ORIGINAL ARTICLE

A deep fusion model for stock market prediction with news headlines and time series data

Pinyu Chen[1] · Zois Boukouvalas[2] · Roberto Corizzo[1]

Abstract

Time series forecasting models are essential decision support tools in real-world domains. Stock market is a remarkably complex domain, due to its quickly evolving temporal nature, as well as the multiple factors having an impact on stock prices. To date, a number of machine learning-based approaches have been proposed in the literature to tackle stock trend prediction. However, they typically tend to analyze a single data source or modality, or consider multiple modalities in isolation and rely on simple combination strategies, with a potential reduction in their modeling power. In this paper, we propose a multimodal deep fusion model to predict stock trends, leveraging daily stock prices, technical indicators, and sentiment in daily news headlines published by media outlets. The proposed architecture leverages a BERT-based model branch fine-tuned on financial news and a long short-term memory (LSTM) branch that captures relevant temporal patterns in multivariate data, including stock prices and technical indicators. Our experiments on 12 different stock datasets with prices and news headlines demonstrate that our proposed model is more effective than popular baseline approaches, both in terms of accuracy and trading performance in a portfolio analysis simulation, highlighting the positive impact of multi-modal deep learning for stock trend prediction.

Keywords Stock market analysis · Portfolio analysis · Deep learning · Multimodal learning · Sentiment analysis · Time series prediction

1 Introduction

From the start of the 20th century, the financial sector has made consistent investments in researching price prediction and market dynamics models [1]. For the purpose of forecasting stock price trends, conventional quantitative approaches rely on historical time series price data [2]. In recent days, leveraging models to analyze financial time series has become essential for managing market risks and making informed investment decisions [3].

An escalating quantity of progressively advanced models is being introduced in research works to address the inherent intricacies of time series data within specific domains. Notably, stock market data is quite complex, as it is characterized by a multi-dimensional, volatile, and dynamically evolving nature. Furthermore, stock market data displays interconnections with various external factors, such as macroeconomic events and news disseminated by media sources. Consequently, appropriately integrating these factors is of paramount importance when developing predictive models that yield satisfactory levels of accuracy.

Autoregressive models [4–10] are proficient in addressing prediction tasks that take into account temporal auto-correlation, but fall short in adequately accounting for the multivariate nature of the data and the intricacies of nonlinear feature interactions. On the other hand, machine learning and deep learning models tailored for temporal

Pinyu Chen and Roberto Corizzo have equally contributed to this work.

✉ Roberto Corizzo
rcorizzo@american.edu

Pinyu Chen
pc1694a@american.edu

Zois Boukouvalas
boukouva@american.edu

1 Department of Computer Science, American University, 4400 Massachusetts Avenue, NW, Washington, DC 20016, USA

2 Department of Mathematics and Statistics, American University, 4400 Massachusetts Avenue, NW, Washington, DC 20016, USA

data, exemplified by long short-term memory models and their various adaptations [11–21], are capable of mitigating these limitations. However, their utility is typically confined to the analysis of a single data source, thereby diminishing their effectiveness in highly volatile and challenging-to-predict domains. Therefore, it is advisable to explore ensemble-based and hybrid combination methods [22–32] as they hold greater promise compared to other approaches. These methods can encompass multiple data sources and harness the diversity of multiple predictors, offering a more comprehensive solution to address the challenges of data analysis in such complex and dynamic domains.

A significant drawback of many works in the literature is that they are limited to the analysis of a single source of data (or modality). Some studies highlighted that the analysis of financial news in addition to stock prices may play a key role in stock market prediction [33, 34]. For this reason, researchers are focusing on devising new and more sophisticated ways to integrate different relevant sources of data that may impact stock prices, resulting in more accurate models.

However, several methods that consider multiple sources of data, address them in isolation, relying on simple combination strategies to perform a joint analysis. Moreover, the specialized terminology and scarcity of labeled data in the financial industry exacerbates the difficulty of accurately performing sentiment analysis, making general-purpose text-based models insufficient. To this aim, language models fine-tuned on financial textual data present provide new exciting opportunities for the integration of accurate textual analysis in stock market prediction models [35].

In this paper, we propose a multimodal deep learning model for stock market trend prediction that consists of two branches: a FinBERT branch which specializes on the analysis of the textual content in financial news and accurately model market sentiment, and a LSTM branch which captures temporal market dynamics in complex multivariate data, including stock prices and technical indicators. Our deep fusion approach allows us to effectively leverage multiple modalities leading to improved generalizability, reduced bias, and increased efficiency compared to single-modality approaches.

In summary, the main contributions of our paper can be summarized as follows:

- We propose a deep fusion model architecture for stock market trend prediction that seamlessly considers and integrates multiple modalities (stock prices, technical indicators, news headlines) in a joint feature space with multiple specialized branches, empowering the model with a more comprehensive understanding of patterns and complex nonlinear relationships in stock market dynamics that leads to the extraction of more robust and trustworthy next-day trend predictions;
- We devise an end-to-end optimization and hyperparameter tuning workflow which allows us to identify and select highly effective configurations for each branch of the fusion model, resulting in a competitive stock prediction performance tailored to the characteristics of a specific stock under analysis;
- We perform an extensive evaluation of 12 real-world stocks from different sectors in two different evaluation periods and with different market conditions (uptrend, downtrend). This evaluation encompasses two analytical perspectives: model accuracy and portfolio performance in a realistic simulation, where model predictions are leveraged for practical automated trading decisions. Our experimental results show that our approach can outperform state-of-the-art methods for stock market prediction.

