Multi-Modal-TRANS: A Transformer-based Framework for Multi-Modal Financial Risk Forecasting on Temporal and Static Data

Michael Shell
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: http://www.michaelshell.org/contact.html

Xiang Fu
School of Information Science and Technology
ShanghaiTech University
Email: fuxiang2022@shanghaitech.edu.cn

James Kirk
and Montgomery Scott
Starfleet Academy
San Francisco, California 96678-2391
Telephone: (800) 555–1212
Fax: (888) 555–1212

1. Abstract

Accurate forecasting and correlation analysis of financial risks are crucial for maintaining market stability and guiding regulatory decision-making. However, traditional econometric and statistical models often fail to capture the nonlinear, non-stationary, and interdependent nature of modern financial systems. This project proposes a unified multimodal Transformer-based framework for the dynamic prediction and correlation analysis of credit risk, market risk, and liquidity risk. The model integrates dynamic time-series data (e.g., daily K-line, returns, volatility, and trading volume) with static structural data (e.g., corporate financial statements, leverage ratios, and macroeconomic indicators) to jointly learn temporal evolution and firm-level fundamentals. A cross-modal attention mechanism is employed to fuse heterogeneous features, while frequency decomposition techniques (such as EMD and CEEMDAN) are used to extract multi-scale components and suppress noise. The resulting hybrid architecture enables long-term dependency modeling and interpretable risk estimation across multiple dimensions. Empirical evaluations on large-scale market datasets will assess predictive accuracy, robustness, and interpretability compared with benchmark models such as LSTM and VAR. This research aims to provide a data-driven, explainable framework for systemic financial risk assessment, offering new insights into the dynamic interrelations among credit, market, and liquidity risks.

2. Introduction

The stability of modern financial systems depends heavily on the accurate prediction and effective management of financial risks. These risks—primarily credit risk, market risk, and liquidity risk—are interdependent and evolve dynamically under rapidly changing market environments. With the growing complexity of global markets and the massive availability of real-time financial data, there is an increasing need for intelligent and data-driven methods that can monitor and anticipate risk fluctuations in a timely and interpretable manner.

Traditional econometric models, such as Vector Autoregression (VAR) or GARCH-type frameworks, often struggle to capture the nonlinear dependencies and high-frequency dynamics inherent in financial markets. Recent advances in machine learning have introduced deep neural models capable of extracting complex patterns from large-scale financial data. However, most existing models still treat different data modalities—such as time-series trading data (e.g., daily returns, price volatility, transaction volume) and static fundamental indicators (e.g., leverage ratio, profitability, cash flow, macroeconomic metrics)—as isolated information sources. This limitation hinders the model's ability to reflect the joint effects of short-term market turbulence and long-term structural fundamentals.

In this context, Transformer-based architectures have emerged as a powerful paradigm for sequential data modeling, offering superior capability in capturing long-range temporal dependencies through self-attention mechanisms. Their non-recurrent structure allows efficient parallel computation and interpretable feature weighting, making them particularly suitable for high-frequency and heterogeneous financial data. Nevertheless, integrating static and dynamic information streams within a unified Transformer framework remains an open challenge due to differences in sampling frequency, feature representation, and statistical characteristics.

To address these challenges, this project proposes a multimodal Transformer-based framework that integrates dynamic market signals with static financial fundamentals for comprehensive financial risk prediction and cross-risk correlation analysis. The model aims to jointly evaluate credit, market, and liquidity risks by fusing information from diverse temporal and structural data sources. By leveraging multi-head attention mechanisms, the system will learn hierarchical feature dependencies across modalities and risk dimensions, enabling interpretable identification of systemic interactions. This integrated approach is expected to enhance the robustness and diagnostic capability of financial risk assessment, providing valuable insights for market stability, regulatory supervision, and investment decision-making.

