

Fourier-Domain Features and Machine Learning for Maximum Drawdown Prediction in the CSI 1000 Index

Hao Wang

*Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong, China
25117594g@connect.polyu.hk*

Abstract. In the field of investment decision-making and risk management, maximum drawdown is a key measure of downside risk. Time-domain variables such as returns, volatility, historical drawdowns, trading volume, and turnover are relied on by existing studies, but frequency-domain structures receive less attention. This study aims to examine, for the CSI 1000 Index, whether frequency-domain features can improve machine learning forecasts of future maximum drawdown, and the data is from Eastmoney. Based on the conventional risk characteristics, three Fourier indicators are constructed, namely Frequency Energy Shift (FES), Spectral Entropy (SE) and Stock Market Spectral Synchronization (SSC). The samples are divided into training sets, validation sets, and test sets according to time. The model used for prediction is LightGBM. In drawdown observations, MAE, RMSE and Tail MAE are used to evaluate the overall prediction accuracy and performance. The baseline model only uses conventional features, and the full model adds these three frequency-domain indicators. Finally, the results of the two models are compared. Compared with the baseline model, the MAE of the full model has decreased by 12.67%, RMSE by 11.56%, and Tail MAE by 14.71%. The results show that the frequency-domain features can improve the overall prediction accuracy and provide some information not captured by traditional risk variables. This research framework provides a perspective for the prediction of future maximum drawdown and is also helpful for monitoring the downside risk of the stock market.

Keywords: Fourier Analysis, Machine Learning, Maximum Drawdown, LightGBM, CSI 1000

1. Introduction

The financial market itself has the characteristics of uncertainty and nonlinearity, so downside risk prediction has always been very important in quantitative finance and risk management. Maximum drawdown measures the largest loss from a local high point to the subsequent low point over a given period of time. Recently, many studies use machine learning models to predict extreme equity risks, such as maximum drawdown and stock market crashes [1-6]. Accurately predicting maximum drawdown can help investors find unfavorable market conditions earlier.

Machine learning methods are widely used in financial time series prediction, and the predictors include time-domain variables such as returns, volatility, transaction activity and liquidity [7-10].

Research on China's stock market also shows that nonlinear models are useful for predicting yield and price [11-14]. In addition, financial asset prices fluctuate across multiple frequencies. Some recent studies show that Fourier, wavelet, entropy, and some other time-frequency features can effectively complement traditional variables [15-19]. This study selects energy distribution, complexity, and market synchronization as research subjects, these features may contain information not captured by traditional statistical methods.

This study examines the constituent stocks of the CSI 1000 Index, and the trading data used is from Eastmoney. In addition to conventional time-domain risk features, it constructs three Fourier indicators. Frequency Energy Shift (FES) measures changes in the relative strength of high- and low-frequency components, Spectral Entropy (SE) describes spectral energy dispersion and complexity, Stock Market Spectral Synchronization (SSC) captures the frequency-domain relationships between the CSI 1000 Index and two market benchmarks: the CSI 300 Index and the CSI 500 Index. Chronological training, validation, and test sets are input into LightGBM to predict future maximum drawdown. The study will compare two models: a baseline model with conventional features and a full model that adds FES, SE, and SSC to this baseline, the results will be used to determine whether frequency-domain structures provide additional predictive information and, in turn, whether this is a useful feature-engineering approach for downside risk prediction.

2. Methodology

Figure 1 summarizes the whole research workflow, including data processing, feature building, model parameter selection and experimental evaluation.

