Journal of Sustainable Development; Vol. 18, No. 2; 2025
ISSN 1913-9063 E-ISSN 1913-9071
Published by Canadian Center of Science and Education

Revealing the Driving Factors of the Chinese Baijiu Stock Market Based on Machine Learning

Ruiguang Yao[1]

[1] College of Physical and Electronics Engineering, Sichuan Normal University, Chengdu, China

Correspondence: Ruiguang Yao, College of Physical and Electronics Engineering, Sichuan Normal University, Chengdu, China. E-mail: 857295776@qq.com

Received: January 31, 2025 Accepted: February 27, 2025 Online Published: March 9, 2025

doi:10.5539/jsd.v18n2p29 URL: https://doi.org/10.5539/jsd.v18n2p29

Abstract

Stock market volatility significantly impacts investors, policymakers, and industry development. While previous studies have identified key influencing factors, they have largely overlooked the unique volatility of the Chinese Baijiu stock market. This study adopts a complex network perspective, integrating transfer entropy, Peter and Clark momentary conditional independence (PCMCI), and interpretable machine learning methods to reveal the key drivers and mechanisms behind the market's volatility. The research identifies 10 critical factors spanning four dimensions: resources and agriculture, industry and manufacturing, services and consumption, and cross-domain sustainable development. Our findings indicate that the volatility of the Chinese Baijiu stock market is driven by a combination of agricultural, industrial, and sustainable development factors, highlighting the importance of industrial synergies. In addition to confirming the significance of traditional factors, this study also reveals the direct positive causal relationships of emerging industry variables, such as construction decoration, environmental protection, and energy, with the Chinese Baijiu stock market. These findings not only enhance the understanding of the dynamics of the Chinese Baijiu stock market but also provide a transferable research framework and methodology for other industries.

Keywords: key factors, mixed causal model, stock market volatility, Chinese Baijiu stock

1. Introduction

The rapid growth of China's capital markets has positioned Chinese Baijiu stocks as key contributors to the A-share market. Accurately identifying the key drivers of their volatility is crucial for both financial investors and policymakers(Bin, 2024). The volatility observed in the Chinese Baijiu stock market is driven by multiple determinants, including fluctuations in the prices of key raw materials, industry-specific dynamics, and rapid shifts in investor sentiment, each contributing to the market's overall instability. These factors intertwine to create a complex and dynamic market environment, further amplifying the high dynamism and uncertainty of the Chinese Baijiu stock market. However, research on the volatility characteristics and key drivers of the Chinese Baijiu stock market at the capital market level remains scarce, leaving a gap in understanding this complex system.

In recent years, significant improvements in data availability and rapid advancements in computational technologies have greatly facilitated the widespread application of artificial intelligence across various fields (Shi et al., 2021; Zeng et al., 2024). Notably, artificial intelligence has shown immense potential in the financial sector, offering novel perspectives for stock market analysis and helping uncover the dynamics of complex markets (Jiang et al., 2024; Ren et al., 2024; Szczygielski et al., 2024; Weng et al., 2022). Existing research primarily reveals the volatility and underlying mechanisms of the stock market through two main approaches. One approach involves traditional economic models, such as the Panel Autoregressive Distributed Lag (ARDL) model (Xiong et al., 2024), Time-varying Parameter Vector Autoregression(Tita et al., 2025), econometric frameworks like the VARMA-DCC-GARCH-in-mean model(Wang et al., 2025), Autoregressive models(Zheng et al., 2024), data modelling techniques(Demirer and Yuksel, 2024), and the Conditional Threshold Autoregressive model(Ardakani, 2024). The other approach utilizes machine learning methods, including Multi-Layer Perceptron (MLP) (Abolmakarem et al., 2022), XGBoost(Caparrini et al., 2024; Li et al., 2022), Support Vector Machines (SVM)(Yu et al., 2014), Random Forest (RF)(Meher et al., 2024), LASSO(Ellington et al., 2022), LightGBM(Li et al., 2022), and several hybrid models(Cheng et al., 2024).

Traditional economic models often fail to capture the nonlinear and intricate relationships inherent in stock market

behavior. In contrast, machine learning techniques, through the construction of tailored scenarios, can effectively uncover these nonlinear dynamics and account for the complex, multidimensional factors at play. Although previous studies have made significant advancements in stock market prediction, selecting relevant features continues to play a pivotal role in optimizing the accuracy of machine learning models. Transfer entropy, a method for quantifying mutual information in systems with temporal delays, has demonstrated its robustness and utility in constructing information flow networks(Gao et al., 2023). This approach effectively captures the dynamic relationship between characteristic factors and the stock market (Peng et al., 2022). Traditional multivariate time series analysis methods are constrained by strict assumptions, making them inadequate for capturing complex nonlinear causal relationships. PCMCI, a flexible causal discovery method based on graphical causal models, has been successfully applied to climate systems (Menegozzo et al., 2020). By employing conditional independence tests, it extracts complex time-lagged causal relationships even in the presence of autocorrelation. These characteristics make PCMCI a powerful tool for exploring nonlinear causal relationships. Although most machine learning models achieve high predictive accuracy in the financial stock domain, they often lack the ability to explain the multidimensional relationships of features in industry-specific markets (Nie et al., 2021). Interpretable machine learning techniques have been shown to be a powerful tool for identifying the key drivers of volatility in the Russian stock market (Teplova et al., 2023).

Despite significant progress in identifying the influencing factors, the key driving forces behind the volatility of the Chinese Baijiu stock market still face several challenges. Firstly, the decentralization of cross-industry demand, coupled with the influence of raw material suppliers such as wheat and sorghum, gives rise to a multifaceted network that spans the entire market. Consequently, uncovering the underlying interactions among the diverse factors within this network is essential. However, the lack of model interpretability and the limitations in automatically inferring complex nonlinear indirect relationships hinder our in-depth understanding of how and why these key drivers affect the market. These challenges have prompted us to explore the dynamics across industries further and uncover the complex driving mechanisms of the Chinese Baijiu stock market. To bridge this knowledge gap, a potential solution is proposed: revealing the key drivers of the Chinese Baijiu stock market. First, by providing multidimensional feature data, we can gain a comprehensive understanding of the overall trends. Second, a hybrid model combining transfer entropy, PCMCI, and interpretable machine learning is introduced, which can more effectively identify the key drivers of the market. This study not only fills the academic gap in research on the Chinese Baijiu stock market but also provides a valuable paradigm for other financial industries. The findings offer crucial support for policymakers in optimizing industry policies, stabilizing market fluctuations, and assisting investors in formulating effective strategies.

Figure 1 is the overall research framework of the paper. The structure of the paper is as follows: Section 2 outlines the experimental methodologies, Section 3 details the data collection and preprocessing procedures, Section 4 discusses the experimental findings and their underlying factors, and Section 5 concludes with a summary and suggestions for future research.

