International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I6P102

Volume 72 Issue 6, 8-18, June 2024

Original Article

Predicting Stock Market Movements Through Multi-source Data Fusion Graphs: An Approach Employing Graph Convolutional Neural Network

John Ranjith[1], S. Kumar Chandar[2]

[1,2]*School of Business and Management, CHRIST (Deemed to be University), Bengaluru, Karnataka, India.*

[2]*Corresponding Author : kumar.chandarbgl@gmail.com*

Received: 11 March 2024 Revised: 15 April 2024 Accepted: 21 May 2024 Published: 29 June 2024

Abstract *- The stock market plays an important role in the capital market, and investigating price fluctuations in the stock market has consistently been a prominent subject for researchers. The application of soft computing techniques to predict and categorize stock market movements is a significant research challenge that has gathered considerable attention from researchers. Although several studies highlight the significance of incorporating information from two sources in stock movement prediction, the potential of advanced graphical techniques for modeling and analyzing multi-source data remains an unattended research area. This study aims to address this gap by introducing a novel model that utilizes multi-source data fusion graphs to predict future market movements. The primary challenge involves establishing a model that can effectively gather the relationships among various data sources and employ this understanding to improve prediction performance. Compared to several existing methods relying only on historical data or sentiment data, which show limited predictive power and lack generality, the proposed approach seeks to overcome these limitations. The proposed model integrates various information sources, including historical prices, news data, Twitter data, and technical indicators for predicting future stock market trends. This presented method involves constructing a subgraph map for each data type to capture events from both rising and falling markets. Then, a Gated Recurrent Unit (GRU) is employed to aggregate the subgraph nodes. These aggregated nodes are then integrated with a Graph Convolutional Neural Network (GCNN) to classify the multi-source graph, therefore achieving stock market trend prediction effectively. To further validate its effectiveness, the presented model is applied to Indian stock market data, demonstrating its feasibility in fusing multi-source stock data and establishing its suitability for effectively predicting stock market movements.*

Keywords *- Data fusion, Multi-source data, Graph Convolution Neural Network, Gated Recurrent Unit, Multi-source data fusion graphs, Stock market trends, Stock trend prediction.*

1. Introduction

The dynamics of stock market movements are characterized by their inherent chaos, non-stationary, nonlinearity, and fluctuations. For the past three decades, the estimation of stock market trends has been a focal point of many researchers, especially for traders and investors seeking valuable insights for informed decision-making [1]. Two approaches, specifically fundamental analysis and technical analysis, have emerged as essential methodologies in the area of stock market prediction. Fundamental analysis centers on the intrinsic value of stock and is utilized by human professionals, whereas technical analysis uses historical data to predict future trends [2]. Over the past years, the domain of stock market prediction has undergone a transformative development, marked by the prominent role of technical analysis. The utilization of technical analysis has been fundamental in constructing stock market prediction systems [3].

However, the evolution of this field has witnessed a paradigm shift with the integration of Machine Learning (ML), particularly leveraging the recent advancements of Deep Learning (DL) techniques. These advancements in ML and DL have led to the emergence of sophisticated models that exceed the capabilities of their traditional counterparts, notably in the domain of stock market prediction [4]. Despite notable advancements, many existing methods heavily depend on historical prices and technical indicators, leaving a noticeable gap in accuracy and comprehensiveness. This gap necessitates the development of prediction models that are only robust and capable of incorporating a broader spectrum of information for enhanced accuracy and reliability. The evolution of technology has introduced a new era where the stock market is not only influenced by traditional factors but also by the attitude of social media, exerting a substantial impact on stock trends within the social world [7]-[11]. Daily news and Twitter blogs have emerged as influential factors in shaping stock market trends. Recognizing this shift, there arises an irresistible need for a more comprehensive and diverse set of variables to facilitate accurate predictions. Within this dynamic context, the integration of multi-source data, namely historical prices, technical indicators, as well as insights from news and Twitter blogs offers a valuable foundation for improving stock market movement prediction models. Despite these promising possibilities, developing an effective and reliable stock prediction model remains a challenging task.

Numerous methods have been devoted to stock market movement prediction. Most of these methods have relied on technical indicators to anticipate stock market trends. Recently, some methods have begun to incorporate social media data, such as news data or Twitter data, to forecast stock market movements. Only a few methods have used the fusion of social media data with technical indicators to predict stock market trends. Each data has a distinct impact on stock market trends. The main goal of this paper is to anticipate stock market trends based on multisource data. The novelty of this work lies in its new approach of combining a graph network with fused multisource data to make accurate predictions.

To address this challenge, this study adopts a forward-thinking approach by fusing multi-source data to increase stock market movement prediction performance. The proposed methodology introduces an innovative method to predict stock market movements through the application of a Graph Convolutional Neural Network (GCNN). This approach considers a set of variables to form a comprehensive foundation for robust and accurate predictions. This holistic approach intends to contribute to the advancement of prediction models by embracing the richness of diverse information sources and paving the way for more accurate and insightful stock market trend predictions.

The foremost objectives of this present study are as follows:

- To develop an innovative model that integrates multi-source data with a Graph Convolutional Neural Network (GCNN) for predicting future stock price trends.
- To improve prediction accuracy by overcoming the limitation of single source data through the proposed model.
- To incorporate a varied range of inputs, including historical data, news data, Twitter data, and technical indicators, to provide a comprehensive and sophisticated approach to predicting future stock price movements.
- To assess the efficiency of the presented method through experimental studies conducted on the Indian stock prices spanning from 2018 to 2023.
- To evaluate the potentials of different categories of input data on the predictive power of the model, identifying which sources contribute most significantly to accurate prediction.
- Provide insights into the practical implications of the proposed model for traders, highlighting its benefits for decision-making in the stock market.

The remaining sections of the paper have been organized as follows. In Section 2, a comprehensive examination of the existing literature review has been enumerated. The functionalities of the proposed model have been discussed in Section 3. The simulation outcomes with their discussions are elaborated in Section 4. Lastly, Section 5 concludes the paper with some future recommendations.