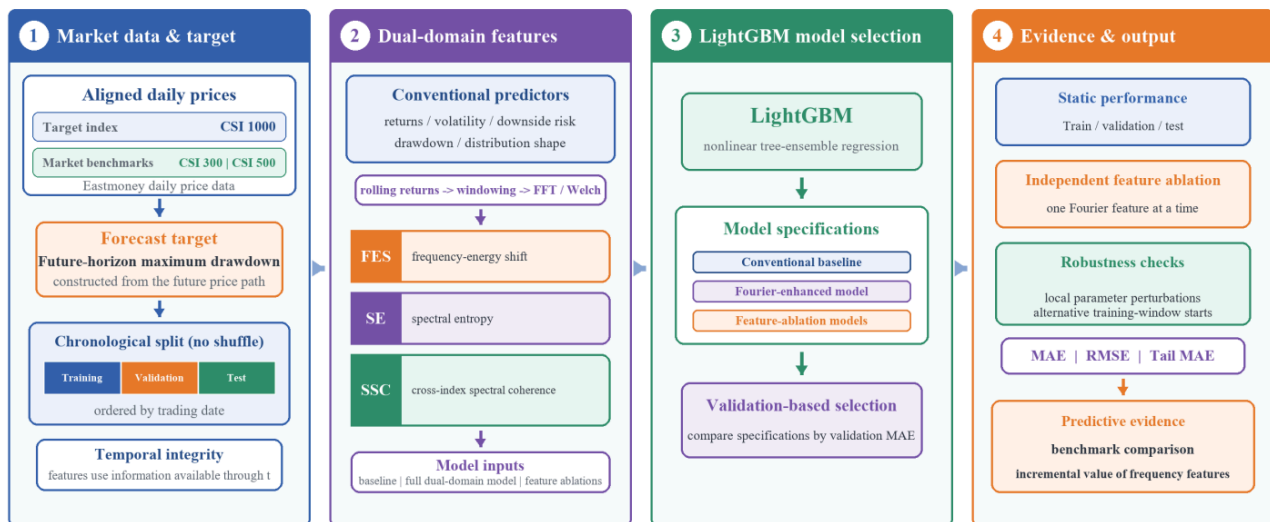


Figure 1. Overall research framework for forecasting the maximum drawdown of the CSI 1000 Index using Fourier features

2.1. Data source, sample design and prediction target

The daily price data of the CSI 1000, CSI 300 and CSI 500 indices all come from Eastmoney. The time range of the sample is from April 4, 2019 to July 3, 2026.

Referring to the general method of time series evaluation [20], samples are divided into training set, validation set and test set according to time, accounting for 70%, 15% and 15%, respectively.

For each trading day t , the prediction target is the maximum drawdown over the subsequent 20 trading days, defined as follows:

$$MDD_t^{(20)} = \min_{1 \leq a \leq b \leq 20} \left(\frac{P_{t+b}}{P_{t+a}} - 1 \right) \quad (1)$$

In this formula, P_{t+j} represents the index price at trading day $t+j$.

This target measures the maximum percentage decline from a peak to a subsequent trough during the next 20 trading days.

2.2. Traditional time-domain features

The baseline model uses seven traditional time-domain features, including the 5-day return, 20-day return, 20-day volatility, downside volatility, realized maximum drawdown, skewness and excess kurtosis.

2.3. Frequency-domain feature construction

In the processing for each trading day, the return series is first mean-centered and then multiplied by a Hann window. Subsequently, the discrete Fourier transform is applied, and the direct current component is removed [15-19]. The meaning of Frequency Energy Shift (FES) is to capture changes in the strength of high-frequency components relative to low-frequency components, the definition of FES is shown below:

$$FES_t = \Delta \log \left[\frac{\sum_{k \in H} S_{t,k} + \varepsilon}{\sum_{k \in L} S_{t,k} + \varepsilon} \right] \quad (2)$$

In this formula, $S_{t,k}$ represents the spectral energy at frequency k on trading day t , H represents the high-frequency set, ε is a small positive constant used to ensure numerical stability.

The rolling window for FES is 50 days, and the boundary between high-frequency and low-frequency components is a 10-day period. A positive value indicates that the relative strength of rapid market fluctuations is increasing.

Spectral Entropy (SE) is calculated using normalized entropy and describes the degree of dispersion and complexity of the spectral energy distribution, the definition of SE is shown below [17, 19]:

$$SE_t = - \frac{\sum_{k=1}^K p_{t,k} \log(p_{t,k})}{\log(K)} \quad (3)$$

In the above equation, $p_{t,k}$ is the normalized spectral energy at frequency k on trading day t , and the meaning of K is the number of frequency components. The range of SE is 0-1, and it is calculated using a 60-day rolling window. The larger the value, the more dispersed and complex the spectral energy distribution.

The role of Stock Market Spectral Synchronization (SSC) is to capture the frequency-domain relationship between the CSI 1000 Index and two market benchmarks. The study separately calculates the Welch coherence between the CSI 1000 Index and the CSI 300 Index and the Welch coherence between the CSI 1000 Index and the CSI 500 Index, and then averages the two coherence measures [17, 18].

The calculation of SSC uses a 60-day rolling window and 15 trading-day Welch segments. Each adjacent segment overlaps by 8 trading days, which is approximately 50% of the segment length. The analysis period is required to be longer than 15 trading days. SSC is defined below:

$$SSC_t = \frac{1}{2|L|} \sum_{k \in L} [Coh_{1000,300}(k) + Coh_{1000,500}(k)] \quad (4)$$

Here, $Coh_{1000,300}(k)$ denotes the Welch coherence between the CSI 1000 and the CSI 300 at frequency k , while $Coh_{1000,500}(k)$ denotes the corresponding coherence between the CSI 1000 and the CSI 500. L represents the retained set of frequencies, and $|L|$ is the number of frequencies included in that set.

