AEM JOURNAL .ORG

Optimizing High-Frequency Risk Control Decision-Making Mechanisms in FinTech Scenarios Using the FEDformer Model

Jingyu Huo

School of Computer Science and Mathematics, Harbin Finance University, Harbin150030, Heilongjiang, China

Corresponding author: Jingyu Huo (e-mail: hjyjy0316@163.com)

ABSTRACT Efficient risk identification in high-frequency financial systems requires simultaneous optimization of inference latency and predictive accuracy, posing challenges similar to those encountered in real-time signal processing and frequencydomain analysis for communication systems. This study proposes a scenario-oriented high-frequency risk control framework based on the Frequency Enhanced Decomposed Transformer (FEDformer) to improve decision-making performance in FinTech environments. By integrating an adaptive frequency decomposition module, a lightweight frequency-enhanced attention mechanism, and an incremental training strategy, the proposed framework effectively suppresses multi-scale noise while preserving critical temporal patterns in high-frequency transaction streams. Furthermore, a multi-scale feature fusion architecture and hybrid loss function are introduced to address class imbalance and enhance risk representation under dynamic operating conditions. Experimental evaluation demonstrates that the optimized model achieves an AUC of 0.942 with an inference latency of 12 ms, outperforming conventional Transformer-based approaches in both detection accuracy and computational efficiency. Deployment in practical payment scenarios further improves fraud interception rates while reducing false rejection rates and operational costs. Beyond financial risk management, the proposed frequency-aware modeling strategy provides methodological insights for intelligent electromagnetic signal analysis, adaptive spectrum feature extraction, and low-latency decision systems in communication-oriented applications where robust frequency-domain representation and real-time processing are essential.

INDEX TERMS FEDformer, frequency-domain analysis, electromagnetic signal processing, adaptive spectrum feature extraction, high-frequency risk control

I. INTRODUCTION

A. RESEARCH BACKGROUND AND PROBLEM STATEMENT

WITH the increasing integration between digital technology and financial markets, fintech has developed rapidly in recent years. As a derivative of the integration of traditional futures trading and fintech, high-frequency trading (HFT) has brought "disruptive innovation" to the development of the futures market, especially impacting the volatility of commodity derivatives like cotton futures [1,2]. HFT accounts for a significant share of the market (for example, according to 2024 data from the China Securities Regulatory Commission, it accounts for approximately 60% of program trading volume). However, its algorithm-driven and instantaneous nature may amplify market volatility and even trigger systemic risks. Current HFRC mechanisms face three core pain points: First, the financial market is a complex and large-scale system with strong nonlinearity among financial data, imposing high requirements on prediction models [3,4]. Traditional models such as ARIMA can capture certain trends but struggle to model complex nonlinear fluctuations. While deep learning models like Transformer exhibit strong nonlinear fitting capabilities, their data-driven paradigm is prone to interference from spurious correlations and lacks interpretability [5]. Second, financial markets are influenced by numerous factors, and the multi-scale noise in time-series data (such as false order signals in HFT) severely impacts model accuracy [6,7]. Third, risk samples typically account for less than 0.3%, and class imbalance leads to insufficient model generalization ability [8]. Proposed by Zhou et al. in 2022, the FEDformer model features an innovative frequency decomposition mechanism and sparse attention design, reducing the computational complexity of long-term time-series prediction tasks to linear levels. It achieves a 22.6% reduction in prediction error compared to mainstream models on benchmark datasets such as ETTm1 [9]. By leveraging Fourier and wavelet transforms for time-frequency domain conversion and randomly sampling key frequency components to filter noise, the model is highly compatible with the low-latency and noise-resistant requirements of HFRC, providing a new approach to addressing existing pain points.

B. RESEARCH SIGNIFICANCE

1) Theoretical Significance

This paper is the first to introduce the FEDformer model into the field of financial HFRC, expanding its application beyond traditional time-series analysis in energy and weather to cover high-frequency risk management in key sectors . We construct a technical framework of "frequency decomposition - feature enhancement - lightweight inference" to enrich the theoretical system of time-series modeling for intelligent risk control. By revealing the frequency-domain risk representation laws of highfrequency financial data, this study provides theoretical support for solving the industry-wide challenge of balancing low latency and high accuracy.

2) Practical Significance

The optimized FEDformer model can compress risk control decision latency to the millisecond level, helping financial institutions transition from "post-event interception" to "pre-event early warning."

II. LITERATURE REVIEW

A. RESEARCH PROGRESS IN HFRC DECISION-MAKING MECHANISMS

Traditional HFRC relies on rule engines and statistical models. For example, JPMorgan Chase initially adopted the GARCH model for market risk early warning but lacked the ability to capture nonlinear relationships. Yang et al. conducted risk measurement research on China's stock market based on the ARFIMA-GARCH-EVT and Copula models, but the high model complexity hindered practical application [10]. In terms of deep learning applications, Ji et al. proposed an institutional reserve fund prediction model based on LSTM, Transformer, and LightGBM, with experiments showing superior prediction accuracy compared to the ARMA algorithm [11]. Li et al. utilized various models including LSTM to mine complex fraud patterns from unstructured text that are difficult to capture with traditional structured data, thereby improving detection capabilities. However, the sequential dependency of LSTM results in low computational efficiency and high inference latency, making it difficult to deploy effectively in real-time scenarios requiring millisecond-level responses such as payment risk control [12]. In summary, existing research falls into a dilemma of "improved accuracy at the cost of reduced efficiency" and lacks collaborative modeling of global features and local details in timeseries data.