2. Review of Literature

The literature on ML models for predicting stock market behavior has explored various methods and techniques. These methods are broadly categorized into two categories. The first category focuses on improving the prediction model's performance, while the second group focuses on merging features to improve prediction performance [1],[2],[3].

Kara et al. [4] investigated the efficiency of Artificial Neural Network (ANN) and Support Vector Machine (SVM) methodologies in predicting stock market trends. This study concentrated on utilizing ten technical indicators as input features for both ANN and SVM for predicting the movement direction of the Istanbul Stock Exchange (ISE), National 100 index. Their findings revealed that ANN outperformed SVM considerably in predicting the index movement. However, this approach may rely too heavily on technical indicators alone, which may not capture all relevant market dynamics, leading to potential inaccuracies in predictions. Similarly, Qiu et al. [5] studied stock market prediction employing ANN in combination with metaheuristic algorithms. Genetic Algorithm (GA) and Stimulated Annealing (SA) were used to fine-tune the parameters of ANN, followed by Backpropagation (BP) for network training. The hybrid approach demonstrated superior performance in predicting the Nikkei 225 index. Nevertheless, the limitation of this approach could be the complexity and computational resources required for implementing metaheuristic algorithms, which may not be feasible for real-time trading applications.

An interesting research on the influence of feature extraction techniques on the efficiency of stock market prediction models was carried out by Zhong and Enke [6]. This study utilized Principal Component Analysis (PCA) and its two variants for feature extraction. The study utilized ANN to predict the S&P 500 index. The study showed an improvement in prediction accuracy when using features produced by PCA compared to the other variables. However, the limitation of this method may be the potential overlook of other feature extraction methods that could further improve predictive performance. Recognizing the irrationality of financial behaviors influenced by psychological factors, recent research has turned to natural languages, like social media and news articles, as the main indicators in this field [7]. Innovative text embedding approaches and ML procedures are being introduced to enhance stock market research.

Joshi et al. [8] conducted a study regarding the correlation among news articles and stock markets using an analysis of sentiments and ML approaches. Three ML algorithms, such as SVM, Naïve Bayes (NB), and Random Forest (RF), were executed to improve accuracy. The outcomes of these three procedures were assessed, and SVM showed enhanced efficiency when compared to NB and RF. While SVM showed improved efficiency compared to NB and RF, a drawback could be the challenge of accurately capturing sentiment scores from the news, which

may contain nuanced language and context. Vargas et al. [9] utilized a Recurrent convolutional neural network (RCNN) for predicting S&P 500 stock market movements. The model data sourced financial news headlines from the day past the forecast day and used a small number of specialized indicators, which were collected from the primary target. The RCNN used two DL models: Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The CNN was utilized for rule-based data separation, while the RNN was used for stock attribute interpretation for prediction purposes. Although this approach demonstrated better results, a limitation could be the potential biases in news headlines and the challenge of interpreting complex financial information accurately.

Gite et al. [10] combined trading data and news to predict the stock market with higher levels of accuracy. In this approach, a Long Short-Term Memory (LSTM) was adopted to build a better stock prediction model. The experimental results clarified that LSTM achieved a competitive result in predicting the Indian stock market compared to LSTM with trading data or news sentiments alone. While LSTM achieved competitive results, it requires a larger amount of labeled data for training DL models, which may not always be available. Sentiments derived from Twitter or news have a remarkable influence on the selling and buying patterns of traders as they easily get inclined by what they read. Therefore, combining one or more dimensions of sentiments improves the prediction accuracy.

Das et al. [11] studied the utility of Twitter feeds for forecasting trends in the stock market. This study presented the sentiment probability for positive and negative opinions to be measured for predicting the stock market. SVM was utilized to define tweets as positive, negative, or neutral sentiments resulting in increased predictive ability on the stock market. However, the existence of noise and volatility in social media data may affect the accuracy of predictions.

Kia et al. [12] designed a hybrid model to predict the direction of stock trends. The method integrated a graph-based semi-supervised approach, using the ConKruG algorithm for constructing a continuous Kruskal-based graph that effectively models global market interactions. Results showed higher prediction accuracy compared to other models. However, a drawback of this model may lie in the complexity and computational resources required for implementing the ConKruH algorithm.

Wang et al. [16] addressed the significance of studying stock market price fluctuations, considering the dynamic and complex nature of the stock market inclined by various factors. The study presented a model involving graph fusion and embedding for handling the diverse information from the Chinese stock market. A multi-attention graph neural network captured the influence of heterogeneous data on stock market fluctuations. High computational complexity is involved in processing and analyzing large volumes of diverse data, which could obstruct scalability and real-time applicability.

Zhang et al. [21] focused on integrating heterogeneous data to analyze their impact on stock market trends. A tensor was created to combine various data types and capture the inherent relationships between data and sentiments. Tensor decomposition predicted stock market movement by finalizing missing values in the sparse tensor. It is hard to capture and represent the complex interplay between heterogeneous data types, which may introduce inaccuracies in the predictive outcomes.

Kabbani and Usta [26] introduced a novel method aimed at forecasting future stock price trends by the application of ML classifiers. The model integrated technical indicators gathered from historical data with sentiment scores computed from news articles released on a given day, thus enriching the feature set used for prediction. The model operated by considering the combined insights gathered from technical indicators and news data to predict future stock trends. The study assessed the performance of three distinct ML classifiers: Logistic Regression (LR), Random Forest (RF), and Gradient boosting algorithm. The findings showed the potential of combining news sentiment scores with technical indicators to improve prediction accuracy in stock price trend prediction.

Xiao and Ihnaini [7] conducted an in-depth exploration into the collective impact of Twitter data, news data, and historical stock data on the prediction of stock market movements. This study begins with the collection and preprocessing of these data, which were subsequently combined to construct a unified feature vector. Different ML classifiers such as SVM, RF, LR, and Naïve Bayes (NB) were employed to predict future stock market movements. The outcomes revealed that NB outperformed the other classifiers in terms of prediction performance. While NB demonstrated better performance than other classifiers, its lower prediction accuracy suggests potential limitations in capturing the complexities inherent in stock price movements.

