Development Economics of China
2025, VOL. 9, NO. 6, 37-43
DOI: 10.47297/wspdecWSP2515-797307.20250906

A Review of Machine Learning-Driven Multifactor Modeling in Stock Prediction

Yinfei Fu*

Faculty of Science, Kunming University of Science and Technology, Kunming, Yunnan. 650031, China

ABSTRACT
With the rapid development of financial markets and the continuous progress of data acquisition technology, machine learning algorithms have been increasingly applied in the field of stock return prediction. This paper systematically reviews the research progress of machine learning-driven multifactor models in stock forecasting in recent years, focusing on analyzing the research results of different machine learning algorithms (e. g., linear regression, neural networks, random forest, deep learning, etc.) combined with multifactor models. By combing and classifying the existing literature, this paper summarizes the differences in the performance of each method in terms of prediction accuracy, model complexity, interpretability, etc., and explores what feature engineering is applicable under different models. Finally, this paper points out the deficiencies in current research, such as the lack of cross-market generalization capability and the high-frequency and low-frequency data fusion challenges, and proposes potential directions for future research, including the combination of real-time decision-making and reinforcement learning, interpretability and regulatory compliance. The research results in this paper can provide references and lessons for academics and financial practice areas.

KEYWORDS
Machine learning; Multi-factor modeling; Stock return prediction; Feature engineering; Investment strategy

1 Introduction

As a core component of modern financial system, the stock market, its price formation mechanism and return prediction have always been an important topic of finance research. With the rapid development of computer technology and the explosive growth of financial data, the application of machine learning algorithms in the field of financial forecasting presents unprecedented potential. In particular, the combination of machine learning and multi-factor modeling provides a new research paradigm and technical path for stock return forecasting.

In recent years, machine learning-driven quantitative investment strategies have made significant progress in both academia and industry. On the one hand, machine learning algorithms are able to deal with high-dimensional nonlinear relationships and capture complex patterns that are difficult to be recognized by traditional linear models; on the other hand, multifactor models provide a systematic risk-return analysis framework, and the combination of the two can give full play to their respective advantages. Studies have shown that this combination not only improves forecasting accuracy, but also adapts to different market environments and regional characteristics. However, existing studies still face many challenges, including data overfitting, insufficient model interpretations, and limited adaptability to market environments, which urgently require systematic summarization and analysis.

The theoretical significance of this study is to systematically sort out the research progress of combining machine learning and multi-factor modeling to provide a reference framework for subsequent studies; the practical value is reflected in the provision of methodological guidance for strategy development for investment institutions. By analyzing the shortcomings of the existing research, this paper will also propose possible future research directions, including the improvement of cross-market generalization ability and the integration of high-frequency and low-frequency data. These explorations will not only help promote the depth of academic research, but also provide new ideas and tools for financial practice.

2 Research

In order to find the most relevant information, we searched various scientific research databases, such as Web of Science, Springer Link, IEEE Xplore, Engineering Index, and MDPI. a series of keywords were used in the search, including ["Machine Learning"] AND ["Multi-factor Modeling"] AND ["Stock Market Forecasting "OR "stock market prediction" OR "stock market forecast"], and set the time frame to the period 2018-2023 . Since stock market forecasting is an interdisciplinary field, we followed the following inclusion and exclusion criteria to select the studies for final analysis.

Inclusion Criteria.

* **Corresponding Author:** Yinfei Fu, 13310948212@163.com

-Studies need to describe in detail the data sources, factor construction methods, and machine learning model architecture.

-Studies must specify the objectives of the study, theoretical background, graphical modeling techniques, and quantitative analysis.

Exclusion Criteria.

-Presentation papers, short papers and opinion papers are excluded.

-Studies that do not provide a clear statement or sufficient insights into their methodology are excluded.

