Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

Stock Market Data Using Data Mining For Feature Extraction

[1]P. Dineshkumar, [2]Dr. B. Subramani

[1]Research Scholar, SNMV College of Arts and Science ,
Coimbatore, Tamilnadu, India.
[2]Principal , SNMV College of Arts and Science ,
Coimbatore, Tamilnadu, India.

Abstract: This paper presents a robust approach for feature extraction from stock market data by combining Principal Component Analysis (IPCA) and Moving Averages (MA). IPCA reduces dimensionality, capturing underlying patterns, while MAs identify trends and cyclic behaviors. The synergistic integration of these techniques enhances the extraction of essential features for stock market analysis. Research method effectively uncovers relevant information, offering valuable insights for trading and investment decisions. It addresses dimensionality challenges and identifies meaningful patterns, promoting a deeper understanding of market dynamics.

Keywords: Feature Extraction, Stock Market Data, Principal Component Analysis (IPCA), Moving Averages (MA), Dimensionality Reduction, Market Trends, Investment Insights;

1. Introduction

The stock market, a dynamic and complex financial ecosystem, fills in as a landmark where financial backers try to use wise judgment that will expand returns and really oversee risk. In this perpetual journey for progress, the utilization of data mining for feature extraction has arisen as an indispensable device. Stock market data is a voluminous and complex substance, incorporating verifiable cost developments, exchanging volumes, organization financials, and a plethora of macroeconomic indicators. The ability to extricate significant features from this colossal dataset is the linchpin of gaining actionable insights that drive intelligent investment decisions. This examination sets out on an investigation of the pivotal pretended by data mining in the extraction of features with regards to stock market data investigation. At its center, this try is energized by the objectives of financial decision-making, risk management, and prescient demonstrating.

Financial Decision-Making becomes the overwhelming focus, by which financial backers go to the archives of authentic market data to create all around informed investment systems. By taking apart past stock execution, knowing patterns, and examining organization explicit data, financial backers can tailor their capital allocations to align with their benefit goals. Furthermore, this interaction works with the development of diversified portfolios and illuminates exact timing for market passage and leave focuses. In the domain of Hazard Management, stock market data analysis arises as an important compass for navigating the labyrinth of potential pitfalls. This encompasses the assessment and mitigation of different kinds of hazard, crossing the range from market risk, affected by outer elements and macroeconomic circumstances, to organization explicit gamble, which comes from poor financial wellbeing or administration issues.

The curtain ascends on Predictive Demonstrating, the crown gem of stock market data analysis. Data mining interweaved with authentic data, turns into the linchpin for predicting market patterns and future cost developments. Predictive models act as the directing light, molding investment systems that optimize returns and limit risk. This exploration is a beacon, illuminating the methodologies and techniques supporting the course of data mining for feature extraction in stock market data analysis. It divulges the significant importance and viable use of this transformative device in the realm of money. With stock market data analysis progressively interlaced with data mining techniques, the financial business is outfitted with a strong weapons store for making data-driven decisions, overseeing risk, and working on predictive displaying. Financial backers and financial foundations, both huge and little, rely upon the insights outfit from this cooperative energy to flourish in the consistently advancing scene of the stock market.

1.1 The Significance of Stock Market Data Analysis

- **Financial Decision-Making:** The basis for investors' critical financial decisions is stock market data, which is a crucial resource. Investors can develop the insights necessary to create well-informed and strategic investment plans by digging through historical data, spotting trends, and extracting relevant features. Research can better comprehend market dynamics and past performance through this analytical method, enabling them to traverse the complex stock market landscape and improve their capacity to make decisions that are in line with their financial goals.
- **Risk Management:** An important aspect of stock market data analysis is risk assessment. By enabling accurate feature extraction, data mining plays a crucial part in understanding and successfully managing the range of risks associated with different investments. Investors can better comprehend the complex web of risks connected with their portfolios by utilizing data-driven insights, whether the risks are caused by market volatility, company-specific weaknesses, or macroeconomic changes. Through better decision-making and the application of risk-reduction techniques, this increased risk awareness eventually promotes a more adaptable and resilient approach to investing in the volatile stock market.
- **Predictive Modeling:** When predicting stock market fluctuations, predictive modeling is a vital tool. Data mining plays a key part in this process at its core, pulling important elements from the enormous pool of historical and current market data. By using these extracted attributes as the foundation for predictive models, investors are able to foresee upcoming shifts and changes in the stock market environment. This empowers investors to take proactive actions and modify their tactics.