2.4. Model specification and hyperparameter selection

The baseline model contains seven conventional time-domain predictors, and the full model further adds FES, SE, and SSC on this basis. Both the baseline model and the full model are LightGBM models that use DART as the boosting type. Recent studies have demonstrated that LightGBM and related boosted-tree ensemble methods are effective in financial prediction [21-24]. The parameter space is an 81-combination grid covering `num_leaves` {10, 15, 25}, `max_depth` {4, 5, 6}, `learning_rate` {0.02, 0.04, 0.06}, and `n_estimators` {120, 140, 160}. The parameters for the baseline model, full model, and three 7+1 ablation experiments are all selected according to their respective lowest MAE values in the validation set.

In the order of `num_leaves`, `max_depth`, `learning_rate`, and `n_estimators`, the selected values for the baseline model are (25, 6, 0.04, 140), the selected values for the full model are (15, 5, 0.06, 120), the selected values for the 7+FES model are (25, 6, 0.06, 140), the selected values for the 7+SE model are (25, 6, 0.06, 120), and the selected values for the 7+SSC model are (15, 5, 0.06, 160).

Fourier settings remain fixed: FES uses a 50-day window and a 10-day boundary, SE uses a 60-day window, SSC uses a 60-day window with 15-trading-day Welch segments. Adjacent segments overlap by 8 trading days, and only frequency components with periods longer than 15 trading days are retained.

2.5. Evaluation metrics

In this study, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Tail Mean Absolute Error (Tail MAE) are used to evaluate predictive performance [20, 25]. MAE is mainly used to reflect the overall magnitude of the prediction errors. RMSE pays more attention to the impact of large errors, while Tail MAE is used to specifically measure errors associated with severe drawdowns.

The formulas for the three metrics are presented below, where y_i and $\hat{y}_i$ represent the actual and predicted maximum drawdowns, respectively, n denotes the number of observations, and T represents the set of tail observations. For Tail MAE, the threshold for tail observations is the first quartile, that is, tail observations are those whose actual maximum drawdowns are less than or equal to the first quartile of the maximum drawdowns in the training set.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$\text{MAE}_{\text{tail}} = \frac{1}{|T|} \sum_{i \in T} |y_i - \hat{y}_i| \quad (7)$$

Following comparative forecast-evaluation practice [20, 25], the percentage improvement relative to the baseline model is calculated using the formula below, a positive value indicates a lower prediction error:

$$\text{Improvement}(\%) = \frac{M_{\text{baseline}} - M_{\text{model}}}{M_{\text{baseline}}} \times 100 \quad (8)$$

3. Results

3.1. Baseline and full-model performance

Table 1 reports the performance of the baseline model and the full model on the training, validation, and test sets. Percentages in parentheses indicate the improvements over the baseline.

Table 1. Prediction errors of the baseline and full models

Split	Model	MAE	RMSE	Tail MAE
Train	7-feature baseline	0.0193	0.0280	0.0446
Train	7 + FES + SE + SSC	0.0178 (+7.87%)	0.0249 (+10.95%)	0.0381 (+14.54%)
Validation	7-feature baseline	0.0328	0.0439	0.0604
Validation	7 + FES + SE + SSC	0.0313 (+4.61%)	0.0420 (+4.45%)	0.0537 (+11.10%)
Test	7-feature baseline	0.0321	0.0483	0.0840
Test	7 + FES + SE + SSC	0.0280 (+12.67%)	0.0427 (+11.56%)	0.0716 (+14.71%)

Across all three datasets and all three metrics, the full model outperforms the baseline model. Figure 2 summarizes the percentage reductions, Table 1 provides the corresponding error values.

On the test set, the MAE, RMSE, and Tail MAE of the full model are reduced by 12.67%, 11.56%, and 14.71%, respectively.