3 Methods of combining machine learning algorithms with multifactor modeling

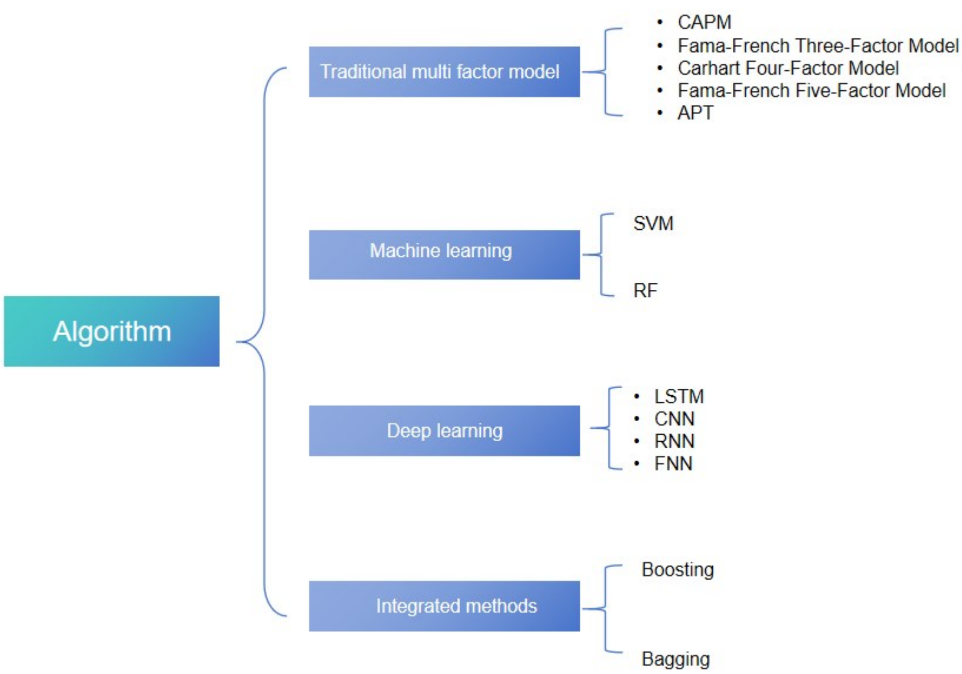

Figure1 Multi-factor modeling algorithms now relevant to stock prediction

3.1 Methods of combining different machine learning algorithms with multifactor models

3.1.1 Traditional multifactor modeling

The theoretical foundation of traditional multifactor model is rooted in Capital Asset Pricing Model (CAPM) [1] and Arbitrage Pricing Theory (APT) [2], whose core idea is to construct a return forecasting framework by identifying and quantifying systematic factors affecting asset returns. Foreign scholars Fama and French (1993) [3] proposed the classic three-factor model (market factor, size factor, and value factor) and verified its effectiveness in explaining stock returns. Subsequently, Carhart (1997)[4] introduced the momentum factor on this basis and constructed a four-factor model, which further enhanced the explanatory power of the model. In recent years, Fama and French (2015)[5] added profitability and investment factors to form a five-factor model, which was validated in the global market.

Domestic scholars Li Zhibing et al. (2017)[6] conducted an empirical study on the Chinese A-share market and found that the five-factor model still has limitations in explaining Chinese stock returns, indicating that market heterogeneity has a significant impact on factor performance. In addition, Chengcheng Xu (2020)[7] pointed out that traditional multi-factor models have weak predictive ability in nonlinear market environments, and there is an urgent need to introduce more sophisticated modeling methods.

