International Journal of Innovative Computing, Information and Control
Volume **21**, Number **3**, June **2025**

ICIC International ©2025 ISSN 1349-4198
pp. 731–746

A MULTI-FACTOR QUANTITATIVE STOCK SELECTION STRATEGY BASED ON TABNET-BOOST

Jianming Li[1], Yuhan Wang[1], Yang Wang[2] and Xiangpei Hu[3]

[1]School of Computer Science
[2]Audit Office
[3]School of Management and Economics
Dalian University of Technology
No. 2, Linggong Road, Ganjingzi District, Dalian 116023, P. R. China
{ lijm; wangyuhan418; wangyang1512; drhxp }@dlut.edu.cn

Received September 2024; revised December 2024

Abstract. *In this paper, we address issues in the field of quantitative stock selection research, where traditional neural network methods face challenges such as overfitting, difficulty in feature selection, and low training efficiency when dealing with high-dimensional heterogeneous financial stock data. We propose a multi-factor quantitative stock selection algorithm based on TabNet-Boost. Our methodology involves two primary stages: feature extraction and model enhancement. In the feature extraction stage, we employ the TabNet tabular neural network to perform sparse selection and representation learning, efficiently identifying and capturing key features from the high-dimensional data. In the model enhancement stage, we integrate the extracted features with multiple weak learners using the Boost ensemble learning algorithm to improve predictive accuracy and generalization performance. Experimental results from training on stock pools and factor pools in the A-share market demonstrate that our algorithm achieves an annualized return significantly higher than the CSI 300 Index over the same period, reaching 29.42%. Additionally, the proposed algorithm shows advantages in terms of strategy stability and model interpretability, evidenced by reduced volatility and improved explanatory power of the model.*

Keywords: Multi-factor stock selection, TabNet, Ensemble learning, Tabular neural network, Feature extraction

1. **Introduction.** Quantitative multi-factor stock selection is an investment strategy based on statistical and mathematical models. It aims to systematically analyze and explore multiple factors (or features) in the stock market to identify and select individual stocks with high profit potential. Currently, the field of quantitative multi-factor stock selection is undergoing rapid development and evolution. It benefits from the rapid advancement of data science and machine learning technologies, enabling investors to better utilize techniques such as big data analysis, machine learning, and deep learning to discover effective stock selection factors and build precise investment models.

Currently, there are many machine learning and deep learning models used in multi-factor quantitative stock selection. Unlike in stock price prediction, LSTM (Long Short-Term Memory), CNN (Convolutional Neural Networks), and other neural networks do not perform optimally in the field of multi-factor stock selection based on heterogeneous tabular data. This is because the correlations between tabular data are complex, and neural networks are not adept at handling data where predictions are heavily influenced by individual features. Ensemble learning based on decision trees mitigates model generalization

DOI: 10.24507/ijicic.21.03.731

errors and enhances overall prediction accuracy by constructing and combining predictions from multiple base learners. It effectively handles diverse features and demonstrates outstanding performance in the field of tabular data. It is widely utilized in practical quantitative trading due to its ability to handle complex data structures and improve model robustness [1]. In addition to ensemble learning, there are also neural networks specifically designed to handle tabular data, such as TabNet [2]. It has found applications in the finance domain. TabNet inherits the sparse selection advantage of tree models and the representation learning capability of DNN (Deep Neural Networks), enabling it to handle tabular data better without any preprocessing, especially in multi-factor quantitative trading data. However, traditional TabNet may overlook or lose some features when dealing with high-dimensional dense tabular data [3,4]. When the data dimensionality is high, such as when there are more than thirty factor data, its prediction accuracy may decrease, making it unable to adapt to the dynamic and complex nature of the stock market.

To address this issue, this paper integrates the traditional TabNet model with Boost ensemble learning models through stacking to construct the TabNet-Boost model. This integration resolves the traditional TabNet model's poor performance on high-dimensional data and alleviates data overfitting. By combining our constructed model with stock selection trading methods, we conduct training and backtesting on stock pools and factor pools, ultimately forming a comprehensive and reliable multi-factor quantitative stock selection strategy.

The organization structure of this paper is as follows. The first part provides a brief introduction to the relevant context of the study. The second part presents an overview of the relevant research in the field of multi-factor quantitative stock selection, including the work of scholars and our own contributions. The third part details the data construction process used in this study. The fourth part elaborates on the construction of the multi-factor quantitative stock selection strategy based on TabNet-Boost. The fifth part conducts comparative experiments. The sixth part presents the conclusion of the study.

2. **Related Work.**

2.1. **Literature review.**

2.1.1. *Factor mining and screening.* In quantitative investing, factors refer to various quantifiable indicators that influence stock prices or market trends. They are used to construct investment strategies and models to assist investors in decision-making. Currently, there are hundreds of available factors in the field of multi-factor quantitative stock selection. They are essential components of stock market prediction. Factors related to the stock market typically include style factors, fundamental financial factors [5], volume-price factors [6], technical factors [7], industry factors [8], and alternative data (such as news, emails, and social media) [9-11], among others. These mature market factors have become important metrics and trading signals in stock trading. Selecting appropriate factors is crucial for developing a successful multi-factor quantitative stock selection strategy. Therefore, how to scientifically and efficiently select factors has become an important research direction in the field of quantitative investment [12].

In traditional stock trading, investors often prioritize fundamental and financial information of stocks. However, with the evolution of machine learning technology and quantitative trading concepts, market trends often exhibit certain patterns of volume and price changes [13]. Through in-depth analysis of the relationship between price and trading volume, investors can uncover potential market trends and reversal points, which helps them capture market opportunities and mitigate risks.

2.1.2. *Ensemble learning and multi-factor investing.* Ensemble learning is a machine learning method that combines multiple individual learning models (called base learners or weak learners) to accomplish learning tasks. These base learners can be generated by the same or different learning algorithms, and they can be generated in parallel or sequentially. The core idea of ensemble learning is to reduce model generalization error and improve overall prediction accuracy by combining the predictions of multiple base learners. The key to this approach lies in the diversity among the learners, which allows them to make different predictions on the training data. These diverse predictions complement each other when integrated, resulting in better performance than a single learner. Common ensemble methods include bagging, boosting, and stacking. Due to its ability to compensate for the limitations of individual models and mitigate the risk of overfitting, ensemble learning techniques are widely used in practical quantitative trading, particularly in the field of multi-factor stock selection. Its effectiveness is sometimes even superior to neural network models. Zhang and Chen have developed a multi-factor stock selection model based on XGBoost, which has demonstrated good performance in the Shanghai and Shenzhen 300 stock pool [14]. Li et al. have proposed a training and optimization model called LightGBM-Bayes, which has been tested on the constituents of the Shanghai and Shenzhen 300 index through backtesting. The results have outperformed the benchmark index [15]. Shi et al. proposed a hybrid model based on attention mechanism, combining CNN-LSTM and XGBoost, to predict stock prices. This model can effectively exploit historical information from multiple periods in the stock market [16]. Ren constructed a short-term stock selection strategy using the Gradient Boosting Decision Tree (GBDT) and the Gradient Boosting Rank (GBRank) algorithm [17]. Zhang combined Feedforward Neural Network (FFNN) with LightGBM to establish a hybrid model, which achieved good performance on real data for the next three months [18]. Ampomah et al. compared the effectiveness of tree-based ensemble learning models (Random Forest, XGBoost, Bagging, AdaBoost, and Extreme Random Forest) in predicting stock price trends. They applied these models to eight different stock datasets from the New York Stock Exchange and the NASDAQ Stock Exchange, evaluating them based on metrics such as accuracy, precision, recall, F1 score, and AUC-ROC curve. The results showed that the AdaBoost model performed the best [19]. In this study, we deeply explored the advantages and role of ensemble learning, reducing the complexity of complex neural network models. Additionally, by using the boosting idea, we alleviated the burden of processing high-dimensional tabular data, which complements well with tabular neural networks. This provides a solid theoretical foundation for our subsequent use of tabular neural networks as the base model for integration.