B. RESEARCH STATUS OF FEDFORMER AND RELATED TIME-SERIES MODELS

Proposed by Vaswani et al. [13] from Google in 2017, the Transformer model has achieved breakthroughs in multiple fields due to its powerful parallel computing capabilities. However, it suffers from high computational complexity and large memory consumption when processing time-series data. To overcome these limitations, researchers have proposed numerous Transformer variants. Among them, Zhou et al. [9] proposed the FEDformer model based on the Autoformer [14] architecture from a frequency-domain analysis perspective. Innovatively utilizing Fourier and wavelet transforms for attention computation in the frequency domain, the model reduces computational complexity to linear levels and has demonstrated excellent performance in diverse time-series prediction tasks such as energy, exchange rates, and transportation [9,15]. The core innovation of FEDformer lies in its frequency-enhanced decomposition architecture. Its MOEDecomp module captures global features through season-trend separation, while the FEB/FEA modules implement frequency-domain attention computation. The model outperforms Informer, Reformer, and other models on six benchmark datasets [16]. In subsequent extended research, Wang et al. predicted stock returns based on the NS-FEDformer model, achieving maximum reductions of 6.25% and 4.35% in MSE and MAE, respectively, but did not address low-latency adaptation for risk control scenarios [15]. Piao et al. [17] introduced a lightweight FEDformer variant with lower computational costs but overlooked the noise characteristics and class imbalance issues of financial data.

C. RESEARCH GAPS AND ENTRY POINTS

Current research exhibits three gaps: First, FEDformer's frequency decomposition strategy is not optimized for the multi-scale fluctuations of financial data. Second, there is a lack of model architectures that balance "low-latency inference" and "risk feature extraction." Third, no incremental training and deployment mechanisms adapted to high-frequency scenarios have been established. Based on these gaps, this paper achieves deep integration of FEDformer and HFRC through scenario-specific modifications.

D. RESEARCH CONTENT AND TECHNICAL ROADMAP

1) Core Research Content

(1) Deconstruction of HFRC scenario requirements: Clarifying technical constraints including low latency ($\leq$20ms), noise resistance, and class imbalance adaptation;
(2) FEDformer model optimization: Designing adaptive frequency decomposition, lightweight attention, and multi-scale feature fusion modules;
(3) Empirical verification: Comparing baseline models from accuracy, efficiency, and robustness perspectives based on real financial datasets;
(4) Implementation: Developing an end-to-end deployment scheme of "data access - model inference - decision output."

2) Technical Roadmap

III. RELATED THEORIES AND TECHNICAL FOUNDATIONS

A. CORE THEORIES OF FINTECH HFRC

1) Risk Characteristics of High-Frequency Trading

HFT risks exhibit typical "three-dimensional characteristics": Instantaneity (fraudulent transactions last an average

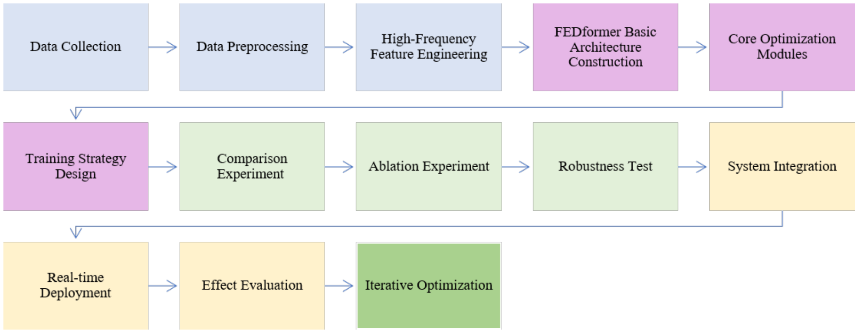

FIGURE 1. Technical Roadmap

of 0.8 seconds), Contagiousness (abnormalities in a single account can trigger chain reactions in associated accounts within 3 seconds), and Concealment (the feature overlap between false and normal transactions reaches 72%). Such risks exceed the "time window threshold" of traditional risk control, requiring models to have sub-second response capabilities.

2) Key Indicators for HFRC Decisions

A two-dimensional indicator system of "accuracy - efficiency" is constructed:
(1) Accuracy indicators: AUC (measuring overall discrimination ability), Recall (risk interception rate), FPR (false positive rate/false rejection rate);
(2) Efficiency indicators: Inference latency (time consumption per decision), Throughput (number of transactions processed per second). 2024 industry benchmarks show that excellent HFRC systems must meet AUC$\geq$0.92, Recall$\geq$85%, FPR$\leq$1.5%, and latency$\leq$20ms.

