks, providing a narrative alignment that supports human reasoning.

Explainable AI (XAI) has become a critical field in finance, driven by regulatory initiatives such as the European Union's General Data Protection Regulation and the Basel Committee's emphasis on model risk management [10]. Techniques such as SHAP values, LIME, and integrated gradients have been applied to credit scoring and fraud detection, but they often struggle with temporal dependencies [11]. Attention-based explanations are particularly appealing because they are inherent to the model architecture and do not require post-hoc approximation. However, they are not without limitations: attention weights can be fragile and may not always correspond to causal relationships [12]. This paper acknowledges these limitations and proposes a governance framework that treats attention-based explanations as one input among many in a deliberative process.

3. Conceptual Framework: Residual Stress Factors and Temporal Attention

The proposed framework operates on the premise that financial fragility is not a sudden event but a gradual accumulation of stress that manifests as persistent, small deviations from equilibrium relationships. Residual stress factors capture this accumulation by measuring the distance of a multi-asset system from a cointegrating surface. These factors are constructed from a set of asset prices that share a long-run equilibrium, often identified through principal component analysis or canonical correlation techniques [3]. The residual for each asset is the

part of its return that cannot be explained by the common factors, and the pooled residual variance across a large cross-section of assets forms the stress signal. Importantly, the signal is designed to be leakage-safe: it avoids relying on forward-looking information or illiquid derivative prices, making it suitable for real-time monitoring [15].

Temporal attention models process these stress signals over a rolling window of historical observations. The input to the model is a multivariate time series consisting of the residual stress factor for each asset class or sector, along with auxiliary variables such as interest rate spreads, credit default swap indices, and macro announcements. The transformer encoder maps this input into a sequence of hidden representations, and the final hidden state is used to predict a binary or continuous fragility indicator (e.g., the probability of a systemic crisis within the next three months). The attention mechanism produces a set of weights for each time step in the input window, indicating how much the model focused on that point when forming its output prediction.

The key advantage of this combination is that residual stress factors provide a theoretically grounded, low-noise input that is directly interpretable as a measure of market misalignment. The temporal attention model then learns which historical intervals of misalignment are most predictive of future breakdowns. For example, the model might learn that a rapid widening of cross-asset residuals during a period of central bank tightening is a stronger signal than a slow drift during a period of loose policy. The attention weights can be visualized as a heatmap over time, allowing analysts to see that the model is "looking" at, say, the three months preceding the 2008 Lehman collapse or the COVID-19 liquidity crisis.

This framework also addresses the problem of non-stationarity in financial data. Financial relationships evolve over time due to regulatory changes, market structure shifts, and technological innovations. A static cointegration relationship may break down. However, the temporal attention model can adapt to regime changes by attending more to recent data or by learning different attention patterns during different macroeconomic states. The model can be trained with a sliding window approach, and the residual stress factors can be re-estimated periodically to reflect updated equilibrium relationships. This dynamic re-estimation is computationally intensive but necessary for maintaining signal quality.

4. System Architecture and Structural Trade-offs

Designing a production-ready system for macro-financial fragility detection involves numerous architectural choices that entail trade-offs among accuracy, speed, interpretability, and sustainability. The first major choice is the data input layer. Residual stress factors require a constantly updated universe of asset prices across equities, fixed income, currencies, and commodities. The breadth of coverage increases the dimensionality of the input, which in turn increases the computational cost of the transformer's self-attention operation, which scales quadratically with sequence length. To manage this, the architecture may incorporate sparse attention mechanisms or hierarchical pooling that first compresses sector-level residuals before feeding them into the cross-sector attention layer [13]. The trade-off is a loss of granular information that might be important for capturing contagion across specific markets.

A second trade-off concerns the length of the historical window. Longer windows provide more context for attention to identify distant precursors, but they also require more training data and increase the risk of overfitting to past regimes. Shorter windows reduce compute and may improve generalization to structural breaks, but they may miss slowly developing fragility signals. A common design choice is to use a window of two to five years, updated

daily, with the attention mechanism free to assign negligible weight to older periods if they are irrelevant. The model's hyperparameters, such as the number of attention heads and the depth of the encoder, must be tuned on historical crisis episodes, which are rare events. This leads to a data imbalance problem that can be addressed through synthetic oversampling, cost-sensitive learning, or anomaly detection formulations [14]. Each approach introduces its own bias: oversampling can create unrealistic patterns, while anomaly detection may fail to capture early warning signals that resemble normal volatility.