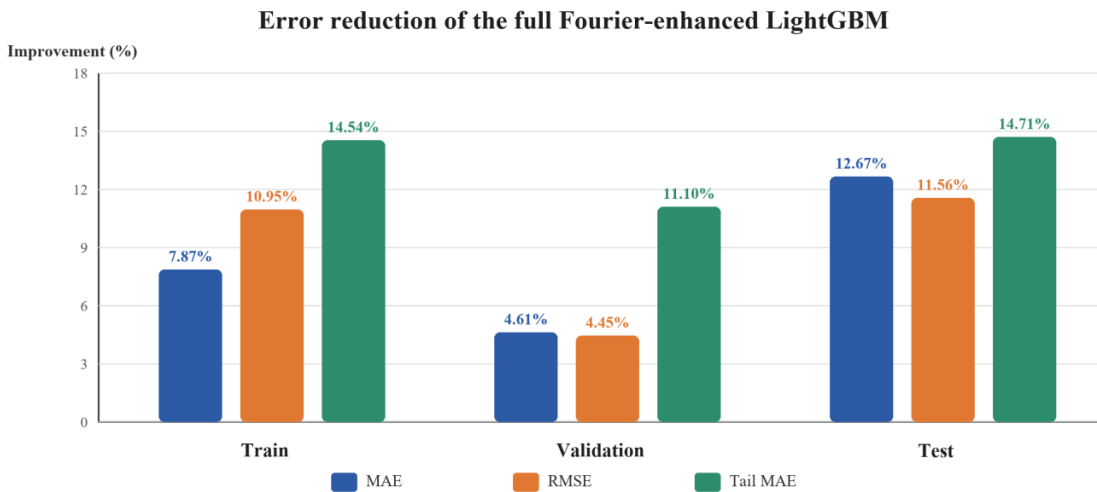


Figure 2. Percentage reductions in prediction errors of the full model relative to the baseline model using seven traditional features

3.2. Independent ablation results for individual fourier features

Table 2. Performance of independently tuned models combining the seven traditional features with one Fourier feature

Variant	Split	MAE	RMSE	Tail MAE
7 + FES	Train	0.0167 (+13.41%)	0.0239 (+14.52%)	0.0365 (+18.12%)
	Validation	0.0316 (+3.68%)	0.0424 (+3.33%)	0.0535 (+11.49%)
	Test	0.0317 (+1.19%)	0.0468 (+3.20%)	0.0755 (+10.12%)
7 + SE	Train	0.0163 (+15.42%)	0.0232 (+17.01%)	0.0356 (+20.17%)
	Validation	0.0320 (+2.38%)	0.0422 (+3.81%)	0.0556 (+7.93%)
	Test	0.0313 (+2.40%)	0.0469 (+2.94%)	0.0800 (+4.71%)
7 + SSC	Train	0.0189 (+1.94%)	0.0266 (+4.73%)	0.0415 (+6.92%)
	Validation	0.0305 (+6.84%)	0.0422 (+3.82%)	0.0553 (+8.39%)
	Test	0.0307 (+4.30%)	0.0455 (+5.92%)	0.0735 (+12.47%)

Relative to the baseline model, each independently tuned 7+1 model improves all three-test metrics. 7 + FES reduces MAE, RMSE, and Tail MAE by 1.19%, 3.20%, and 10.12%, respectively. 7 + SE reduces MAE, RMSE, and Tail MAE by 2.40%, 2.94%, and 4.71%. 7 + SSC reduces MAE, RMSE, and Tail MAE by 4.30%, 5.92%, and 12.47%.

The full model achieves the largest reductions in all three-test metrics. Every ablation is tuned independently based on validation MAE, Table 2 isolates the contribution of each feature.

3.3. Robustness test under fixed sample division

Using the original method of dividing samples by time, under the condition of 15 local hyperparameter perturbations, the performance improvement of the model is still positive: the reduction in MAE is between 5.27% and 12.67%, the reduction in RMSE is between 8.41% and 11.56%, and the reduction in Tail MAE is between 9.35% and 29.30%.

In addition, after 10 tests using shortened training windows, the reduction in MAE is still maintained between 6.87% and 12.67%, the reduction in RMSE is maintained between 6.14% and 11.56%, and the reduction in Tail MAE is maintained between 6.41% and 14.71%. This shows that model performance does not fluctuate greatly due to changes in the length of the training history and remains stable.

4. Conclusion

This study attempts to add three Fourier-enhanced features to the seven basic features within the LightGBM framework to predict the maximum drawdown of the CSI 1000 Index over the next 20 trading days. The data were obtained from Eastmoney. The training set, validation set and test set are divided according to the ratio of 70%/15%/15%. The full model combines seven basic time-domain features with FES, SE, and SSC, the combination of these three indicators is mainly used to verify whether frequency-domain information can provide predictive value beyond conventional indicators.

Under the same experimental settings, the baseline model, the full model and three ablation experiments were analyzed. The selected full model has 15 leaves, a maximum depth of 5, a learning rate of 0.06, and 120 estimators.

On the test set, the MAE, RMSE, and Tail MAE of the full model were reduced by 12.67%, 11.56%, and 14.71%, respectively. The results of the ablation experiments and robustness tests show that all three-evaluation metrics improve when each Fourier feature is added separately, and the performance improvement remains positive under local hyperparameter perturbations and shortened training windows. These results indicate that frequency-domain features can provide information that cannot be captured by the baseline model.