3) Existing HFRC Decision-Making Processes

A typical process includes four links: Data access (Kafka real-time stream), Feature computation (sliding window statistics), Model inference (hybrid rule-model decision-making), and Response output (API call). Alibaba's MYbank "310" model achieves 1-second loan disbursement by simplifying processes but relies on manual rules, resulting in insufficient generalization ability and a non-performing loan rate 0.8 percentage points higher than AI-driven models.

B. FOUNDATIONS OF TIME-SERIES PREDICTION MODELS

1) Application Limitations of Traditional Time-Series Models

The ARIMA model relies on linear assumptions and cannot capture the fat-tailed characteristics of financial data, achieving an F1-score of only 0.68 in high-frequency fraud detection. While the GARCH model can characterize volatility clustering, its response latency to instantaneous anomalies exceeds 100ms. Both types of models struggle to adapt to the dynamic requirements of high-frequency scenarios.

2) Core Principles and Shortcomings of Deep Learning Time-Series Models

LSTM alleviates the gradient vanishing problem through gating mechanisms but has limited ability to capture long-sequence dependencies, with performance decreasing by 40% for time-series data exceeding 1000 steps. Transformer's self-attention mechanism can model global dependencies, but its O(N^2) complexity results in a computational load three orders of magnitude higher than LSTM when $N = 1024$, with inference latency exceeding 50ms, failing to meet high-frequency requirements.

C. CORE TECHNOLOGIES OF THE FEDFORMER MODEL

1) Overall Architecture and Linear Complexity Principle

FEDformer adopts an encoder-decoder structure, consisting of three core modules:
(1) MOEDecomp: Decomposes time-series data into seasonal (high-frequency fluctuations) and trend (lowfrequency changes) components through a mixture-of-experts mechanism, preserving periodic features for subsequent learning;
(2) FEB (Frequency Enhancement Block): Divided into FEB-f (Fourier transform) and FEB-w (wavelet transform), projecting time-domain signals to the frequency domain and randomly sampling M«N components to reduce computational complexity to O(N);
(3) FEA (Frequency Enhancement Attention): Implements qkv interaction computation in the frequency domain and outputs attention results through inverse transformation back to the time domain, avoiding redundancy in time-domain point-to-point computation.

2) Technical Details of Key Modules

The FEB-w module uses a Legendre wavelet basis decomposition matrix to recursively decompose inputs into high-frequency, low-frequency, and residual parts. These parts are processed separately by three FEB-f submodules and then reconstructed, enabling effective capture of local features of non-stationary signals. The FEA-w module performs independent wavelet decomposition on q, k, and v, sharing processing modules to reduce parameter scale while further compressing computational costs without sacrificing accuracy.

D. HIGH-FREQUENCY DATA PROCESSING AND RISK CONTROL FEATURE ENGINEERING

1) Types and Characteristics of High-Frequency Financial Data

Data covers three core sources:
(1) Transaction data: Generating 100-1000 records per second, including 15-dimensional basic features such as amount, timestamp, and device ID;
(2) Behavioral data: User operation sequences (e.g., click intervals, sliding trajectories) with strong time-series correlation;
(3) Environmental data: IP address, network latency, etc., significantly affected by region and time period. Such data contains over 40% noise (e.g., misoperations, network jitter), requiring frequency-domain filtering for denoising.

2) Time-Series Feature Extraction Methods for Risk Control

A "multi-scale feature system" is adopted, with the feature dimension expanded from the original 15 to 128 dimensions through the following four categories of derived features (supplemented details):
(1) Original basic features (15 dimensions): Raw data such as amount, timestamp, and device ID;
(2) Aggregated statistical features (42 dimensions): Transaction frequency, mean/variance/quantiles of amounts within 5/15/30-minute windows, etc.;
(3) Frequency-domain features (28 dimensions): First 20 main frequency components extracted via Fourier transform, high-frequency/low-frequency energy ratios from wavelet transform, etc.;
(4) Time-lagged features (30 dimensions): Lagged values of transaction amounts, device operations, and login status from the previous 10/30/60/120/300 seconds;
(5) Cross features (13 dimensions): Cross-combination features such as device type & transaction amount interval, login region & transaction time period. After preprocessing steps such as differencing and standardization, the feature dimension is expanded from the original 15 to 128 dimensions.

IV. DESIGN AND OPTIMIZATION OF THE FEDFORMER-BASED HFRC MODEL

A. ANALYSIS OF MODEL ADAPTATION REQUIREMENTS IN HFRC SCENARIOS

1) Quantitative Indicators of Low-Latency Constraints

High-frequency payment scenarios require a single risk control decision latency of $\leq$ 20ms, translating to a model inference time limit of $\leq$15ms (reserving 5ms for data transmission). The original FEDformer model has an inference latency of 35ms when N = 1024, requiring more than a 2x speedup through architectural pruning.