The interpretability of attention weights is a central claim of the framework, but it requires careful validation. Attention weights do not necessarily reflect causal importance; they may simply indicate that certain time steps are correlated with the model's output, not that they caused the prediction. To mitigate this, the system can incorporate a secondary explainability module that computes integrated gradients or Shapley values on the input residual stress factors themselves, providing a second layer of attribution that can be cross-checked against the attention pattern [11]. This dual-explanatory approach increases computational overhead but provides more robust evidence for regulatory scrutiny.

Another structural trade-off is the decision between a centralized system architecture and a federated one. Centralized systems, where all data flows to a single model instance, are easier to train and maintain but raise data privacy and sovereignty issues, especially when cross-border data sharing is required. Federated learning, where local models are trained on data that remain within national borders and only gradient updates are shared, offers a way to respect jurisdictional boundaries while still capturing global network effects [16]. However, federated training of temporal attention models is challenging due to the need for synchronized sequence alignment and the risk of communication bottlenecks during training. A hybrid architecture that uses a central model trained on public or aggregated data and then fine-tuned locally with private data may offer a pragmatic middle ground.

5. Deployment and Infrastructure Considerations

Deploying a macro-financial fragility detection system at scale requires robust data pipelines capable of ingesting terabytes of tick-level and end-of-day data from multiple exchanges and data vendors, cleaning and aligning the data across time zones, and computing residual stress factors in near real time. The latency requirements vary by use case: for central bank surveillance, daily updates may be sufficient, whereas for high-frequency trading firms, millisecond-level updates might be needed. The system we describe is intended for supervisory and policy applications, so a daily or even weekly cadence is acceptable, but the infrastructure must still handle the volume of globally listed assets.

The computation of residual stress factors involves repeatedly estimating cointegrating vectors over a rolling window. This is a computationally expensive operation that can be parallelized across asset pairs or sectors using distributed computing frameworks [17]. The transformer model itself, when trained on a five-year window of daily data for thousands of assets, can have tens of millions of parameters. Training such a model from scratch is resource-intensive, but transfer learning can reduce the burden: a base transformer trained on a large corpus of synthetic or historical data can be fine-tuned on the specific target fragility signal. The inference cost is lower but still significant if the model is run daily for a large number of portfolios or risk factors.

Sustainability considerations are increasingly important in machine learning system design. The energy consumption of training and running large transformer models contributes to

carbon emissions, and financial institutions are under pressure to align their operations with climate goals [18]. The system can be optimized by using model pruning, quantization, and knowledge distillation to reduce model size without significant loss of predictive accuracy. Alternatively, the temporal attention model can be replaced by a lightweight variant, such as a linear transformer with linear complexity, at the cost of slightly reduced expressiveness [19]. The choice must be weighed against the need for high accuracy in detecting rare but catastrophic events.

Data governance is another critical infrastructure challenge. Financial data are subject to strict regulatory requirements regarding accuracy, completeness, and auditability. The residual stress factor construction must be fully documented and replicable, with version control for the cointegration estimates. The attention weights and model predictions must be stored in an immutable log that can be produced during regulatory examinations. The system should include a feedback loop that allows human analysts to flag false positives or false negatives and to update the model through online learning or retraining. This feedback loop is essential for maintaining trust and ensuring that the model adapts to evolving market structure without drifting away from its stability mandate.

6. Governance, Fairness, and Policy Implications

The deployment of an explainable fragility detection system raises important governance questions about who is responsible for interpreting the model's outputs and for acting on them. Central banks and financial regulators typically employ a committee-based decision-making process, and the attention-based explanations provided by the model can serve as a discussion tool rather than a deterministic rule. The model's predictions and attention heatmaps can be presented to the Financial Stability Committee alongside other indicators, allowing members to debate whether the model's "attention" to a particular historical period (e.g., the 2013 taper tantrum) is relevant to the current situation. This deliberation helps avoid automation bias and ensures that the model is a support tool, not a replacement for human judgment.

Fairness concerns arise in multiple dimensions. First, the residual stress factors are derived from market prices, which reflect the behavior of large institutional investors and may not capture the financial health of smaller or less liquid markets. This could lead to a systematic underestimation of fragility in emerging economies or in sectors dominated by small and medium enterprises [20]. Second, the temporal attention model may learn to attend disproportionately to periods that are historically unique to the financial centers of the United States and Europe, thereby biasing the fragility signal against regions with different market structures. To mitigate this, the model should be trained on a globally representative dataset, and the attention weights should be analyzed for regional disparities. If the model consistently ignores data from a particular region, the system should flag that as a potential fairness issue.