Theissler et al. [28] conducted an in-depth examination and outlined future research directions concerning the utilization of XAI in time series forecasting. This study involved an analysis of literature in this domain, categorizing various approaches and identifying limitations of the methods. Maqbool et al. [29] constructed an ML approach for stock market prediction. The study aimed to improve stock prediction accuracy by incorporating sentiment scores extracted from financial news with a Multilayer Perceptron (MLP). The method showed better results than other methods. However, one limitation of this approach may be its dependence on sentiment analysis of financial news, which can be subjective and prone to inaccuracies.

Koukaras et al. [30] explored the application of microblogging sentiment analysis combined with ML classifiers for stock market movement prediction. This study investigated the potential of using Twitter data to predict stock market movements. Twitter data can be prone

to misinformation, manipulation, and sudden shifts in sentiment, which may introduce challenges in accurately capturing market movements.

Chen et al. [31] developed a model fusion methodology aimed at predicting stock market trends by using multi-source data. This method integrated a varied array of data types, including weighted unstructured data and structured data. To predict stock market trends, various ML algorithms, including RF, LSTM, and MLP, were employed. However, it is important to note that this method may encounter challenges related to the dynamic nature of financial markets.

Li et al. [32] assessed the theoretical models employed at different levels of data fusion, such as data-level, feature-level, and decision-level fusion, to examine the progression of stock market prediction from a data fusion point and offered a comprehensive overview. The study highlighted the effective application of data fusion methods in the area of stock market prediction.

Ma et al. [27] designed a multi-source aggregated model for predicting stock market trends. This model integrated past data with sentiment analysis of news related to selected stocks. To improve the representation of sentiments from news, the study pre-trained an embedding feature generator to align news sentiments with actual stock market movements.

Additionally, this model used a graph convolutional neural network to gather news effects of associated companies on the target stock. However, the complexity involved in pretraining the embedding feature generator and incorporating the graph network requires significant computational resources, which could limit the accessibility of the system.

3. Proposed Model

In this study, a predictive model using GCNN is proposed to predict future movements in the stock market. The structure of the presented model is depicted in Figure 1. The model constructed in this study defines historical prices, news data, Twitter data, and technical indicators as different types of sub-graphs. It integrates the features of different node types by considering edges and edge weights, using a Gated Recurrent Unit (GRU) for subgraph node aggregation.

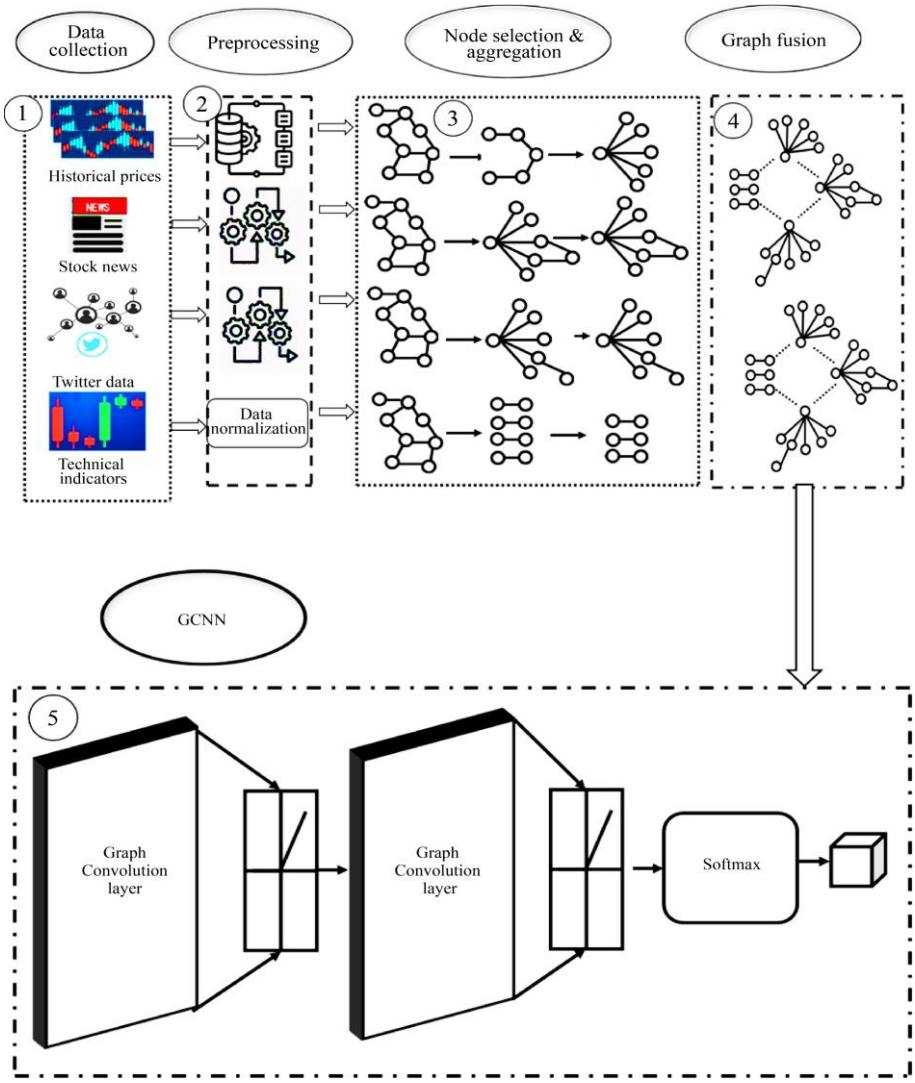
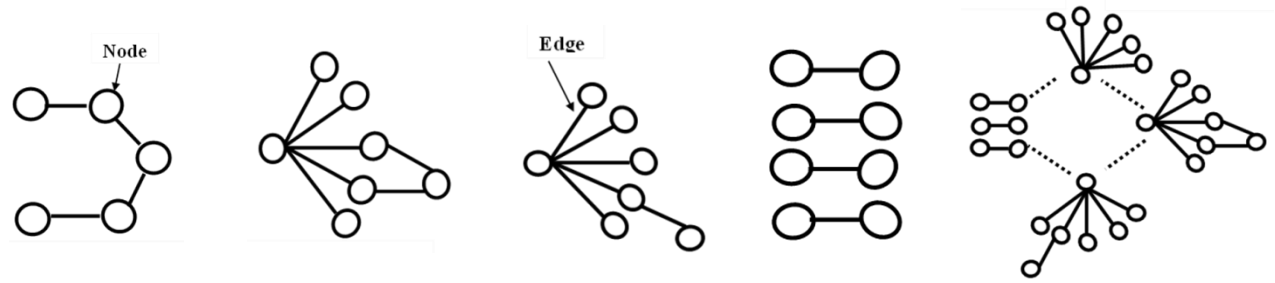


