Finance Research Letters 62 (2024) 105195

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Stock price crash prediction based on multimodal data machine learning models

Yankai Sheng [a], Yuanyu Qu [a], Ding Ma [b,*]

[a] School of Banking & Finance, University of International Business and Economics, Beijing 100029, China
[b] School of Economics, Wuhan University of Technology, Wuhan 430070, China

ARTICLE INFO

Keywords:
Stock price crash
Multimodal data
Machine learning
Graph data
LightGBM

ABSTRACT

This study introduces multimodal data machine learning framework to predict stock crashes. It encapsulates market data, graph data cultivated from industry affiliations through node2vec, and text data derived from sentiment analysis. The LightGBM is utilized, marking an improvement by 7.13% over preceding studies, achieving 75.85% balanced accuracy. An innovative long-short portfolio construction approach is articulated, demonstrating the practical significance of the predictions, with a 4.75% portfolio return in 2022 — a 27.26% advancement over the CSI 300. This endeavour in leveraging multimodal data machine learning for stock crash prediction offers a promising performance, serving as a valuable reference for investors.

1. Introduction

Abrupt index crashes or individual stock collapses may interrupt the stable operation of capital markets. The deleterious repercussions of such instances are vividly illustrated by the mid-2015 stock market crash in China and circuit breakers in the U.S. market in 2020. These extreme phenomena exert multifaceted negative impact, manifesting in diminished resource allocation efficiency (An et al., 2015) and eroded investor confidence (Farmer, 2012). The accurate prediction of stock market crashes can facilitate mitigating the risk of investors, improving regulatory efficiency, and promoting the development of capital markets.

The existing literature offers a spectrum of interpretations to elucidate stock price crashes. From a corporate finance perspective, information asymmetry resulting from the principal-agent problem can lead to the accumulation and sudden release of negative news (Hutton et al., 2009; Kim et al., 2011), triggering stock price crashes. From a capital market viewpoint, the temporal characteristics of stock price are highly correlated with crash risk (Boyer et al., 2010; Chen et al., 2001; Jang and Kang, 2019), leading to a consensus on the value of incorporating timely and transparent market data. Meanwhile, behavioural finance experts have highlighted the impact of investor behavior and sentiment on market stability, positing textual data mining as a potential predictive tool (Brown et al., 2014; Yao et al., 2021). However, risk contagion and stock price synchronicity have not been taken account of in the current stock crash prediction models despite the abundance of stylized facts and empirical verifications (Chan and Chan, 2014; Moskowitz and Grinblatt, 1999; Piotroski and Roulstone, 2004). Given the inherently heterogeneous and multidimensional nature of the information needed to predict stock price crashes, it becomes evident that a more comprehensive approach is necessary. Traditional single-modal (Chatzis et al., 2018) or double-modal prediction models (Kaya et al., 2023), while offering valuable insights, may not fully capture the complex dynamics at play.

* Corresponding author.
E-mail address: mading@whut.edu.cn (D. Ma).

https://doi.org/10.1016/j.frl.2024.105195
Received 16 November 2023; Received in revised form 17 February 2024; Accepted 29 February 2024
Available online 1 March 2024