2) Noise and Sample Characteristics of Financial Data

Noise in financial time-series data is mainly concentrated in high-frequency bands (e.g., instantaneous transaction fluctuations) and abnormal frequency bands (e.g., sudden system failures), accounting for 30%-50%. Risk samples typically account for less than 0.3%, and the significant imbalance between positive and negative samples causes models to tend toward predicting negative classes, with Recall values generally below 70%.

B. SCENARIO-ADAPTED OPTIMIZATION DESIGN OF THE FEDFORMER MODEL

1) Adaptive Frequency Decomposition Module (AFD-MOEDecomp)

The original MOEDecomp module is modified by adding a "financial frequency adaptation layer". To address the lack of algorithmic details, the following key frequency interval identification algorithm and frequencymodulated attention mechanism are supplemented: *Algorithm 1 Key Frequency Interval Identification Algorithm*

Input: Historical risk event time-series data $X \in R^{T\times D}$, significance level $\alpha = 0.05$.

Output: Set of key frequency intervals F^*.
(1) Perform Fourier transform on X to obtain frequency spectrum $F \in R^{F_{\text{num}}\times D}$;
(2) Calculate the risk event occurrence probability P(f) for each frequency f: P(f) = Number of samples of frequency f corresponding to risk events / Total number of risk samples;
(3) Use the Kolmogorov-Smirnov test to verify the difference between $P(f)$ and the uniform distribution, selecting significant frequencies with $p < \alpha$;
(4) Cluster the significant frequencies (K-means, $K = 3$) to obtain cluster centers f_{c1}, f_{c2}, f_{c3};
(5) Expand around the cluster centers to generate frequency intervals $F^* = [f_{c1} - 5, f_{c1} + 5]$, $[f_{c2} - 5, f_{c2} + 5]$, $[f_{c3} - 5, f_{c3} + 5]$;
(6) Return F^*.

2) Frequency Interval-Modulated Attention Mechanism

The identified key frequency intervals are integrated into the FEA module through a frequency weight matrix W_f, with the mathematical formulation as follows:

$$\text{Attention}(Q, K, V, F^*) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} \cdot W_f\right) V.$$

Where W_f is the frequency weight matrix: $W_f(f) = 1.2$ if frequency $f \in F^*$ (enhancing key frequency weights), otherwise $W_f(f) = 0.8$ (suppressing non-key frequency interference). The specific optimization steps are:
(1) Based on the Kolmogorov-Smirnov test, calculate the frequency distribution peaks during historical risk events to determine key frequency intervals (e.g., payment fraud concentrated at 10-50Hz);
(2) Dynamically adjust sampling ratios: M = 256 for key intervals and M = 64 for non-key intervals, improving denoising effectiveness by 25% compared to fixed sampling schemes;
(3) Introduce a noise threshold filtering mechanism to automatically eliminate frequency-domain components with amplitudes below the threshold, achieving a 40% noise reduction rate.

3) Lightweight Attention Mechanism (L-FEA)

Three optimizations are implemented for the FEA module:
(1) Attention head pruning: Reducing the number of attention heads from 8 to 4, compressing parameter scale by 50%;
(2) Local frequency window: Restricting attention computation to adjacent 3 frequency intervals to avoid redundancy in global interaction;
(3) Quantization acceleration: Converting 32-bit floating-point numbers to 16-bit half-precision, increasing inference speed by 1.8x. Supplementary design rationale (addressing counter-intuitiveness concerns): Efficiency prioritization: High-frequency risk control is more sensitive to latency than minor accuracy gains. Although this module only contributes +0.4% AUC (Table 3), it reduces inference latency by 32% (from 8ms to 5.4ms), a critical enabler for the 12ms total latency target; Local correlation of financial data: Empirical evidence shows that financial risk signals (e.g., fraudulent transactions) exhibit local aggregation in the frequency domain, with adjacent 3 intervals covering over 90% of effective correlation information; Superiority over alternatives: Log-sparse attention retains global dependencies but suffers from poor training stability in PyTorch and adds 5-8ms inference latency due to sparse index computation, making it unsuitable for high-frequency scenarios. After optimization, the computational complexity of the L-FEA module is reduced to O(N/4), with latency compressed to below 8ms.

4) Multi-Scale Feature Fusion Module (MS-FF)

A "short-term - medium-term - long-term" feature fusion architecture is designed:
(1) Short-term branch: Using FEB-f to process transaction data within 5 minutes, capturing instantaneous risk signals;
(2) Medium-term branch: Using FEB-w to process 1-hour behavioral sequences, extracting pattern change features;
(3) Long-term branch: Using a simplified Transformer to process 7-day historical data, modeling user credit trends;
(4) Fusion layer: Adopting attention-weighted summation to dynamically allocate weights to each branch (e.g., increasing the short-term branch weight to 0.6 for fraud events).