2. Literature Survey

2.1 Name Entity Recognition (NER)

E. T. Khaing (2019) et.al proposed Stock Trend Extraction using Rule-based and Syntactic Feature-based Relationships between Named Entities. Trend extraction is a basic part of financial data analysis, including the ID of named substances and their connections in text records, for example, news stories or website pages. This interaction uncovers connections between named elements and related words inside stock data. The test lies in managing unstructured, time-subordinate, and different word ranges without a syntactic design. Past exploration frequently neglected trend extraction in view of named elements and their connections. The proposed framework contains two key parts: Name Entity Recognition (NER), which characterizes exceptional words into classes like stocks and dates, and connection extraction to distinguish and analyze connections among these elements. This strategy can deal with unstructured text data with fluctuating lengths and reaches, promising significant insights for financial analysis. Further testing will improve its accuracy.

2.2 Independent Component Analysis (ICA)

A. Wijitcharoen (2016) et.al proposed ICA-DEODA: An independent feature extraction model for stock index forecasting. ICADEODA, a cutting-edge data engineering technique that combines Independent Component Analysis (ICA) with an unsupervised feature selection method, is introduced in this study. Higher-order analysis is made possible by ICADEODA, boosting knowledge discovery methods. ICA separates independent components from deciding variables associated with the Security Exchange of Thailand (SET) in the context of stock forecasting. With the aid of DEODA, an unsupervised feature selection method, the most insightful independent components are found to be used as forecasting inputs for Support Vector Regression (SVR). According to the findings of the experiments, ICADEODA works better than conventional techniques like IPCA-SVR, SVR, and IPCA-DEODA-SVR. This method is a reliable instrument for effective feature selection and information discovery in financial analysis since it not only lowers prediction errors but also demonstrates superior direction determination.

2.3 Latent Dirichlet Allocation (LDA)

N. Kanungsukkasem (2019) et.al proposed Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction. This It was created to find features in a variety of text data, mainly news articles and financial time series. These FinLDA-derived characteristics are added-on inputs for machine learning algorithms, improving the predictability of financial time series. The article provides information on the posterior distributions employed in Gibbs sampling for two FinLDA variants and presents a thorough framework for integrating. The Efficient Market Hypothesis (EMH) assumes that markets are informational efficient, but FinLDA shows that in the constantly changing field of financial research, better prediction models may be possible.

2.4 K-Means Clustering Approach

S. Sangsawad (2017) et.al proposed extracting significant features based on candlestick patterns using unsupervised approach. This exploration acquaints calculations planned with separate fundamental features from candle designs, essential for the specialized analysis of stock lists. These features include the candle's heading, the hole among OPEN and CLOSE costs of adjoining candles, the body level of current and past candles, and the length of the candles. To address the test of equivocally characterized candle parts, the exploration applies K-Means grouping. Exploratory data from the Thai SET list, traversing from 1990 to 2017, approves the methodology's viability in protecting the closeness among crude and decoded candle graphs. The separated outcomes can act as contributions for different machine learning techniques like Artificial Neural Networks, Reinforcement Learning, or Content-Based Image Retrieval, offering a promising avenue for additional exploration while potentially lessening data input for image-related approaches.

2.5 Semi-Supervised Road Centerline Extraction

R. Liu (2020) et.al proposed A Semi-Supervised High-Level Feature Selection Framework for Road Centerline Extraction. However, getting labeled data for road extraction takes time, thus there aren't many labeled samples available. It combines the ridge transversal approach, Markov random field (MRF), and high-level feature selection. Three crucial processes make up this method: the extraction of various features, the semi-supervised delineation of the road area, and the extraction of the road centerline. Road extraction from remote sensing photos is important; however this work is still difficult. Currently used methods include knowledge-based approaches, mathematical morphology, differential geometry, and supervised classification, the latter of which mainly relies on labeled samples.