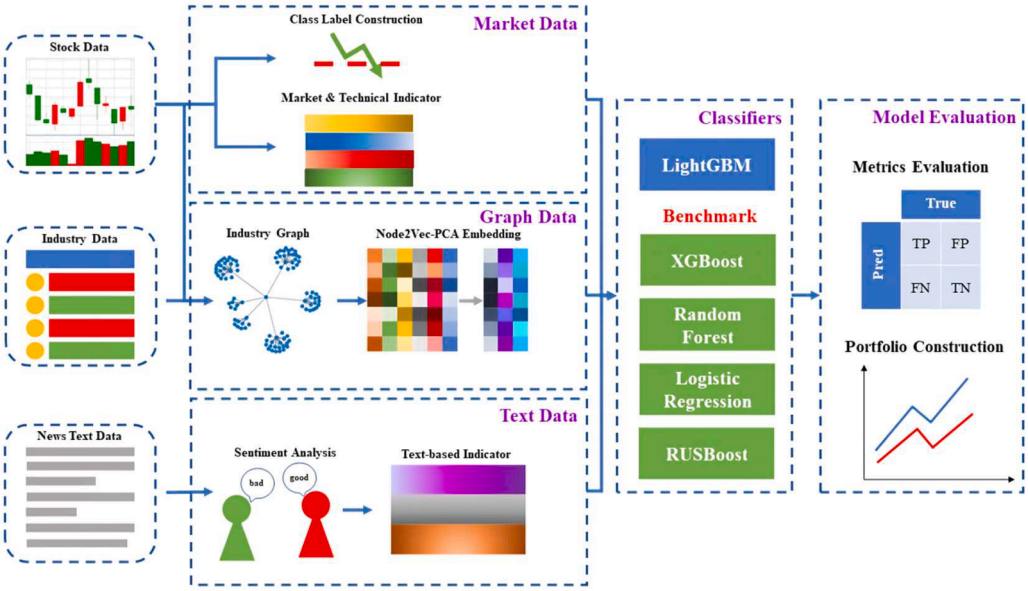


Fig. 1. Proposed framework.

On the basis of mature market and text data analysis, adding graph data to generate multimodal data may be a solution to depict risk contagion and improve prediction accuracy. Characterized as a combination of disparate data types that provides a holistic view, multimodal data has proven its applicability in diverse disciplines such as multisensory systems, medical diagnostics, and meteorological monitoring (Lahat et al., 2015). Despite its promising applications, its adoption in financial analysis is still nascent. Pioneering efforts include Cheng et al. (2022) and Zhou et al. (2023), with the former constructed a heterogeneous graph based on market data, events linkage and industrial relations to forecast stock price, while the latter integrated trading data, technical analysis, and sentiment scores for price trend prediction.

Our study ventures further into the untapped potential of multimodal data, including a novel graphical modality to enhance stock crash predictions by capturing risk contagion and stock synchronicity alongside traditional market and textual data. Utilizing the Light Gradient Boosting Machine (LightGBM), an advanced and interpretable classifier, we attain a notably improved classification performance. This refined approach to data mining may assist regulators in risk management and potentially aid investors in mitigating losses. Furthermore, we crafted an approach to verify the practical usefulness of the stock crash prediction model. Aside from accuracy-centric metrics (Chatzis et al., 2018), we propose a portfolio return indicator, generated from long-short portfolios based on the probabilities of price crashes predicted, thereby testing whether the stock crash prediction model can guide investment and acquire considerable return.

The potential contribution of this study is threefold. Firstly, our framework introduced an innovative modality into existing research. An intra-industry graph can comprehensively reflect risk contagion and stock price synchronicity. Secondly, our multimodal framework for stock price crash prediction achieved 75.85 % balanced accuracy, outperforming existing models by 7.13%. Thirdly, we introduced a portfolio return approach for performance assessment, which verify the practical application of the model. The LightGBM model can achieve an over 4.7% portfolio return, 27.26% higher than the CSI 300 index during the same period. This enhancement in accuracy and notable portfolio return could potentially offer benefits to both regulators and investors.

The remainder of this paper is organized as follows: Section 2 presents the technical route of our model. Section 3 shows the empirical results and further discussion. The conclusion and further research directions are stated in Section 4.

2. Proposed framework

The technical route of the multimodal machine learning framework is shown in Fig. 1, including three main steps, namely multimodal data construction, classifier specification, and model evaluation.

2.1. Data

The dataset includes all the composite stocks of Shen Wan secondary industries, which amount to a total number of 2928. Market data and news data were both obtained from the China Stock Market & Accounting Research (CSMAR) database. Index market data and industry affiliation information were obtained from Tushare (https://www.tushare.pro/). Since the classification standards for ShenWan industries were proposed in 2014, we believe that two years of operation were necessary. Thus, our timeframe was set from 2016 to 2022. The samples of the first 6 years were chosen as the training set, and the samples from the remaining years were chosen as

Table 1
Description and indicators of different modalities.