2.1.3. *Research on tabular data models and TabNet.* Heterogeneous tabular data is the most commonly used data format in the fields of multi-factor investing and financial risk management [20]. In this domain, improving predictive performance and robustness is highly beneficial for end-users and fund securities companies providing solutions [21].

Many scholars have conducted research on why traditional neural network models perform poorly on financial tabular data. One of the reasons is attributed to poor data quality, which includes missing values and outliers, resulting in imbalanced tables. This makes most machine learning methods ineffective. However, algorithms based on decision trees can handle internal missing values or extreme variable ranges by finding appropriate approximation and splitting values [22-24]. Secondly, the feature correlations between tabular data are complex and irregular. The inductive bias of convolutional neural networks to learn the structure and relationships between features from scratch is significant and time-consuming engineering [25-27]. Lastly, the prediction results of tabular data are

heavily influenced by individual important features. Compared to neural networks, decision tree models can effectively handle different feature importance by selecting individual features and appropriate thresholds while ignoring the rest of the data samples [28].

This shows that despite the increasing complexity and richness of parameters in neural network structures, ensemble learning based on decision trees often exhibits superior performance in training and prediction tasks for tabular data in quantitative multi-factor stock selection [29].

With the tremendous success of attention mechanisms in Computer Vision (CV) and Natural Language Processing (NLP) domains, some scholars have recently introduced attention mechanisms into the processing of heterogeneous tabular data. TabNet is one of the earliest table data models based on attention mechanisms. It employs sequential attention to select features inferred at each step, thereby achieving interpretability and efficient learning effects. The research on TabNet in the field of quantitative trading or other fields has been rich in recent years. Wei et al. used the feature selection function of TabNet in combination with LSTM to predict the stock price and obtained good prediction results [30]. Wang used the TabNet feature selection importance method to enhance feature gain in stock data, facilitating better training for downstream methods [31]. McDonnell et al. leveraged TabNet's interpretability in the insurance sector to provide a comprehensive and scientific explanation of the risk factors influencing insurance predictions [32]. Joseph et al. combined optimized Bayesian methods with TabNet to predict the probability of diabetes, enhancing accuracy while improving interpretability, which holds significant practical value [33].

It can be said that the sequential attention mechanism of tabular neural networks, or TabNet, is well-suited for high-dimensional multi-factor quantitative stock selection scenarios. In such scenarios, a large amount of training and tabular data generation is required, and the tree-based manifold mechanism can provide highly accurate predictions.

2.2. **Our contributions.** In terms of factor mining, this paper combines traditional methods with machine learning approaches for multiple rounds of factor screening and mining. It aims to reduce the dimensionality and intercorrelation of factor features while maintaining their effectiveness. Traditional factor analysis provides interpretability to the factors, while machine learning feature importance selection maximizes the factors' ability to represent the causes of stock price changes.

In terms of algorithmic models, we propose the TabNet-Boost model. For the specific context of multi-factor quantitative stock selection, we construct high-dimensional heterogeneous training data of stocks and factors using scientific methods. This enables the TabNet model to focus on different factor features in each round of feature selection, thereby not only exploring the reasons affecting stock price trends but also allowing clear observation of the model training process and feature importance using masks, which enhances interpretability and increases investor trust in the model. In our approach, the TabNet model, trained on heterogeneous tabular data, serves as the base model and is integrated with Boost family models using Stacking. The Boost family models mitigate the burden of high-dimensional data processing on TabNet, thereby leveraging TabNet's strengths in feature selection. We employ TabNet, XGBoost, LightGBM, and CatBoost as the first-layer models, each trained on the initial dataset. The reason for using Boosting ensemble learning algorithms is that they offer better stability and prediction performance. Moreover, for prediction tasks involving heterogeneous tabular data, these algorithms can leverage the characteristics of decision tree manifolds for precise predictions and provide better interpretability compared to other neural networks. XGBoost, LightGBM, and CatBoost have become the most widely used boosting models

due to their efficient training processes, built-in handling of missing values, support for large-scale data and categorical features, strong generalization capabilities, and extensive tuning options. These advantages enable them to excel in various practical applications. In contrast, traditional GBDT and AdaBoost are less commonly used because they fall short in terms of speed, memory efficiency, and sensitivity to noise. The outputs of the first-layer models, along with the original dataset labels, are used to generate a new dataset to train a second-layer logistic regression model to obtain the final result. This approach addresses the issue of performance degradation in the original TabNet model when handling high-dimensional dense tabular data, harnesses the advantages of each base model, reduces the risk of overfitting, enhances interpretability, and improves overall model performance and stability. It makes the model more flexible and adaptable to different data scenarios. We train and backtest our constructed model combined with stock selection and trading methods on the stock pool and factor pool, ultimately forming a comprehensive and reliable multi-factor quantitative stock selection strategy. Experimental results demonstrate that our strategy outperforms the standalone TabNet model, with improvements in both accuracy and annualized return.

3. Data Construction.

3.1. Construction of the stock pool. The A-share market has a large number of retail investors and is significantly affected by policy fluctuations. Therefore, it is crucial to select stocks with sound fundamentals before training the machine learning model. Based on the "CITIC Securities Industry Classification Standard 2.0 and Revision Instructions" published by the CITIC Securities Research Department in 2020, this paper selects 68 A-share stocks from various sectors classified under the first-level industries of CITIC. The selection criteria are as follows.

1) Cover as many as possible of the 30 first-level industries to ensure the stock pool is broadly representative and resilient to sudden financial events.
2) Exclude stocks that are about to be delisted.
3) Exclude ST (Special Treatment) stocks.
4) Filter out Growth Enterprise Market (GEM) stocks.
5) Select stocks with solid fundamentals.
6) Ensure the stock pool covers a variety of styles to cope with complex market conditions.