3. Proposed Methodology

3.1 Feature Extraction

The combination of Improved Principal Component Analysis (IPCA) and Moving Averages (MA) for feature extraction from stock data is a potent strategy. This approach efficiently uncovers underlying patterns in stock prices while mitigating dimensionality challenges. The process involves selecting moving average types (e.g., SMA or EMA) and their respective time periods, capturing short to long-term trends. Feature engineering further enriches the data, incorporating trading volume, technical indicators, and macroeconomic data. Standardization ensures data readiness for IPCA, where principal components are identified to explain significant variance. The choice of retained components depends on dimensionality reduction goals or a predefined explained variance threshold. IPCA-MA offers a robust feature extraction method, empowering financial market analysis and decision-making with multifaceted insights.

3.2 Improved Principal Component Analysis (IPCA)

Principal Component Analysis, a dimensionality reduction technique, plays a critical role in the key information extraction from complex stock market data. We want to use improved principal component analysis (IPCA) to change the original data into a new set of variables known as primary components that account for the majority of the variability in the dataset. These elements offer a more streamlined depiction of the data, making it simpler to use and comprehend. A critical decision must be made on how many major components to keep, and this decision is frequently based on a threshold of explained variance. This process guarantees that we record the most pertinent data while minimizing dimensionality.

Data's dimensionality is decreased using IPCA while maintaining as much variation as possible. There are linear transformations involved.

Covariance Matrix Calculation

The covariance matrix is calculated from the original data.

$$\varSigma = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)\,(X_i - \mu)^T$$

Where:

- Σ is the covariance matrix.
- *n* is the number of data points
- *Xi* represents the data point.
- μ is the mean of the data

Eigenvalue and Eigenvector Calculation

IPCA then calculates the eigenvalues and eigenvectors of the covariance matrix.

$$\varSigma v = \lambda v$$

Where:

- Σ is the covariance matrix.
- *v* is the eigenvector
- λ is the eigenvalue

Selecting Principal Components:

The top *k* eigenvectors corresponding to the largest eigenvalues are selected as the principal components.

Moving Averages (MA):

Moving Averages are fundamental in identifying trends and patterns in stock prices and trading volumes. By calculating various types of moving averages, such as simple or exponential, we seek to capture short-term fluctuations and behavior in the stock market data. Experimentation with different time windows for these moving averages allows us to discover which timeframes are most informative for research specific analysis. This step is essential for extracting meaningful insights from the data and providing valuable input to the subsequent modeling stages.

Simple Moving Average (SMA):

By averaging the data points in a moving frame, the SMA is determined,

$$SMA_t = \frac{X_{t-1} + X_{t-2} + \cdots + X_{t-n}}{n}$$

Where:

- SMA_t is the Simple Moving Average at time *t*
- X_{t-i} are the data points in the *n*-period window

Exponential Moving Average (EMA):

The EMA gives more weight to recent data, and it is calculated recursively.

$$EM\,A_t = \alpha \cdot X_t + (1-\alpha) \cdot EM\,A_{t-1}$$

Where:

- $EM\,A_t$ is the Exponential Moving Average at time *t*
- X_t is the data point at time *t*
- α is the smoothing factor, typically between 0 and 1

These equations represent the core concepts behind IPCA and Moving Averages for feature extraction in stock market data. Actual implementations may involve more complex computations and additional considerations.

Here's a proposed algorithm for feature extraction from stock market data using Improved Principal

Component Analysis (IPCA) and Moving Averages (MA):

```
Algorithm: IPCA-MA
Step 1: Create a covariance matrix from the preprocessed data.
Step 2: Compute the eigenvalues and eigenvectors of the covariance matrix.
Step 3: Sort the eigenvalues in descending order and select the top 'k' eigenvectors, where 'k' is the desired number of principal components.
Step 4: Project the data onto the 'k' selected eigenvectors to obtain the new feature space.
Step 5: Evaluate the importance of each principal component by examining the explained variance.
Step 6: Choose the number of principal components that capture a significant portion of the total variance (e.g., 95%).
Step 7: Select the most relevant MAs and other original features that complement the IPCA components.
Step 8: Utilize the reduced feature set, including IPCA components and selected MAs, to build predictive models or perform other analyses, such as trend analysis, volatility prediction, or anomaly detection.
Step 9: Assess the performance of your models using appropriate evaluation metrics (e.g., Mean Squared Error for regression tasks or classification accuracy for classification tasks).
Step 10: Fine-tune your models by adjusting hyperparameters or exploring different algorithms as needed.
Step 11: Interpret the feature importance and the relationships between the selected features to gain insights into stock market trends and behaviors.
Step 12: Visualize the results using plots, charts, and dashboards to present findings effectively.
Step 13: Create a detailed report summarizing the feature extraction process, model performance, and insights gained.
```