Fig. 1 The general framework of stock market prediction using multi-source data graph



(a) Trading data graph (b) News data graph (c) Twitter data graph (d) Technical indicator graph (e) Embedding graphs with different elements

Fig. 2 Multiple graph fusion

The aggregated results in heterogeneous graph data and attention mechanisms for both homogeneous and heterogeneous nodes are developed by fusing node semantics. This ensures a precise representation of message overflow resulting from information correlation among different semantics. The model uses classification on the heterogeneous graph for predicting the movement in the stock market.

In Phase 1, this study focuses on data collection, illustrating four distinct categories of stock data for input. Phase 2 involves preprocessing tasks like cleaning, handling missing values, and normalization. Subsequently, in phase 3, subgraph construction is performed for the four dissimilar categories of edges in subgraphs, connecting aggregated nodes from various subgraphs and creating distinct weight matrices for various edge types. Phase 4 fuses the sub-graphs, establishing heterogeneous edges based on node properties and constructing a sophisticated network that incorporates multi-source data from the stock market, providing an accurate representation of semantic correlations among different data. Phase 5 of Figure 1 employs a cross-entropy loss function for training the GCNN, facilitating the prediction and classifications of stock market movement data, and combining different indicator features for node aggregation.

3.1. Preprocessing of Heterogeneous Data

Due to the intrinsic nature of stock data, the essential stages of constructing the graph involve acquiring and preprocessing data. This study places significant emphasis on these processes, including both the collection of stock data and textural data. The stock data utilized in this study is sourced from Yahoo Finance [13], spanning the years 2018 to 2023. Each sample includes the opening, closing, lowest and highest price, and volume. Standardization of the historical prices is achieved by means of the application of a min-max approach, bringing all stock data within a uniform range between 0 and 1. Preprocessing of news data and Twitter data is carried out with the NLTK library in Python 3 [14],[15].

Furthermore, technical indicators are computed from the historical data using Equations (1) and (2).

$$\text{Simple Moving Average (SMA)} = \frac{C_1+C_2+C_3....C_n}{n} \quad (1)$$

$$\text{Exponential Moving Average (EMA)} = c_t * \frac{2}{1+N} + \text{PREVIOUS EMA} * \left(1-\frac{2}{1+N}\right) \quad (2)$$

Where c is the closing price, N is the number of samples, and n signifies the number of days. Using Equations (1) and (2), SMA10, SMA20, EMA10, and EMA20 are computed. The obtained indicators and historical data are normalized to the range [0,1] using the min-max method, as expressed in Equation (3),

$$z_{norm} = \frac{z-z_{min}}{z_{max}-z_{min}} \quad (3)$$

Where Z_{norm} represents the normalized value, z is the actual value, z_{min} corresponds to the minimum value, and z_{max} signifies the maximum value.

3.2. Subgraph Construction and Heterogeneous Graph Embedding

Figure 2 provides a graphical representation of the subgraph data embedding outlined in Figure 1. In Figure 2(a), the historical prices subgraph is depicted, where nodes represent five consecutive trading days. Edges connect nodes corresponding to the stability of the data of adjacent trading days. Each node in this subgraph includes five indicators, including opening, closing, lowest, highest price, and volume. Figure 2(b) depicts the stock market news subgraph, treating each daily stock market news as an indicator subgraph. All the news item functions as a subgraph node with a sentiment quantization vector as a node feature. In Figure 2(c), the Twitter data graph is shown, where each Twitter item is taken as a node of the corresponding subgraph, and the associated word is taken as the node feature. Figure 2(d) displays the technical indicators subgraph, including SMA10, SMA20, EMA10, and EMA20. Figure 2(e) shows the heterogeneous network formed by combining these four categories of subgraphs, with edges established between various types of nodes.

The heterogeneity of the various types contributes to the creation of corresponding edge types. The first edge type involves constructing edges between historical price nodes and news nodes; the second edge type connects news data nodes with Twitter data nodes; the third edge type links Twitter data nodes with technical indicator nodes; and the fourth edge type establishes connections between technical indicator nodes and historical prices nodes. To incorporate

each node index attribute with the embedding technique, a node filtering and aggregation technique is utilized during node preprocessing. This step significantly enhances the effectiveness of subsequent processes involving subgraph data fusion and analysis. The GRU presented in [16] is utilized to finalize the aggregation procedure.

3.3. Label Assignment

The foremost aim of this study is to predict the stock market trend on a specified day, D, for a specific bank stock, C. This is quantified by the parameter, $\Delta_{D,C}$, calculated using the formula,

$$\Delta_{\mathrm{D,C}} = \frac{|\mathrm{Close_D} - \mathrm{Close_{D-1}}|}{\mathrm{Close_{D-1}}} \quad (4)$$

In this study, the stock trend prediction problem is framed as a classification task. Each sample represents a pair of specific days and company (D, C). For this pair, the system is equipped with a collection of trading data $T_{D-1,C}$, Twitter blogs $T_{D-1,C}$, news articles $N_{D-1,C}$ published on the previous day (D-1), and technical Indicators $TI_{D-1,C}$ related to the company, C. The classification aims to assign the two labels, 'rise' or 'fall', as described by the following Equation (5),

$$\mathrm{Close_{D,C}} = \begin{cases} \text{rise} & \text{if } |\Delta_{\mathrm{D,C}}| > \lambda \\ \text{fall} & \text{otherwise} \end{cases} \quad (5)$$

Where λ is the threshold, set to 0.01, corresponding to 1% a daily variation in absolute value.