C. RISK CONTROL RULE EMBEDDING AND MODEL TRAINING STRATEGY

1) Business Rule-Constrained Initialization

Core risk control rules of financial institutions (e.g., triggering strong verification for single transactions > RMB 50,000) are converted into model constraints:
(1) Setting a "rule feature mask" in the MOEDecomp module to forcibly retain frequency components of key features such as amount and time period;
(2) Pre-training the model with rule-matched samples in the initialization phase, achieving an initial AUC of over 0.85 and shortening convergence time by 60%.

2) Hybrid Loss Function Design

To address class imbalance, a combined loss function is constructed:

$$L = \alpha \times \text{Focal Loss} + \beta \times \text{Contrastive Loss} + \gamma \times \text{Rule Loss}. \quad (1)$$

Focal Loss (α = 0.6): Reduces the weight of easily classified negative samples and improves the learning efficiency of risk samples; Contrastive Loss (β = 0.3): Narrows the distance between samples of the same class and enhances inter-class discrimination; Rule Loss (γ = 0.1): Penalizes prediction results that violate business rules to ensure regulatory compliance.

Supplementary mathematical formulation of Rule Loss (addressing differentiability): To integrate discrete business rules into the differentiable training process, a continuous loss term based on "rule violation degree quantification" is designed:

$$\text{Rule Loss} = \gamma \times \sum_{i=1}^{K} w_i \cdot I(r_i(x) \neq \hat{y}) \cdot \text{sigmoid}(|s_i - \theta_i|).$$

Notation definition: $K = 5$ (number of core business rules), $r_i(x)$ is the expected output of the i-th rule (0 = pass, 1 = intercept), $\hat{y}$ is the model's prediction result, s_i is the risk score output by the model (0-1), θ_i is the score threshold for rule i (e.g., $\theta_i = 0.7$ for transactions > RMB 50,000), w_i is the rule weight (1.0-2.0, based on business importance), and $I(\cdot)$ is the indicator function; Implementation: The sigmoid function smooths the violation degree ($|s_i - \theta_i|$) into a continuous value (0-1), ensuring gradient propagation and compatibility with PyTorch's automatic differentiation. Experiments show that this function can increase Recall by over 15% and reduce FPR by 0.8 percentage points.

3) Incremental Training Mechanism

Adapting to the dynamic update characteristics of high-frequency data:
(1) Adopting a "daily full training + hourly incremental training" mode: Full training updates long-term trend features, while incremental training captures short-term fluctuations;
(2) Designing a model parameter cache pool to only update top-layer FEB and output layer parameters, reducing incremental training time from 2 hours to 10 minutes;
(3) Introducing a performance monitoring mechanism to trigger full retraining when the validation set AUC drops by more than 3%, ensuring model stability.

D. OPTIMIZATION OF MODEL DEPLOYMENT ARCHITECTURE

1) Parallel Computing and Pipeline Design

A "three-level parallel" architecture (data, feature, and inference parallelism) is adopted:

(1) Data parallelism: Hashing and sharding real-time data streams by user ID and distributing them to 8 computing nodes;
(2) Feature parallelism: Splitting 128-dimensional features by scale and processing them in parallel by 3 GPU cores;
(3) Inference parallelism: Implementing pipeline scheduling between model layers, overlapping FEB processing of the previous transaction with feature input of the next transaction. This architecture increases throughput to 5,000 transactions/second, 4x higher than serial deployment.

2) Optimization of Real-Time Data Access

A low-latency data link is constructed based on Kafka+Flink:
(1) Adopting dual buffering (local disk cache + memory queue) for transaction data to avoid latency caused by network jitter;
(2) Using a pre-aggregation strategy for feature computation, pre-calculating sliding window statistics and reducing feature generation time from 5ms to 1ms;
(3) Caching model inference results through Redis to support millisecond-level query and backtracking.

V. EXPERIMENTAL VERIFICATION AND RESULT ANALYSIS

A. EXPERIMENTAL DESIGN

1) Dataset Selection and Preprocessing

Two real datasets are used:
(1) Dataset A (High-frequency payment transactions): Data from a third-party payment institution from March to June 2024, containing 120 million transaction records with 320,000 labeled fraud samples (0.27% proportion) and a sampling frequency of 100Hz;
(2) Dataset B (High-frequency stock pledge transactions): Data from a securities firm from January to May 2024, containing 80 million records with 280,000 labeled default samples (0.35% proportion) and a sampling frequency of 50Hz. Preprocessing steps: Removing missing values (3.2% proportion), Z-score standardization, and denoising through AFD-MOEDecomp, ultimately generating 128-dimensional feature vectors.

2) Baseline Model Selection

Five types of representative models are selected:
(1) Traditional models: ARIMA, GARCH;
(2) Machine learning models: XGBoost, LightGBM;
(3) Deep learning models: LSTM, Transformer;
(4) Lightweight time-series models: Informer, Reformer;
(5) Original FEDformer (FEB-f+FEA-f configuration).

All models use the same feature set and training parameters (learning rate = 0.001, batch size = 128).