This algorithm combines IPCA for dimensionality reduction and MA for capturing trends and patterns in stock market data. The selected features can be used for various stock market analysis tasks, such as forecasting, risk assessment, and decision support.

4. Experimental Result

4.1 Accuracy

Accuracy is the degree of closeness between a measurement and its true value. The formula for accuracy is:

$$\boldsymbol{Accuracy = \frac{(true\ value\ -\ measured\ value)}{true\ value} * 100}$$

Table 1: Comparison Table of Accuracy

Dataset	NER	ICA	Proposed IPCA-MA
100	69	76	85
200	75	79	83
300	80	74	89
400	88	78	94
500	89	70	95

The Comparison table 1 of Accuracy demonstrates the different values of existing NER, ICA and Proposed IPCA-MA. While comparing the Existing algorithm and Proposed IPCA-MA, provides the better results. The existing algorithm values start from 69 to 89, 70 to 79 and Proposed IPCA-MA values starts from 83 to 95. The proposed method provides the great results.

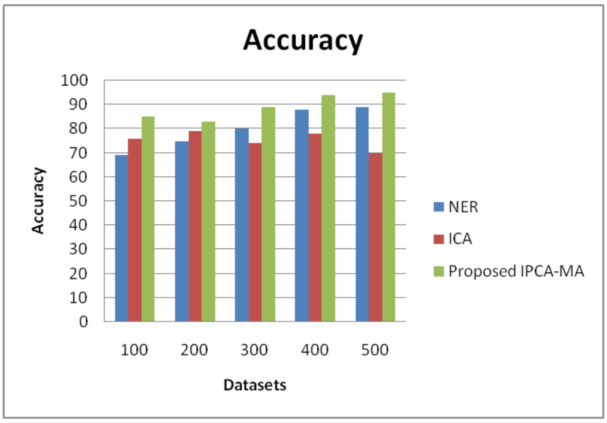

Fig 1: Comparison chart of Accuracy

The Figure 1 Shows the comparison chart of Accuracy demonstrates the existing NER, ICA and Proposed IPCA-MA. X axis denote the Dataset and y axis denotes the Accuracy. The Proposed IPCA-MA values are better than the existing algorithm. The existing algorithm values start from 69 to 89, 70 to 79 and Proposed IPCA-MA values starts from 83 to 95. The proposed method provides the great results.

4.2 Precision

Precision is a measure of how well a model can predict a value based on a given input.

$$\boldsymbol{Precision} = \frac{\boldsymbol{true\ positive}}{(\boldsymbol{true\ positive}\ +\ \boldsymbol{false\ positive})}$$

Table 2: Comparison Table of Precision

Dataset	NER	ICA	Proposed IPCA-MA
100	90.12	87.37	97.67
200	83.69	89.82	99.26
300	80.62	88.54	98.21
400	75.55	85.63	95.58
500	78.94	83.72	90.87

The Comparison table 2 of Precision demonstrates the different values of existing NER, ICA and Proposed IPCA-MA. While comparing the Existing algorithm and Proposed IPCA-MA, provides the better results. The existing algorithm values start from 75.55 to 90.12, 83.72 to 89.82 and Proposed IPCA-MA values starts from 90.87 to 99.26. The proposed method provides the great results.