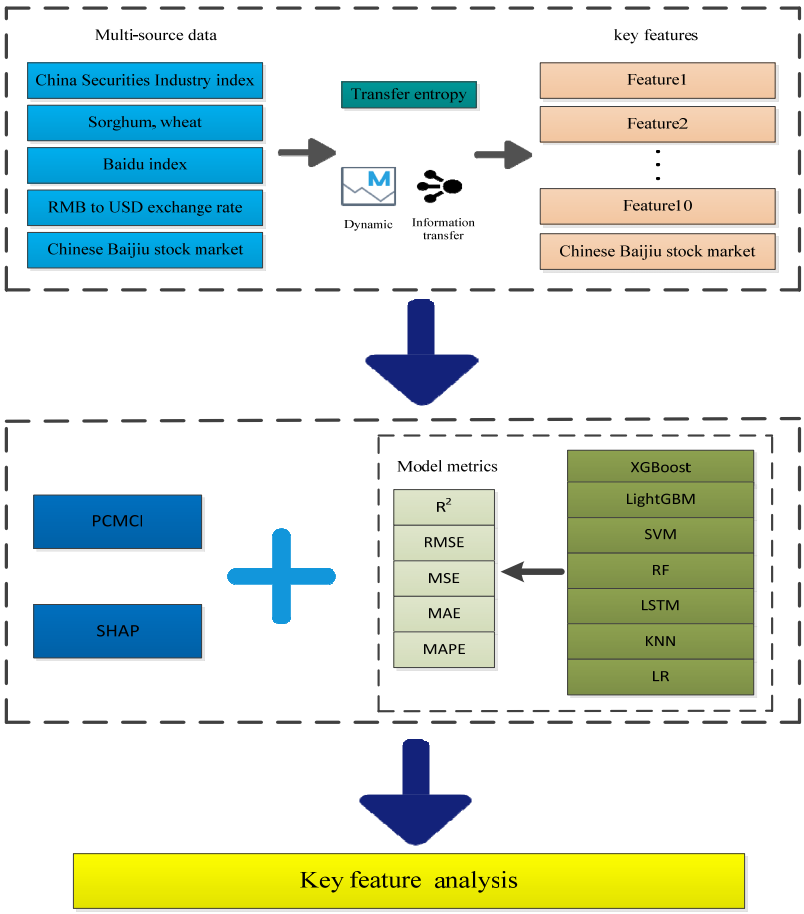

Figure 1. The research framework

2. Method

2.1 Transfer Entropy

Transfer entropy has been proven to be an effective scalar measure for capturing the information flow and directional characteristics between time-series data. It not only quantifies correlations between variables but also identifies non-correlation information transfer relationships. Compared to information entropy, which measures the correlation between two random variables, transfer entropy further reveals the directional nature of information propagation. Figure 2 presents a schematic representation of transfer entropy.

Transfer entropy is more practical and nonlinear than Granger causality(Yin et al., 2023). The reciprocities have directivity, $TE_{x\to y}$ or $TE_{y\to x}$ (Sensoy et al., 2014). The transition probabilities can be defined as follows:

$$p(x_{i+1}\left|x_i^{(k)},y_i^{(l)}\right.)=p(X_{i+1}=x_{i+1}\left|X_i^{(k)}=x_i^{(k)},Y_i^{(l)}=y_i^{(l)}\right.) \tag{1}$$

Where time series of X could be treated as a Markov process of degree k . Likewise, Y is j degree Markov process. $X_j^{(k)}=(X_i,X_{i-1},.....,X_{i-k+1})$, $Y_i^{(l)}=(Y_i,Y_{i-1},....,Y_{i-l+1})$, $x_i^{(k)}$ and $y_i^{(l)}$ are the state of $X_i^{(k)}$ and $Y_i^{(l)}$ respectively(Ardakani, 2024). The Transfer entropy from a variable Y to the other variable X is defined as follows:

$$TE_{Y\to X}(k,l)=H(X_{i+1}\mid X_i^{(k)})-H(X_{i+1}|\mathrm{X}_i^{(k)},Y_i^{(l)})=\sum p(x_{i+1},x_i^{(k)},y_i^{(l)})\log\frac{p(x_{i+1},x_i^{(k)},y_i^{(l)})}{p(x_{i+1},x_i^{(k)})} \tag{2}$$

where i_n is element n of the time series of variable X and j_n is element n of the time series of variable Y (Hudec et al., 2021). To facilitate the calculation of Transfer entropy, we take $k=l=1$, thus, the normal formula for Transfer entropy of Y to X is as follows(Qiu and Yang, 2020):

$$NTE_{Y\to X} = \sum_{i_{n+1},i_n,j_n} p(i_{n+1},i_n,j_n)\log\frac{p(i_{n+1},i_n,j_n)p(i_n)}{p(i_{n+1},i_n)p(i_n,j_n)} \tag{3}$$

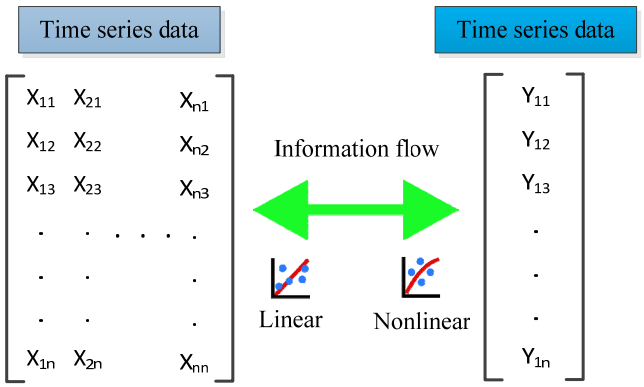

Figure 2. Illustrative diagram of transfer entropy

2.2 Peter and Clark Momentary Conditional Independence (PCMCI)

PCMCI is an innovative technique for causality analysis of time series data(Shen et al., 2024). The model system allows for a clearer identification of key variables associated with the conditions. Unlike Granger causality, PCMCI circumvents the conditioning on irrelevant variables, leading to more significant effects in both higher and lower dimensions(Guo et al., 2021). Figure 3 illustrates a schematic representation of PCMCI.

$$X_t^j = f_j(p(X_t^j),\eta_t^j) \tag{4}$$

Where f_j is some potentially nonlinear functional dependency and η_t^j represents mutually independent dynamical noise. The nodes in a time series graph represent the variables X_t^j at different lag times, and $\dot{\mathrm{P}}(X_t^j)\subset X_t^- = (X_{t-1},X_{t-2},...)$ denote the causal parents of variables X_t^j among the past of all N variables(Runge et al., 2019).

$$\tilde{\mathrm{P}}(X_t^j) = X_t^- = \{X_{t-\tau}^i : i=1,...,N,\tau=1,...,\tau_{\max}\} \tag{5}$$

$$\rho_{X_{t-\tau}^i\to X_t^j}^{MCI} = \frac{c\sigma_{X_{t-\tau}^i}}{\sqrt{\sigma_{X_t^j}^2 + c^2\sigma_{X_{t-\tau}^i}^2}} \tag{6}$$

Where $\sigma_{X_{t-\tau}^i,X_t^j}$ are the variances of the noise terms driving $X_{t-\tau}^i$ and X_t^j respectively, and c is their coupling coefficient(Krich et al., 2020).

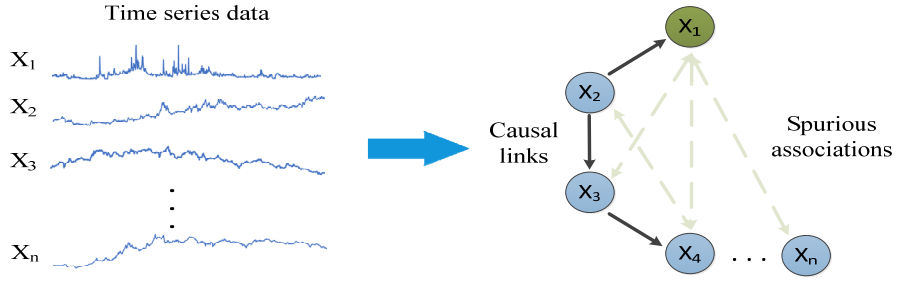

Figure 3. Illustrative diagram of PCMCI

2.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost), introduced by Chen and Guestrin in 2016, extends gradient-boosted decision trees by incorporating parallel processing, making it suitable for both classification and regression tasks(Engbers, 2019). Indeed, it represents an advanced version of the widely adopted Gradient Boosted Decision Tree (GBDT) algorithm, effectively overcoming its computational constraints (Homafar et al., 2022).

$$\Omega(\theta) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \tag{7}$$

Within the regularization term, T denotes the number of leaf nodes, and w represents the weight of each leaf. γ and λ are regularization parameters that govern the penalty associated with T and w (Homafar et al., 2022).

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{8}$$

Where $L(\theta)$ is the loss function and $\Omega(\theta)$ is a regularization function, controlling the model's complexity(Chen and Jia, 2024; Homafar et al., 2022).