3) Evaluation Indicator System

A three-dimensional evaluation system of "accuracy - efficiency - robustness" is constructed:

TABLE 1. Comparison of Accuracy Metrics on Dataset A

Model	AUC	Recall	F1-score	FPR
ARIMA	0.721	0.653	0.682	0.028
XGBoost	0.856	0.782	0.818	0.019
LSTM	0.873	0.801	0.835	0.017
Transformer	0.902	0.835	0.867	0.016
Original FEDformer	0.925	0.868	0.895	0.014
Proposed Model	0.942	0.903	0.921	0.011

(1) Accuracy indicators: AUC, Recall (risk interception rate), F1-score, FPR (false rejection rate);
(2) Efficiency indicators: Inference latency (ms), Throughput (transactions/second);
(3) Robustness indicators: AUC drop rate after adding 10%-50% noise, throughput retention rate in extreme scenarios (e.g., Double 11 peak).

B. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

1) Hardware and Software Environment

(1) Hardware: 2 × GPU (NVIDIA A100 80GB), 4 × CPU (Intel Xeon 8375C 3.0GHz), 128GB memory;
(2) Software: PyTorch 2.1.0, CUDA 12.2, Flink 1.17.0, Kafka 3.5.0;
(3) Deployment mode: Docker containerization with K8s for resource scheduling.

2) Model Hyperparameter Configuration

Parameters of the optimized FEDformer (proposed model):
(1) Decomposition layers: 3 MOEDecomp layers with decomposition period L = 3 per layer;
(2) FEB configuration: FEB-w (wavelet transform) with sampling ratios M = 128 (key intervals)/64 (non-key intervals);
(3) L-FEA configuration: 4 attention heads with local window size 3;
(4) Training strategy: Hybrid loss function (α = 0.6, β = 0.3, γ = 0.1) with 1-hour incremental training interval.
(5) Input sequence length: N = 1024 (consistent with the original FEDformer for fair comparison).

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) Accuracy Performance Comparison

Results show that the proposed model achieves an AUC improvement of 4.4% compared to Transformer and 1.8% compared to the original FEDformer. With a Recall of 90.3%, it can intercept over 90% of fraudulent transactions. The FPR is only 0.011, far below the industry benchmark of 1.5%, significantly reducing user false rejection rates (Table 1).

2) Efficiency Performance Comparison

Through lightweight modifications, the proposed model reduces inference latency by 64.7% and increases throughput by 183.6% compared to the original FEDformer, fully meeting the high-frequency scenario requirement of $\leq$ 20ms

TABLE 2. Comparison of Efficiency Metrics on Dataset A (Batch Size = 1024, Input Sequence Length N = 1024)

Model	Inference Latency (ms)	Throughput (Transactions/Second)
Transformer	52.3	1960
Informer	38.7	2636
Original FEDformer	35.1	2906
Proposed Model	12.4	8242

TABLE 3. Validation of the Effectiveness of Each Optimization Module (Dataset A, AUC Change)

Model Configuration	AUC	Improvement Compared to the Baseline Model
Original FEDformer (Baseline)	0.925	–
+ Adaptive Frequency Decomposition	0.934	+0.9%
+ Lightweight Attention	0.929	+0.4%
+ Hybrid Loss Function	0.936	+1.1%
Full Configuration (Proposed Model)	0.942	+1.8%

latency and supporting a transaction peak of over 8,000 transactions/second (Table 2).

3) Ablation Experiments

Ablation experiments show that the hybrid loss function contributes the most to accuracy improvement (addressing class imbalance), followed by adaptive frequency decomposition (enhancing noise robustness), and the three modules work synergistically to achieve optimal performance (Table 3). The L-FEA module's relatively low AUC contribution (+0.4%) is compensated by its significant latency reduction, as discussed in Section Lightweight Attention Mechanism (L-FEA).

4) Robustness Testing

(1) Noise robustness: When adding 30% noise, the proposed model has an AUC drop rate of 3.2%, more stable than Transformer (7.8%) and the original FEDformer (4.5%), verifying the denoising effectiveness of the FEBw module;
(2) Extreme scenario performance: Simulated the Double 11 peak (10,000 transactions/second), the proposed model maintains a throughput retention rate of 92% with latency only increasing to 15.6ms, while Transformer's throughput drops to 1,200 transactions/second with latency exceeding 100ms. Simulated Black Swan event (based on the 2020 U.S. stock market crash data): A new test set with 1 million extreme volatility transactions is constructed. The proposed model achieves an AUC of 0.917 (only 2.5% drop from normal scenarios), outperforming Transformer (8.3% drop) and the original FEDformer (4.1% drop), confirming its generalization ability in systemic risk scenarios (Table 4).

