Enhancing Predictive Accuracy in Financial Machine Learning Through Gaussian Mixture Decomposition

Xingyou Li[1,a*] , Bo Yu[2,b]

{lininglixingyou@qq.com[a*],1912842012@qq.com[b] }

School of International Business, Zhejiang International Studies University, Hangzhou, China[1&2]
Department of Economics, Ghent University, Ghent, Belgium[1]

Abstract. When machine learning is applied to financial markets, one of the persistent challenges is improving the accuracy of model predictions due to the presence of substantial noise in financial data. This study aims to enhance data quality through a novel approach: decomposing data into distinct groups using Gaussian mixture distributions and independently training machine learning models on each group. Three different models were empirically validated: Convolutional Neural Networks, Support Vector Machines, and Attention Models. The results indicate that this method significantly enhances the predictive accuracy of the models. In addition, a majority voting approach was employed to select prediction results, demonstrating promising results in terms of accuracy and enhancing the practical utility of the method. The efficacy of this approach lies in separating various influencing factors in the data, making relatively simple data distributions more conducive to training machine learning models. This research provides new insights for modeling stock markets using machine learning, offering a means to mitigate the adverse impact of data noise on predictive performance and improving prediction accuracy.

Keywords: Gaussian , Mixture Distribution , CNN , SVM , Attention

1 Introduction

Currently, machine learning technology has made significant progress in fields such as image processing and natural language processing. However, in comparison to the rapid advancements in other domains, the effectiveness of various machine learning models applied to financial markets is still in need of improvement. One of the main reasons for this is the complexity and variability of financial market data, which contains a substantial amount of noise, posing significant challenges for model training. Unlike structured data such as images and speech, financial markets involve unstructured, diverse information, with prices constantly changing. How to handle the high dimensionality, heterogeneity, dynamic nature, and severe noise contamination in financial data is one of the core challenges that machine learning faces when applied to financial markets.This paragraph discusses the modeling of features based on stock market data and the persistent issue of noise in stock market data. Many studies have delved into the sources and characteristics of data noise.Zhang, Li, and Chen et al. (2023) [1] argue that the noisiness of stock market time series inevitably affects the classification accuracy of predictive models.Data quality is one of the challenges that machine learning models face in the application of finance. Noise leads to significant non-stationarity and uncertainty in the data,

MSIEID 2023, December 08-10, Guangzhou, People's Republic of China

DOI 10.4108/eai.8-12-2023.2344476

making it difficult for models to capture critical market trends and signals, thereby reducing prediction accuracy. Therefore, in-depth research into methods for improving data quality is crucial for machine learning in the finance domain. Several studies have been dedicated to this: Wu, Chen, and Wang et al. (2020) [2] used technical indicators to represent stock data, thus mitigating the impact of data noise on deep reinforcement learning models.

2 Research Design and Empirical Study

2.1GMM and Decomposing Stock Data

Gaussian Mixture Model (GMM) offers unique advantages when applied to fitting stock index data compared to other methods. Stock data often exhibit complex distributions and various behavioral patterns, making it challenging for a single distribution model to capture this diversity. GMM, by combining multiple Gaussian distributions, provides flexibility in modeling the variations associated with different market states, thus better characterizing the characteristics of stock data. Due to the advantages of Gaussian Mixture Distributions, they have gained wide application in financial market research. Nasir, Sheraz, and Dedu (2022) [3] employed the Gaussian Mixture Model to study the returns of the Pakistan Stock Exchange index (PSX-100) and the Pakistani Rupee exchange rate.

Decomposing stock data into Gaussian Mixture Distribution (GMM) components can improve the quality of data input into machine learning models for several reasons: Firstly, stock prices often exhibit multi-modal distributions, indicating the presence of multiple potential market states or trends. GMM's multiple components are well-suited for representing this multi-modality. Secondly, decomposing stock data into different Gaussian distribution groups offers the advantage of segmenting the data into different states or patterns. Each Gaussian distribution represents a distinct market feature or trend. For example, post-decomposition market trends can be understood as stable growth, high volatility, or the influence of specific events among different groups. This decomposition aids in a more accurate understanding of different behaviors and dynamics in the market, providing better data for model training. Lastly, the benefit of inputting data into machine learning models in different groups is the ability to further explore the characteristics between market states. Machine learning models can focus on the patterns in different market states, thus improving the prediction of future stock price trends.