3.4. Graph Convolutional Neural Network

GCNNs have emerged as an influential class of models handling data with complex relational structures. Unlike traditional CNNs designed for grid-like data, GCNNs are specifically designed for graph-structured data, enabling them to capture intricate relationships among interconnected entities [17],[18],[19]. GCNNs typically consist of multiple layers, each performing graph convolutions to iteratively refine the node representations.

The GCNN model takes two main inputs: features, X and an adjacency matrix, A. The role of the GCNN layer is to distribute information from each node to its neighbors, resulting in an updated representation for each node by integrating neighbor information. The formulation for the p^{th} GCNN layer is expressed as follows,

$$H_p = ReLU\left(A.H_{p-1}.W_{p-1}\right) \quad (6)$$

$$\hat{A} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}} \quad (7)$$

$$\tilde{A} = A + I_N \quad (8)$$

$$\widetilde{D_{ii}} = \sum_j \tilde{A}_{ij} \quad (9)$$

Where I is the Identity matrix, $\tilde{A}$ represents the Adjacency matrix of the graph with added self-connections, and W_{p-1} is the weight matrix. In this study, the Rectified Linear Unit (ReLU) is used. A three-layer GCNN model is constructed for predicting stock market movements, represented by the following formula,

$$y = softmax\left(ReLU\left(A.ReLU\left(A.ReLU(AX.W_0\,.W_1\,.W_2)\right)\right)\right) \quad (10)$$

Where W_0 is the input to Graph Convolution Layer 1 (GCL1) weight matrix, W_1 corresponds to the GCL1 to GCL2, and W_2- GCL2 to output weight matrix. The softmax activation function is further applied.

$$softmax(x_i) = \frac{1}{\sum_i \exp(x_i)} \exp(x_i) \quad (11)$$

The GCNN is trained using cross entropy function,

$$Loss = -\sum_{i \in S_N} \sum_{r=1}^{F} S_{ir} ln y_{ir} \quad (12)$$

4. Experiments and Result Analysis

This section initially provides a summary of the stock data used in the experiments and subsequently presents the simulation results of the presented model. In addition to this, the presented model is compared with other standard approaches.

4.1. Dataset Description

To evaluate the potential of the proposed GCNN, a dataset comprising five prominent bank stocks, namely Axis Bank Limited (AXISBANK.NS), HDFC Bank Limited (HDFCBANK.NS), ICICI Bank Limited (ICICIBANK.NS), IndusInd Bank Limited (INDUS INDBK.NS), and Kotak Mahindra Bank Limited (KOTAKBANK.NS), characterized by the maximum market capitalization in the Indian stock market, are utilized. The dataset contains historical market data from the Indian market spanning from 01/01/2018 to 29/09/2023, sourced from Yahoo Finance. The price data for each stock includes open, low, high, close, and volume prices. Additionally, the news data is gathered from press sources, and Twitter data is collected from the Twitter platform. A statistical depiction of the stock trading data is given in Table 1. The data is divided into three subsets. Set 1, covering the interval from 01/01/2018 to 11/01/2022, is used for training; set 2, spanning from 12/01/2022 to 05/08/2022, is employed for validation; and set 3, covering the interval 08/08/2022 to 29/09/2023, serves as the testing dataset.

Table 1. Trading data distribution

Indicators	Open	High	Low	Close	Volume
Min	260.5	264.4	256.5	259.2	2274010
Max	1008	1008.7	990.45	998.3	288408519
Mean	574.9	581.6	567.9	574.9	21926185
Std	224.90	225.7	223.9	224.9	15830701
Range	747.5	744.3	733.9	729.1	286134509
Skewness	0.324	0.319	0.330	0.323	4.981

Table 2. Confusion matrix

		Predicted value	
		Rise	Fall
True value	Rise	TP	FN
	Fall	FP	TN

4.2. Evaluation Criteria

Table 2 displays the confusion matrix of the obtained results, using the rise and fall of stock market closing prices as the basis for classification predictions. In this matrix, a True Positive (TP) is recorded when both the actual and predicted outcomes indicate a rising trend. If the prediction suggests a rise while the actual value is a fall, it is classified as False Positive(FP). On the other hand, a False Negative (FN) happens when the prediction indicates a fall, but the actual value is a rise. If both the actual and predicted results align with a falling trend, it is considered a True Negative. To assess the effectiveness of the proposed stock market prediction model, metrics such as accuracy, recall, precision, and F1-score are considered as expressed in Equations (13)-(16).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{15}$$

$$\text{F1} - \text{score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision+Recall})} \tag{16}$$

Accuracy serves as an indicator of the model's predictive capability, while recall and precision offer insights into its performance. Low recall indicates that samples predicted as negative are likely TN, while high prediction suggests that when the model predicted a TP, it is often accurate. However, there exists a trade-off between recall and precision, leading to the consideration of the F1 score to optimize the model by striking a balance between the two.

4.3. Simulation Results and Discussions

This Section discussed the extraction of features from different sources: historical data, news data, Twitter data, and technical indicators. The subsequent evaluation focused on predicting future stock movements by considering these four feature types individually and in various combinations. The obtained outcomes are detailed in Table 3.

Initially, the effectiveness of the presented model was evaluated using individual data sources. Technical indicators derived from historical prices demonstrated superior performance across all metrics, as indicated by experiments 1-4. Historical prices only achieved moderate accuracy of 53.52% and recall of 60.13%, suggesting that relying solely on historical data might capture some movements, but precision and F1-score are not optimal. News data yielded the lowest accuracy of 49.30% but relatively higher precision, indicating that predictions based on news data are more likely to be correct but do not capture enough positive instances.

Twitter data presented a modest overall performance. Technical indicators provided better results with accuracy, recall, precision, and F1-score of 59.51%, 63.29%, 63.69%, and 63.49%, respectively. Higher accuracy and balanced precision and recall compared to other individual features. This superiority is attributed to the inclusion of information on the rise and fall of the stock market during the computation of technical indicators from past prices, making them more informative compared to other data.

Second, performance with two different types of features was examined, revealing significant accuracy improvement in experiments 5-10 compared to experiments 1-4. Combining historical and news data resulted in improved performance compared to individual features, with accuracy of 55.28%, recall of 62.03%, precision of 59.39%, and F1-score of 60.68%. Additionally, there is a 1.76% and 5.98% improvement in accuracy compared to using only trading data and news data, respectively.