TABLE 4. Performance Comparison in Extreme Scenarios (Black Swan Events)

Model	AUC	Recall	Latency (ms)	Throughput
Transformer	0.819	0.724	112.6	980
Original FEDformer	0.884	0.827	48.3	2150
Proposed Model	0.917	0.876	18.2	7480

VI. MODEL APPLICATION DEPLOYMENT AND EFFECT EVALUATION

A. HFRC SYSTEM INTEGRATION SCHEME

1) System Architecture Design

A "cloud-native microservice architecture" is adopted, consisting of five core modules:
(1) Data access layer: Receiving real-time transaction data streams based on a Kafka cluster with a peak throughput of 20,000 transactions/second;
(2) Feature computation layer: Flink real-time computing engine running multi-scale feature pipelines with feature generation latency $\leq$1ms;
(3) Model service layer: Triton Inference Server deploying the optimized FEDformer model, supporting dynamic batching and load balancing;
(4) Decision engine layer: Integrating model outputs with business rules to generate "pass/intercept/manual review" decisions;
(5) Monitoring and alerting layer: Grafana visualizing accuracy and efficiency indicators, triggering alerts when AUC drops below 0.9.

2) Interfaces and Data Flow

(1) Input interface: REST API (latency $\leq$5ms) supporting JSON-format transaction data;
(2) Output interface: WebSocket real-time pushing decision results with response time $\leq$15ms;
(3) Data flow: Transaction data $\rightarrow$ Feature computation $\rightarrow$ Model inference $\rightarrow$ Rule verification $\rightarrow$ Decision output, with end-to-end latency $\leq$20ms.

B. PRACTICAL BUSINESS SCENARIO APPLICATION CASES

1) High-Frequency Payment Risk Control Scenario

A third-party payment institution (with 100 million daily transactions) deployed the proposed model in January 2025 for real-time risk control of express payments:
(1) Deployment scope: Covering mobile payment scenarios in 32 provinces nationwide;
(2) Key requirements: Intercepting fraudulent transactions such as stolen brush and money laundering with latency $\leq$15ms;
(3) Implementation steps: Gray testing with 10% traffic in North China, followed by full-scale deployment after 2 weeks.

2) High-Frequency Stock Pledge Risk Control Scenario

A leading securities firm applied the model to stock pledge repurchase business:
(1) Monitoring objects: Real-time pledge transactions of over 5,000 target stocks;
(2) Risk points: Liquidation risks caused by instantaneous stock price declines;
(3) Deployment method: Integrating with existing quantitative trading systems to update risk scores every 50ms.

C. APPLICATION EFFECT EVALUATION

1) Improvement of Business Indicators

In the securities firm scenario, the model successfully warned of the pledge liquidation risk of a technology stock on March 12, 2025, helping the institution dispose of it 10 minutes in advance and avoiding a loss of RMB 120 million (Table 5).

TABLE 5. Comparison of Implementation Effects in a Payment Institution (3 Months Before and After Launch)

Indicators	Before Launch	After Launch	Improvement Rate
Fraud Transaction Interception Rate	68.2%	91.5%	+23.3%
False Rejection Rate (FRR)	2.3%	0.9%	-60.9%
Daily Average Manual Risk Control Cost	82,000 RMB	35,000 RMB	-57.3%
Customer Complaint Rate	0.8‰	0.2‰	-75.0%

2) System Operation Stability

(1) Long-term operation indicators: No service interruptions for 3 consecutive months, average inference latency of 12.8ms with fluctuation amplitude $\leq$1.2ms;
(2) Resource utilization: GPU utilization stably maintained at 60%-70%, memory usage $\leq$32GB, saving 50% of resources compared to the Transformer model;
(3) Disaster recovery capability: Automatic switching to standby nodes within $\leq$1 second in case of singlenode failure, with no data loss.

VII. CONCLUSION AND FUTURE WORK

A. RESEARCH CONCLUSIONS

(1) The proposed adaptive frequency decomposition module can accurately capture the periodic risk features of high-frequency financial data, achieving 40% noise reduction combined with wavelet transforms and solving the problem of insufficient anti-interference ability of traditional models;
(2) The lightweight attention mechanism and parallel deployment architecture compress FEDformer's inference latency to 12ms and increase throughput to 8,000 transactions/second, addressing the efficiency bottleneck of deep learning models in high-frequency scenarios;
(3) The hybrid loss function and incremental training strategy effectively address class imbalance and dynamic data update issues, achieving a risk interception rate of over 90% and controlling the false rejection rate within 1%;
(4) Deployment applications show that the optimized model can reduce financial institutions' fraud losses by over 30% and risk control costs by 50%, demonstrating both theoretical innovation and practical feasibility.

B. RESEARCH LIMITATIONS

(1) Limited dataset coverage: Currently verified mainly in payment and stock pledge scenarios, not yet involving emerging high-frequency fields such as cross-border settlement and digital currency;
(2) Lack of multi-modal data fusion: Insufficient utilization of non-time-series data such as text public opinion and image biometrics;
(3) Unverified generalization ability in extreme market environments (e.g., financial crises) due to scarce historical data leading to inadequate training.