2.2Obtaining Distributions and Data Segmentation

This paragraph mentions the selection of the Shanghai Stock Exchange SME (Small and Medium Enterprises) Index and its constituent stocks as empirical data. The standardized price change data is fitted using the GaussianMixture module in Python. The specific results are as shown in Figure 1:

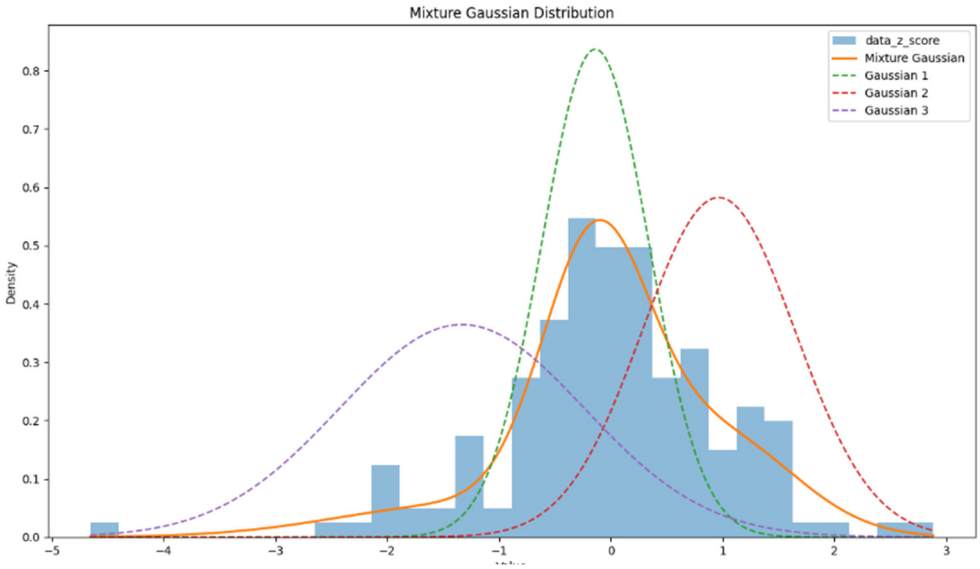

Figure 1: Gaussian Mixture Distribution Fitting Results

The time span covered is from April 20, 2022, to April 20, 2023, comprising a total of 245 trading days. Based on the daily price changes of the Shanghai Stock Exchange SME Index, a Gaussian Mixture Distribution is constructed. It is assumed that the mixture distribution consists of three Gaussian distributions.

After removing stocks with severe data gaps, a total of 430 constituent stocks from the Shanghai Stock Exchange SME Index are retained for analysis. The data is divided into three sets for training a machine learning model: the first 160 days are used for the training set, the following 40 days constitute the validation set, and the last 45 days make up the testing set. The process of grouping individual stock data is as follows: Based on the number of sub-distributions in the Gaussian Mixture Distribution, the processed data is divided into three groups. Individual stock price change data is input into the Gaussian Mixture Distribution of the index, and the group corresponding to the sub-distribution with the highest weight is selected, while the data for the other two groups is set to 0 for that period. A schematic illustration is as shown in Figure 2:

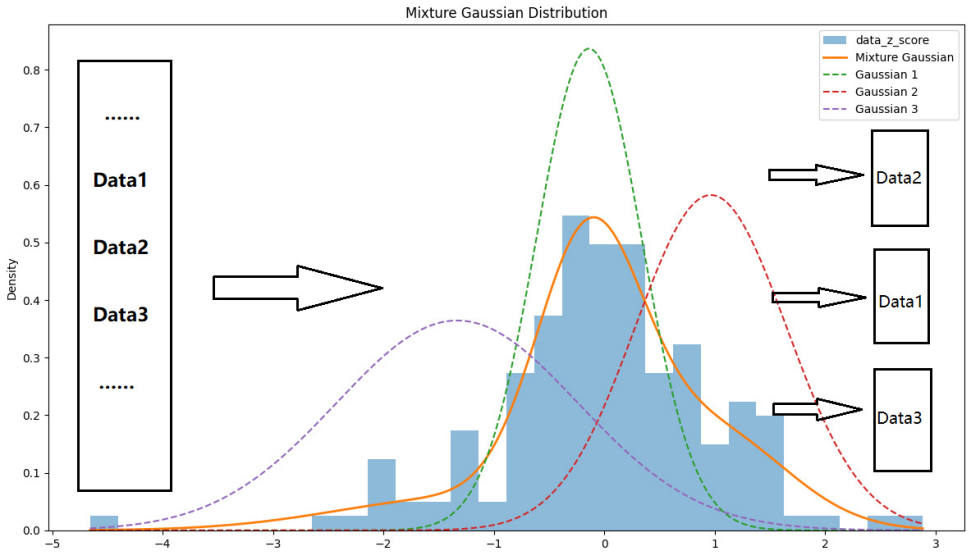

Figure 2: Data Grouping Illustration

The machine learning model is designed for the following use case: It considers a group of five consecutive trading days, and uses all nine features of individual stocks as input variables. The goal is to predict whether a stock's opening price on the next Monday will be greater than or equal to the closing price on the current Friday.The baseline accuracy for simple guessing without relying on any additional information is 50%. The effectiveness of model training is measured by the model's prediction accuracy. During the testing phase, there are a total of 45 days, with no subsequent data available for the last week. Therefore, the validation results cover 40 days, which amounts to 8 weeks. There are 3440 data groups for testing (8 weeks × 430 individual stocks). Among these, 1883 groups show that the opening price next week is greater than or equal to the closing price this week, while the remaining 1557 groups do not exhibit this pattern.

2.3 Empirical Study and Results

In this article, the CNN model used consists of three sets of convolutional layers combined with Dropout layers to extract features. After feature extraction, there are two fully connected layers for classification. In between these fully connected layers, Dropout layers are added to enhance the model's generalization ability. The specific structure is illustrated as shown in Figure 3:

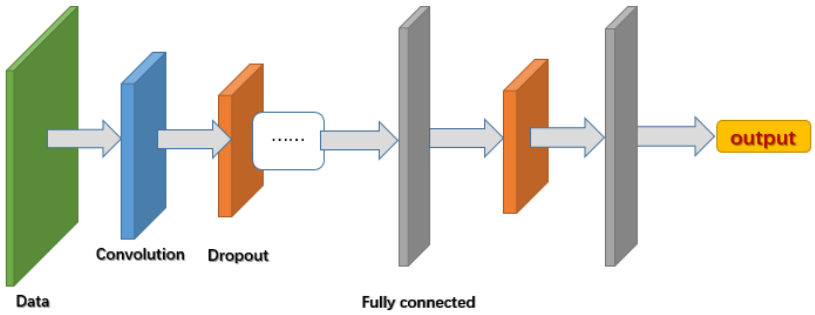

Figure 3: CNN Structure Illustration

The CNN predictions are detailed in Table 1:

Table 1. CNN Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2	Gaussian Distribution 3
Positive	1925	313	2247	1992
Negative	1515	3127	1193	1448
Correct Count	1718	1510	1818	1795
Accuracy Rate	0.4994	0.4390	0.5285	0.5218

From the table, it can be observed that the original data and the data from Gaussian Distribution 1 did not perform well in the CNN model, with both having accuracy rates below 50%. However, the data from Gaussian Distribution 2 and 3 showed some improvement, achieving accuracy rates of 52.85% and 52.18%, respectively. This suggests that data preprocessing has had a positive impact on the model's performance.

SVM exhibits a degree of sensitivity to data noise, where noisy data can potentially cause issues during the determination of the separating hyperplane. SVM aims to find the optimal hyperplane, and noisy data may result in incorrect support vectors near the separation boundary, impacting the performance of classification or regression. Therefore, the data preprocessing approach in this article may have a certain effect. The specific results are presented in Table 2:

Table 2. SVM Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2	Gaussian Distribution 3
Positive	1697	1809	2614	2989
Negative	1743	1631	826	451
Correct Count	1778	1764	1813	1924
Accuracy Rate	0.5167	0.5128	0.5270	0.5593

From the table, it is evident that all prediction results have accuracy rates above 50%. Particularly, the predictions for Gaussian Distribution Groups 2 and 3 have higher accuracy rates than the original data. This indicates that the data preprocessing approach has indeed had a positive impact on the results.