3.1.2 Traditional machine learning

Random Forest (RF) and Support Vector Machine (SVM) in machine learning are both rooted in statistical learning theory, the former belongs to the integrated learning framework of Bagging idea, and the latter originates from the principle of Structural Risk Minimization (SRM) put forward by Vapnik [8]. The former belongs to the integrated learning framework of Bagging idea, while the latter originates from the principle of structural risk minimization (SRM) proposed by Vapnik [8]. Random Forest constructs a large number of decision trees by Bootstrap resampling, and randomly extracts a subset of features to increase the differences between trees when nodes split, and finally integrates the outputs by voting or averaging; its convergence is guaranteed by the theory of "strength-correlation" of Breiman [9], and when the correlation between trees decreases and a single tree becomes less correlated, it is possible to integrate the outputs by voting or averaging. SVM maps the original features to the high-dimensional Hilbert space through the kernel function, and searches for the maximal spacing hyperplane in this space, the optimization goal of which is to minimize the empirical risk and the complexity of the model (in the VC-dimension), so that it has a good ability of generalization even under the condition of small samples. The introduction of the kernel trick enables SVM to implicitly handle nonlinear classification or regression tasks without explicitly computing high-dimensional coordinates.

In existing studies, RF and SVM are widely used for stock direction or yield prediction.Krauss et al. (2017)[10] compared deep neural networks, gradient boosted trees and random forests in a high-frequency trading scenario of the S&P500, and found that random forests have an annualized excess return of 43% after deducting the transaction cost, which highlights their robustness in noisy financial data. The robustness of Random Forest in noisy financial data is emphasized. As for support vector machines, Amin (2000)[11] verified that the accuracy of RBF kernel SVM in predicting the direction of next day's stock price is better than that of traditional feed-forward neural networks in the NASDAQ market. In recent years, research on fusing feature engineering and machine learning has further improved model performance. For example, Patel et al. (2015) [12] proposed the Trend Deterministic Data Preparation (TDDP) framework: firstly, technical indicators (RSI, MACD, OBV, etc.) are utilized to construct trend-discriminative features, and subsequently, the features with trend-discriminative power are constructed. Firstly, technical indicators (RSI, MACD, OBV, etc.) are used to construct features with trend discriminatory power, then the features are downscaled by principal component analysis, and finally, the refined features are fed into RF and SVM to predict the direction of the Nifty 50 index respectively. The experimental results show that the accuracy of the TDDP-processed SVM model reaches 89.13%, which is significantly higher than that of the benchmark model without feature processing (75.25%), confirming the synergistic effect of combining the feature engineering and kernel methods.

However, the existing results also reveal some limitations: although RF can output feature contributions via Gini importance or ranking importance, the inter-tree interactions are complicated and difficult to be mapped to economic explanations; the kernel function and penalty parameters C and γ of SVM are highly sensitive to the results, and the quadratic programming solution under large-scale training sets is time-consuming. In addition, both of them do not explicitly deal with time dependence, and need to adapt to the dynamic evolution of financial time series with the help of rolling window or state space model. Future research can incorporate the local feature importance of RF into the regularization path of SVM under the framework of Interpretable Integrated Kernel Learning (IKL), with the aim of simultaneously improving the prediction accuracy, reducing the risk of overfitting, and maintaining the necessary economic interpretability.

3.1.3 LSTM neural networks

Compared to the "shallow" machine learning represented by Random Forests and Support Vector Machines mentioned above, Long Short-Term Memory Networks (LSTMs) take a step forward in modeling principles: instead of relying on artificially constructed features and kernel function mappings, LSTMs are based on a Recurrent Neural Network (RNN) framework and adaptively learn long-range dependencies in a time series by means of a gating mechanism (forgetting gates, input gates, output gates), output gates) to adaptively learn long-range dependencies in time series. This end-to-end deep structure retains the traditional statistical model's portrayal of time-series correlations, but also has the ability to automatically extract nonlinear dynamic features, thus serving as a natural transition bridging the classical machine learning and deep learning paradigms.