No.	Name	Description	Indicators
I	Market Data	Descriptive and technical indicators to capture the characteristics of the price fluctuation (Ma and Yan, 2022)	• Descriptive indicators, including moving averages, standard deviations and weekly maximums of returns and turnovers. • Technical indicators including MACD, MOM, OBV, RSI, ROC, CMO, PPO.
II	Graph Data	Industry graph $G^{industry}$ is built as follows: $G^{industry} = (V^{ind}, E^{ind}, W^{ind})$ where V^{ind} are network nodes represented by listed companies. E^{ind} denotes co-belongs of companies in the same ShenWan industries. W^{ind} stands for the weight assigned for each edge by the correlation coefficient of the weekly return of the past 20 weeks inter-companies.	• Node2vec (Grover and Leskovec, 2016) graph embedding technique, transforming complicated graphs into high-dimensional tabular data, with 128 dimensions. • PCA (Principal Component Analysis) for data dimensionality reduction (Zhou et al., 2023), with 32 principal components and over 85 % cumulative contribution degree.
III	Textual Data	Acquire the whole news article dataset. Filter news whose title contains the name of the listed company. Merge the formal and informal text lists proposed by Yao et al. (2021), eliminating duplicate words, and create the financial sentiment dictionary.	• Positivity and disagreement indexes are calculated by $index_{ind} = \ln\left(\frac{1+pos}{1+neg}\right)$ $disagreement_{ind} = 1 - \left\|\frac{pos-neg}{pos+neg}\right\|$ where *pos* and *neg* are the word frequencies of positive and negative words from a news article respectively. • Conduct a weekly average of these two indicators for an individual stock. • *Media_Atten*, denoted by the number of news articles related to each stock in the sample, which measures the prevalence of a stock.

the test set. After the elimination of missing values, we acquired 774,953 samples in total, with 642,755 training samples and 132,198 test samples.

2.2. Multimodal data construction

We summarize the description and indicators of different modalities in Table 1.

2.3. Label definition and classifiers

The widely accepted measure for stock price crashes was proposed by Hutton et al. (2009) built on the assumption of normal distribution of returns, the stock is deemed as crash if its return falls into the [0, 0.1%) confidence interval of the past returns. Following their approach, for an individual stock in week *t*, we denoted the average weekly return of the past 20 weeks as MA_Return_t and its standard deviation as STD_Return_t. For a [0, 0.1%) confidence interval, $MA\_Return_t - 3.09\ STD\_Return_t$ is the boundary of classification, as shown in Eq. (1).

$$Label_t = \begin{cases} 0, & Return_t \geq MA\_Return_t - 3.09\ STD\_Return_t \\ 1, & Return_t < MA\_Return_t - 3.09\ STD\_Return_t \end{cases} \tag{1}$$

However, such classification leads to an extremely imbalanced dataset, which requires model training with caution. Therefore, we chose the following classifiers to compare their performance. We compared among the following classifiers, including a basic classification model LR (Logistic Regression), an ensemble learning model RF (Random Forest), an imbalanced data machine learning model RUB (Random Under-sampling Boosting, Seiffert et al. (2010)), and two gradient boosting decision tree models, namely XGB (eXtreme Gradient Boosting) and LGB (Light Gradient Boosting Machine). Considering its superior efficiency (Ke et al., 2017) and performance (Sun et al., 2020), we employed LGB as the main classifier, which adopt several state-of-the-art mechanisms including histogram-based decision trees, leaf-wise algorithm, gradient-based one-side sampling and exclusive feature bundling. A potential concern is that the neural network models should be introduced into the analysis due to their promising performance. The reason is that we place emphasis on the interpretability of the model, as our subsequent discussions and analyses are contingent upon this aspect. However, the interpretability of neural networks remains a challenging and actively researched issue within the academic community (Zhang et al., 2021). In short, we used 5 classifiers and a dataset with 48 features.

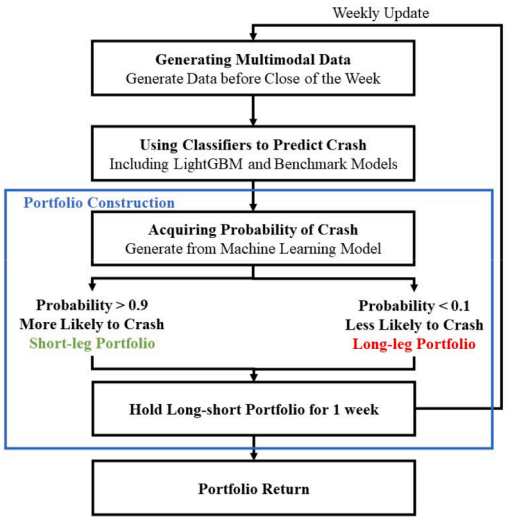
2.4. Evaluation metrics

2.4.1. Metrics for traditional imbalanced data

Referring to the evaluation metrics of unbalanced data defined by Johnson and Khoshgoftaar (2019), we chose *Sensitivity*, *Specificity*, and *Balanced Accuracy* (*BA*) as the evaluation metrics, which can be defined by Eqs. (2)–(4).