2.4 Light Gradient Boosting Machine (LightGBM)

LightGBM is a supervised model. For a given dataset $D = \{(x_i, y_i)\}$, the objective of LightGBM is to identify an approximate function $f(x)$ of the function $f^*(x)$ for minimizing the loss function $L(y, f(x))$ (Zhang et al., 2022).

$$f = \arg\min_{f}[E_{y,x} L(y, f(x))] \tag{9}$$

Leveraging the principles discussed above, LightGBM effectively captures the statistical properties of the samples, enabling precise classification and regression of data. A threshold θ was introduced for LightGBM in the rough classification, which significantly enhanced the TPR of LightGBM, while the cost of FPR was acceptable. p_{Benign} denotes the probability that LightGBM determines event d as benign(Wang et al., 2022). And p_{Malice} illuminates the probability that d is malicious, as determined by LightGBM. The model tends to favor the majority (benign) class in the presence of class imbalance. That is to say, the probability p_{Benign} will, in general, be much greater than p_{Malice}. And θ is gauged by the quantitative relationship between p_{Benign} and p_{Malice}.

$$\theta_{best} = \left\langle \max \left\{ \frac{p_{Malice}}{p_{Benign}} \geq \theta \right\}_{TPR} ; \left\{ \frac{p_{Malice}}{p_{Benign}} \geq 0 \right\}_{FPR} \leq FPR_{\downarrow} \right\rangle \tag{10}$$

The ratio of the probabilities of malicious and benign events, $\mathrm{P}Malice / \mathrm{P}Benign$, θ. When $\mathrm{P}Malice / \mathrm{P}Benign \geq \theta$, the final category is malicious; otherwise, it is benign. Based on TPR and FPR, θ is optimized for LightGBM through the training and validation sets. The optimal $\theta best$ corresponds to the value of θ at which the true positive rate (TPR) is maximized, while the false positive rate (FPR) remains within an acceptable range.

2.5 Support Vector Machine (SVM)

The SVM algorithm was initially employed to construct a linear classifier(Dahal et al., 2021). Its objective is to define a decision boundary that separates two classes, enabling the prediction of the label vector from one or more

feature vectors. This boundary, known as the hyperplane, is positioned to maximize the distance from the nearest data points of each class. The closest data points in each class are called support vectors, given a labeled training data set(Lawal and Kwon, 2021).

$$(x_1, y_1), ..., (x_n, y_n), x_i \in R^d \qquad y_i \in (-1, +1) \tag{11}$$

Among them, x_i represents a feature vector, y_i is the class label of the training compound i. Subsequently, the optimal hyperplane can be defined as.

$$wx^T + b = 0 \tag{12}$$

w denotes the weight vector, x represents the input feature vector, and b signifies the bias(Graf et al., 2005).

Another application of SVM is the kernel method, which is capable of establishing high-dimensional and nonlinear models. In a nonlinear issue, the kernel function can be utilized to add extra dimensions to the original data, thereby transforming it into a linear problem in the resulting high-dimensional space(Zhang et al., 2023). The kernel function facilitates efficient computation, which would otherwise require operations in the high-dimensional space.

$$K(x, y) = < f(x), f(y) > \tag{13}$$

K represents the kernel function, and x, y is an n-dimensional input. f is employed to map the input from the n-dimensional space to the m-dimensional space. $< x, y >$ represents the dot product.

2.6 Random Forests (RF)

Random Forest utilizes CART decision trees as base learners. Unlike traditional decision trees, it selects an optimal feature from a subset of randomly chosen sample features at each node to construct the tree (Makariou et al., 2021). The left and right subtrees are split accordingly, with the best feature from the selected subset being used for tree construction (Gregorutti et al., 2017). This approach enhances the model's generalization capability by reducing overfitting.

The input is taken as the sample set $D = \{(x_1, y_1), (x_2, y_2), ...(x_m, y_m)\}$,and the weak classifier is iterated T times. The output is the result $f(x)$.

For $t = 1, 2..., T$, the training set is randomly sampled t times, and a total of m times are conducted, obtaining a sampling set m containing D_t samples(Lei, 2021). Employing the sample set D_t to train t decision tree models $G_t(x)$. During the training of decision tree nodes, a subset of features is randomly selected from the available features at each node. Among these selected features, the most optimal one is chosen to partition the left and right subtrees of the decision tree.(Jiménez-Carvelo et al., 2019).

2.7 Long Short-Term Memory Network (LSTM)

LSTM is a specialized variant of the recurrent neural network (RNN) architecture, designed to capture both long-term and short-term dependencies(Mutinda and Langat, 2024)，It addresses the gradient vanishing issue in the processing of long sequence data by introducing memory units (Cell State) for ordinary RNNs. LSTM can capture dependencies in long sequences and is a highly suitable deep learning model for tasks such as time series, natural language processing, and speech recognition(Saeed and Yin, n.d.; Wu et al., 2024). A single LSTM unit comprises an input gate, an output gate, a forget gate, and a cell candidate layer. Each gate has an activation function with two weights.

Each hidden state h_{t-1} of the LSTM unit is weighted by the recurrent weights, and the current input X_t is weighted by the input weights IW (Belagoune et al., 2021). The forget, input, and output gates employ a Sigmoid activation function, whereas the cell candidate gate utilizes a hyperbolic tangent function, including a bias term.

The calculation method of the value of the output gate at time t(Chen et al., 2024):

$$i_t = Sigmoid(IW_i \times X_t + RW_i \times h_{t-1} + b_i) \tag{14}$$

$$f_t = Sigmoid(IW_f \times X_t + RW_f \times h_{t-1} + b_f) \tag{15}$$

$$o_t = Sigmoid(IW_o \times X_t + RW_o \times h_{t-1} + b_o) \tag{16}$$

$$c_t^* = \mathrm{Tan}\, H(IW_c \times X_t + RW_c \times h_{t-1} + b_c) \tag{17}$$

Where: X_t represents the input feature at the timestamp t, h_{t-1} indicates the previous hidden state, and RW, IW, and b respectively represent the recurrent weights, input weights, and biases(Gao and Zhang, 2023).

2.8 K-Nearest Neighbor (KNN)

The KNN algorithm is a supervised learning method that can be applied to both classification and regression

problems. KNN is applied in multiple domains, such as text classification, agriculture, medicine, finance, face recognition, economic prediction, and heart disease diagnosis(Xing and Bei, 2020). KNN assigns labels to unlabeled data points by calculating the distances between each unlabeled point and all other points in the dataset (Kück and Freitag, 2021). Then, by discovering patterns in the dataset, each unlabeled data point is allocated to the category of the most similar labeled data(Sinhashthita and Jearanaitanakij, 2020). The straight-line distance formula is the most frequently employed approach for computing distances in KNN.

$$d(x, X) = \sqrt{\sum_{i=1}^{n} (x_i - X_i)^2} \tag{18}$$

Classification is performed by considering the k nearest neighbors (minimum distance), where k denotes the number of neighbors involved in the majority voting process. The class label of the test sample is then assigned based on the majority vote from its k nearest neighbors, with the category having the most representatives among the k neighbors determining the label(Haddadi et al., 2024).

2.9 SHapley Additive exPlanations (SHAP)

SHAP is a game-theory-based method used to explain the output of machine learning models (Zeng et al., 2024). To produce an interpretable model, SHAP utilizes an additive feature attribution method, where the model output is represented as a linear combination of the input variables(Homafar et al., 2022).

$$f(x) = g(x') = \Phi_0 + \sum_{i=1}^{M} \Phi_i x_i' \tag{19}$$

Where M denotes the quantity of input features, and Φ_0 represents the constant value when all inputs are absent, inputs x' and x are correlated via a mapping function, $X = h_x(x')$ (Mendonça et al., 2022).

As is shown in, where Φ_0, Φ_1, Φ_2, and Φ_3 increase the predicted value of $g()$, while Φ_4 decreases the value of $g()$ (Chelgani, 2021).

SHAP values attribute to each feature its contribution to the change in the model's predicted output (Baniasad et al., 2021). They delineate the manner in which the current output can be obtained from the starting value $E[f(z)]$, which would be anticipated if no features related to the current output $f(x)$ were known.

$$f(h_x(z')) = E[f(z) \mid z_s] \approx f([z_s, E) \tag{20}$$

SHAP values quantify the impact of each feature on the model's prediction by attributing the change in the expected output when conditioning on that feature. In cases involving interdependent nonlinear models or when the contribution of individual features to the output f(x) is unknown, the feature contributions can be determined based on assumptions of efficiency, virtuality, additivity, and symmetry, with X representing the auxiliary variable for the i-th feature(Nesa and Yoon, 2024).