Table 3. Evaluation of the proposed model using different combinations of input features

Expt. No.	Trading data	News data	Twitter data	Technical indicator	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
1	•				53.52	60.13	57.93	59.01
2		•			49.30	47.47	55.15	51.02
3			•		47.54	46.20	53.28	49.49
4				•	59.51	63.29	63.69	63.49
5	•	•			55.28	62.03	59.39	60.68
6	•		•		56.69	62.66	60.74	61.68
7	•			•	62.32	66.46	66.04	66.25
8		•	•		50.35	49.37	56.12	52.53
9		•		•	54.36	60.76	58.18	59.44
10			•	•	55.56	61.39	60.63	61.01
11	•	•	•		57.75	56.96	63.38	60.00
12	•		•	•	68.66	72.78	71.43	72.10
13		•	•	•	66.20	69.62	69.62	69.62
14	•	•	•	•	77.46	82.28	78.31	80.25

Similarly, for trading and Twitter data (experiment 6), the proposed model showed similar performance to experiment 6, indicating that combining historical data with either news or Twitter data yields comparable outcomes. Experiment 6 showed a 3.17%, 2.67%, and 9.15%,12.19% accuracy recall improvement compared to using experiment 1 and experiment 3 data. Experiment 7, combining historical data and technical indicators, exhibited outstanding performance by yielding a higher accuracy of 62.32% compared to experiments 1-4 as well as experiments 5, 6, 8-10. These improvements underscore the influence of technical indicators and trading data in comparison to Twitter and news data. Combining news and Twitter data demonstrated performance like using either news or Twitter data independently, with a slight improvement in precision. It showed 1.05% accuracy improvement for news data and 2.81% for Twitter data. Combining news data with technical indicators (experiment 9) showed improved accuracy and precision, while the combination of Twitter and technical indicators (experiment 10) showed modest improvements. Experiments 9 and 10 demonstrated improved accuracy of 5.06% and 8.02% for news and twitter data, respectively.

Third, the model's effectiveness with combinations of three feature types was investigated. A noticeable enhancement in prediction accuracy was evident in experiments 11-13 compared to experiments 1-10. The presented model showed modest performance when combining past data, news, and Twitter data (experiment 11), suggesting that including all three types of data may not significantly improve predictive capabilities. Substantial improvement in accuracy, recall, and precision was achieved by integrating past data, news data, and technical indicators (experiment 12). Experiment 12 exhibited improvements of 15.14% for Experiment 1, 21.12% for Experiment 3, 9.15% for Experiment 4, 11.97% for Experiment 6, 6.34% for Experiment 7, and 13.1% for Experiment 10.

The combination of past data, Twitter data, and technical indicators showed results comparable to the combination of past and news data, indicating that either news or Twitter data with past data and technical indicators is effective. Furthermore, experiment 13 demonstrated improved accuracy of 16.9%, 18.66%, 6.69%, 15.85%, 11.84%, and 10.64% for experiments 13, 2, 3, 4, 8, 9, and 10, respectively.

Finally, the model's performance with combinations of all four feature types (experiment 14) yielded notable results compared to experiments 1-13. The presented model attained 77.46% accuracy, 82.28% recall, 78.31% precision, and 80.25% F1 score. The presented model showed the highest performance overall metrics, emphasizing the importance of considering all available data sources for stock market prediction. The empirical findings highlight that combining multiple data sources, especially historical data, news data, Twitter data, and technical indicators, leads to improved prediction. The comprehensive utilization of all available data yielded the best results.

4.4. Comparison with Baseline Models

To further validate the superiority, the proposed model was compared with existing models such as SVM [20], Hidden Markov Model (HMM) [22], LSTM [23], TeSIA [24], and Graph Neural Network (GNN) [25], MLP [31], and BiLSTM [27], all of which could be applied for multi-source data. Chen and Hao [20] built a feature-weighted SVM for stock market prediction and reported better prediction outcomes. Chai et al. [22] devised a multisource heterogeneous data analysis technique for predicting forthcoming stock prices by integrating various sources of information.

Utilizing domain-specific emotional dictionaries and relationship diagrams, features were generated from multiple data sources. A multivariate Gaussian mixture model was applied to signify these features and then combined into an HMM. An LSTM model with an attention mechanism was presented for predicting stock market trends based on heterogeneous data [23]. Li et al. [24] employed multi-source information to create a foundation for predictions and demonstrated success in various multi-source data prediction projects. The third-order tensor employed had iterations set at 5000. A stock market prediction model by combining multisource data and GNN was presented by Li et al. [25]. Multiple sources of information were collected, and subgraphs were constructed for each data type, which was then aggregated with GRU and LSTM. A fully connected classifier was utilized to make predictions.

Table 4 lists the simulation parameters in the compared model. All the models were validated with the same data to ensure a more realistic and accurate comparison of outcomes. Figure 3 shows the comparative outcomes of the experimental outcomes between the presented model and other existing methods. As noticed, the proposed model consistently outperformed other models over all metrics, demonstrating the maximum accuracy, recall, precision, and F1-score. The GNN model reached closest to the proposed model in terms of accuracy but still falls short in recall, precision, and F1-score.

GNN showed a slightly higher precision and F1-score, indicating its ability to make accurate positive predictions, while LSTM showed a slightly higher recall. HMM and TeSIA exhibited moderate performance, with HMM showing slightly better results in accuracy, recall, and precision. Traditional models SVM and MLP demonstrated the lowest performance compared to LSTM, BiLSTM, and GNN.

The proposed model achieved notably high recall, indicating its effectiveness in correctly identifying positive instances. Precision and F1-score also showed enhanced performance, suggesting a balanced trade-off between precision and recall. It is proved that the proposed model revealed higher performance, providing higher accuracy and a better balance between recall and precision compared to the baseline models in the comparison.