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{3}$$



Basi Element

Stock Pool	Shen Wan Secondary Industry Composite Stocks
Benchmark	CSI 300 Index
Sample Period	Whole Time Period of 2022
Update Frequency	Weekly Update
Weighting Strategy	Equal Weighting

Basic Assumption

- The algorithm does not take time to run.
- Transaction cost is negligible.
- Short selling is unlimited.

Fig. 2. Construction of long-short portfolio.

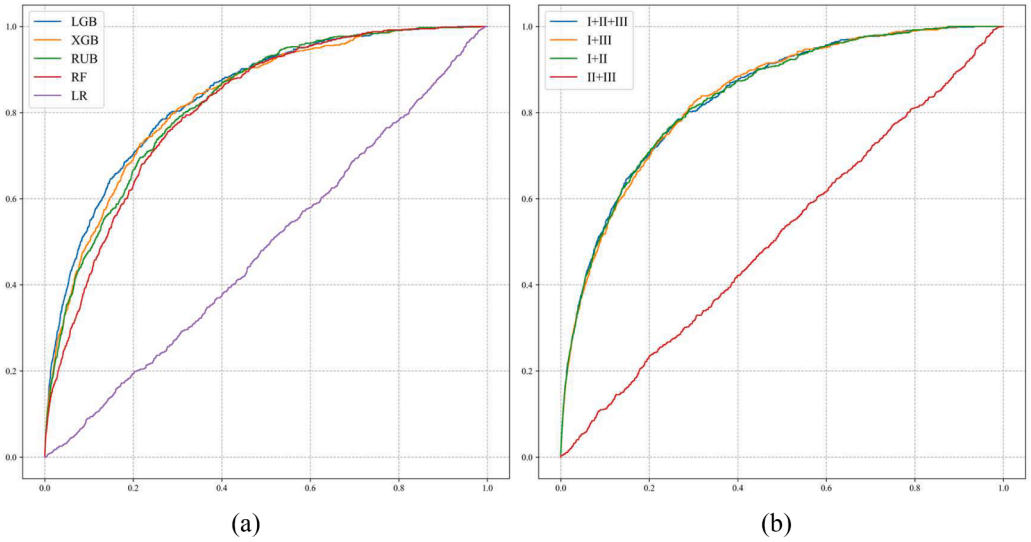


Fig. 3. ROC curves across classifiers and modality combinations.

$$BA = (Sensitivity + Specificity)/2 \tag{4}$$

In addition, the Receiver Operating Characteristic (ROC) curve was utilized to visualize the overall performance.

2.4.2. Long-short portfolio return

Furthermore, we proposed a long-short portfolio construction approach to test model usefulness in a practice. The main process is shown in Fig. 2.We opted for fixed thresholds of 10% and 90% to delineate between crash and non-crash samples, diverging from the classical practice in portfolio construction literature which typically utilizes ranking based on certain indicator (Fama and French, 2015).

3. Empirical results

3.1. Performance comparison

After generating data and training classifiers, we evaluated all models' performance, detailed in this section. ROC curves for each model are displayed in Fig. 3(a), revealing LR's underperformance. RF slightly lags behind RUB, while LGB and XGB show comparable

Table 2
Model performance comparison (%).

Algorithm	*Sensitivity*	*Specificity*	*BA*	*Portfolio Return*
Panel A: Performance Comparison across Models				
LGB	77.02	74.68	**75.85**	**4.7491**
XGB	79.30	71.82	75.56	1.1746
RUB	62.90	**81.37**	72.14	0.0000
RF	67.88	78.17	73.01	0.0000
LR	**87.63**	11.29	49.46	0.0000
Panel B: Performance Comparison across Modality Combinations				
I+II	77.15	**74.74**	75.95	1.7706
I+III	**83.06**	69.11	**76.09**	-3.8552
II+III	66.80	34.49	50.65	0.0000
I+II+III	77.02	74.68	75.85	**4.7491**