3. Dataset

This study includes the China Securities Index (CSI) 300 Index and its constituent stocks, as the CSI 300 Index represents more than 70% of the total market capitalization of Chinese stocks, serving as a robust indicator of the overall performance of China's A-share market(Chen and Xu, 2023). Thirty-three major industry indices were selected from the China Securities Index (https://www.csindex.com.cn/) as features to represent comprehensive information across various industries. The most important raw materials for Chinese Baijiu brewing are sorghum and wheat (Liu et al., 2023). Due to the continuous price fluctuations of sorghum and wheat across different regions and time periods, we collected the daily average selling prices of sorghum and wheat in China from Mysteel (https://www.mysteel.com/). Previous studies have shown that the exchange rate between the Chinese yuan (CNY) and the US dollar (USD) impacts China's stock market (Xiong et al., 2024). Therefore, we downloaded the CNY-USD exchange rate data from Investing.com (https://cn.investing.com/). The Baijiu stock data were obtained from Eastmoney (https://www.eastmoney.com/).

The Baidu Index platform (https://index.baidu.com/) is a tool for tracking and analyzing the search volume of specific keywords, provided by Baidu, China's largest search engine. It reflects investor sentiment to some extent (Fang et al., 2020). We downloaded the PC and mobile Baidu Index data for high-frequency keywords related to Baijiu stocks (Baijiu, Chinese Baijiu, Baijiu brands) from Baidu Index as an additional data source to reflect investor sentiment.

Previous studies in financial economics have focused on the role of entropy and mutual information in quantifying the relationships between financial assets, providing a theoretical foundation for the study of market interconnectedness (Ardakani, 2024). Table 1 summarizes the transfer entropy information between 37 features

and Baijiu sector stocks. In using transfer entropy, we discretized the raw data using KDTREE. After the coarse screening with transfer entropy, 10 features were selected and organized, containing sufficient information. As shown in Table 1, the importance and magnitude of information exchange between Baijiu sector stocks and the features are evident. Clearly, wheat, Baidu Index, environmental protection, construction decoration, semiconductors, transportation, energy, non-metallic materials, agriculture, livestock and fisheries, and banking play a dominant role in the Baijiu sector stocks. Figure 4 presents the transfer entropy with an increased time sliding window, highlighting the time-varying influence of these ten features on the information flow within the Chinese Baijiu stock, thus reflecting the complex dynamic characteristics of the Chinese Baijiu stock market. All data were processed into daily data from November 2019 to October 2024 using mean imputation.

Table 1. Characteristics and entropy of stocks in the white Baijiu sector

	Transfer entropy		Transfer entropy		Transfer entropy		Transfer entropy		Transfer entropy
X1	-0.8695	X9	0.0825	X17	0.1127	X25	0.0299	X33	-0.0159
X2	0.6331	X10	-0.0701	X18	-0.0822	X26	-0.0361	X34	0.0447
X3	0.1926	X11	0.0302	X19	0.0229	X27	0.0044	X35	0.0133
X4	-1.7913	X12	0.1297	X20	0.0292	X28	0.0764	X36	-0.0181
X5	0.0973	X13	-0.0068	X21	-0.0077	X29	0.0191	X37	0.0362
X6	0.0379	X14	-0.0261	X22	0.0436	X30	-0.0082		
X7	0.0216	X15	0.1618	X23	0.0181	X31	0.0074		
X8	-0.0128	X16	0.0668	X24	0.0815	X32	0.1235		

Key: X1= Sorghum; X2= Wheat; X3= Baidu index; X4= Exchange rate between RMB and US dollar; X5= Energy; X6= Chemical industry; X7= Non-ferrous metal; X8= Steel; X9= Non-metallic material; X 10= Paper and packaging;; X11= Aerospace & Defense; X12= Architectural decoration; X13= Electrical equipment; X14= Mechanical manufacturing; X15= Environmental protection; X16= Business Services an d Supplies; X17= Transportation; X18= Passenger cars and parts; X19= Consumer durables; X20= Texti le Clothing and jewelry; X21= Consumer service; X22= Retail; X23= Food, Beverage and tobacco; X24 = Agriculture, animal husbandry and fishing; X25= Household and personal goods; X26= Medical treat ment; X27= Medicine; X28= Bank; X29= Insurance; X30= Computer; X31= Electrons; X32= Semicond uctor; X33= Telecommunications services; X34= Communication equipment and technical services; X35= Media; X36= Utilities; X37= Real estate

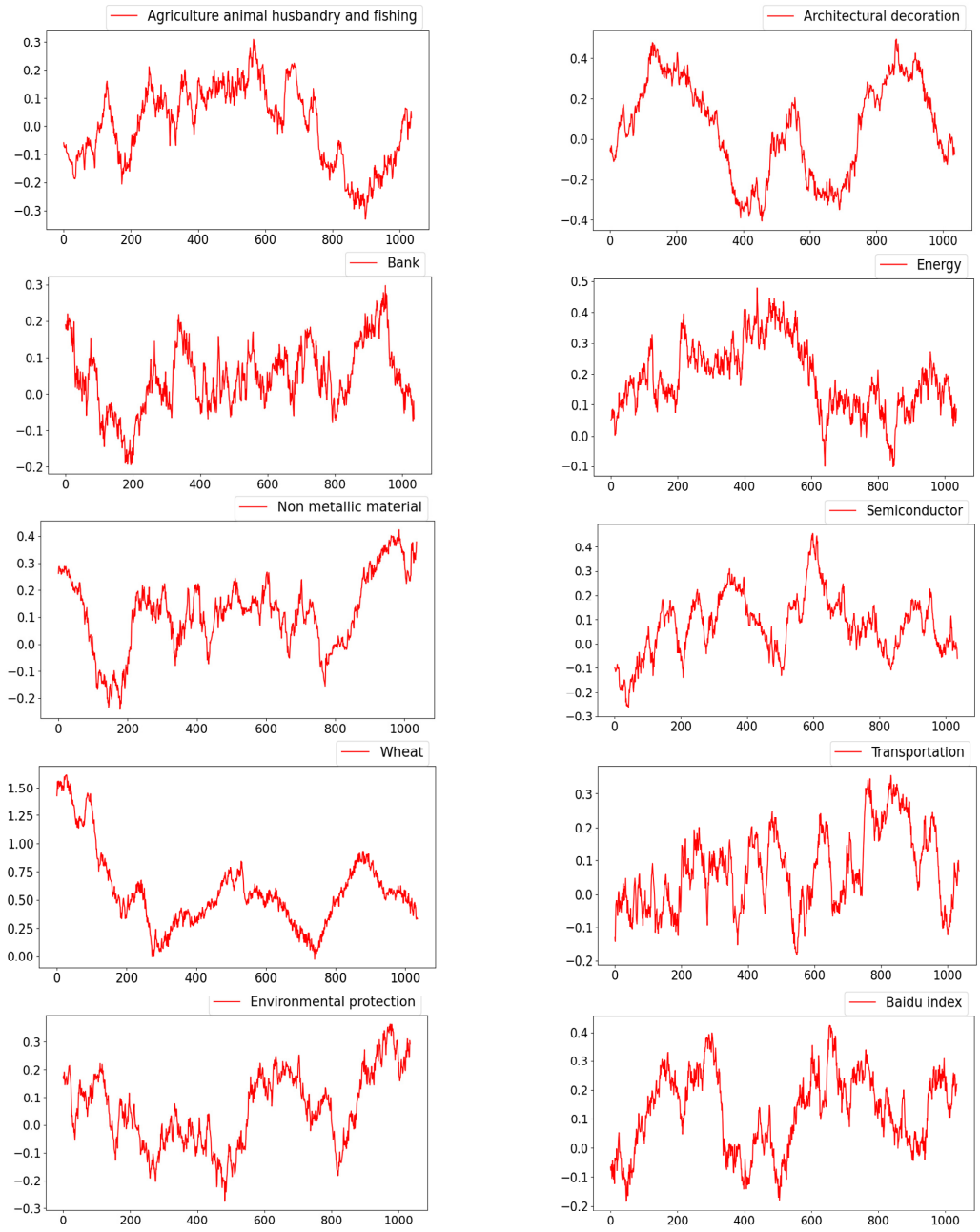

Figure 4. Transfer entropy diagram

4. Experiment Result and Discussion

4.1 Prediction

Transfer entropy initially screened the key factors influencing Chinese Baijiu stock, revealing significant differences and highlighting the linear interactions between various features and Chinese Baijiu stock. Among the 37 features, 10 of the most prominent influencing factors were selected. These factors are distributed across four different dimensions: Resources and Agriculture (wheat and agriculture, animal husbandry and fishing), Industry and Manufacturing (semiconductors, non-metallic materials, and construction decoration), Services and Consumption (banking, transportation, and Baidu Index), and Cross-Sector and Sustainable Development (environmental protection and energy). These features were stratified into two levels: direct factors and indirect factors, effectively filling the research gap in the Chinese Baijiu stock market. However, since transfer entropy cannot reveal the nonlinear relationships between features, there are limitations in interpreting the output of

complex models for this phenomenon.