3. Background

3.1. Financial Risk

The accurate assessment and proactive mitigation of financial risk are paramount for ensuring market stability and sound investment strategies. Financial risk prediction is fundamentally defined as the quantitative process of estimating the likelihood of an adverse financial outcome (e.g., asset impairment, corporate default) within a given period. Academic frameworks classify this risk into several distinct categories, essential for comprehensive risk management [2]. Market Risk. This quantifies potential losses due to adverse price fluctuations (e.g., equity, interest rate, or foreign exchange rates). It is characterized by high-frequency dynamics and sensitivity to real-time market events and investor sentiment. Liquidity Risk. This is the risk that a firm cannot meet its short-term debt obligations without incurring substantial losses, typically due to an inability to easily sell assets or raise cash. It has a mid-to-high frequency component, driven by trading volume and market depth. Credit Risk. This addresses a counterparty's fundamental inability or unwillingness to meet its contractual debt obligations. It is characterized by its low-frequency nature, often correlating with statutory financial reporting cycles, and reflects the counterparty's intrinsic solvency. Business Risk. This refers to the fundamental risk associated with a company's operations, reflecting the uncertainty of future profits. It is typically assessed via low-frequency fundamental metrics like cost structure and revenue volatility. Investment Risk. This broadly encompasses the chance that actual investment returns will be different from expected returns. It often integrates aspects of the risks mentioned above when evaluating portfolio performance. This project is specifically designed to enhance the prediction of firm-level risk by capturing signals related to Credit Risk and Market Risk from diverse, multi-modal data sources.

3.2. MultiModal Data

Effective risk modeling necessitates the integration of heterogeneous data, a core challenge this project addresses. We categorize the inputs used in this research based on their temporal nature: Static Data and Time Series Data. Static Data. This modality primarily comprises Fundamental Data derived from corporate financial statements. These inputs, such as debt-to-equity ratios and profitability metrics, are updated at a low frequency (quarterly or annually). They provide a deep, structural snapshot of a firm's intrinsic value and long-term operational health. Time Series Data. This modality includes Historical Trading Data sourced from stock exchanges (prices, volume, volatility). Characterized by its high frequency (tick, minute, or daily), this data captures continuous market dynamics, real-time sentiment, and instantaneous liquidity. The fundamental challenge in utilizing these streams jointly is the issue of fusion: successfully bridging the gap between the static, structural nature of fundamental data and the dynamic, high-frequency nature of time series data to achieve a unified, comprehensive risk profile.

3.3. Transformer

The development of the Transformer architecture marks a paradigm shift in sequential data processing, offering significant advantages over traditional recurrent neural networks (RNNs) for complex, long-range financial dependencies. Central to its superior performance is the Self-Attention Mechanism, which allows the model to globally weigh the relevance of all input elements (e.g., historical data points) to a single element being processed, regardless of their distance. This non-sequential processing capability is crucial for financial data, which is often characterized by non-linear, multi-scale, and long-range temporal correlations. Consequently, the Transformer has emerged as a state-of-the-art solution across various FinTech applications, including: high-frequency algorithmic trading, time series forecasting, and automated credit scoring. In the context of risk prediction, the Transformer's ability to efficiently capture complex relationships between distant historical market events and current fundamental data offers a robust foundation for building high-accuracy, context-aware risk models.

4. Related Work

The application of machine learning (ML) in financial risk prediction has led to two primary research trajectories: Fundamental-based Machine Learning Research, which focuses on analyzing structured accounting data for solvency assessment, and Market-based Time Series Modeling Research, which models high-frequency data for short-term dynamics.

4.1. Fundamental-Based Method

This research trajectory centers on Fundamental-based Machine Learning Research, aiming to leverage detailed financial statement data (e.g., balance sheets and income

statements) to assess corporate failure risk. The analysis focuses on metrics reflecting operations, revenue, profit, and profit margins [3] [1] [6]. These studies demonstrate superior classification performance compared to traditional statistical models. However, the predictive horizons of these models are inevitably constrained by the low frequency of the input data (quarterly reports). Consequently, models based purely on fundamental data are slow to react to rapidly emerging risk factors or sudden shifts in market confidence, limiting their utility for instantaneous risk management.