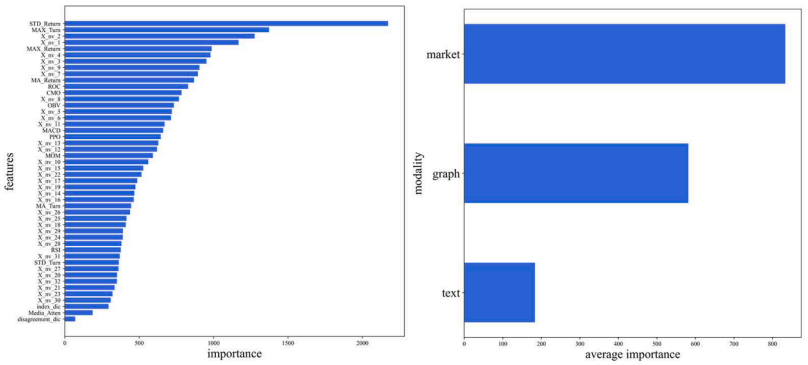


Fig. 4. Feature importance plot.

effectiveness.

Evaluation metrics for all algorithms mentioned in Section 2.4 are shown in Panel A of Table 2. It is evident that although LR achieves the highest *Sensitivity*, it exhibits the minimum *Specificity*, resulting in the lowest *BA*. On the other hand, LGB may not excel in terms of *Sensitivity* and *Specificity* individually, but it attains the highest *BA* and significantly outperforms other algorithms in terms of *Portfolio Return*. XGB has a similar *BA*, but its *Portfolio Return* is relatively low. It should be noted that in our portfolio construction approach, RF (Random Forest), RUB (Rubric), and LR (Logistic Regression) are unable to effectively filter any stocks. This leads to a *Portfolio Return* of zero, which is obtained following the calculation steps in Fig. 2. We analysed the crash probabilities generated by each model, and it is apparent that a majority of the probabilities fall within the (49, 51) range for RUB and LR and (40, 60) for RF, indicating their inability to clearly distinguish between crash and non-crash samples.

3.2. Further discussion

3.2.1. Feature and modality importance

To solve the interpretability problem in machine learning, we exported the feature importance of LightGBM, which is generated from numbers of times the feature is used in a model, as shown in Fig. 4(a). The importance of different feature types is computed by the average importance of features in the type, as shown in Fig. 4(b). It is obvious that market data, especially descriptive indicators, play a crucial role. Graph data also exhibit high importance. The features extracted from textual data seem less significant, as their importance ranks in the bottom three. A possible explanation lies in the sparsity of news data compared with market and graph data.

From a macro perspective, we hope to discover whether different modality combinations can lead to different performance. We first displayed the ROC curves of different modality combinations as shown in Fig. 3(b), where I, II, and III denote market, graph, and textual data respectively. Intuitively, the combination of II and III leads to poor model performance, further demonstrating the significance of market data. However, the best combination cannot be distinguished. Hence, we displayed a further comparison in Panel B of Table 2 where the highest *BA* is achieved by modality I+III. However, *BA* of modality I+II+III is not significantly lower than I+III and the ROC curves cannot distinguish which one is more accurate. Thus, the notably high Portfolio Return of modality I+II+III is more crucial. Additionally, modality I+II acquired higher portfolio return (1.7706%) than modality I+III (-3.8552%) and model trained with all three modalities can acquire the highest portfolio return (4.7491%). Such evidence can further prove that graph data is crucial.

Table 3
Monthly balanced accuracy and portfolio return of different models (%).

Month	*BA*		*Portfolio Return*		*Volatility*		CSI 300
	LGB	XGB	LGB	XGB	LGB	XGB	
Jan	**76.08**	75.64	**-11.85**	-12.81	3.99	**3.15**	-7.93
Feb	74.72	**82.39**	4.03	**9.08**	**2.51**	2.60	0.21
Mar	**77.53**	76.03	-2.60	**-2.42**	**1.62**	2.01	-9.13
Apr	**73.80**	73.28	**-0.35**	-1.75	**2.77**	3.11	-3.87
May	**80.46**	69.92	**-1.29**	-1.67	3.56	**2.74**	-0.37
Jun	**64.31**	62.76	**6.46**	-1.36	**3.37**	4.40	9.38
Jul	60.42	**63.77**	-0.63	**0.01**	**1.42**	2.02	-5.25
Aug	**82.37**	79.20	**3.76**	3.40	**0.58**	0.68	-1.51
Sep	76.10	**76.56**	5.53	**10.24**	**1.11**	1.34	-7.65
Oct	**66.94**	66.32	5.41	**7.34**	**1.84**	2.50	-7.18
Nov	**63.38**	61.52	**-1.57**	-6.74	**3.20**	4.68	6.41
Dec	**77.93**	73.50	-2.26	**-2.14**	1.78	**1.10**	1.38
Average	**72.84**	71.74	**0.39**	0.10	**2.31**	2.53	-2.13