3.2. Construction and mining of the factor pool. The construction of the factor pool emphasizes creating a comprehensive library of factors from both macro and micro perspectives, minimizing data preprocessing to ensure that critical information is not lost. Currently, there are numerous methods for factor analysis and selection in academia, ranging from simple statistical methods and regression methods to single-factor analysis. In past research, scholars often aimed to build extensive and broad factor pools to ensure comprehensive coverage of factor types, followed by manual factor selection. This selection from a financial investment perspective often results in a factor reserve of 50 to 100 factors in the pool. However, over time, many of these factors may become ineffective, leading to high-dimensional data that can cause overfitting, unnecessary memory consumption, increased computational costs, and reduced performance of learning algorithms. There are two methods to address this issue: feature extraction and feature selection. Feature extraction constructs new factors from combinations of original factors but sacrifices the interpretability of the original features. Feature selection, on the other hand, retains the interpretability of the original features and preserves important features while removing irrelevant and redundant ones. This improves learning efficiency and predictive accuracy [34].

To enhance the interpretability of the learning results, this paper innovatively combines traditional factor analysis with ensemble learning feature importance ranking, proposing a sequential three-stage factor selection method. This approach not only addresses the overfitting and potential factor failure issues associated with traditional factor analysis methods but also improves factor interpretability, aiming to uncover the various factors influencing stock price movements as comprehensively as possible. The specific approach involves obtaining a comprehensive set of stock factors from JQData, BigQuant, and Wind databases. This includes 69 individual volume-price factors, turnover rate factors, technical factors, per-share factors, capital flow factors, shareholder factors, valuation factors, volatility factors, and financial factors, as well as 50 complex volume-price factors, to construct the initial factor pool. The factors are then subjected to a three-stage selection process:

1) First Stage: Initial screening through factor analysis;
2) Second Stage: Calculation of Pearson correlation coefficients for further screening;
3) Third Stage: Feature selection using ensemble learning algorithms, integrating volume-price factors to identify the most impactful factors on returns.

After these three rounds of screening, the final factor pool is formed. The selection process is illustrated in Figure 1. As illustrated in Figure 1, the factor data is initially divided into two parts: platform-acquired factors and volume-price factors. For the platform-acquired factors, the process starts with the first stage, where 69 factors undergo factor analysis. These factors are evaluated and scored based on multiple criteria, including return performance, return curve divergence, IC mean, IC standard deviation, IR value, the ratio of $|\mathrm{IC}| > 0.02$, and turnover rate. This comprehensive evaluation narrows down the list to 49 factors. In the second stage, Pearson correlation analysis further refines this to 31 factors. Finally, in the third stage, ensemble learning feature importance ranking selects the top 12 factors. The volume-price factors, on the other hand, directly undergo the third stage of feature importance ranking, resulting in the selection of 18 factors. Together with the previously selected 12 factors, these form the final factor pool of 30 factors. This sequential feature selection approach enables the identification of significant features without losing interpretability. Selecting from the set of important features enhances the robustness of these key features and reduces the dimensionality of the data.

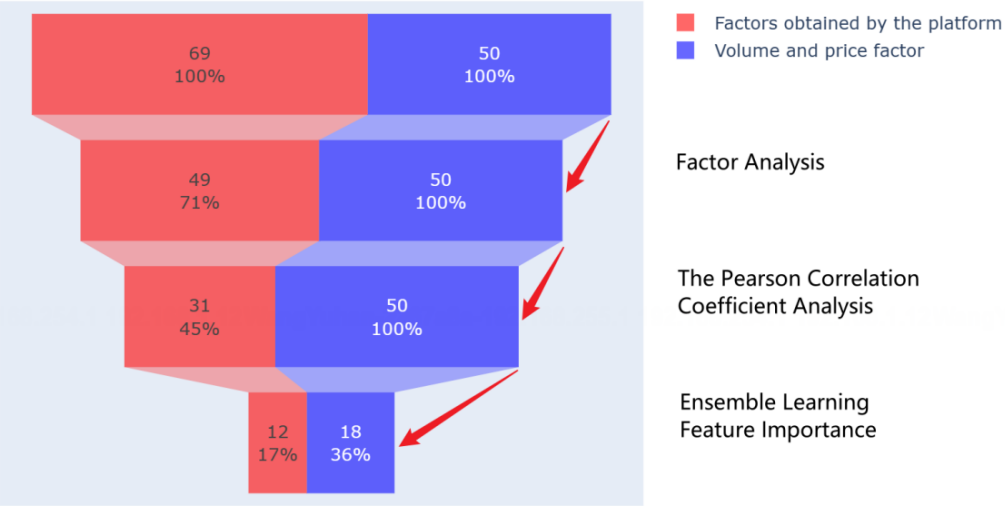


FIGURE 1. Factor screening process

4. Construction of a Multi-Factor Quantitative Stock Selection Strategy Based on TabNet-Boost Ensemble Learning.

4.1. **Algorithm design.** For the specific context of multi-factor quantitative stock selection, we designed a TabNet framework suited for high-dimensional data prediction by integrating factor pools and stock pools. This framework employs a sequential attention mechanism to select which features to infer at each decision step. The design of the attention mechanism allows the classifier to focus on important parts of the input, such that each step's prediction concentrates on changes in specific factors (with different types of factors being focused on in each round). This not only enables exploration of the reasons influencing stock price trends but also allows clear observation of the model training process and feature importance using masks, enhancing interpretability and increasing investor trust in the model. As illustrated in Figure 2, TabNet selects features from stock factor data. As shown, TabNet uses multiple decision blocks, each focusing on processing subsets of input features used for inference. According to the sequential attention mechanism, the first decision selects factor features related to volume and price, and processes them as input. The second decision focuses on valuation-related factor features, and processes them as input. The third decision selects financial-related factor features, and processes them as input. The aggregated information from these predictions is then output as the final prediction result.

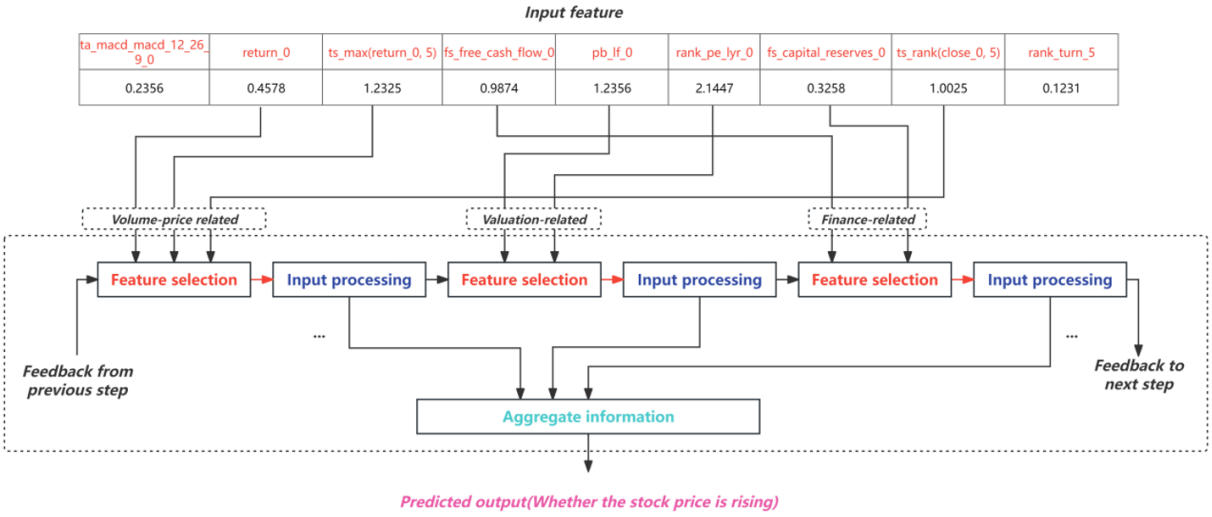


FIGURE 2. TabNet's feature selection in stock factor data

Figure 3 shows the overall structure of the TabNet encoder, where the Feature Transformer and Attentive Transformer modules, illustrated in Figures 4 and 5 respectively, are the two most crucial components. The Feature Transformer module is responsible for feature computation, while the Attentive Transformer module handles feature selection.