The specific structure of the attention mechanism model in this article is as follows: Firstly, it obtains the representation vector of the input sequence based on the attention algorithm. The representation vector is then passed through a fully connected layer and a Softmax layer for classification. The role of the attention mechanism in the model is to capture the features of the input data. The specific structure is as shown in Figure 4:

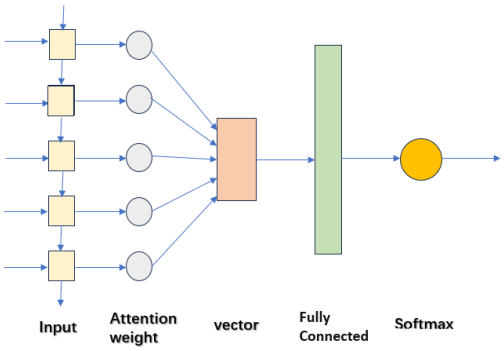

Figure 4: Attention Illustration

The specific attention mechanism model results are shown in the following Table 3:

Table 3. Attention Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2	Gaussian Distribution 3
Positive	1952	2891	1678	2117
Negative	1488	549	1762	1323
Correct Count	1809	1914	1769	1824
Accuracy Rate	0.5061	0.5241	0.4965	0.5552

The results of the attention mechanism show that the majority of the data have prediction accuracy greater than 50%. In particular, the prediction performance of Gaussian distribution groups 1 and 3 is better than the original data without processing. This also demonstrates that data processing has achieved some effectiveness.Combining the results of all three models, it can be observed that in each of the three models, the prediction accuracy of grouped data is stronger than the original data without processing. Furthermore, the prediction accuracy in all cases is above 50%. These preliminary results indicate that data processing based on Gaussian mixture distribution demonstrates certain effectiveness.

3 Further Analysis

3.1 Analysis of the Effectiveness of Data Processing

Gaussian Mixture Models can simulate situations where multiple subpopulations interact. Specifically, Gaussian Mixture Models assume that data comes from a mixture of multiple Gaussian distributions, with each Gaussian distribution representing a subpopulation. When modeling stock indices, each Gaussian distribution can be seen as a different force acting on the

market. For example, they can be viewed as different types of investors: institutional investors may exhibit relatively stable Gaussian distributions, while individual investors' behavior may show more volatility and higher variance Gaussian distributions. In fact, many studies have proposed the presence of multiple different forces in the stock market. For instance, Chu and Song (2023) [4] suggested that the intraday reversal phenomenon in the Chinese A-share market is related to heterogeneous investors. They used different trading volumes and order imbalances to identify different investors. Ng, Wu, and Yu (2016) [5] studied 27,828 samples from 39 countries worldwide, and found that foreign investor heterogeneity plays a role in stock liquidity.

From the above, it is evident that financial data's characteristics stem from the combined effects of multiple forces. This interaction makes it challenging to describe the data distribution using a single simple distribution. In the field of machine learning, a concept closely related to this characteristic of data distribution is 'distribution adaptation,' 'domain adaptation,' or 'data shift.' Current research in this domain primarily focuses on how to handle differences in data distribution between the test set and training set and how to maintain model performance when transferring a trained model to a data scenario with a different distribution.For example, Farahani, Voghoei, and Rasheed (2021) [6] pointed out that training data and testing data may come from different distributions, and in such cases, differences between these distributions may lead to a decrease in model performance. Subbaswamy and Saria (2020) [7] conducted research in the medical field and found that models sometimes struggle to generalize effectively (i.e., make accurate predictions). Such variations commonly occur when models transition from the training phase to the deployment phase, and these changes are attributed to differences in data, including patient characteristics, disease prevalence, measurement timing, equipment, treatment modes, and more.

The above research indicates that the data distribution has a significant impact on model performance. While most current research focuses on differences in the distribution of training and testing data, this paper believes that the impact of data distribution differences also applies to training data in machine learning. When the data follows a mixed distribution, the model may capture features from the training data. However, the various market forces not only constantly change but also exert various influences on each other. This makes it challenging to maintain a stable relationship between stock data trends and their features. The model needs to deal with continuously changing scenarios during the learning and prediction process, leading to an inevitable decrease in performance. By grouping the original data, the new data distribution becomes relatively uniform, and the relationship between data trends and features becomes relatively simple. The scenarios that machine learning models need to learn become more stable, mitigating some of the aforementioned drawbacks to some extent.