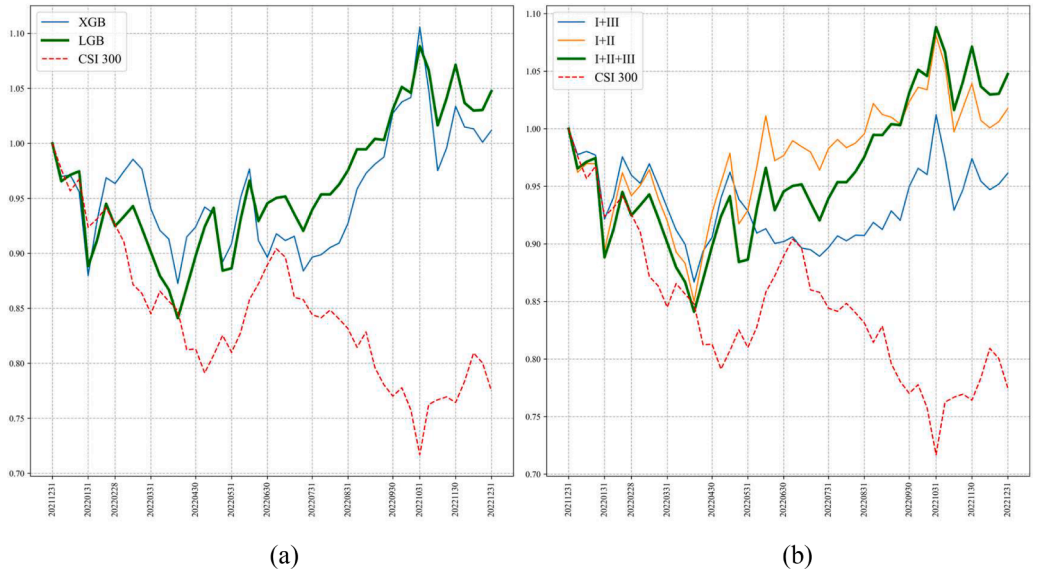


Fig. 5. Trend of portfolio net value across models and modality combinations.

Table 4
Monthly balanced accuracy and portfolio return of different modality combinations (%).

Month	*BA*			*Portfolio Return*			*Volatility*			CSI 300
	I+II+III	I+II	I+III	I+II+III	I+II	I+III	I+II+III	I+II	I+III	
Jan	76.08	**77.44**	75.86	-11.85	-11.2	**-8.18**	3.99	3.53	**2.40**	-7.93
Feb	74.72	**74.80**	75.40	4.03	**5.21**	4.07	2.51	2.72	**2.23**	0.21
Mar	**77.53**	76.12	73.69	-2.60	**-2.34**	-3.01	1.62	1.79	**1.55**	-9.13
Apr	73.80	**73.85**	73.29	-0.35	**0.76**	-2.80	2.77	3.53	**2.43**	-3.87
May	80.46	**80.47**	68.03	-1.29	0.23	**2.52**	3.56	3.81	**2.50**	-0.37
Jun	64.31	**64.61**	61.27	**6.46**	5.00	-2.91	3.37	3.36	**1.06**	9.38
Jul	60.42	60.53	**61.93**	-0.63	**0.62**	-0.57	1.42	1.29	**0.70**	-5.25
Aug	**82.37**	82.31	80.26	**3.76**	1.28	1.16	**0.58**	0.62	0.60	-1.51
Sep	76.10	**77.21**	76.80	**5.53**	2.76	4.56	**1.11**	1.43	1.52	-7.65
Oct	66.94	65.60	**72.58**	5.41	5.47	**6.38**	**1.84**	1.95	2.41	-7.18
Nov	**63.38**	63.13	62.05	**-1.57**	-3.92	-3.83	**3.20**	3.25	3.36	6.41
Dec	77.93	75.65	**79.32**	-2.26	-2.11	**-1.32**	1.78	1.64	**1.18**	1.38
Average	**72.84**	72.64	71.71	**0.39**	0.15	-0.33	2.31	2.41	**1.83**	-2.13