It can be seen that the Feature Transformer consists of two parts: the parameter-shared part (shared across decision steps) and the parameter-independent part (decision step dependent). Each part is composed of fully connected layers, batch normalization layers, and gated linear units.

The function of the Attentive Transformer is to perform feature selection by learning a "soft" mask, ensuring that each decision step focuses on relevant features rather than wasting effort on irrelevant ones, thereby enhancing the model's efficiency. Prior scales are used to inform the model about the extent to which a feature has been utilized in historical decisions. For instance, if the previous round primarily focused on features related

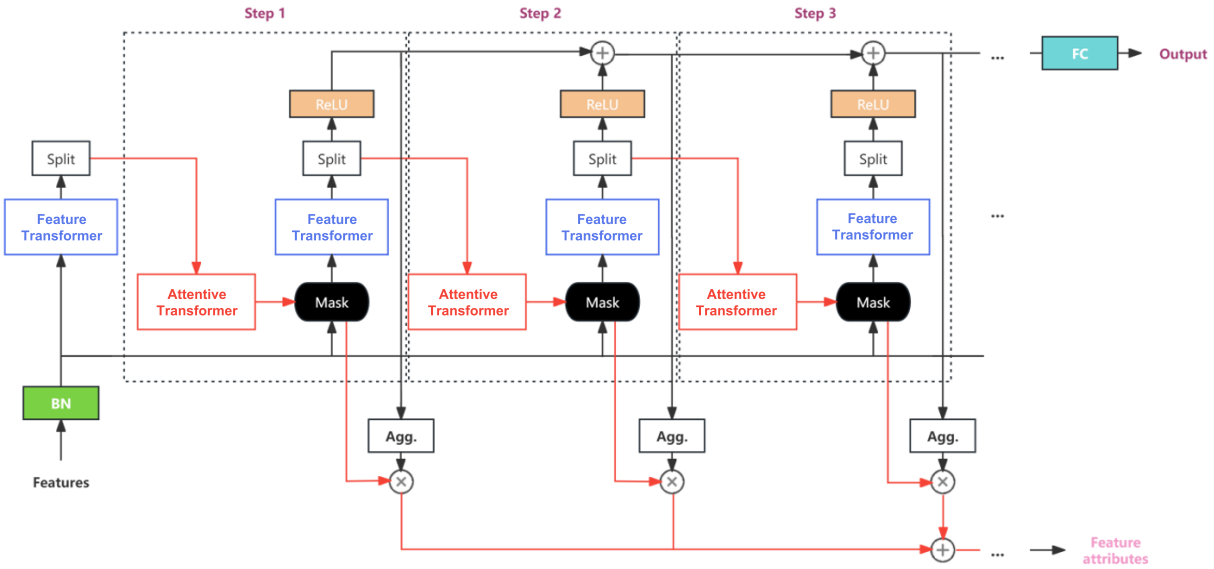


FIGURE 3. The overall structure of the TabNet encoder

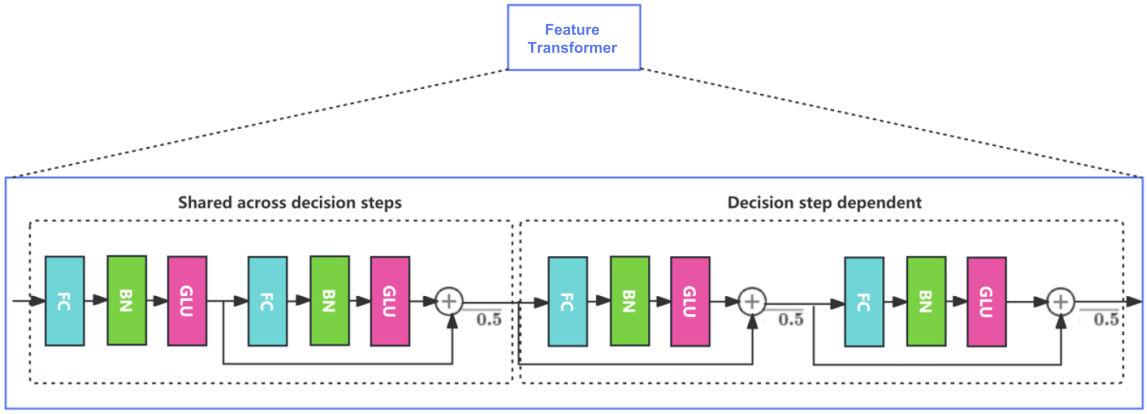


FIGURE 4. The main structure of Feature Transformer

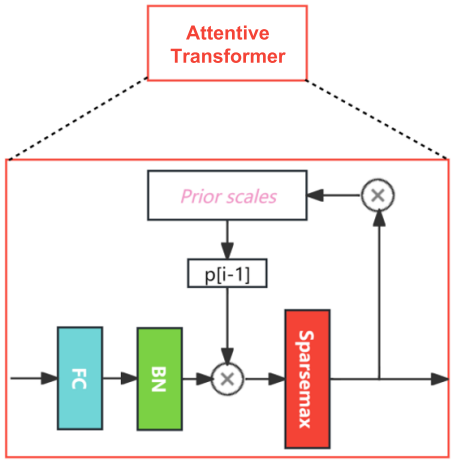


FIGURE 5. The main structure of Attentive Transformer

to financial factors, the current round might focus on technical or other dimensional features. Since different stocks may be more influenced by macroeconomic factors or financial factors at different times, this characteristic of TabNet allows for "adaptive" training of sample features. In contrast, the feature selection in a decision tree of a tree model is based on all samples and cannot adapt to individual samples.

The TabNet model designed for the field of multi-factor quantitative stock selection balances interpretability, which may limit its performance in handling high-dimensional dense tabular data. When dealing with such data, the model's feature selection mechanism needs to filter a large number of features, potentially leading to the oversight or loss of some important features. Consequently, as the data dimensionality increases, there is a decline in prediction accuracy.