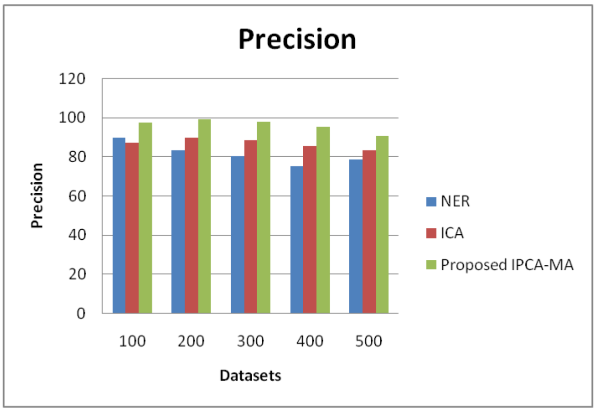

Fig 2: Comparison chart of Precision

The Figure 2 Shows the comparison chart of Precision demonstrates the existing NER, ICA and Proposed IPCA-MA. X axis denote the Dataset and y axis denotes the Precision ratio. The Proposed IPCA-MA values are better than the existing algorithm. The existing algorithm values start from 75.55 to 90.12, 83.72 to 89.82 and Proposed IPCA-MA values starts from 90.87 to 99.26.The proposed method provides the great results.

4.3 Recall

Recall is a measure of a model's ability to correctly identify positive examples from the test set:

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positives}}{(\boldsymbol{True\ Positives} + \boldsymbol{False\ Negatives})}$$

Table 3: Comparison Table of Recall

Dataset	NER	ICA	Proposed IPCA-MA
100	0.80	0.86	0.87
200	0.77	0.80	0.98
300	0.84	0.70	0.90
400	0.85	0.75	0.95
500	0.86	0.76	0.96

The Comparison table 3 of Recall demonstrates the different values of existing NER, ICA and Proposed IPCA-MA. While comparing the Existing algorithm and Proposed IPCA-MA, provides the better results. The existing algorithm values start from 0.77 to 0.86, 0.76 to 0.86 and Proposed IPCA-MA values starts from 0.87 to 0.98. The proposed method provides the great results.

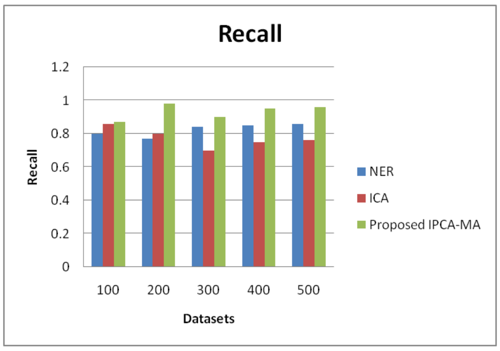

Fig 3: Comparison chart of Recall

The Figure 3 Shows the comparison chart of Recall demonstrates the existing NER, ICA and Proposed IPCA-MA. X axis denote the Dataset and y axis denotes the Recall ratio. The Proposed IPCA-MA values are better than the existing algorithm. The existing algorithm values start from 0.77 to 0.86, 0.76 to 0.86 and Proposed IPCA-MA values starts from 0.87 to 0.98. The proposed method provides the great results.

4.4 F-Measure

F-measure is a test's accuracy that combines precision and recall. It is calculated by taking the harmonic mean of precision and recall.

$$\boldsymbol{F-Measure} = \frac{(\boldsymbol{2} * \boldsymbol{Precision} * \boldsymbol{Recall})}{(\boldsymbol{Precision} + \boldsymbol{Recall})}$$

Table 4: Comparison Table of F-Measure

Dataset	NER	ICA	Proposed IPCA-MA
100	0.86	0.76	0.96
200	0.84	0.74	0.94
300	0.80	0.60	0.90
400	0.79	0.69	0.99
500	0.77	0.66	0.93

The Comparison table 4 of F-Measure Values explains the different values of existing NER, ICA and Proposed IPCA-MA. While comparing the Existing algorithm and Proposed IPCA-MA, provides the better results. The existing algorithm values start from 0.77 to 0.86, 0.66 to 0.76 and Proposed IPCA-MA values starts from 0.93 to 0.99. The proposed method provides the great results.