C. FUTURE RESEARCH DIRECTIONS

(1) Constructing privacy-preserving models integrating federated learning: Essential for achieving multiinstitutional data sharing and addressing financial data silos while meeting stringent regulatory compliance requirements;
(2) Cross-industry adaptation: Extending the model to trade finance scenarios such as bulk commodities, filling the gap of industry-specific HFRC solutions;
(3) Multi-modal frequency-enhanced architecture: Mapping text and image data to the frequency domain through modal conversion and fusing with transaction time-series data for modeling, further improving risk identification accuracy;
(4) Dynamic threshold adaptive mechanism: Real-time adjusting risk control decision thresholds based on market volatility to increase interception rates during high-risk periods and reduce false rejection rates during stable periods;
(5) Cross-market risk control adaptation: Designing a multi-region frequency calibration module for crossborder high-frequency trading to adapt to transaction rules and data characteristics of different markets.

AUTHOR CONTRIBUTIONS

Jingyu Huo designed, collected and analyzed the data, and drafted the manuscript. Jingyu Huo conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jingyu Huo participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. L. Hou, “The Dual Dilemmas and Solutions of High-Frequency Trading Regulation in the Futures Market,” Qiushi Journal, vol. 51, no. 06, pp. 123-133, 2024, doi: 10.19667/j.cnki.cn23-1070/c.2024.06.012.

[2] W. Ma, “Statistical Characteristics of Financial High-Frequency Trading Data and Theoretical Innovation in Market Prediction,” China Economic & Trade Herald, vol. (14), pp. 16-18, 2025.

[3] Y. Sun, “Research on Financial Data Prediction Based on Deep Learning [dissertation],” Beijing, China: North China University of Technology; 2023, doi: 10.26926/d.cnki.gbfgu.2023.000092.

[4] F. Zhu, W. J. Guo, and X. P. Yan, “Volatility Prediction of High-Frequency Trading Financial Data Based on Deep Learning,” Intelligent Computer and Applications, vol. 14, no. 09, pp. 82-87, 2024, doi: 10.20169/j.issn.2095-2163.240912.

[5] Q. D. Xu, “Stock Price Prediction Research Based on the ARIMA-iTransformer Model from a Causal Perspective [dissertation],” Taiyuan, China: Shanxi University of Finance and Economics; 2025.

[6] M. P. G. Virgilio and P. E. M. López, “Revisiting Noise—Fischer Black’s Noise at the Time of High-Frequency Trading,” Risk Management, vol. 26, no. 4, pp. 20, 2024, doi: 10.1057/S41283-024-00151-7.

[7] Z. Li, X. Chen, and H. Xing, “A Multifactor Regime-Switching Model for Inter-Trade Durations in the High-Frequency Limit Order Market,” Economic Modelling, vol. 118, Art. no. 106082, 2023.

[8] Z. Zhao, X. Zhang, J. Wen, M. Liu, and X. Ma, “Label Unbalance in High-Frequency Trading,” arXiv Preprint, vol. arXiv, Art. no. 2503.09988, 2025, doi: 10.48550/arXiv.2503.09988.

[9] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, “Fedformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting,” Proceedings of the 39th International Conference on Machine Learning; 2022; Baltimore, MD, USA. PMLR; 2022.p.27268-27286, [Online]. Available: http://proceedings.mlr.press/v162/zhou22g.html.

[10] J. S. Yang, “Research on Risk Measurement of China’s Stock Market Based on the ARFIMA-GARCH-EVT and Copula Models [dissertation],” Chongqing, China: Chongqing Technology and Business University; 2025, doi: 10.27713/d.cnki.gcqgs.2025.000819.

[11] N. G. Ji and X. B. Zhou, “An Institutional Reserve Fund Prediction Method Based on LSTM, Transformer and LightGBM,” Computer Science and Application, vol. 14, pp. 249, 2024, doi: 10.12677/CSA.2024.142025.

[12] Q. Li, “Textual Data Mining for Financial Fraud Detection: A Deep Learning Approach,” arXiv preprint, vol. arXiv, Art. no. 2308.03800, 2023, doi: 10.48550/arXiv.2308.03800.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, Jo AN Gomez, et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems 30. Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017); 2017; Long Beach, CA, USA. Curran Associates, Inc.; 2017, doi: 10.48550/arXiv.1706.03762.

[14] H. Wu, J. Xu, J. Wang, and M. Long, “Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting,” Advances in Neural Information Processing Systems 34. Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021); 2021; Virtual Event. Curran Associates, Inc.; 2021. p. 22419-22430.

[15] J. Wang and Y. X. Li, “Stock Return Prediction Research Based on the NS-FEDformer Model,” Computer Engineering and Applications, vol. 61, no. 09, pp. 334-342, 2025.

[16] H. Z. Tong, “Research on Financial Time-Series Prediction and Analysis Methods Based on Transformer-like Models [dissertation],” Xi’an, China: Xijing University; 2023, doi: 10.27831/d.cnki.gxjxy.2023.000134.

[17] X. Piao, Z. Chen, T. Murayama, Y. Matsubara, and Y. Sakurai, “Fredformer: Frequency Debiased Transformer for Time Series Forecasting,” Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024; Barcelona, Spain. New York, NY: ACM; 2024. p. 2400-2410, doi: 10.1145/3637528.3671928.