In the specific context of multi-factor quantitative stock selection, the dimensionality and interaction correlation of factor features significantly impact TabNet's predictive performance. Despite employing innovative factor mining and screening techniques to minimize the dimensionality and interaction correlation of factor features, high-dimensional feature scenarios are still inevitable. Furthermore, consistently spending extensive time and effort on factor feature dimensionality reduction and other feature engineering tasks contradicts the ensemble learning principle of "reducing the burden of feature engineering".

In this paper, we innovatively integrate TabNet with Boost models using a stacking ensemble approach. The reason for this integration is their complementary strengths. Boost models combine multiple weak classifiers to flexibly handle feature selection and combination, reduce the risk of overfitting, classify based on decision paths, and effectively process high-dimensional, dense tabular data. They are well-suited for structured data and help minimize feature engineering while reducing overfitting. TabNet, on the other hand, excels in feature selection and dimensionality reduction. By leveraging the strengths of both models, the overall predictive accuracy is improved.

The stacking ensemble method involves layering the TabNet model and Boost models. First, base models such as TabNet, XGBoost, LightGBM, and CatBoost are used as the first layer models. These models are trained on the initial dataset. The outputs of these first-layer models, along with the original dataset labels, are then used to create a new dataset for training the second-layer model. This approach takes advantage of the strengths of each base model, reduces overfitting, enhances interpretability, and further improves the model's performance and stability. It makes the model more flexible and adaptable to different data scenarios, yielding promising results in the multi-factor quantitative stock selection context. Figure 6 shows the model structure.

The first-layer models consist of TabNet and Boost family models, including XGBoost, LightGBM, and CatBoost. Processed high-dimensional heterogeneous tabular data constructed from stock and factor data are input into the TabNet model for training. Within TabNet, the feature computation module calculates shared features among stock factors, such as volume-price features, at the shared parameter layer, and computes different features for each decision round, such as financial factors, at the non-shared parameter layer. Additionally, the feature selection module learns different feature factors influencing stock price trends under the influence of different factors in each round. The trained data and labels are further used for joint training with XGBoost, LightGBM, and CatBoost models, respectively, to obtain the stock price movements under each ensemble learning model.

The second-layer model is a logistic regression model. It linearly combines the outputs from the first-layer model Stacking. The regression model analyzes the predictions of the first-layer models on stock price movements, thus achieving better classification results.

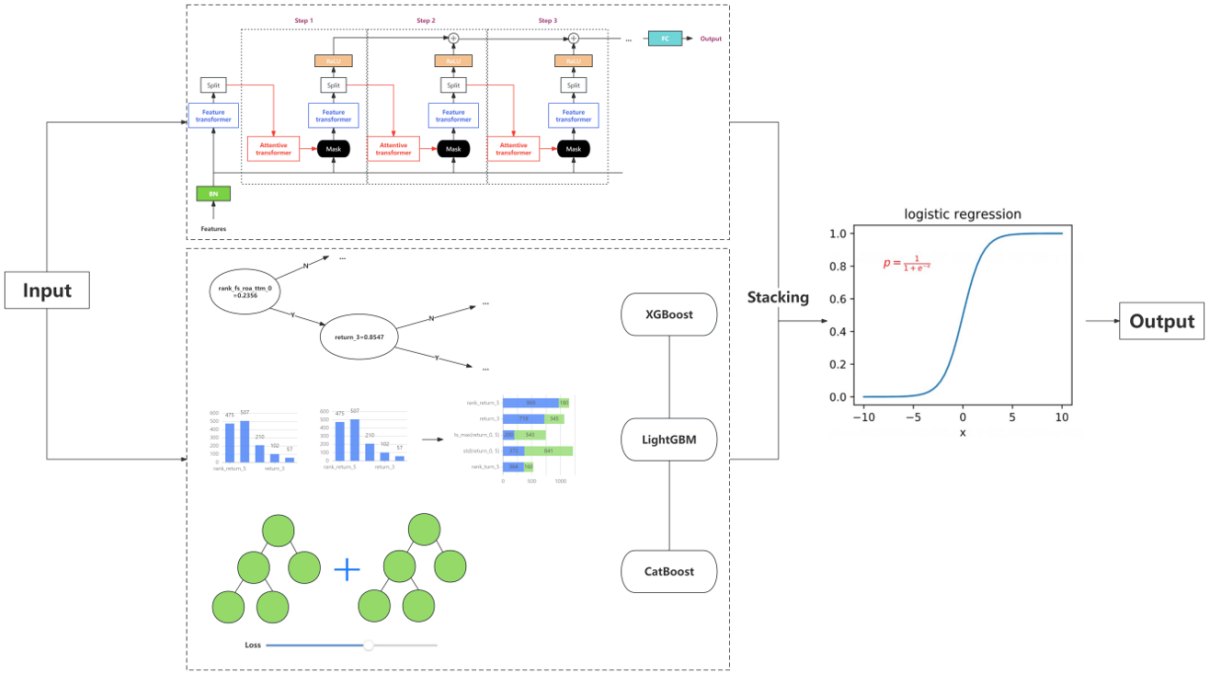


FIGURE 6. Model structure diagram

It possesses the characteristics of simplicity, speed, and stability. Table 1 depicts the TabNet-Boost Stacking algorithm process.

This completes the design of the quantitative multi-factor stock selection model based on TabNet-Boost ensemble learning method.

4.2. **Trading strategy design.** A scientific and comprehensive multi-factor stock selection model includes not only research on machine learning model methods but also the formulation of trading strategies. An excellent trading strategy can not only bring excess returns but also guide future trading styles. Combining some excellent trading strategies with A-share trading rules, this paper formulates the following trading strategy.

1) First, collect data on stocks in the stock pool and factors in the factor pool. Specify the start time of the training set: March 1, 2017, to March 1, 2022, and the start time of the test set (backtesting time): March 1, 2022, to March 1, 2024.

2) Label the stock data and calculate the returns: Calculate the 5-day closing price (as the selling price) divided by the opening price of the next day (as the buying price). If the future 5-day return is positive, set the label to 1; otherwise, set the label to 0. Use quantiles to handle outliers. Filter out cases of limit-up trading, as buying at limit-up prices is difficult, and the subsequent trend is unpredictable, posing high risks.

3) Collect and extract basic factors and derivative factors for the corresponding time.

4) Handle missing data.

5) Train the stock data and factor data using machine learning models.

6) Sort the predicted probabilities of the trained data in descending order.

The rebalancing period is set to 5 days, with buying at the opening price and selling at the closing price. During rebalancing, buy the top 7 stocks ranked by probability, and buy them with equal weights. If the stocks that need to be bought according to the algorithm are already in the current holdings, continue to hold; otherwise, sell. Other trading details are shown in Table 2.