4.2. Technical-Based Method

In contrast to the low-frequency constraints of fundamental analysis, technical-based method focus on time series modeling research to capture the temporal dynamics of high-frequency trading data [4] [5]. This approach utilizes advanced sequence models, including architectures such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU), which have proven effective in forecasting short-term volatility, market momentum, and aggregate sentiment. While these models offer high real-time responsiveness, they typically lack the critical constraint of a company's fundamental value. Predictions derived solely from market signals are prone to amplifying transient noise, technical biases, or herd behavior, often failing to distinguish between fleeting market overreaction and genuine, fundamental solvency deterioration.

4.3. Multi-modal Method

While both research paths offer valuable, yet incomplete, insights, the logical evolution involves combining these heterogeneous data sources. Existing Multi-modal Fusion Attempts have been explored, but these studies reveal a Critical Gap in the literature. Early fusion attempts are often limited to simple feature concatenation, where extracted features from both modalities are merely joined before being fed into a final classification layer. This approach is shallow; it fails to learn the complex, non-linear interaction and synergistic effects between the two data streams . Specifically, these models struggle with: (1) Developing a sophisticated deep fusion mechanism to align and weigh multi-scale features appropriately, and (2) Achieving adequate model interpretability , which is crucial in regulated financial environments, making it difficult to ascertain whether a risk signal originates from underlying financial distress or panic-driven market dynamics.

The absence of an innovative, deeply integrated, and interpretable multi-modal deep learning architecture capable of harmonizing high-frequency market dynamics with low-frequency fundamental value. This project aims to fill this void by developing such a framework to significantly enhance the robustness and diagnostic capability of financial risk prediction.

5. Problem Statement

Financial markets are inherently nonlinear, dynamic, and interdependent systems in which multiple forms of risk—particularly credit risk, market risk, and liquidity risk—evolve simultaneously and interact through complex feedback mechanisms. Accurately forecasting these risks and quantifying their correlations is critical for maintaining market stability, optimizing investment strategies, and preventing systemic crises.

However, several fundamental challenges remain unsolved. First, existing econometric and machine learning models often treat each risk dimension independently and fail to capture cross-risk dependencies and temporal coupling among them. Second, financial data are inherently multimodal: they include dynamic, high-frequency time series (e.g., daily price returns, volatility, trading volumes) and static, low-frequency structural data (e.g., financial statements, leverage ratios, macroeconomic indicators). Traditional models lack mechanisms to effectively integrate these heterogeneous data sources. Third, most time-series prediction models, such as RNN or LSTM, struggle with long-term dependencies, non-stationarity, and noise sensitivity, resulting in unstable predictions under regime shifts or high-volatility conditions.

Consequently, there is a pressing need for a unified framework capable of:

(1) fusing static and dynamic data modalities,
(2) modeling complex temporal relationships across multiple risk dimensions, and
(3) providing interpretable correlation analysis to reveal systemic interactions.

This project addresses these gaps by developing a multimodal Transformer-based financial risk prediction framework that integrates dynamic (daily K-line and market indicators) and static (firm-level financial and macroeconomic) data. By leveraging self-attention and cross-modal fusion, the model aims to jointly predict credit, market, and liquidity risks, quantify their interdependencies, and improve both accuracy and interpretability in systemic risk assessment.a

6. Experiment Methodology

6.1. Overview

This project aims to develop a multimodal Transformer-based framework for comprehensive financial risk prediction and correlation analysis. The core methodology involves the integration of dynamic time-series data (e.g., daily K-line, returns, volatility, trading volume) and static structural data (e.g., corporate financial statements, leverage ratios, profitability, and macroeconomic indicators). By leveraging deep attention mechanisms, the proposed model will capture both temporal dependencies and cross-modal relationships, allowing for interpretable and robust forecasting of credit, market, and liquidity risks.