In recent studies, deep learning models have been shown to significantly complement and even surpass traditional machine learning algorithms. Recent studies have shown that the fusion of deep learning and multi-factor modeling significantly improves stock price prediction accuracy. Li Bin et al. (2019)[13] found that the average monthly long-short return of LSTM multi-factor model reaches 2.78% by systematically testing 96 A-share anomalous factors, which is 1.2 percentage points higher than that of XGBoost, and the maximum retraction is reduced by 34%, which is the first time to confirm the advantage of the time-series neural network in capturing the nonlinear relationship of factors. The MDT-CNN-LSTM framework proposed by Chaofan Cao et al. (2022) [14] innovatively employs the Hankel tensor to process multifactor time series data of 48 cross-industry stocks, and its out-of-sample annualized return reaches 29.7%, which is a 41.5% improvement over the traditional LSTM model, and the volatility is reduced by 22%, which highlights the efficacy of the convolutional layer in extracting local factor features.

Despite the outstanding performance of LSTM in processing time series, its inherent shortcomings should not be ignored: firstly, the model's capture of long-distance dependence is still limited by gradient propagation paths, and gradient attenuation or explosion may still occur when the sequence is too long; secondly, LSTM has a complex structure and a large number of parameters, which makes it very easy to overfitting and time-consuming to train in small-sample scenarios; thirdly, its unidirectional memory mechanism ignores the potential value of future information to the current prediction, resulting in the loss of the potential value of future information to the current prediction, which leads to a loss of the potential value of future information to the future. Third, its unidirectional memory mechanism ignores the potential value of future information for current prediction, resulting in a lag in response to local mutations or swing trends; finally, as a black-box model, LSTM lacks interpretability and is difficult to visually present the specific contribution of each factor to the price movement, which limits its transparency and trust in practical investment.

3.1.4 Integrated algorithms

After a single model (be it SVM, Random Forest or LSTM) exposed the "overfitting-interpretability" trade-off problem, Ensemble Learning has been proposed to improve the overall performance with the core idea of "combining weak learners". The core idea of Ensemble Learning is to improve the overall performance. The theoretical basis of Ensemble Learning is based on the bias-variance decomposition: Bagging methods reduce the variance by bootstrap resampling; Boosting methods (e.g., AdaBoost, GBDT, XGBoost) reduce the bias by iteratively weighting the residuals; and Stacking realizes the nonlinear fusion of heterogeneous models under the framework of meta-learning to compressed the bias and variance at the same time. Theoretically, it can compress both bias and variance.

In principle, Bagging parallel trains the base-learner and then votes/averages; Boosting serially weights the error samples; Stacking is divided into two layers: the first layer outputs probabilities or predictions using different algorithms (SVM, RF, GBDT, etc.), and the second layer trains a "meta-learner" featuring these outputs. ".Bao et al. (2017) [15] achieved a 23.6% annualized return in S&P500 index forecasting by stacking a hybrid framework of auto-encoder and LSTM, significantly outperforming both a single LSTM model (16.2%) and a traditional time series approach (9.8%).Gyamerah et al. (2021) [16] used AdaBoost+KNN as base and GBM as meta-classifier in Nairobi Exchange and obtained 78.10% accuracy with 0.8238 AUC. Chen and Wei (2020) [17] constructed a heterogeneous integrated learning framework (RF+XGBoost+SVM →LR) in CSI 300 constituents and achieved 38.2% accuracy with the Stacking method. Stacking method achieves 38.2% annualized return, Sharpe ratio of 2.71, and maximum retracement is controlled within 7.3%, which is significantly better than a single machine learning model.

However, the shortcomings of the integrated algorithm should not be ignored: first, high computational cost and complexity of parameter tuning, Stacking needs to optimize two layers of hyper-parameters at the same time, and the number of parameters grows exponentially; second, the explanatory further degradation, the meta-learning machine treats the output of the base model as a "black-box feature", and it is difficult to trace the contribution of a single factor; third, the risk of leakage, traditional cross-validation may lead to the introduction of the time series. Third, there is the risk of time series leakage, which can be mitigated by the use of rolling windows or time slices, as traditional cross-validation may introduce future information.