Table 4. Prediction models' parameters settings

Author names	Models	Parameters
Chen et al. [20]	SVM	C=0.01, Kernel=RBF, max. iteration = 500
Chai et al. [22]	HMM	Components=5, max_ iteration=500
Zhang et al. [23]	LSTM	Hidden neurons=200, dropout=0.1, max_iteration=500
Li et al. [24]	TeSIA	Tensor_ order=3, max_ iteration=5000
Li et al. [25]	GNN	Dropout= 0.1, max_ iteration=1000
Chen et al. [31]	MLP	Hidden layers=3, max_ iteration=500
Ma et al. [27]	BiLSTM	Hidden neurons =200, max_ iteration=1000

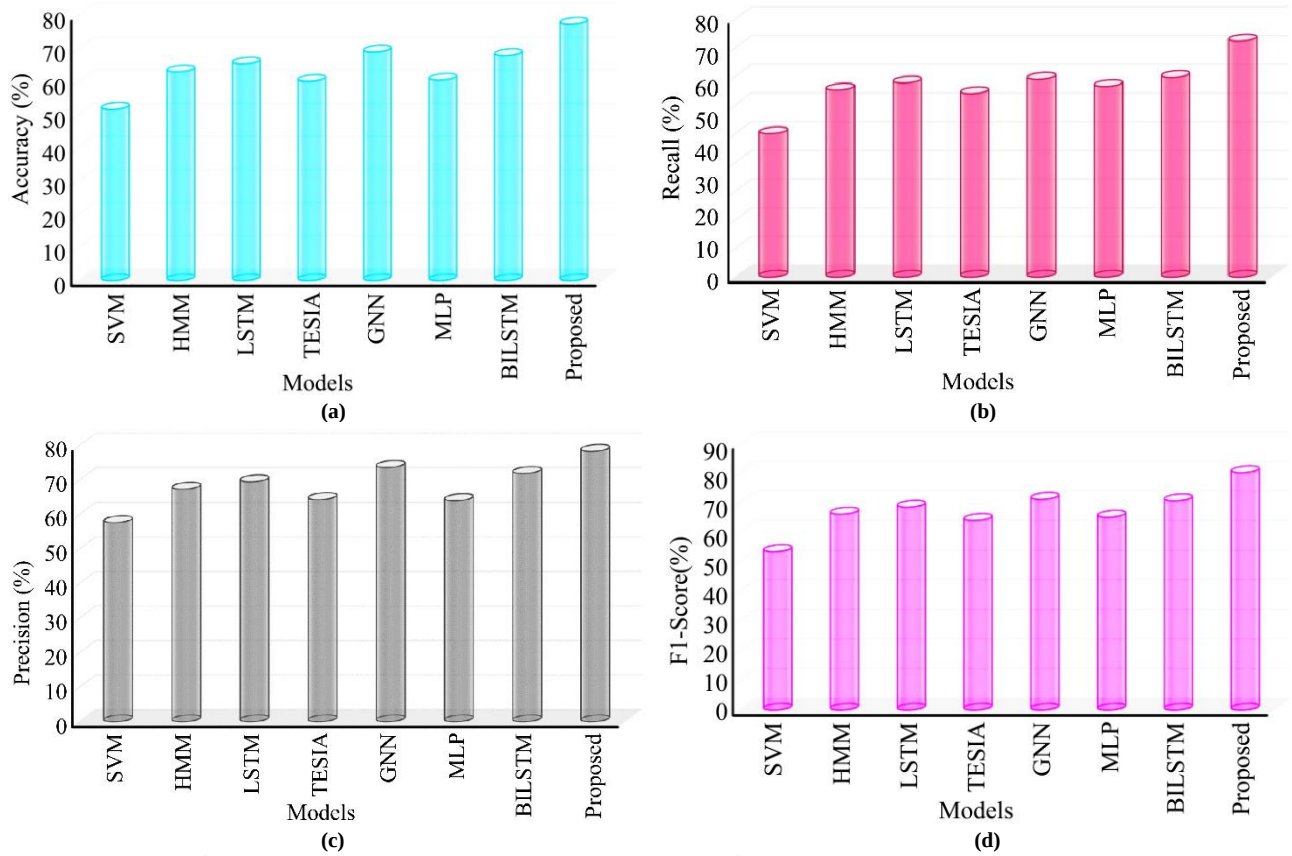


Fig. 3 Performance comparison with the existing methods in terms of (a) Accuracy, (b) Recall, (c) Precision (d) F1-score

4.5. Discussions

Stock price trend prediction has become an important topic of research for both investors and researchers. Over the years, numerous methods have been presented for stock price prediction, predominantly relying on historical data or technical indicators. However, with the advent of social media, its significant influence on stock market movements has gathered considerable attention. While some studies have attempted to predict future stock price trends solely using social media data, it has been observed that such an approach alone may not fully capture the complexities of stock market dynamics. Hence, there is a growing demand for integrating social media data with quantitative data to enhance prediction accuracy.

Motivated by these shortcomings, this study proposes a novel stock price trend prediction model utilizing multi-source data fusion. Specifically, a GCN was designed to predict future stock trends using multisource data. Multisource data was collected, and graphs were generated to represent the relationships between different data points. These graphs serve as input to the GCN for training. A series of experiments were conducted using data from the banking sector to assess the effectiveness of the presented model. Various combinations of input data were explored to validate the model's effectiveness. The outcomes displayed that the proposed model attained promising results with an accuracy of 77.46% along with a high recall of 82.28%, precision of 78.31%, and F1-score of 80.25%, indicating its robust performance in predicting stock trends when utilizing multisource data as input.

Furthermore, the effectiveness of the presented model was assessed with existing techniques. The proposed model demonstrated more accurate results compared to other existing approaches. It was also seen that the presented model outperformed other traditional methods, exhibiting higher accurate predictions. This highlights the effectiveness of the multi-source data fusion approach in enhancing stock trend prediction accuracy and underscores

the potential of GCN in utilizing different data sources for enhanced predictive modeling in the domain of stock market analysis.

5. Conclusion and Future Recommendations

In this study, GCNN is utilized to develop a system for predicting stock market movements utilizing multi-source data fusion graphs. The motivation for this approach originated from the idea that the fusion of heterogeneous data could achieve precise predictions of stock market trends. Unlike other existing studies, this study proposed a GCNN and multi-source data fusion for predicting futures in the stock market. The novel predictive model presented in this study, incorporating multisource data graph fusion, surpasses traditional models, thereby expanding the application of multi-source data fusion and GCNN in the area of predicting fluctuations in the stock market.