6.2. Data Mining Techniques to Be Used

The project will employ a systematic pipeline of advanced data mining and deep learning techniques to transform raw, heterogeneous data into action able risk insights, including the following three parts: data preprocessing, multimodal fusion architecture, and frequency decomposition for noise reduction.

In terms of data preprocessing, we obtain stock data of SSE, SZSE and financial reports. To address the differing scales of features, a Z-score normalization will be applied to time-series data, and Min-Max scaling will be used for features destined for attention mechanisms. Meanwhile, We will generate a comprehensive set of technical indicators from the daily K-line data to enrich the temporal representation. This includes: trend indicators (Simple Moving Average, Exponential Moving Average), momentum indicators (Relative Strength Index, Moving Average Convergence Divergence), volatility indicators (Bollinger Bands, standard deviation of returns), and liquidity indicators (average trading volume, volume-weighted average price).

In terms of multimodal fusion architecture, the core of our data mining approach is the novel fusion mechanism within the transformer framework, consisting of a temporal encoder, a static feature encoder, and a cross-attention for fusion. The preprocessed time-series of technical indicators and raw prices will be passed through a Transformer encoder. Self-attention layers will capture long-range dependencies and patterns across time. Then, the vector of static fundamental and macroeconomic indicators will be projected into a latent space using a dense neural network. Following that, we will use a cross-attention mechanism where the static feature vector serves as a query to attend to the encoded temporal sequence (keys and values) instead of simple concatenation. This allows the model to dynamically weigh the importance of different historical time points based on the company's current fundamental profile, enabling a deep, context-aware fusion.

In terms of frequency decomposition for noise reduction, we may employ Empirical Mode Decomposition (EMD) or its variant CEEMDAN to improve robustness against market noise. These techniques will decompose price series into intrinsic mode functions (IMFs) of different frequencies, potentially allowing the model to focus on trend and cyclical components while filtering out high-frequency noise.

6.3. Feasibility

The successful execution of this project is highly feasible, supported by four key pillars:

- **Data Feasibility:** The primary dataset required for this project is the "Shanghai and Shenzhen Stock Exchange Data," which we already possess. This dataset comprehensively includes daily K-line data, basic company information, balance sheets, and risk warning lists, perfectly aligning with our requirements for both dynamic time-series and static fundamental data. Its structured nature minimizes data collection overhead.
- **Technical Feasibility:** The proposed model is built on the Transformer architecture, which is well-established in deep learning libraries like PyTorch and TensorFlow. Team members have foundational knowledge in deep learning and financial data analysis. The implementation of encoders and attention mechanisms, while complex, is a well-documented process with ample open-source resources and research papers available for guidance.
- **Computational Feasibility:** Training a Transformer model on financial time-series data is computationally demanding but manageable. The data size, while large, is not on the scale of internet-scale datasets. The project can be conducted using high-performance computing resources available through the university or cloud computing credits (e.g., AWS, GCP), making the computational requirements feasible.
- **Methodological Feasibility:** The core concept of using attention for multimodal fusion is a recognized and active research area. Our approach of applying it to financial risk prediction is innovative yet methodologically sound, building directly on advancements in NLP and computer vision. The project's scope, from data preprocessing to model evaluation, follows a standard and achievable machine learning lifecycle.

6.4. Evaluation Plan

A rigorous and multi-faceted evaluation plan will be implemented to validate the performance, robustness, and utility of the proposed model.

- **Evaluation Metrics:** Model performance will be assessed using a suite of metrics to provide a comprehensive view including Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the F1-Score.
- **Baseline Models for Comparison:** The proposed Multi-Modal Transformer will be benchmarked against a range of strong baseline models to isolate the contribution of its architectural innovations, including traditional models (Logistic Regression, Gradient Boosting Machines) and deep learning models (LST and GRU).
- **Validation Methodology:** A time-series split (e.g., a rolling window or expanding window approach) will be used for all experiments to prevent data leakage and respect the temporal order of financial data. The dataset will be divided into sequential training, validation, and test sets.
- **Ablation Studies:** We will conduct ablation studies to quantify the contribution of each modality. This involves training the model with (1) only time-series data, (2) only static data, and (3) both, to demonstrate the necessity of multi-modal integration.