3.2 Model demonstration - Table 1

Table 1 Review on stock models

Model	Dataset	Data Source Type	Baseline Models	Duration
Effective Transfer Entropy, Integrated Machine Learning Model[18]	US stocks (S&P 500 constituents)	Technical+Fundamental+S entiment + ETE network indicators	- LR - LR - XGB	2000-2018
[19]	All A-shares listed on the HoSE (Hanoi Stock Exchange)	Fundamental + Momentum + Valuation	- OLS regression - CAPM - FF3	Jan 2015 - Jun 2023
LightGBM[20]	A-share daily	Style factors+ technical indicators	-Traditional technical- indicator voting- Random Forest - AdaBoost - XGBoost	Sep 2000 - Dec 2021
[21]	Chinese Equity	Fundamental + Holdings	OLS Linear Regression	2005-2021

3.3 Analysis and discussion - tabular analysis

We summarize the following four comparative tables for different stock markets, namely, Vietnam stock market (HoSE), China A-share market (CSI 300/500/800, HS300, A-shares), U.S. stock market (S&P 500), and Chinese Equity Fund, and draw relevant conclusions.

3.3.1 Vietnam Stock Market (HoSE)

Table 2 HoSE-oriented prediction models

	Model Name	Feature Engineering Methods	Evaluation Metric	Metric Value
[19]	XGBoost	-Data cleaning and variable calculation -outlier removal,missing value handling -z-score normalization	AUC Annualized Return (only choose the first group) Max Drawdown Sharpe Ratio	>0.5 34.81% 29.08% 1.38
	LightGBM	same as above	AUC Annualized Return Max Drawdown Sharpe Ratio	>0.5 36.78% 34.81% 1.4

3.3.2 China A-share Market (CSI 300/500/800, HS300, A-shares)

Table 3 A-share market models

	Model Name	Feature Engineering Methods	Evaluation Metric	Metric Value
[20]	Random Forest	- Data preprocessing;outlier removal, -missing value handling,standardization	Cumulative relative return annualized return max drawdown Sharpe ratio Volatility	-45.23% -62.35% -11.73% 59.29% -43.07% 22.85%
	Adaboost	Same as above	Same as above	-8.44% -25.56% -1.81% -56.99% -4.49% 23.99%
	XGBoost	Same as above	Same as above	-7.59% -24.71% -1.62% 53.69% 5.21% 23.88

Return category (cumulative, relative, annualized): negative sign = loss or benchmark underperformance, Risk category (maximum retracement, Sharpe ratio): negative sign = risk not compensated by return (e.g., negative Sharpe) or emphasized loss (e.g., retracement)

3.3.3 U.S. Stock Market (S&P 500)

Table 4 U.S. Stock Models

	Model Name	Feature Engineering Methods	Evaluation Metric	Metric Value
[18]	Logistic Regression (LR)	-Log-returns of stock prices (3M,6M,1Y lags) -ETE network indicators	Accuracy	~50–60%
	Multilayer Perceptron (MLP	-Log-returns -ETE network indicators -Normalized features		~54–60%
	Random Forest (RF)	-LOG-returns -ETE network indicators -Random feature subsets		~54–63%
	XGBoost (XGB)	-LOG-returns -ETE network indicators •Tree-based feature importance	same as above	~53-60%
	LSTM Network	-LOG-returns (sequential) -ETE network indicators -TIME-series normalization		~53-60%

3.3.4 China's equity funds

By summarizing the above part of the applied models for different stock markets, we find that: the machine learning model significantly outperforms the traditional linear model.

3.4 Feature engineering

3.4.1 Stock market prediction (short-term trading)

Applicable scenarios: high-frequency trading, short-term strategies (e.g., intraday or intraweek trading)
Recommended feature engineering:
a. Technical indicators: momentum (MOM), volatility (VOL), RSI, MACD, Bollinger bands, etc.
b. Volume and price data: volume, bid-ask spread, order book depth.