TABLE 1. TabNet-Boost Stacking algorithmic process

Algorithm: TabNet-Boost Stacking

Input: X_train, y_train
Output: final_predictions

```
1. # Define basic learners
base_learners = [clf1: TabNet, clf2: XGBoost, clf3: LightGBM, clf4: CatBoost]
2. # Define meta learner
meta_learner = meta_clf
3. # Divide the data set into training set and test set
X_train_base, X_meta_train, y_train_base, y_meta_train = train_test_split(X_train,
y_train, test_size = 0.5, random_state = 42)
4. # Initialize element feature matrix
meta_features = np.zeros((X_meta_train.shape[0], len(base_learners)))
5. # For each basic learner
   for i, TabNet, XGBoost, LightGBM, CatBoost in enumerate(base_learners):
      # Perform K-fold cross validation on the training set
      for each fold in KFold(X_train_base):
         # Train a basic learner on the current fold
         train_base_learner(clf, X_train_fold, y_train_fold)
         # Predict and fill the meta-feature matrix on the current fold
         meta_features[test_index, i] = predict(clf, X_test_fold)
6. # Training a meta-learner using a meta-feature matrix
train_meta_learner(meta_learner, meta_features, y_meta_train)
7. # Make predictions on the test set
meta_features_test = np.zeros((X_test.shape[0], len(base_learners)))
for i, clf in enumerate(base_learners):
     meta_features_test[:, i] = predict(clf, X_test)
8. # Use a meta-learner to make final predictions
final_predictions = predict(meta_learner, meta_features_test)
```

TABLE 2. Trading strategy design parameters

Indicator name	Indicator value
Initial funding	¥1000000
Position rebalancing cycle	5 days
Backtest price type	Restoration of rights later
Position limit	Total account value $*$ 0.2
Transaction rate limit	0.025
Whether to calculate handling fees and slippage	Yes
Backtest data frequency	Daily

Thus, a scientifically complete quantitative multi-factor stock selection strategy based on TabNet-Boost ensemble learning, in accordance with A-share rules, has been constructed. In order to provide readers with a clearer understanding of the strategy construction process, the strategy's construction flowchart is shown in Figure 7.

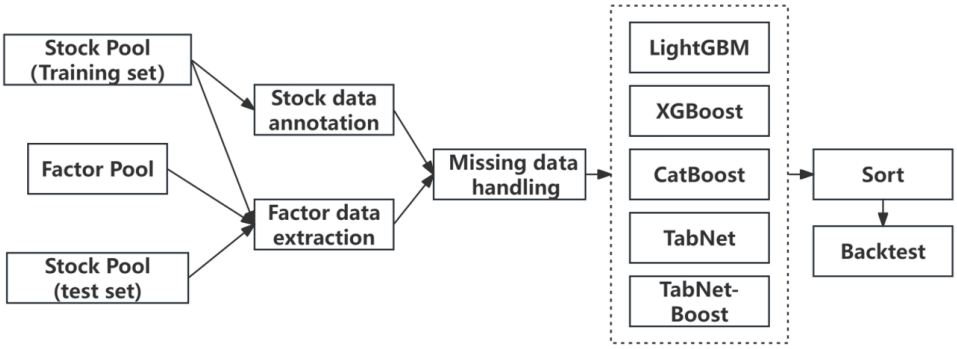


FIGURE 7. Flowchart for trading strategy building

5. Experimental Section.

5.1. **Experimental design and results.** Based on the algorithm design and trading strategy outlined in Section 4, six control experiments were designed: XGBoost, LightGBM, CatBoost, AdaBoost, TabNet, and Buy and Hold.

We constructed seven quantitative trading strategies based on TabNet-Boost, TabNet, XGBoost, LightGBM, CatBoost, AdaBoost algorithms, and Buy and Hold strategy respectively. Table 3 shows the specific returns and risks of these seven quantitative trading strategies, Figure 8 displays the profit curves of the seven strategies, and Figure 9 presents the profit curves of TabNet and TabNet-Boost strategies. In an environment where the Shanghai and Shenzhen 300 Index only achieved a return of −22.78% from March 1, 2022, to March 1, 2024, quantitative trading strategies based on Boost algorithms demonstrated excellent performance. Particularly, our proposed TabNet-Boost quantitative multi-factor trading strategy achieved an annualized rate of return of 29.42%, exceeding the benchmark return by 52.2%, and outperforming the Buy and Hold strategy return by 30.04%. Moreover, the Sharpe ratio of the TabNet-Boost strategy is the highest, indicating its superior risk-adjusted performance compared to other strategies. Additionally, the TabNet-Boost strategy has the smallest maximum drawdown among all strategies except for the Buy and Hold strategy, demonstrating lower risk and smaller asset price volatility compared to other strategies.

Combining the other indicators in the table, and comprehensively analyzing from various dimensions, in comparison to the other strategies, the multi-factor quantitative stock

TABLE 3. The specific benefits and risks of quantitative trading strategies

Static strategy	TabNet-Boost	TabNet	XGBoost	LightGBM	CatBoost	AdaBoost	Buy and Hold
Strategy rate of return	**69.61%**	43.99%	53.07%	41.06%	45.0%	19.19%	−1.2%
Annualized rate of return	**29.42%**	20.76%	24.65%	19.49%	21.2%	9.51%	−0.62%
Benchmark rate of return (CSI 300)				−22.78%			
Sharpe ratio	**1.09**	0.75	0.9	0.71	0.75	0.37	−0.15
Information ratio	**0.17**	0.12	0.15	0.12	0.11	0.09	0.08
Maximum drawdown	23.33%	26.39%	27.47%	24.6%	31.09%	31.78%	**17.63%**