3.2.2. Explanation for superiority of the multimodal framework

Although LGB achieves the highest *BA*, it does not substantially surpass that of XGB. To interpret why the *Portfolio Return* of LGB is almost four times that of XGB, we dismantled the *BA* and *Portfolio Return* and calculated the *Volatility* of return month by month and reported the results in Table 3. Overall, the investment portfolio generated by LGB is more stable, and compared with that of XGB, even if the LGB investment portfolio falls, the decline is not so large. In other words, its stability of profitability is relatively high. The *Portfolio Return* of LGB and XGB are respectively 27.26% and 23.69% higher than CSI 300 index return in 2022, which validate the profitability of the multimodal framework.

Fig. 5(a) depicts the weekly net value trends of various model portfolios alongside the CSI 300 trend. During the initial four months of sharp market decline, XGB exhibited remarkable stability. In contrast, during the May-June market rebound, LGB outperformed XGB. Throughout the second half, despite market downturns, our portfolios remained profitable. LGB maintained steady performance, whereas XGB experienced significant fluctuations, resulting in lower returns.

The model incorporating all modalities did not achieve the highest *BA*. This contrasts with its superior *Portfolio Return*, as detailed in Table 4. The three-modal training approach seems to offer a more robust portfolio. Fig. 5(b) illustrates the weekly net value trends of portfolios from different modality combinations. Initially, no significant difference was evident among these combinations. Post-April, the I+III modality combination showed the weakest performance, possibly due to its mediocre *BA*. In May and June, the I+II combination yielded the highest returns, but lacked the consistency observed in the three-modality approach. During the whole time period, the I+II+III displayed a more stable profitability and finally achieved the highest return. In general, by adding a graphical modality, the model can be more profitable.

4. Conclusion

We innovatively proposed a multimodal data framework for stock price crash prediction, with the introduction of a stock graphical modality along with the adoption of market and text modality. From the perspective of industry correlation, this analyzing approach leveraging complex network to uncover the impact of stock risk contagion and price synchronicity in forecasting stock crashes. This impact is further corroborated by enhanced portfolio returns observed in the training outcomes incorporating graph data. Among the various machine learning classifiers evaluated, LightGBM exhibits outstanding performance over benchmark models, with over 75% balanced accuracy and 4.7% yearly portfolio return. Notably, our research represents a pioneering effort in applying multimodal data to stock market crash prediction. It outperforms single-modal data in terms of accuracy and generates outstanding and stable returns compared to double-modal data. Moreover, through in-depth analysis of monthly balanced accuracy and portfolio returns, we provided further insights into the key factors contributing to the superior performance of the multimodal model. LightGBM model trained with all three modalities achieves superior performance due its relatively high continuity of profitability.

Our work improves stock crash prediction accuracy and expands understanding of multimodal data's utility in financial predictive modeling. This has significant implications for investment risk management and decision-making. Moreover, it opens further research avenues, such as incorporating additional data sources and employing advanced machine learning or deep learning techniques for enhanced predictions. Future research could explore these areas to strengthen the multimodal data machine learning framework's robustness and applicability.

CRediT authorship contribution statement

Yankai Sheng: Writing – original draft, Software, Methodology. **Yuanyu Qu:** Writing – review & editing, Supervision. **Ding Ma:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data can be acquired from Tushare at https://www.tushare.pro/ and CSMAR at https://data.csmar.com/.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No's. 72004173 & 72373020).

References

An, Z., Li, D., Yu, J., 2015. Firm crash risk, information environment, and speed of leverage adjustment. J. Corp. Finance 31, 132–151. https://doi.org/10.1016/j.jcorpfin.2015.01.015.

Boyer, B., Mitton, T., Vorkink, K., 2010. Expected idiosyncratic skewness. Rev. Financ. Stud 23, 169–202. https://doi.org/10.1093/rfs/hhp041.