7. Timetable and Workload Division

The project will be conducted from **October 2024 to January 2025**, following the institutional schedule for proposal, interim, and final submissions. It is organized into

five major stages, each with distinct objectives and responsibilities distributed among the four team members (A–D). The division of work ensures balanced contribution across literature review, data preparation, modeling, evaluation, and reporting.

7.1. Stage 1: Literature Review and Project Planning (Oct 9 – Oct 20)

This stage focuses on developing a comprehensive understanding of financial risk prediction, multi-modal data fusion, and Transformer-based time-series analysis. The team will define the project scope, objectives, and expected outcomes while finalizing data sources and workflow for subsequent stages.

Member responsibilities:

- **A:** Research background, theoretical framework, and risk classification (credit, market, liquidity).
- **B:** Literature review on deep learning and Transformer architectures.
- **C:** Drafting project objectives, defining scope, and formulating the problem statement.
- **D:** Proposal formatting and coordination of the presentation.

7.2. Stage 2: Data Collection and Preprocessing (Oct 21 – Nov 23)

The focus is on acquiring and preparing the data. Both dynamic (daily K-line, returns, volatility, liquidity indices) and static (financial statements, leverage ratios, macroeconomic indicators) data will be collected, cleaned, normalized, and aligned. Exploratory analysis will be conducted to extract informative features.

Member responsibilities:

- **A:** Dynamic data acquisition and preprocessing.
- **B:** Static data collection and integration.
- **C:** Data normalization, imputation, and feature engineering.
- **D:** Exploratory data visualization and documentation.

7.3. Stage 3: Model Development and Interim Report (Nov 24 – Dec 2)

During this phase, the multi-modal Transformer framework will be implemented, combining temporal and static encoders via a cross-attention mechanism. The team will complete preliminary training, evaluate early results, and prepare the interim report and presentation.

Member responsibilities:

- **A:** Transformer encoder and attention fusion implementation.
- **B:** Integration of preprocessing modules and debugging.
- **C:** Experimental setup and baseline model comparison (LSTM, GRU).
- **D:** Report writing and interim presentation preparation.

7.4. Stage 4: Model Optimization and Evaluation (Dec 3 – Dec 30)

This stage involves hyperparameter tuning, model robustness tests, and interpretability analysis. Performance will be evaluated using standard regression and classification metrics, and comparative experiments will be conducted against baseline models.

Member responsibilities:

- **A:** Hyperparameter tuning and robustness validation.
- **B:** Baseline benchmarking and error analysis.
- **C:** Interpretability analysis (attention visualization, SHAP analysis).
- **D:** Report writing and presentation coordination.

7.5. Stage 5: Paper Writing and Submission (Jan 1 – Jan 9)

In the final stage, the team will prepare the academic paper or extended report summarizing the findings, finalize figures and references, and submit the complete work for evaluation or publication.

Member responsibilities:

- **A:** Writing of introduction and related work.
- **B:** Methodology and results sections.
- **C:** Figures, tables, and supplementary content.
- **D:** Editing, proofreading, and submission coordination.

References

[1] Noujoud Ahbali, Xinyuan Liu, Albert Nanda, Jamie Stark, Ashit Talukder, and Rupinder Paul Khandpur. Identifying corporate credit risk sentiments from financial news. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 362–370, 2022.

[2] Micha Bender and Sven Panz. A general framework for the identification and categorization of risks-an application to the context of financial markets. *Available at SSRN 3738273*, 2020.

[3] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[4] Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*, 2016.

[5] Young-Chan Lee. Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, 33(1):67–74, 2007.

[6] Dan Wang, Zhi Chen, Ionuţ Florescu, and Bingyang Wen. A sparsity algorithm for finding optimal counterfactual explanations: Application to corporate credit rating. *Research in International Business and Finance*, 64:101869, 2023.