In general, frequency-domain information can provide predictive value beyond traditional variables. The three indicators respectively correspond to changes in the relative intensity of high- and low-frequency components, the dispersion and complexity of spectral energy, and spectral synchronization between the CSI 1000 Index and the two market benchmarks. These properties also provide an interpretable basis for the improvement of model performance. This research framework shows how frequency-domain representations complement conventional risk indicators without replacing the useful information already contained in conventional indicators. These findings provide a new perspective on downside risk management in the financial market and also offer a more rational basis for investment decisions.

References

- [1] Goldberg, L. R., & Mouti, S. (2022). Sustainable investing and the cross-section of returns and maximum drawdown. *The Journal of Finance and Data Science*, 8, 353–387.
- [2] Dichtl, H., Drobetz, W., & Otto, T. (2023). Forecasting stock market crashes via machine learning. *Journal of Financial Stability*, 65, 101099.
- [3] Malladi, R. K. (2024). Application of supervised machine learning techniques to forecast the COVID-19 U.S. recession and stock market crash. *Computational Economics*, 63, 1021–1045.
- [4] Ren, T., Li, S., & Zhang, S. (2024). Stock market extreme risk prediction based on machine learning: Evidence from the American market. *The North American Journal of Economics and Finance*, 74, 102241.
- [5] Sheng, Y., Qu, Y., & Ma, D. (2024). Stock price crash prediction based on multimodal data machine learning models. *Finance Research Letters*, 62, 105195.
- [6] Deng, S., Luo, Q., Zhu, Y., Ning, H., Yu, Y., Gao, Y., Shen, Q., & Shimada, T. (2024). Multi-sentiment fusion for stock price crash risk prediction using an interpretable ensemble learning method. *Engineering Applications of Artificial Intelligence*, 135, 108842.
- [7] Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, 197, 116659.
- [8] Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., Tang, C., & Li, J. (2022). A survey on machine learning models for financial time series forecasting. *Neurocomputing*, 512, 363–380.
- [9] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379, 20200209.
- [10] Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Aubet, F. X., Callot, L., & Januschowski, T. (2023). Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55, 1–36.
- [11] Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. *Journal of Financial Economics*, 145, 64–82.
- [12] Liu, Q., Tao, Z., Tse, Y., & Wang, C. (2022). Stock market prediction with deep learning: The case of China. *Finance Research Letters*, 46, 102209.
- [13] Ma, C., & Yan, S. (2022). Deep learning in the Chinese stock market: The role of technical indicators. *Finance Research Letters*, 49, 103025.
- [14] Murray, S., Xia, Y., & Xiao, H. (2024). Charting by machines. *Journal of Financial Economics*, 153, 103791.
- [15] Song, D., Baek, A. M. C., & Kim, N. (2021). Forecasting stock market indices using padding-based Fourier transform denoising and time series deep learning models. *IEEE Access*, 9, 83786–83796.

- [16] Wang, P., Li, X., Qin, Z., Qu, Y., & Zhang, Z. (2022). Stock price forecasting based on wavelet filtering and ensemble machine learning model. *Mathematical Problems in Engineering*, 2022: 4024953.
- [17] Jiang, Z., Huang, Y., Li, T., & Cai, F. (2022). Fusion of time-frequency-based convolutional neural network in financial time series forecasting. *Journal of Zhengzhou University (Natural Science Edition)*, 54: 81-88.
- [18] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. *Proceedings of Machine Learning Research*, 162: 27268-27286.
- [19] Grilli, L., & Santoro, D. (2022). Forecasting financial time series with Boltzmann entropy through neural networks. *Computational Management Science*, 19: 665-681.
- [20] Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37: 788-832.
- [21] Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, 602, 259–268.
- [22] Deng, S., Xiao, C., Zhu, Y., Tian, Y., Liu, Z., & Yang, T. (2022). Dynamic forecasting of the Shanghai Stock Exchange index movement using multiple types of investor sentiment. *Applied Soft Computing*, 125, 109132.
- [23] Rashidpoor Toocheai, M., & Moeini, F. (2023). Evaluating the performance of ensemble classifiers in stock returns prediction using effective features. *Expert Systems with Applications*, 213, 119186.
- [24] Januschowski, T., Wang, Y., Torkkola, K., Erkkila, T., Hasson, H., & Gasthaus, J. (2022). Forecasting with trees. *International Journal of Forecasting*, 38, 1473–1481.
- [25] Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293, 116983.