We first compared the predictive performance of different models. Seven models were selected for this comparison: XGBoost, LightGBM, SVM, RF, LSTM, KNN, and LR. The models demonstrated similar performance on both the training and test datasets, suggesting that overfitting was not present. Here, we combined five-fold cross-validation with Bayesian optimization to enhance the model's generalization ability and efficiently find the optimal hyperparameters in the high-dimensional hyperparameter space, thereby improving model performance. To provide comprehensive information, we calculated R^2, RMSE, MSE, MAE, and MAPE. Table 2 shows the distribution of prediction metrics for different models. As shown in Table 2, the LightGBM model performed the best overall, so we chose to combine the LightGBM model with SHAP. Figure 5 presents the predictive performance of LightGBM.

Table 2. Presents a comparison of the prediction outcomes from various models

Model	R^2	RMSE	MSE	MAE	MAPE
XGBoost	0.9838	74.4697	5545.7408	56.4945	0.0266
LightGBM	0.9925	50.5746	2557.7931	35.4537	1.7071
SVM	0.9207	178.8166	31975.4080	151.9801	0.0778
RF	0.9855	77.3348	5980.6772	46.7452	0.0238
LSTM	0.7501	282.4719	79790.4062	224.7991	11.3154
KNN	0.9893	60.4853	3658.4726	41.3822	0.0204
LR	0.7932	266.3244	70928.6771	215.0116	0.1136

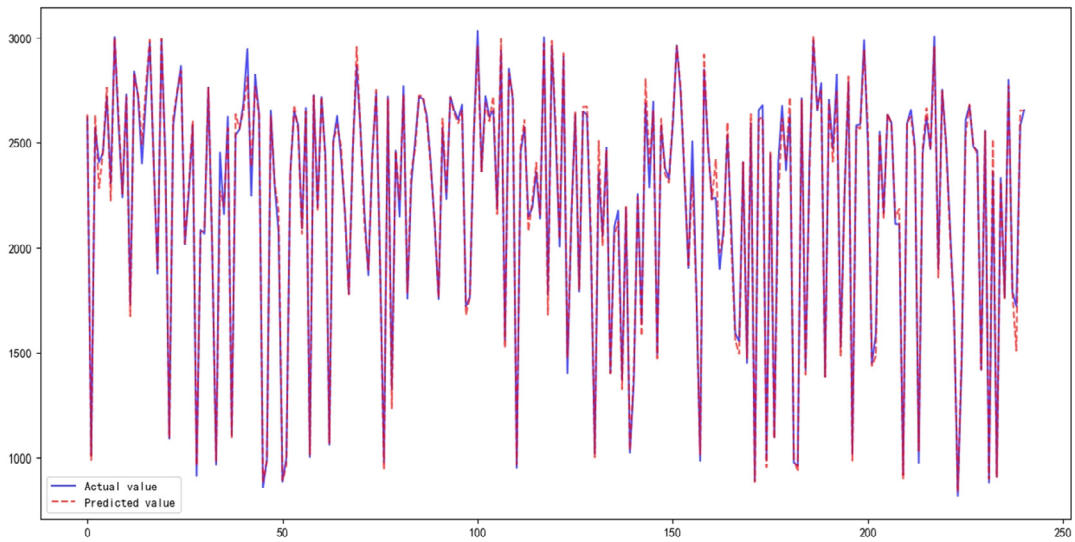

Figure 5. The measured value of Baijiu stock differs from the predicted value

4.2 Causality Analysis

As shown in Figure 6, the variables analyzed in this study interact with each other, rather than behaving independently. The inherent interconnectedness within the Baijiu sector highlights the intricate nature of stock dynamics. These features do not exist independently; rather, they influence each other directly or indirectly. Due to the diversity of features, this study focuses on Chinese Baijiu stock rather than discussing each individual feature. In the stock market environment, almost all features are interconnected in complex ways. This complexity arises from the interactions between different industry characteristics, and ordinary linear relationships cannot capture these connections.

The causal relationship diagram shows that Chinese Baijiu stock exhibits a significant positive causal relationship with both wheat and environmental protection. This relationship indicates a high degree of dependence of Chinese Baijiu stock on wheat, reflecting the close connection between certain agricultural sectors and Chinese Baijiu stock at the supply chain level. Additionally, Chinese Baijiu stock shows a strong positive causal relationship with environmental protection, suggesting a correlation with the growth of the Chinese Baijiu stock market, and further supporting the influence of climate on the stock market(Gan et al., 2024).

There is a significant indirect negative causal relationship between Chinese Baijiu stock and Baidu Index. This negative correlation may reflect a reverse volatility effect between investor sentiment and actual stock market performance, supporting previous findings on the partial role of investor sentiment in the stock market (Cheema and Fianto, 2024). Moreover, Chinese Baijiu stock exhibits a direct causal relationship with both the energy and transportation sectors, offering a new perspective for investment and research in Chinese Baijiu stock. These causal relationships provide important empirical support for further research into the economic dynamics of Chinese Baijiu stock and highlight the importance of industry synergies and interdependence in the broader market environment.

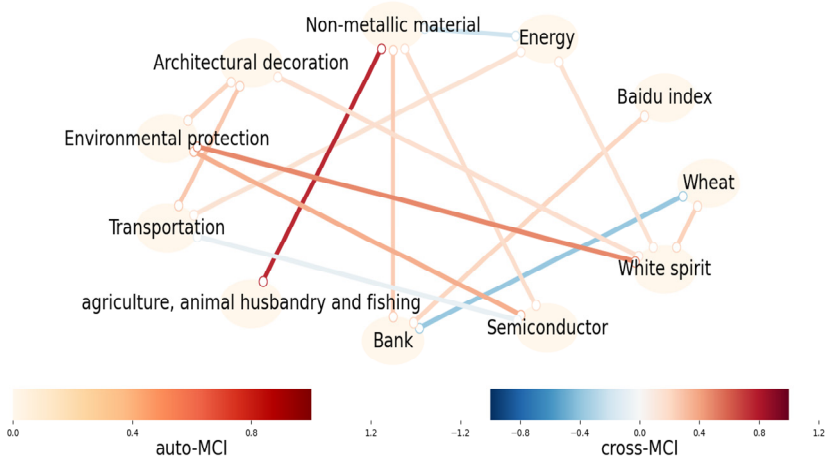

Figure 6. Causal dependence

4.3 SHAP

As shown in Figure 7 and 8, this study utilizes the SHAP method to analyze the key features influencing Chinese Baijiu stock. Figure 7 shows the SHAP distribution plot (violin diagram), and Figure 8 presents the SHAP mean absolute value plot (bar chart), both illustrating the impact of different features on the model output and their importance ranking.

First, the results of the study indicate that wheat is the dominant factor influencing the fluctuation of Chinese Baijiu stock, which is not surprising, as previous research has identified a linkage between the stock market and agricultural products (Woode et al., 2024), although no studies have specifically addressed Chinese Baijiu stock. The SHAP value for wheat not only ranks first but also exhibits a significantly wide distribution range, indicating that its impact on the Chinese Baijiu stock market is highly volatile in both direction and magnitude. The color distribution shows that high wheat values significantly drive up the Chinese Baijiu stock market, while low values are primarily concentrated in the negative impact region. This may be closely related to the supply and demand fluctuations in the agricultural market and changes in grain prices.