Table 5 China's equity funds models

	Model Name	Feature Engineering Methods	Evaluation Metric	Metric Value
[21]	Linear Regression (OLS)	-30 fund anomaly factors -Standardized cross -sectional factors	Monthly return FF3-alpha Sharpe ratio	1.575% (NW-t=2.50) 0.580%(NW-t=2.44) 0.783
	LightGBM	-Same 30 anomaly factors -Sliding window training (12 months)	Same as above	1.688% (NW-t=2.65) 0.727% (NW-t=3.07) 0.854
	Random Forest (RF)	-Same 30 anomaly factors -Bagging-based ensemble learning -Non-linear feature interactions		1.641% (NW-t=2.62) 0.664% (NW-t=2.70) 0.833

c. Market Sentiment: News Sentiment Analysis (NLP), Social Media Heat (e.g. Reddit/Twitter discussion volume).
d. Time series processing: sliding window normalization, difference processing (removing non-stationarity).
Recommendation modeling:
LSTM (capturing sequence dependencies), LightGBM/XGBoost (handling high-dimensional features).

3.4.2 Fund return prediction (medium and long-term investment)

Applicable scenarios: public/private fund portfolio optimization, pension target fund screening
Recommended Characterization Engineering:
a. Fund Anomaly Factors: active share (Active Share), Sharpe Ratio, maximum retracement, money flow (Flow).
b. Position Derivatives: Industry Concentration, Position Valuation (PE/PB), Position Adjustment Frequency.
c. Macroeconomic correlations: interest rate, inflation rate, GDP growth rate (to be standardized and included).
Recommendation Model:
LightGBM/RF (dealing with non-linear relationships), Fama-Macbeth regression (factor significance test).

3.4.3 Multi-factor stock selection (quantitative investment)

Applicable scenarios: CSI 300/CSI 500 constituent stock screening
Recommended feature engineering:
a. Fundamental factors: ROE, revenue growth rate, gearing ratio.
b. Technical factors: momentum, volatility, turnover rate in the past 12 months.
c. Alternative data: changes in institutional holdings, signals from executives to increase/reduce holdings.
d. Dimensionality reduction processing: PCA (covariance reduction), factor orthogonalization.
Recommendation modeling:
XGBoost (feature importance ranking), Stacking integration (combining SVM/RF prediction results).

3.4.4 Market Crisis Early Warning (Risk Management)

Applicable scenarios: financial crisis, black swan event monitoring
Recommended feature engineering:
a. Volatility indicators: VIX index, tail risk (VaR).
b. Liquidity indicators: bid-ask spread, market depth.
c. Correlation networks: correlation breakdown between stocks/industries (with Granger causality or ETE).
Recommended models:
Random Forest (detecting nonlinear anomalies), Isolated Forest (unsupervised anomaly detection).

3.4.5 Cross-market asset allocation (global macro strategy)

Applicable scenarios: stock bond commodity multi-asset portfolio
Recommended feature engineering:
a. Cross-market spreads: equity and bond correlation, commodity and exchange rate linkage.
b. Macroeconomic cycle: Merrill Lynch Clock Indicators (inflation/growth quadrant).
c. Geopolitics: event dummy variables (e.g., wars, trade frictions).
Recommended Models:
VAR models (multivariate time series), Graph Neural Networks (GNN) (capturing topological relationships between

markets).