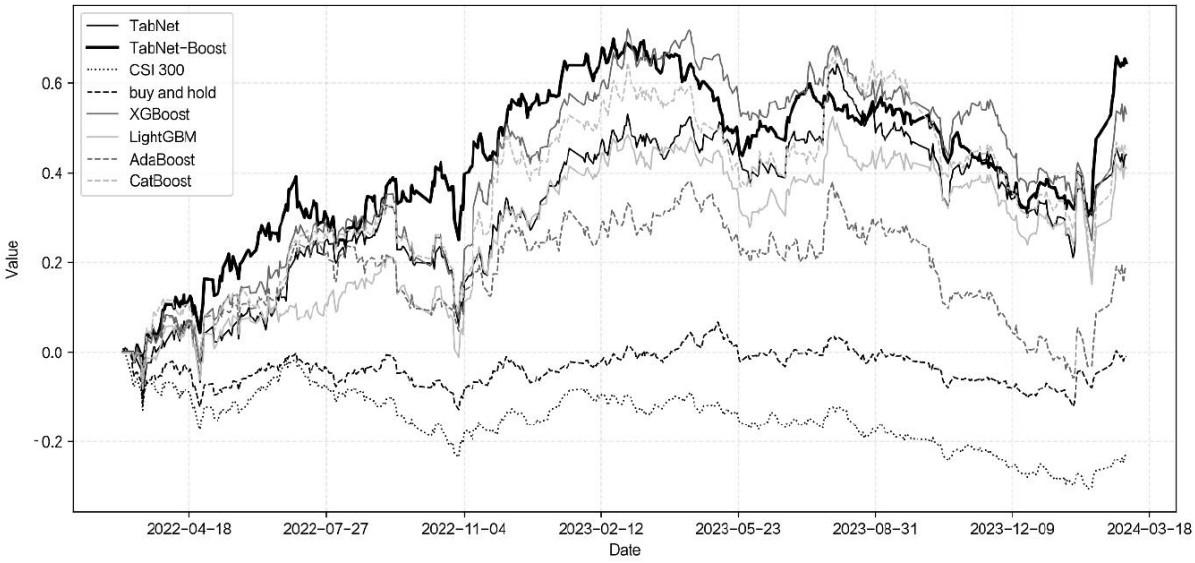


FIGURE 8. Yield curves for seven strategies

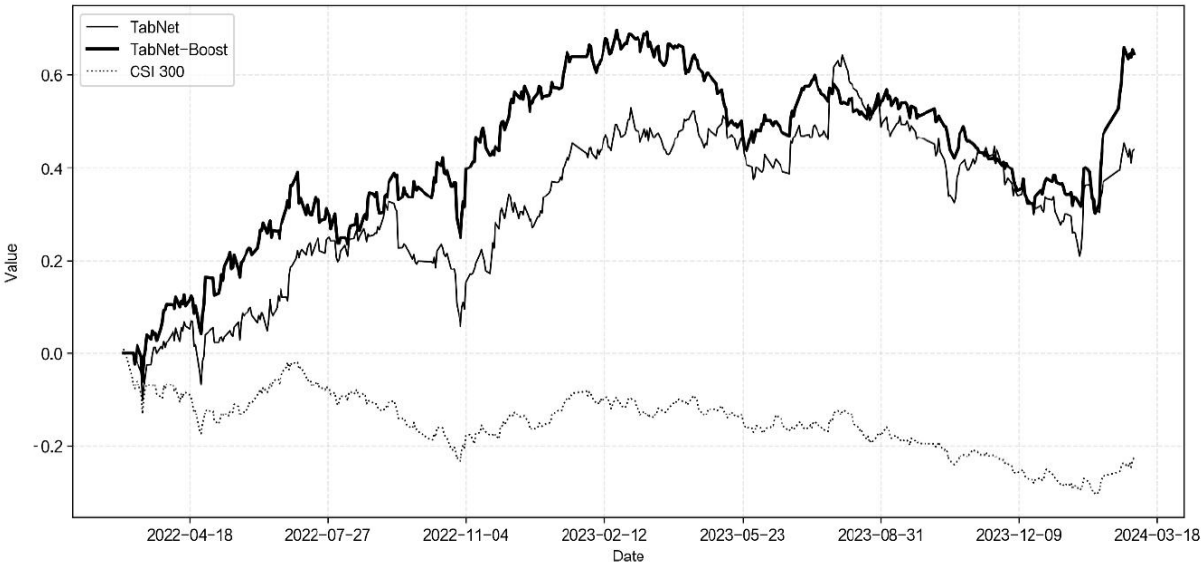


FIGURE 9. Yield curves for TabNet strategy and TabNet-Boost strategy

selection strategy based on TabNet-Boost exhibits characteristics such as high returns, stability, strong risk resistance, and profitability.

5.2. **Model and strategy evaluation.** In addition to focusing on strategy returns and profitability, as a design of machine learning algorithms, attention should also be paid to certain performance metrics and evaluations of the model. In previous research on multi-factor quantitative trading strategies, most scholars often overlooked this part. Indeed, the performance of a strategy is most intuitively reflected in the annualized rate of return and Sharpe ratio. After all, even a model with excellent performance may perform poorly in backtesting scenarios. This paper comprehensively focuses on the returns of the strategy and the performance of the model mainly based on the following considerations. Firstly, the performance of the model reflects its accuracy and stability in prediction. Although the performance of the model may fluctuate with changes in data and factors in the ever-changing stock market, the performance indicators of the model trained over a period of

time can effectively evaluate the strategy's ability to assess specific stock pools. Secondly, certain specific metrics such as Recall and F1-Score can effectively assess the performance of true positive cases, thereby inferring the strategy algorithm's ability to identify risks.

This paper evaluates the algorithm models primarily based on Precision, Recall, F1-Score, AUC, and Accuracy. From Table 4, it can be observed that the TabNet-Boost model scores the highest in Precision, Recall, F1-Score, and Accuracy metrics, while the XGBoost model scores the highest in the AUC metric. The TabNet-Boost model outperforms TabNet and other Boost models in terms of Recall and F1-Score, which are important metrics for quantitative stock selection. A high recall rate can assess the performance in terms of true positive cases. Identifying a rising stock as falling carries less risk than identifying a falling stock as rising, demonstrating the ability of our strategy to recognize risk indicators.

Table 4. Performance metrics for each model

Model	Precision	Recall	F1-Score	AUC	Accuracy
TabNet-Boost	**0.69**	**0.64**	**0.66**	0.80	**0.75**
TabNet	0.65	0.62	0.63	0.79	0.72
XGBoost	0.55	0.59	0.57	**0.82**	0.65
LightGBM	0.53	0.62	0.57	0.75	0.64
CatBoost	0.56	0.56	0.56	0.69	0.66
AdaBoost	0.41	0.54	0.47	0.68	0.52

6. **Conclusion.** This paper proposed a multi-factor quantitative trading and stock selection strategy based on the traditional TabNet model and the Boost family models through Stacking integration: TabNet-Boost. It utilized a combination of traditional methods and machine learning feature importance for factor selection analysis, conducted model training in the stock pool and factor pool, and then designed trading strategies. Finally, the TabNet-Boost model and other models were back-tested for nearly two years. The back-testing results demonstrated that the multi-factor quantitative stock selection strategy based on TabNet-Boost achieved an annualized rate of return of 29.42% even in unfavorable market conditions. The overall model performance reached the optimal level among its peers. Whether it is strategy returns, Sharpe ratio, or model performance indicators, they outperformed traditional TabNet and other Boost ensemble learning models, confirming the effectiveness and stability of the TabNet-Boost model. This has significant implications for research in multi-factor quantitative stock selection trading and optimization of table data models. This paper introduces a new approach to multifactor quantitative stock selection: by applying dimensionality reduction to feature factors and integrating ensemble learning, it achieves better predictive performance on tabular data. In the future, this method holds significant research value and implications in the field of quantitative trading, especially for multifactor quantitative stock selection using high-dimensional heterogeneous tabular data as training input.

REFERENCES

[1] V. Borisov, T. Leemann, K. Seßler et al., Deep neural networks and tabular data: A survey, *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[2] S. Ö. Arik and T. Pfister, TabNet: Attentive interpretable tabular learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.35, no.8, pp.6679-6687, 2021.

[3] C. Shah, Q. Du and Y. Xu, Enhanced TabNet: Attentive interpretable tabular learning for hyperspectral image classification, *Remote Sensing*, vol.14, no.3, 716, 2022.