Secondly, environmental protection and energy are also key factors, closely linked to the price fluctuations of Chinese Baijiu stock. The SHAP values for environmental protection are mainly concentrated in the positive region, indicating that it has a positive driving effect on the Baijiu stock market, while the impact of energy on Chinese Baijiu stock is more balanced in direction. Thirdly, the SHAP mean absolute values in Figure 6 further reveal the relative importance ranking of other features. Features such as architectural decoration, transportation, and non-metallic materials rank next in importance but still significantly influence Chinese Baijiu stock. These features reflect the indirect impact of infrastructure construction and materials supply and demand. The mean absolute values for Baidu index and banking are relatively small, indicating that their impact on the Baijiu stock market is limited, which may be related to their limited information transmission capabilities or weak correlation with the target variable.

The SHAP distribution plot shows that red stripes indicate higher feature values drive the Baijiu stock market upward, while blue stripes show that lower feature values have a negative impact. The width of the feature stripes reflects the extent of the feature's contribution to the Baijiu stock market, with wider stripes indicating a higher

contribution. The stripes for wheat, environmental protection, and energy are significantly wider than those of other features, highlighting their key role in the model's interpretability. Moreover, the results align with the PCMCI observations, suggesting that some causal relationships identified by PCMCI are validated through SHAP analysis. Both wheat and environmental protection show strong driving effects in both methods, further supporting the critical role of food and environmental factors in the dynamics of the Baijiu stock market.

It is worth noting that, compared to traditional influencing factors emphasized in previous studies (such as stock returns, trade prices, and financial policies) (Zhang et al., 2025), this study identifies emerging catalytic factors, such as architectural decoration, transportation, and non-metallic materials, which play significant roles in the structure of the Baijiu stock market. This may be closely related to the acceleration of China's urbanization process and the upgrading of its industrial structure. The importance of architectural decoration and non-metallic materials in downstream industries has gradually become more prominent, with wide applications in construction, infrastructure development, and the use of renewable materials. Lastly, the study also finds that the Baidu Index, as an indicator of investor sentiment, while overall of lower importance, may provide valuable supplementary information during specific time periods.

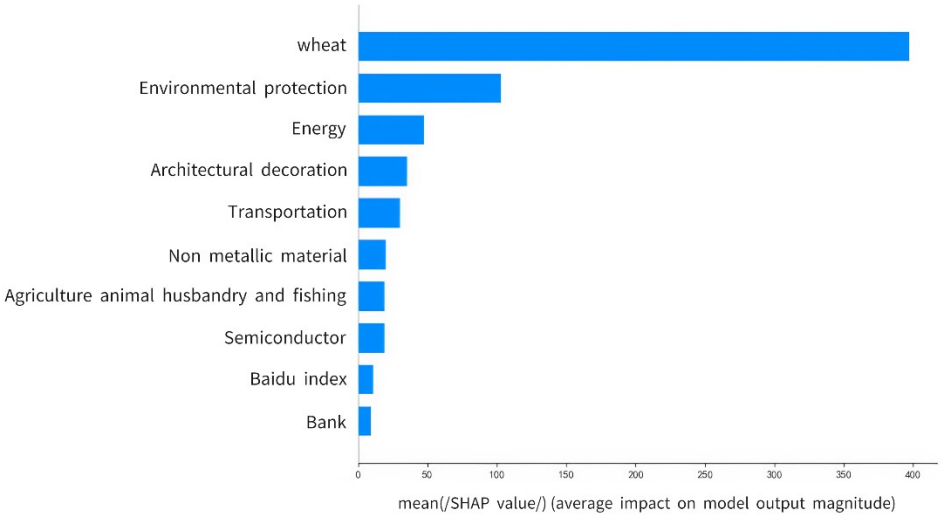

Figure 7. SHAP summary plot

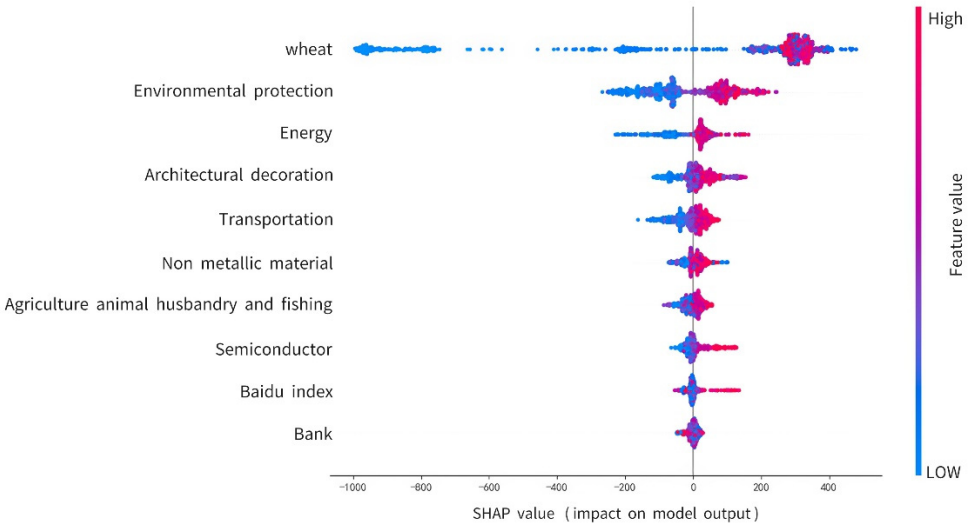

Figure 8. SHAP feature diagram

5. Conclusion

This study introduces a novel approach to explore the key factors and mechanisms driving fluctuations in the Chinese Baijiu stock market from the perspectives of feature selection and causal analysis, combining Transfer Entropy, PCMCI, and interpretable machine learning methods. For the first time, this research offers systematic theoretical and empirical insights into the volatility of the Chinese Baijiu stock market. Using a complex network perspective, 10 key features were identified from 37 variables, covering agricultural resources, industrial manufacturing, and the service and sustainability sectors. These factors were categorized into two levels: direct and indirect factors. Direct factors, including wheat, architectural decoration, environmental protection, and energy, have a significant nonlinear positive causal relationship impact on Chinese Baijiu stock. The Baidu Index indirectly impacts the market through the banking sector and wheat prices, demonstrating a nonlinear negative causal relationship.

From a theoretical perspective, this study introduces a novel framework for understanding financial market fluctuations, particularly by examining the dynamics of the Chinese Baijiu stock market and the nonlinear causal relationships inherent in industry-specific characteristics. Practically, the study provides actionable tools for risk managers and investors, assisting in portfolio optimization, risk management, and strategic decision-making. Policy-wise, the findings suggest that strategic adjustments in agricultural price policies, energy conservation, and emission reduction measures could not only foster a healthier development of the Chinese Baijiu stock market but also promote the sustainable development of the Baijiu industry as a whole.

Future research could enhance model robustness by integrating real-time socio-economic data and macroeconomic variables, while incorporating cross-validation across multiple deep learning modeling approaches, to mitigate the impact of confounding effects on causal inference. This framework not only offers a unique analytical perspective for the Chinese liquor stock market but can also be adapted to other industries and markets by appropriately adjusting data variables and methods. However, different industries and markets have their own characteristics and data structures, so when applying this framework, it is necessary to select suitable machine learning models and data processing methods based on the specific context. Subsequent research will integrate this framework with cloud computing technology to develop a real-time monitoring system, thereby providing more intuitive and timely decision-making support for investors, policymakers, and industry development.