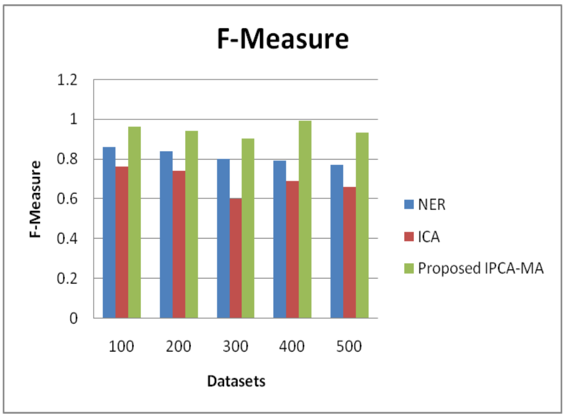

Figure 4: Comparison Chart of F-Measure

The Figure 4.4 Shows the comparison chart of F-Measure demonstrates the existing NER, ICA and Proposed IPCA-MA. X axis denote the Dataset and y axis denotes the F-Measure ratio. The Proposed IPCA-MA values are better than the existing algorithm. The existing algorithm values start from 0.77 to 0.86, 0.66 to 0.76 and Proposed IPCA-MA values starts from 0.93 to 0.99. The proposed method provides the great results.

5. Conclusion

The integration of IPCA with MAs proves to be a powerful feature extraction strategy for stock market data. This approach effectively reduces dimensionality, isolates meaningful patterns, and facilitates informed decision-making in stock trading and investment. By unveiling significant trends and cyclic behaviors, it contributes to a comprehensive understanding of market dynamics, allowing traders and investors to make more informed choices. This methodology holds promise for improving the accuracy and robustness of stock market analysis and prediction.

References

[1] E. T. Khaing, M. M. Thein and M. M. Lwin, "Stock Trend Extraction using Rule-based and Syntactic Feature-based Relationships between Named Entities," 2019 International Conference on Advanced Information Technologies (ICAIT), Yangon, Myanmar, 2019, pp. 78-83, doi: 10.1109/AITC.2019.8920986.

[2] Wijitcharoen, B. Watanapa, P. Padungweang and W. Anantasabkit, "ICA-DEODA: An independent feature extraction model for stock index forecasting," 2016 International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 2016, pp. 1-4, doi: 10.1109/ICSEC.2016.7859924.

[3] N. Kanungsukkasem and T. Leelanupab, "Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction," in IEEE Access, vol. 7, pp. 71645-71664, 2019, doi: 10.1109/ACCESS.2019.2919993.

[4] S. Sangsawad and C. C. Fung, "Extracting significant features based on candlestick patterns using unsupervised approach," 2017 2nd International Conference on Information Technology (INCIT), Nakhonpathom, Thailand, 2017, pp. 1-5, doi: 10.1109/INCIT.2017.8257862.

[5] R. Liu, Q. Miao, Y. Zhang, M. Gong and P. Xu, "A Semi-Supervised High-Level Feature Selection Framework for Road Centerline Extraction," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 5, pp. 894-898, May 2020, doi: 10.1109/LGRS.2019.2931928.

[6] H. Li, X. Fei, M. Yang, K. -M. Chao and C. He, "From Music Information Retrieval to Stock Market Analysis: Theoretical Discussion on Feature Extraction Transfer," 2021 IEEE International Conference on e-Business Engineering (ICEBE), Guangzhou, China, 2021, pp. 54-58, doi: 10.1109/ICEBE52470.2021.00027.

[7] M. Joshi and G. Goel, "Stock Market Prediction Approach: An Analysis," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 945-948, doi: 10.1109/AISC56616.2023.10085262.

[8] D. Peng, "Analysis of Investor Sentiment and Stock Market Volatility Trend Based on Big Data Strategy," 2019 International Conference on Robots & Intelligent System (ICRIS), Haikou, China, 2019, pp. 269-272, doi: 10.1109/ICRIS.2019.00077.

[9] Z. Pu, "Multidimensional Market Unstructured Information Quantification and Feature Extraction," 2022 3rd International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2022, pp. 527-530, doi: 10.1109/ICCSMT58129.2022.00117.

[10] A. Wijitcharoen, B. Watanapa, P. Padungweang and W. Anantasabkit, "ICA-DEODA: An independent feature extraction model for stock index forecasting," 2016 International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 2016, pp. 1-4, doi: 10.1109 /ICSEC.2016.7859924.