[4] Q. Cai and J. He, Credit payment fraud detection model based on TabNet and XGBOOT, *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pp.823-826, 2022.
[5] S. Barak, A. Arjmand and S. Ortobelli, Fusion of multiple diverse predictors in stock market, *Information Fusion*, vol.36, pp.90-102, 2017.
[6] E. Hadavandi, H. Shavandi and A. Ghanbari, Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting, *Knowledge-Based Systems*, vol.23, no.8, pp.800-808, 2010.
[7] J. Zhang, S. Cui, Y. Xu et al., A novel data-driven stock price trend prediction system, *Expert Systems with Applications*, vol.97, pp.60-69, 2018.
[8] X. Zhong and D. Enke, Forecasting daily stock market return using dimensionality reduction, *Expert Systems with Applications*, vol.67, pp.126-139, 2017.
[9] D. L. Minh, A. Sadeghi-Niaraki, H. D. Huy et al., Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network, *IEEE Access*, vol.6, pp.55392-55404, 2018.
[10] P. Y. Zhou, K. C. C. Chan and C. X. Ou, Corporate communication network and stock price movements: Insights from data mining, *IEEE Transactions on Computational Social Systems*, vol.5, no.2, pp.391-402, 2018.
[11] L. Shi, Z. Teng, L. Wang et al., DeepClue: Visual interpretation of text-based deep stock prediction, *IEEE Transactions on Knowledge and Data Engineering*, vol.31, no.6, pp.1094-1108, 2018.
[12] Y. Zhang, S. H. Tan, J. Yang, T. Kim and J. Bae, Stock price movement prediction based on re-extract feature LSTM, *ICIC Express Letters*, vol.16, no.2, pp.187-194, 2022.
[13] E. Shaikh, V. Mishra, F. Ahmed et al., Exchange rate, stock price and trade volume in US-China trade war during COVID-19: An empirical study, *Studies of Applied Economics*, vol.39, no.8, 2021.
[14] X. Zhang and W. Chen, Stock selection based on extreme gradient boosting, *2019 Chinese Control Conference (CCC)*, pp.8926-8931, 2019.
[15] Z. Li, W. Xu and A. Li, Research on multi factor stock selection model based on LightGBM and Bayesian Optimization, *Procedia Computer Science*, vol.214, pp.1234-1240, 2022.
[16] Z. Shi, Y. Hu, G. Mo et al., Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction, *arXiv Preprint*, arXiv: 2204.02623, 2022.
[17] Y. Ren, Research on short term stock selection strategy based on machine learning, *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pp.20-23, 2021.
[18] Y. Zhang, Stock volatility prediction with hybrid model of FFNN and LightGBM, *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp.750-754, 2022.
[19] E. K. Ampomah, Z. Qin and G. Nyame, Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement, *Information*, vol.11, no.6, 332, 2020.
[20] R. Shwartz-Ziv and A. Armon, Tabular data: Deep learning is not all you need, *Information Fusion*, vol.81, pp.84-90, 2022.
[21] J. M. Clements, D. Xu, N. Yousefi et al., Sequential deep learning for credit risk monitoring with tabular financial data, *arXiv Preprint*, arXiv: 2012.15330, 2020.
[22] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785-794, 2016.
[23] G. Ke, Q. Meng, T. Finley et al., LightGBM: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, vol.30, 2017.
[24] L. Prokhorenkova, G. Gusev, A. Vorobev et al., CatBoost: Unbiased boosting with categorical features, *Advances in Neural Information Processing Systems*, vol.31, 2018.
[25] L. Katzir, G. Elidan and R. El-Yaniv, Net-DNF: Effective deep modeling of tabular data, *International Conference on Learning Representations*, 2020.
[26] N. Rahaman, A. Baratin, D. Arpit et al., On the spectral bias of neural networks, *International Conference on Machine Learning (PMLR)*, pp.5301-5310, 2019.
[27] B. R. Mitchell, *The Spatial Inductive Bias of Deep Learning*, Ph.D. Thesis, Johns Hopkins University, 2017.
[28] I. Shavitt and E. Segal, Regularization learning networks: Deep learning for tabular datasets, *Advances in Neural Information Processing Systems*, vol.31, 2018.
[29] L. Grinsztajn, E. Oyallon and G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, *Advances in Neural Information Processing Systems*, vol.35, 2022.
[30] X. Wei, H. Ouyang and M. Liu, Stock index trend prediction based on TabNet feature selection and long short-term memory, *PloS One*, vol.17, no.12, e0269195, 2022.

[31] Z. Wang, TabNet with data augmentation apporach in stock return prediction task, *2022 19th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp.1-5, 2022.
[32] K. McDonnell, F. Murphy, B. Sheehan et al., Deep learning in insurance: Accuracy and model interpretability using TabNet, *Expert Systems with Applications*, vol.217, 119543, 2023.
[33] L. P. Joseph, E. A. Joseph and R. Prasad, Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture, *Computers in Biology and Medicine*, vol.151, 106178, 2022.
[34] S. Carta, S. Consoli, A. S. Podda et al., Statistical arbitrage powered by explainable artificial intelligence, *Expert Systems with Applications*, vol.206, 117763, 2022.

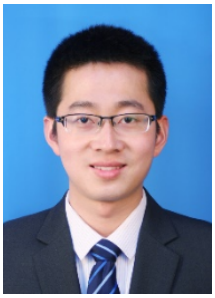
Author Biography

Jianming Li received the bachelor's degree in Ship Engineering from Dalian University of Technology, China, 1999; the M.Sc. degree in Computer Application Technology from Dalian University of Technology, China, 2002; the Ph.D. degree in Computer Application Technology from Dalian University of Technology, China, 2007.

Dr. Li is currently a full-time associate professor at the Dalian University of Technology, China. His main research interests include the machine learning, classification and prediction algorithms of deep learning, software automation, and quantitative analysis and strategy research in the financial field. He has published over 50 papers in journals and conferences.

Yuhan Wang obtained a bachelor's degree in Engineering Degree, majoring in Ship Engineering, from September 2017 to June 2021, Harbin Engineering University.

Mr. Wang is currently studying for a master's degree in Dalian University of Technology. His primary research areas include deep learning, big data analytics, data mining, and the study of multi-factor quantitative trading strategies.



Yang Wang obtained a master's degree in Management, from September 2011 to January 2014, majoring in Accounting at Dongbei University of Finance and Economics.

Mr. Wang is deputy director of the Economic Responsibility Audit Office of the Audit Department of Dalian University of Technology, Senior Accountant, Certified Public Accountant, High end accounting talents of the Ministry of Education. His research direction is financial auditing in universities.

Xiangpei Hu received his B.S. (1983), M.S. (1987) and Ph.D. (1996) degrees from Harbin Institute of Technology, China, respectively. He is a professor of Management Science at Dalian University of Technology, China, "Distinguished Young Scholars" of National Natural Science Foundation of China (NNSFC), "Chang-jiang Scholars Distinguished Professor" of Ministry of Education (MOE) of China.

His research and teaching interests are electronic commerce, supply chain and logistics management, intelligent operations research and the real-time optimization control for dynamic systems. He has published over 200 scholarly papers in refereed journals.