3.2 The Selection of the Number of Sub-Distributions in Gaussian Mixture Models

In the previous empirical analysis, we observed and selected three Gaussian distributions to form a mixture distribution. However, is there a better choice? Here, we choose two Gaussian distributions to compose a Gaussian mixture distribution and conduct relevant tests to verify how the number of sub-distributions should be selected. When there are two sub-distributions, the fit is as shown in the following Figure 5:

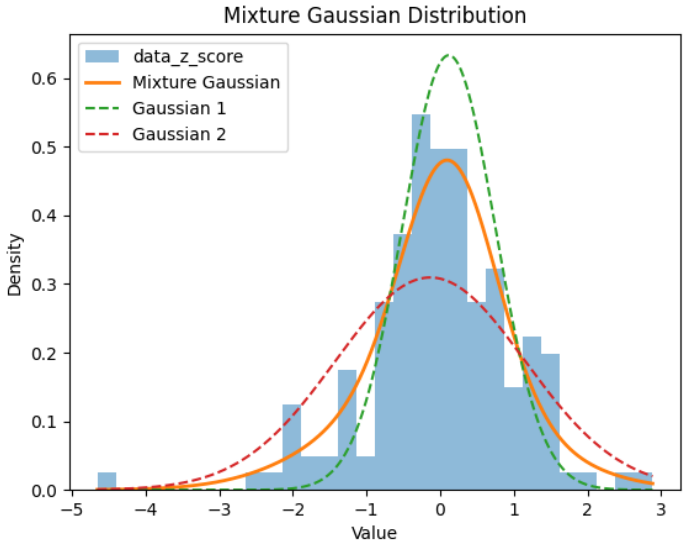

Figure 5: Gaussian Mixture Distribution Fitting Results

The CNN prediction results are shown in the following Table 4:

Table 4. CNN Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2
Positive	1731	1956	2449
Negative	1709	1484	991
Correct Count	1730	1739	1858
Accuracy Rate	0.5029	0.5055	0.5401

At this point, the results are similar to those discussed earlier. The prediction accuracy for the original data is around 50%. The two subgroups of the Gaussian mixture distribution have higher prediction accuracy than the original data, with Gaussian mixture 2 performing the best.And the SVM prediction results are shown in the following Table 5:

Table 5. SVM Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2
Positive	1410	1408	363
Negative	2030	2032	3077
Correct Count	1713	1709	1540
Accuracy Rate	0.4980	0.4968	0.4477

The prediction results here have shown a significant decrease compared to the previous section. None of the three sets of data have an accuracy exceeding 50%, and the grouped data performs worse than the original data.The Attention prediction results are shown in Table 6:

Table 6. Attention Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2
Positive	2119	1611	2618
Negative	1321	1829	822
Correct Count	1838	1766	1861
Accuracy Rate	0.5343	0.5134	0.5410

At this point, the results are close to the previous section, with all prediction accuracies exceeding 50%, and the second Gaussian distribution group performs better than the original data.

In summary, on the one hand, it can be seen that the data processing in the attention and CNN models remains effective, indicating that the data processing method is still valid even when the number of sub-distributions is set to 2. However, on the other hand, the prediction results of SVM are not as good as in the previous section. This may be due to the numerous factors influencing financial data noise. The core idea of data grouping is to separate these influencing factors, allowing the model's training to focus on the impact of a particular factor and reduce interference between different factors. Therefore, if the number of components in the Gaussian mixture distribution is too small, it may not be sufficient to separate these influencing factors. Choosing the right number of components is one of the key aspects of the data processing method in this study, and further research is needed.

3.3 The Robustness of Data Processing Methods

The previous empirical results have shown the effectiveness of data processing methods. In this section, we will further validate the robustness of these methods. Considering the persistence and periodicity of factors affecting financial time series, such as macroeconomics and interest rates, this study does not shuffle the data but exchanges the composition of the training and testing sets. In the previous section, we used data from days 201 to 245 as the testing set. Here, we choose data from days 161 to 200 as the testing set, keeping the order of the remaining data unchanged and splitting it into training and validation sets in the same proportion. This results in a total of 3,440 data points in the testing set, including 1,934 with high opening prices and 1,506 with low opening prices.The specific validation results for CNN are shown in the following Table 7:

Table 7. CNN Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2	Gaussian Distribution 3
Positive	2032	3028	2377	2121
Negative	1408	412	1063	1319
Correct Count	1712	1840	1873	1785
Accuracy Rate	0.4977	0.5349	0.5445	0.5189

The results are similar to those in the previous section, with the accuracy of the original data slightly below 50%, while the data processed with grouping shows a significant improvement in performance. And the results for SVM are shown in the following Table 8:

Table 8. SVM Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2	Gaussian Distribution 3
Positive	2097	2203	2615	3093
Negative	1343	1237	825	347
Correct Count	1783	1811	1829	1873
Accuracy Rate	0.5183	0.5265	0.5317	0.5445

The results are similar to those in the previous section, with all accuracies exceeding 50%, and the accuracy of the data after grouping has increased. And the results for attention mechanism are shown in the following Table 9:

Table 9. Attention Prediction Results

Prediction Results	Original Data	Gaussian Distribution 1	Gaussian Distribution 2	Gaussian Distribution 3
Positive	2733	2983	2592	3058
Negative	707	457	848	382
Correct Count	1847	1837	1854	1888
Accuracy Rate	0.5544	0.5134	0.5538	0.5442

At this point, there is some difference from the results in the previous section. Although all accuracies are above 50%, the accuracy of the data after grouping in the three groups is not higher than that of the original data. In summary, the results in this round of data are similar to those in the previous section but also show some differences. For CNN and SVM, the model's prediction accuracy is improved after grouping. However, the attention mechanism does not show this effect. This indicates that the data processing method proposed in this paper has some robustness, but it is also correlated with specific models. To address this issue, further analysis will be conducted below.

3.4 Group Selection Based on Majority Voting

In the previous empirical analysis, the model's predictive accuracy significantly improved after grouping the data. However, how can we determine which specific group performs better? By summarizing the results above, it is observed that, except for the Attention model in the second round, which exhibited better performance on the raw data, in all other cases, there is always at least one processed group of data whose results surpass those of the raw data.

In comparison, it is evident that in both rounds of experimentation, apart from the attention mechanism, both CNN and SVM exhibit superior predictive performance in at least one Gaussian group compared to the original data. So, which specific grouping demonstrates better predictive accuracy? "Upon summarizing the results presented above, it is observed that the best-performing group among the three models is not fixed across the two rounds of data, and there is no consistent dominance of data from a specific distribution in obtaining optimal results.

This suggests that there isn't a single, determining group that consistently exerts an influence, meaning there isn't a single force that consistently has a decisive impact on the financial market. This aligns with our common understanding that the factors influencing the entire stock market can vary over different periods. Therefore, this paper proposes using a voting method to determine the prediction results for each round. A 'high open' prediction is made only when two or more groups predict a high open; otherwise, it is considered a 'low open' prediction. The accuracy of the group predictions after voting is presented in the Table 10 below:

Table 10. Summary of Results Comparison

Model	First Round Data		Second Round Data	
	Original Data	Vote	Original Data	Vote
CNN	0.4994	0.5131	0.4977	0.5317
SVM	0.5167	0.5387	0.5183	0.5485
Attention	0.5061	0.5276	0.5544	0.5573

From the table, it is evident that the voting method outperformed the original data in both rounds of experimentation. Particularly noteworthy is the attention model in the second round of data, which initially did not outperform the original data but achieved better results after incorporating the voting method. These results indicate that the voting method not only enhances the robustness of the data processing approach in this paper, making it suitable for all models, but also provides practical applicability to this method, allowing it to produce satisfactory prediction results.

4 Conclusions

In conclusion, the use of Gaussian mixture distribution to decompose stock data and apply it to the training of machine learning models is a promising approach. It is based on statistics, breaking down the market into a combination of different influential factors when dealing with financial data. This separation of the effects of various key factors on the market significantly enhances the predictive accuracy of machine learning models. In practice, it provides a more comprehensive and valuable perspective for investment research and decision-making. It also contributes to the research and practical application of machine learning in the field of finance.

References

[1] Zhang L, Li C, Chen L, et al. A Hybrid Forecasting Method for Anticipating Stock Market Trends via a Soft-Thresholding De-noise Model and Support Vector Machine (SVM)[J]. World Basic and Applied Sciences Journal, 2023, 13(2023): 597-602.

[2] Wu X, Chen H, Wang J, et al. Adaptive stock trading strategies with deep reinforcement learning methods[J]. Information Sciences, 2020, 538: 142-158.

[3]Nasir I, Sheraz M, Dedu S. Mixture Models and Modelling Volatility of Returns–a Study on Gaussian and Heterogeneous Heavy Tail Mixtures[J]. Economic Computation & Economic Cybernetics Studies & Research, 2022, 56(4).

[4] Chu X, Song S. Cross-Sectional Reversal of Intraday Returns and Investor Heterogeneity in an Emerging Market[J]. Borsa Istanbul Review, 2023.

[5] Ng L, Wu F, Yu J, et al. Foreign investor heterogeneity and stock liquidity around the world[J]. Review of Finance, 2016, 20(5): 1867-1910.

[6] Farahani A, Voghoei S, Rasheed K, et al. A brief review of domain adaptation[J]. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020, 2021: 877-894.

[7] Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI[J]. Biostatistics, 2020, 21(2): 345-352.