Future research will focus on integrating behavioural finance factors into predictive models for stock price trend prediction. In addition, a more universally applicable future predictive model will be developed to compare similarities and differences between the Indian and other markets.

Likewise, future research will aim to evaluate the proposed model's interpretability and explainability, permitting stakeholders to better understand the factors driving stock market predictions and make informed decisions based on the model's insights.

References

[1] Akhter Mohiuddin Rather, V.N. Sastry, and Arun Agarwal, "Stock Market Prediction and Portfolio Selection Models: A Survey," *OPSEARCH*, vol. 54, pp. 558-579, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[2] O. Bustos, and A. Pomares-Quimbaya, "Stock Market Movement Forecast: A Systematic Review," *Expert Systems with Applications*, vol. 156, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Robert D. Edwards, John Magee, and W.H.C. Bassetti, *Technical Analysis of Stock Trends*, 11th ed., CRC Press, pp. 1-686, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Yakup Kara, Melek Acar Boyacioglu, and Ömer Kaan Baykan, "Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5311-5319, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[5] Mingyue Qiu, Yu Song, and Fumio Akagi, "Application of Artificial Neural Network for the Prediction of Stock Market Returns: The Case of the Japanese Stock Market," *Chaos, Solitons & Fractals*, vol. 85, pp. 1-7, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[6] Xiao Zhong, and David Enke, "Forecasting Daily Stock Market Return Using Dimensionality Reduction," *Expert Systems with Applications*, vol. 67, pp. 126-139, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Qianyi Xiao, and Baha Ihnaini, "Stock Trend Prediction Using Sentiment Analysis," *PeerJ Computer Science*, vol. 9, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Kalyani Joshi, H.N. Bharathi, and Jyothi Rao, "Stock Trend Prediction Using News Sentiment Analysis," *International Journal of Computer Science and Information Technology*, vol. 8, no. 3, pp. 67-76, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] Manuel R. Vargas, Beatriz S.L.P. de Lima, and Alexandre G. Evsukoff, "Deep Learning for Stock Market Prediction from Financial News Articles," *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*, Annecy, France, pp. 60-65, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Shilpa Gite et al., "Explainable Stock Prices Prediction from Financial News Articles Using Sentiment Analysis," *PeerJ Computer Science*, vol. 7, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Sushree Das et al., "Real-Time Sentiment Analysis of Twitter Streaming Data for Stock Prediction," *Procedia Computer Science*, vol. 132, pp. 956-964, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Arash Negahdari Kia, Saman Haratizadeh, and Saeed Bagheri Shouraki, "A Hybrid Supervised Semi-Supervised Graph-Based Model to Predict One-Day Ahead Movement of Global Stock Markets and Commodity Prices," *Expert Systems with Applications*, vol. 105, pp. 159-173, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yahoofinance, Finance Yahoo. [Online]. Available: www.yahoofinance.com

[14] Natural Language Toolkit, NLTK, 2023. [Online]. Available: https://www.nltk.org

[15] Edward Loper, and Steven Bird, "NLTK: The Natural Language Toolkit," *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia Pennsylvania, vol. 1, pp. 63-70, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[16] Jun Wang et al., "A Graph-Based Approach to Multi-Source Heterogeneous Information Fusion in Stock Market," *Plos One*, vol. 17, no. 8, pp. 1-23, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Thomas N. Kipf, and Max Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *arXiv*, pp. 1-14, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[18] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," *arXiv*, pp. 1-9, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[19] Qiang QU, Hongtao YU, and Ruiyang Huang, "Spammer Detection Technology of Social Network Based on Graph Convolution Network," *Chinese Journal of Network and Information Security*, vol. 4, no. 5, pp. 39-46, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] Yingjun Chen, and Yongtao Hao, "A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction," *Expert Systems with Applications*, vol. 80, pp. 340-355, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[21] Xi Zhang et al., "Improving Stock Market Prediction via Heterogeneous Information Fusion," *Knowledge-Based Systems*, vol. 143, pp. 236-247, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Lei Chai et al., "A Multi-Source Heterogeneous Data Analytic Method for Future Price Fluctuation Prediction," *Neurocomputing*, vol. 418, pp. 11-20, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Qun Zhang, Lijun Yang, and Feng Zhou, "Attention Enhanced Long Short-Term Memory Network with Multi-Source Heterogeneous Information Fusion: An Application to BGI Genomics," *Information Sciences*, vol. 553, pp. 305-330, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Qing Li et al., "Tensor-Based Learning for Predicting Stock Movements," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, pp. 1784-1790, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[25] Xiaohan Li et al., "A Graph Neural Network-Based Stock Forecasting Method Utilizing Multi-Source Heterogeneous Data Fusion," *Multimedia Tools and Applications*, vol. 81, no. 30, pp. 43753-43775, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Taylan Kabbani, and Fatih Enes Usta, "Predicting The Stock Trend Using News Sentiment Analysis and Technical Indicators in Spark," *arXiv*, pp. 1-4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Yu Ma et al., "Multi-Source Aggregated Classification for Stock Price Movement Prediction," *Information Fusion*, vol. 91, pp. 515-528, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Andreas Theissler et al., "Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions," *IEEE Access*, vol. 10, pp. 100700-100724, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[29] Junaid Maqbool et al., "Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach," *Procedia Computer Science*, vol. 218, pp. 1067-1078, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[30] Paraskevas Koukaras, Christina Nousi, and Christos Tjortjis, "Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning," *Telecom*, vol. 3, no. 2, pp. 358-378, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Xi Chen et al., "A Model Fusion Method Based on Multi-Source Heterogeneous Data for Stock Trading Signal Prediction," *Soft Computing*, vol. 27, no. 10, pp. 6587-6611, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[32] Aihua Li et al., "Research on Stock Price Prediction from a Data Fusion Perspective," *Data Science in Finance and Economics*, vol. 3, no. 3, pp. 230-250, 2023. [CrossRef] [Google Scholar] [Publisher Link]