Brown, N.C., Wei, K.D., Wermers, R., 2014. Analyst recommendations, mutual fund herding, and overreaction in stock prices. Manag. Sci 60, 1–20. https://doi.org/10.1287/mnsc.2013.1751.
Chan, K., Chan, Y.C., 2014. Price informativeness and stock return synchronicity: evidence from the pricing of seasoned equity offerings. J. Financ. Econ. 114, 36–53. https://doi.org/10.1016/j.jfineco.2014.07.002.
Chatzis, S.P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., Vlachogiannakis, N., 2018. Forecasting stock market crisis events using deep and statistical machine learning techniques. Expert Syst. Appl. 112, 353–371. https://doi.org/10.1016/j.eswa.2018.06.032.
Chen, J., Hong, H., Stein, J.C., 2001. Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices. J. Financ. Econ. 61, 345–381. https://doi.org/10.1016/S0304-405X(01)00066-6.
Cheng, D., Yang, F., Xiang, S., Liu, J., 2022. Financial time series forecasting with multi-modality graph neural network. Pattern Recognit. 121, 108218 https://doi.org/10.1016/j.patcog.2021.108218.
Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116, 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010.
Farmer, R.E., 2012. The stock market crash of 2008 caused the Great Recession: theory and evidence. J. Econ. Dyn. Control 36, 693–707. https://doi.org/10.1016/j.jedc.2012.02.003.
Grover, A., Leskovec, J., 2016. node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864. https://doi.org/10.1145/2939672.2939754.
Hutton, A.P., Marcus, A.J., Tehranian, H., 2009. Opaque financial reports, R2, and crash risk. J. Financ. Econ. 94, 67–86. https://doi.org/10.1016/j.jfineco.2008.10.003.
Jang, J., Kang, J., 2019. Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. J. Financ. Econ. 132, 222–247. https://doi.org/10.1016/j.jfineco.2018.10.005.
Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 1–54. https://doi.org/10.1186/s40537-019-0192-5.
Kaya, D., Reichmann, D., and Reichmann, M., 2023. Out-of-Sample predictability of firm-specific stock price crashes: a machine learning approach. Available at SSRN 4043938. 10.2139/ssrn.4043938.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. In: Proceedings of the Advances Neural Information Processing Systems, 30. Available at. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
Kim, J.B., Li, Y., Zhang, L., 2011. CFOs versus CEOs: equity incentives and crashes. J. Financ. Econ. 101, 713–730. https://doi.org/10.1016/j.jfineco.2011.03.013.
Lahat, D., Adali, T., Jutten, C., 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. Proc. IEEE 103, 1449–1477.
Ma, C., Yan, S., 2022. Deep learning in the Chinese stock market: the role of technical indicators. Finance Res. Lett. 49, 103025 https://doi.org/10.1016/j.frl.2022.103025.
Moskowitz, T.J., Grinblatt, M., 1999. Do industries explain momentum? J. Finance 54, 1249–1290. https://doi.org/10.1111/0022-1082.00146.
Piotroski, J.D., Roulstone, D.T., 2004. The influence of analysts, institutional investors, and insiders on the incorporation of market, industry, and firm?specific information into stock prices. Account. Rev. 79, 1119–1151. https://doi.org/10.2308/accr.2004.79.4.1119.
Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A., 2010. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 40, 185–197. https://doi.org/10.1109/TSMCA.2009.2029559.
Sun, X., Liu, M., Sima, Z., 2020. A novel cryptocurrency price trend forecasting model based on LightGBM. Finance Res. Lett. 32, 101084 https://doi.org/10.1016/j.frl.2018.12.032.
Yao, J., Feng, X., Wang, Z., Ji, R., Zhang, W., 2021. Tone, sentiment and market impacts: the construction of Chinese sentiment dictionary in finance. J. Manag. Sci. China 24, 26–46. https://doi.org/10.19920/j.cnki.jmsc.2021.05.002.
Zhang, Y., Tino, P., Leonardis, A., Tang, K., 2021. A survey on neural network interpretability. IEEE Trans. Emerg. Top. Comput. Intell. 5, 726–742. https://doi.org/10.1109/TETCI.2021.3100641.
Zhou, F., Zhang, Q., Zhu, Y., Li, T., 2023. T2V_TF: an adaptive timing encoding mechanism based Transformer with multi-source heterogeneous information fusion for portfolio management: a case of the Chinese A50 stocks. Expert Syst. Appl 213, 119020. https://doi.org/10.1016/j.eswa.2022.119020.