4 Research challenges and future directions

The challenges and future directions of mainstream methods are as follows: first, the lack of cross-market generalization ability, the need to explore the adaptation mechanism of the model in different market structures and trading rules; second, the integration of high-frequency and low-frequency data, the organic combination of macroeconomic variables and micro trading signals still needs to be broken through; third, the combination of real-time decision-making and reinforcement learning, how to dynamically update the factor weights in the execution of the strategy and control the transaction costs need to be studied; fourth, the combination of real-time decision-making and reinforcement learning, how to dynamically update the factor weights and control the Third, the combination of real-time decision-making and reinforcement learning, how to dynamically update factor weights during strategy execution and control transaction costs need to be studied; fourth, interpretability and regulatory compliance, the need to further develop hybrid models that can take into account both predictive performance and causal inference. In the future, we can continue to improve the theoretical and empirical foundation of the quantitative investment system along the three-step strategy of "intelligent factor screening - nonlinear dynamic regulation - interpretability".

References

[1] Sharpe W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, *19*(3), 425-442.

[2] Ross S. A. (1976). The arbitrage theory of capital asset pricing. Journal of Economic Theory, *13*(3), 341-360.

[3] Fama E. F., & French K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, *33*(1), 3-56.

[4] Carhart M. M. (1997). On persistence in mutual fund performance. Journal of Finance, *52*(1), 57-82.

[5] Fama E. F., & French K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, *116*(1), 1-22.

[6] Li Z., Liu H., & Dong X. (2017). Zhongguo A gu shichang de wu yinzi moxing yanjiu [A five-factor model in China's A-share market]. Journal of Financial Research, *12*, 191-206.

[7] Xu C. C. (2020). Gupiao jiage yuce fangfa zongshu [A review of stock price prediction methods]. Zhongguo Shichang [China Market], (9), 42-43+68.

[8] Vapnik V. N. (1998). Statistical learning theory. Wiley.(For SRM principle specifically, cite this book as it contains the definitive formulation)

[9] Breiman L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[10] Krauss C., Do X. A., & Huck N. (2017). Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. European Journal of Operational Research. 259(2), 689-702.

[11] Amin M. (2000, June). Stock market prediction via support vector machines: A NASDAQ case study. Paper presented at the IEEE/IAFE/ INFORMS Conference on Computational Paper presented at the IEEE/IAFE/INFORMS Conference on Computational Intelligence for Financial Engineering (CIFEr), New York, NY, USA.

[12] Patel J., Shah S., Thakkar P., & Kotecha K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. expert Systems with Applications, 42(1), 259-268.

[13] Li B., Wang X., & Chen Y. (2019). Jiqixuexi qudong de jibenmian lianghua touzi yanjiu [Machine learning-driven fundamental quantitative investment research]. Zhongguo Gongye Jingji [China Industrial Economics], *2019*(8), 45-62.

[14] Cao C.-F., Luo Z.-N., Xie J.-X., & Li L. (2022). MDT-CNN-LSTM moxing de gujia yuce yanjiu [Stock price prediction using the MDT-CNN-LSTM model]. Jisuanji Gongcheng yu Yingyong [Computer Engineering and Applications], *58*(5), 280-286.

[15] Bao W., Yue J., & Rao Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLOS ONE, *12*(7), e0180944.

[16] Gyamerah S. A., Ngare P., & Ikpe D. (2021). Ensemble machine learning for stock market prediction: a case study of the Nairobi Securities Exchange. expert Systems with Applications, *185*,. 115591.

[17] Chen L., & Wei Y. (2020). Heterogeneous ensemble learning for stock selection in China's A-share market. Expert Systems with Applications, * 158*, 113516.

[18] Kim S., Ku S., Chang W., & Song J. W. (2020). Predicting the direction of US stock prices using effective transfer entropy and machine learning techniques. IEEE Access, 8, 111660-111682.

[19] Song L. (2022). Integrated learning model and system implementation for multi-factor mining of stocks [Master's thesis, University of Electronic Science and Technology of China]. China National Knowledge Infrastructure.

[20] Xiang X. (2022). Research on multi-factor quantitative stock selection strategy based on machine learning [Master's thesis, Zhongnan University of Economics and Law]. .

[21] Li R., & Ye Z. (2023). Machine learning-based prediction of fund returns. Statistics & Decision, *39*(11), 156-161.