References

Abolmakarem, S., Abdi, F., Khalili-Damghani, K., & Didehkhani, H. (2022). A multi-stage machine learning approach for stock price prediction: Engineered and derivative indices. *SSRN Electronic Journal*, 24, 200449. https://doi.org/10.2139/ssrn.4074883

Ardakani, O. M. (2024). Portfolio optimization with transfer entropy constraints. *International Review of Financial Analysis, 96*. https://doi.org/10.1016/j.irfa.2024.103644

Baniasad, M., Mofrad, M. G., Bahmanabadi, B., & Jamshidi, S. (2021). COVID-19 in Asia: Transmission factors, re-opening policies, and vaccination simulation. *Environmental Research, 202*, 111657. https://doi.org/10.1016/j.envres.2021.111657

Belagoune, S., Bali, N., Bakdi, A., Baadji, B., & Atif, K. (2021). Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems. *Measurement: Journal of the International Measurement Confederation, 177*, 109330. https://doi.org/10.1016/j.measurement.2021.109330

Bin, L. (2024). Analysis on the investment value of high-end Baijiu stocks. *Academic Journal of Business & Management, 6*, 209–213. https://doi.org/10.25236/ajbm.2024.060230

Caparrini, A., Arroyo, J., & Escayola Mansilla, J. (2024). S&P 500 stock selection using machine learning classifiers: A look into the changing role of factors. *Research in International Business and Finance, 70*, 102336. https://doi.org/10.1016/j.ribaf.2024.102336

Cheema, M. A., & Fianto, B. A. (2024). Investor sentiment and stock market anomalies: Evidence from Islamic countries. *Pacific Basin Finance Journal, 88*, 102557. https://doi.org/10.1016/j.pacfin.2024.102557

Chelgani, S. C. (2021). Estimation of gross calorific value based on coal analysis using an explainable artificial intelligence. *Machine Learning with Applications, 6*, 100116. https://doi.org/10.1016/j.mlwa.2021.100116

Chen, G., & Jia, G. (2024). A hybrid causal machine learning to reveal driving factors responsible coal market: Case of the Chinese industry. *Journal of Cleaner Production, 434*, 140249. https://doi.org/10.1016/j.jclepro.2023.140249

Chen, J., & Xu, L. (2023). Do exchange-traded fund activities destabilize the stock market? Evidence from the China securities index 300 stocks. *Economic Modelling, 127*, 106450. https://doi.org/10.1016/j.econmod.2023.106450

Chen, X., Yang, F., Sun, Q., & Yi, W. (2024). Research on stock prediction based on CED-PSO-StockNet time series model. *Scientific Reports, 14*, 27462. https://doi.org/10.1038/s41598-024-78984-1

Cheng, H., Guo, H., & Shi, Y. (2024). Multifactor conditional equity premium model: Evidence from China's stock market. *Journal of Banking and Finance, 161*, 107117. https://doi.org/10.1016/j.jbankfin.2024.107117

Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics, 11*, 278–289. https://doi.org/10.4236/ojs.2021.112015

Demirer, R., & Yuksel, A. (2024). Do industries lead the stock market? Evidence from an emerging stock market. *Borsa Istanbul Review*. https://doi.org/10.1016/j.bir.2024.11.005

Ellington, M., Stamatogiannis, M. P., & Zheng, Y. (2022). A study of cross-industry return predictability in the Chinese stock market. *International Review of Financial Analysis, 83*, 102249. https://doi.org/10.1016/j.irfa.2022.102249

Fang, J., Gozgor, G., Lau, C. K. M., & Lu, Z. (2020). The impact of Baidu Index sentiment on the volatility of China's stock markets. *Finance Research Letters, 32*, 101099. https://doi.org/10.1016/j.frl.2019.01.011

Gan, K., Li, R., & Zhou, Q. (2024). Climate transition risk, environmental news coverage, and stock price crash risk. *International Review of Financial Analysis, 96*, 103657. https://doi.org/10.1016/j.irfa.2024.103657

Gao, Y. C., Tan, R., Fu, C. J., & Cai, S. M. (2023). Revealing stock market risk from information flow based on transfer entropy: The case of Chinese A-shares. *Physica A: Statistical Mechanics and its Applications, 624*, 1–14. https://doi.org/10.1016/j.physa.2023.128982

Gao, Z., & Zhang, J. (2023). The fluctuation correlation between investor sentiment and stock index using VMD-LSTM: Evidence from China stock market. *North American Journal of Economics and Finance, 66*, 101915. https://doi.org/10.1016/j.najef.2023.101915

Graf, H. P., Cosatto, E., Bottou, L., Durdanovic, I., & Vapnik, V. (2005). Parallel support vector machines: The cascade SVM. *Advances in Neural Information Processing Systems*.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing, 27*, 659–678. https://doi.org/10.1007/s11222-016-9646-1

Guo, P., Huang, Y., & Wang, J. (2021). Scalable and flexible two-phase ensemble algorithms for causality discovery. *Big Data Research, 26*, 100252. https://doi.org/10.1016/j.bdr.2021.100252

Haddadi, S. J., Farshidvard, A., Silva, F. dos S., dos Reis, J. C., & da Silva Reis, M. (2024). Customer churn prediction in imbalanced datasets with resampling methods: A comparative study. *Expert Systems with Applications, 246*. https://doi.org/10.1016/j.eswa.2023.123086

Homafar, A., Nasiri, H., & Chelgani, S. C. (2022). Modeling coking coal indexes by SHAP-XGBoost: Explainable artificial intelligence method. *Fuel Communications, 13*, 100078. https://doi.org/10.1016/j.jfueco.2022.100078

Hudec, M., Mináriková, E., Mesiar, R., Saranti, A., & Holzinger, A. (2021). Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. *Knowledge-Based Systems, 220*, 106916. https://doi.org/10.1016/j.knosys.2021.106916

Jiang, F., Ma, T., & Zhu, F. (2024). Fundamental characteristics, machine learning, and stock price crash risk. *Journal of Financial Markets, 69*. https://doi.org/10.1016/j.finmar.2024.100908

Jiménez-Carvelo, A. M., González-Casado, A., Bagur-González, M. G., & Cuadros-Rodríguez, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International, 122*, 25–39. https://doi.org/10.1016/j.foodres.2019.03.063

Krich, C., Runge, J., Miralles, D. G., Migliavacca, M., Perez-Priego, O., El-Madany, T., Carrara, A., & Mahecha, M. D. (2020). Estimating causal networks in biosphere-atmosphere interaction with the PCMCI approach. *Biogeosciences, 17*, 1033–1061. https://doi.org/10.5194/bg-17-1033-2020

Kück, M., & Freitag, M. (2021). Forecasting of customer demands for production planning by local k-nearest neighbor models. *International Journal of Production Economics, 231*, 107837. https://doi.org/10.1016/j.ijpe.2020.107837

Lawal, A. I., & Kwon, S. (2021). Application of artificial intelligence to rock mechanics: An overview. *Journal of Rock Mechanics and Geotechnical Engineering, 13*(2), 248–266. https://doi.org/10.1016/j.jrmge.2020.05.010

Lei, H. (2021). Financial index data prediction based on improved GBDT model. In *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* (pp. 697-702). https://doi.org/10.1109/ICAICA52286.2021.9498075

Li, Z., Xu, W., & Li, A. (2022). Research on multi-factor stock selection model based on LightGBM and Bayesian optimization. *Procedia Computer Science, 214*, 1234–1240. https://doi.org/10.1016/j.procs.2022.11.301

Liu, M., Li, S., Weiss, T., Li, Y., Wang, D., & Zheng, Y. (2023). Solid-state fermentation of grain sorghum to produce Chinese liquor: Effect of grain properties and fermenting culture. *Journal of Cereal Science, 114*, 103776. https://doi.org/10.1016/j.jcs.2023.103776

Makariou, D., Barrieu, P., & Chen, Y. (2021). A random forest-based approach for predicting spreads in the primary catastrophe bond market. *Insurance: Mathematics and Economics, 101*, 140–162. https://doi.org/10.1016/j.insmatheco.2021.07.003

Meher, B. K., Singh, M., Birau, R., & Anand, A. (2024). Forecasting stock prices of fintech companies of India using random forest with high-frequency data. *Journal of Open Innovation: Technology, Market, and Complexity, 10*, 100180. https://doi.org/10.1016/j.joitmc.2023.100180

Mendonça, Y. V. S., Naranjo, P. G. V., & Pinto, D. C. (2022). The role of technology in the learning process: A decision tree-based model using machine learning. *Emerging Science Journal, 6*(3), 280–295. https://doi.org/10.28991/ESJ-2022-SIED-020