Policy implications of the proposed framework are significant. Macroprudential authorities could use the fragility detection system to adjust capital buffers, loan-to-value ratios, or liquidity requirements in a preemptive manner. However, such use introduces a feedback loop: if the model predicts elevated fragility and regulators tighten conditions, the tightening itself may alter the very market relationships that the residual stress factors measure. This reflexivity issue is well known in financial regulation and requires that the system be periodically recalibrated using out-of-sample periods where the regulatory response was different [21]. The attention-based explainability can help regulators understand whether their own actions have been incorporated into the model's logic, thereby informing a more dynamic policy framework.

Finally, the transparency provided by attention explanations can support public accountability. If a regulator decides not to act on a fragility warning, the model's attention pattern can be disclosed (in aggregated form) to explain why the warning was deemed false. Conversely, if a crisis occurs that the model missed, the absence of attention to certain pre-crisis periods can be examined to identify model gaps. Such transparency must be balanced with the need to protect proprietary trading algorithms and confidential supervisory information. A tiered disclosure mechanism, where general attention heatmaps are made public while detailed inputs remain confidential, could strike an appropriate balance.

7. Conclusion

This paper has presented a framework for explainable macro-financial fragility detection that integrates residual stress factors with temporal attention models. The residual stress factor provides a theoretically grounded, early signal of market misalignment that goes beyond conventional volatility measures. The temporal attention mechanism endows the model with a degree of transparency that is essential for regulatory acceptance, allowing analysts to see which historical periods most influence the model's predictions. We have examined the structural trade-offs inherent in designing such a system, from data dimensionality and window length to interpretability validation and computational sustainability. Deployment considerations include data pipelines, model training costs, and governance mechanisms for feedback and fairness. The socio-technical perspective emphasizes that the model must be embedded in a human decision-making process that respects reflexivity, jurisdictional diversity, and public accountability.

Future research should focus on evaluating the framework on out-of-sample crisis episodes, integrating additional stress sources such as geopolitical and climate risks, and developing more rigorous causal inference methods to complement attention-based explanations. The residual stress factor itself can be extended to incorporate derivative market data and non-traditional data sources such as central bank communications. As macro-financial systems become ever more interconnected and machine learning models become ever more powerful, the need for explainable, sustainable, and fair fragility detection will only grow. The proposed framework offers a step in that direction.

References

1. Danielsson, J., & Zigrand, J. P. (2008). Equilibrium asset pricing with systemic risk. Economic Theory, 35(2), 293–319.
2. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
3. Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. Econometrica, 55(2), 251–276.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
5. Borio, C. (2014). The financial cycle and macroeconomics: What have we learnt? Journal of Banking & Finance, 45, 182–198.

6. Kleinberg, J., Ludwig, J., & Mullainathan, S. (2018). An introduction to the special issue on machine learning and finance. Journal of Financial Economics, 130(3), 449–452.

7. Bondarenko, O. (2004). Statistical arbitrage and securities prices. Review of Financial Studies, 16(3), 875–919.

8. Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. Philosophical Transactions of the Royal Society A, 379(2194), 20200209.

9. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 3543–3556.

10. Basel Committee on Banking Supervision. (2021). Principles for effective risk data aggregation and risk reporting. Bank for International Settlements.

11. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

12. Serrano, S., & Smith, N. A. (2019). Is attention interpretable? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2931–2951.

13. Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509.

14. Le, T. T., & Kim, M. (2020). A cost-sensitive learning approach for rare event prediction in financial time series. Expert Systems with Applications, 145, 113137.

15. Liu, T. (2026). Beyond volatility: A leakage-safe residual-stress signal for drawdown risk monitoring. Available at SSRN 6503179.

16. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

17. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113.

18. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

19. Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast autoregressive transformers with linear attention. Proceedings of the 37th International Conference on Machine Learning, 5156–5165.

20. Stiglitz, J. E. (2010). Risk and global economic architecture: A view from the developing countries. In The Rationale for International Financial Standards (pp. 89–114). Edward Elgar.

21. Soros, G. (2013). Fallibility, reflexivity, and the human uncertainty principle. Journal of Economic Methodology, 20(4), 309–317.