Menegozzo, G., Dall'alba, D., & Fiorini, P. (2020). Causal interaction modeling on ultra-processed food manufacturing. In *2020 IEEE International Conference on Automation Science and Engineering (CASE)* (pp. 200-205). https://doi.org/10.1109/CASE48305.2020.9216973

Mutinda, J. K., & Langat, A. K. (2024). Stock price prediction using combined GARCH-AI models. *Scientific African, 26*, e02374. https://doi.org/10.1016/j.sciaf.2024.e02374

Nesa, M., & Yoon, Y. (2024). Speed prediction and nearby road impact analysis using machine learning and ensemble of explainable AI techniques. *Scientific Reports, 14*, 25208. https://doi.org/10.1038/s41598-024-74545-8

Nie, P., Roccotelli, M., Fanti, M. P., Ming, Z., & Li, Z. (2021). Prediction of home energy consumption based on gradient boosting regression tree. *Energy Reports, 7*, 1246–1255. https://doi.org/10.1016/j.egyr.2021.02.006

Peng, S., Han, W., & Jia, G. (2022). Pearson correlation and transfer entropy in the Chinese stock market with time delay. *Data Science and Management, 5*, 117–123. https://doi.org/10.1016/j.dsm.2022.08.001

Qiu, L., & Yang, H. (2020). Transfer entropy calculation for short time sequences with application to stock markets. *Physica A: Statistical Mechanics and its Applications, 559*, 125121. https://doi.org/10.1016/j.physa.2020.125121

Ren, T., Li, S., & Zhang, S. (2024). Stock market extreme risk prediction based on machine learning: Evidence from the American market. *North American Journal of Economics and Finance, 74*, 102241. https://doi.org/10.1016/j.najef.2024.102241

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances, 5*(11). https://doi.org/10.1126/sciadv.aau4996

Saeed, B., & Yin, W. (n.d.). Optimizing stock price prediction for South Asian markets using LSTM, GRU, CNN with greedy algorithm.

Sensoy, A., Sobaci, C., Sensoy, S., & Alali, F. (2014). Effective transfer entropy approach to information flow between exchange rates and stock markets. *Chaos, Solitons and Fractals, 68*, 180–185. https://doi.org/10.1016/j.chaos.2014.08.007

Shen, B., Yang, S., Hu, J., Zhang, Y., Zhang, L., Ye, S., Yang, Z., Yu, J., Gao, X., & Zhao, E. (2024). Interpretable causal-based temporal graph convolutional network framework in complex spatio-temporal systems for CCUS-EOR. *Energy, 309*, 133129. https://doi.org/10.1016/j.energy.2024.133129

Shi, G. Y., Zhou, Y., Sang, Y. Q., Huang, H., Zhang, J. S., Meng, P., & Cai, L. L. (2021). Modeling the response of negative air ions to environmental factors using multiple linear regression and random forest. *Ecological Informatics, 66*, 101464. https://doi.org/10.1016/j.ecoinf.2021.101464

Sinhashthita, W., & Jearanaitanakij, K. (2020). Improving kNN algorithm based on weighted attributes by Pearson correlation coefficient and PSO fine tuning. In *InCIT 2020 - 5th International Conference on Information Technology* (pp. 27-32). https://doi.org/10.1109/InCIT50588.2020.9310938

Szczygielski, J. J., Charteris, A., Obojska, L., & Brzeszczyński, J. (2024). Recession fears and stock markets: An application of directional wavelet coherence and a machine learning-based economic agent-determined Google fear index. *Research in International Business and Finance, 72*. https://doi.org/10.1016/j.ribaf.2024.102448

Teplova, T., Sokolova, T., & Kissa, D. (2023). Revealing stock liquidity determinants by means of explainable AI: The role of ESG before and during the COVID-19 pandemic. *Resources Policy, 86*. https://doi.org/10.1016/j.resourpol.2023.104253

Tita, A. F., French, J. J., Gurdgiev, C., & Obalade, A. (2025). Does the tail of finance wag the dog of the real economy? Dynamic connectedness of the stock market and business confidence. *International Review of Economics and Finance, 98*, 103856. https://doi.org/10.1016/j.iref.2025.103856

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., & Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management, 301*. https://doi.org/10.1016/j.jenvman.2021.113941

Wang, W., Moffatt, P. G., Zhang, Z., & Raza, M. Y. (2025). Volatility spillovers and conditional correlations between oil, renewables and stock markets: A multivariate GARCH-in-mean analysis. *Energy Strategy Reviews, 57*, 101639. https://doi.org/10.1016/j.esr.2025.101639

Weng, F., Zhu, J., Yang, C., Gao, W., & Zhang, H. (2022). Analysis of financial pressure impacts on the health care industry with an explainable machine learning method: China versus the USA. *Expert Systems with Applications, 210*, 118482. https://doi.org/10.1016/j.eswa.2022.118482

Woode, J. K., Idun, A. A. A., & Kawor, S. (2024). Comovement between agricultural commodities and stock returns of commodity-dependent sub-Saharan Africa countries amidst the COVID-19 pandemic. *Scientific African, 23*, e01972. https://doi.org/10.1016/j.sciaf.2023.e01972

Wu, W., Xu, M., Su, R., & Ullah, K. (2024). Modeling crude oil volatility using economic sentiment analysis and opinion mining of investors via deep learning and machine learning models. *Energy, 289*, 130017. https://doi.org/10.1016/j.energy.2023.130017

Xing, W., & Bei, Y. (2020). Medical health big data classification based on kNN classification algorithm. *IEEE Access, 8*, 28808–28819. https://doi.org/10.1109/ACCESS.2019.2955754

Xiong, Y., Shen, J., Yoon, S. M., & Dong, X. (2024). Macroeconomic determinants of the long-term correlation between stock and exchange rate markets in China: A DCC-MIDAS-X approach considering structural breaks. *Finance Research Letters, 61*, 105020. https://doi.org/10.1016/j.frl.2024.105020

Yin, N., Wang, H., Wang, Z., Feng, K., Xu, G., & Yin, S. (2023). A study of brain networks associated with freezing of gait in Parkinson's disease using transfer entropy analysis. *Brain Research, 1821*, 148610. https://doi.org/10.1016/j.brainres.2023.148610

Yu, H., Chen, R., & Zhang, G. (2014). A SVM stock selection model within PCA. *Procedia Computer Science, 31*, 406–412. https://doi.org/10.1016/j.procs.2014.05.284

Zeng, Q., Lu, X., Xu, J., & Lin, Y. (2024). Macro-driven stock market volatility prediction: Insights from a new hybrid machine learning approach. *International Review of Financial Analysis, 96*, 103711. https://doi.org/10.1016/j.irfa.2024.103711

Zhang, K., Chu, Z., Xing, J., Zhang, H., & Cheng, Q. (2023). Urban traffic flow congestion prediction based on a data-driven model. *Mathematics, 11*(19), 4075. https://doi.org/10.3390/math11194075

Zhang, Y., Zhao, X., & Zhang, Z. (2025). Financial regulatory policy uncertainty: An informative predictor for financial industry stock returns. *North American Journal of Economics and Finance, 75*, 102321. https://doi.org/10.1016/j.najef.2024.102321

Zhang, Z., Wang, L., Chen, G., Gu, Z., Tian, Z., Du, X., & Guizani, M. (2022). STG2P: A two-stage pipeline model for intrusion detection based on improved LightGBM and K-means. *Simulation Modelling Practice and Theory, 120*, 102614. https://doi.org/10.1016/j.simpat.2022.102614

Zheng, Y. L., Yang, M. Y., Liu, K. X., Chen, Y. K., & Wu, X. (2024). Dynamic and asymmetric connectedness among fossil energies and stock markets of the Belt and Road countries under shocks from extreme events. *Transnational Corporations Review, 16*, 200104. https://doi.org/10.1016/j.tncr.2024.200104

Acknowledgments

Not applicable.

Authors contributions

Dr. Yao Ruiguang was responsible for study design, data collection, manuscript drafting, and revision. All authors read and approved the final manuscript.

Funding

Not applicable.

Competing interests

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

Not applicable, as this research does not involve human or animal subjects.

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Data sharing statement

No additional data are available.