The paper is structured as follows. Section 2 summarizes relevant works for stock market prediction. Section 3 describes our proposed method in detail. Section 4 discusses our experimental settings and the results obtained in our experiments. Section 5 wraps up the paper and discusses relevant directions for future work.

2 Related work

In this section, we review relevant works pertaining to time series prediction and forecasting, with particular focus on stock market analysis.

2.1 Autoregressive models

Autoregressive models are recognized for their ability to characterize associations across multiple time steps for a target feature through the learning of coefficients. One prominent autoregressive technique is the autoregressive integrated moving average (ARIMA) model [4]. Renowned for its efficacy in short-term prediction tasks, ARIMA forecasts the future value of a variable by linearly combining past values and errors, following the application of differentiation operations to render the time series stationary. The methodology outlined in [5], for instance, employs an ARIMA model to predict coronavirus cases using Johns Hopkins epidemiological data. Prophet [6] stands out as another prevalent autoregressive forecasting method, grounded in an additive model that accommodates nonlinear trends through yearly, weekly, and daily variations. This approach also addresses seasonality and holiday effects, demonstrating robustness to outliers and missing

data. Prophet aims at enhanced configurability and user-friendliness in comparison with ARIMA. Vector autoregression (VAR) [7] is another noteworthy autoregressive approach that extends beyond predictive tasks for single variables. This method concurrently learns coefficients for multiple variables, considering their temporal correlations. A noteworthy investigation in [8] highlights its effectiveness in forecasting tasks embedded within a spatiotemporal context. In the realm of stock market applications, recent research [9] introduces a moving average heterogeneous autoregressive (MAT-HAR) model, treating thresholds as a moving average-generated, time-varying parameter. This model is employed to forecast the monthly realized volatility of the US stock market. Another study [10] applies univariate ARIMA models to the Amman Stock Exchange. Despite their effectiveness in numerous applications, autoregressive models exhibit certain limitations. Besides being limited to the analysis of single variables or modalities, their simplicity makes them incapable of capturing nonlinear relationships between multiple variables, which are frequently encountered in real-world multivariate data.

2.2 Machine learning and deep learning models

Machine learning and deep learning models tailored for temporal data, such as long short-term memory (LSTM) [36] models and their variations, constitute an advancement over autoregressive models due to their capacity to effectively analyze multivariate data and handle nonlinear feature interactions. In the research by [12], recurrent neural network (RNN) models featuring long short-term memory units are proposed to predict pollutant particle levels at multiple time horizons. In the domain of stock market analysis, [13] employs an LSTM model to predict the next-day closing price of the S &P 500 index, utilizing nine predictors selected from fundamental market data, macroeconomic data, and technical indicators. The authors in [14] introduce an LSTM-based model architecture for forecasting air leaks, assessing its potential within the healthcare sector. The work in [15] devises a tensor decomposition approach for feature extraction, where predictive clustering trees are used for forecasting, and their performance is compared to LSTM models. A comparative investigation of deep neural networks with LSTM networks for stock market analysis is presented by [16], focusing on daily and weekly movements of the Indian BSE Sensex index. Another work by [11] empirically analyzed LSTM networks leveraging a diverse set of real-world datasets, and identified that such models are quite effective in predicting stock market prices. The study in [17] conducts a comparative analysis involving LSTM, gated recurrent unit (GRU), and drop-GRU models in the context of power consumption forecasting, demonstrating the satisfactory performance of the devised models in this application. Combinations of GRU and convolutional neural networks (CNN) have also been explored. For instance, the GRU-CNN model proposed in [37] has shown to be effective for stock market prediction. A decision support system reinforced with LSTM for swing trading is proposed in [18], where predictions and reports that incorporate forecasted values of company stock for the next 30 days are extracted, alongside technical indicators. In the research by [19], bidirectional and stacked LSTM predictive models are benchmarked against shallow neural networks and simplified forms of LSTM networks, with analyses conducted on publicly available stock market data.

The work in [20] demonstrates that LSTM networks combined with bidirectional gated recurrent unit (BiGRU) can accurately predict the closing price of stock market, offering a more competitive performance than simpler models. In [21], a bidirectional LSTM model (Bi-LSTM), proposed for the first time for speech recognition tasks [38], is adopted and optimized by particle swarm optimization (PSO), giving place to a PSO-Bi-LSTM approach to predict useful long-, mid-, and short-term investment strategies. A CNN-LSTM model complemented by an attention mechanism was proposed in [39]. Dilated convolutions have been explored in [40] and have shown great success in extracting multi-scale patterns at different time granularities. A common limitation of these approaches is their confinement to the analysis of a single data type or modality, which constrains their effectiveness in the presence of highly volatile phenomena that depends on multiple factors.

2.3 Ensemble-based and hybrid models

Ensemble-based and combination methodologies involving hybrid models offer a robust approach to address these intricacies, encompassing the utilization of multiple data sources and the amalgamation of various predictors. An AI platform, as proposed by [22], leverages four machine learning ensemble methods, namely neural network regression ensemble, support vector regression ensemble, boosted regression trees, and random forest. The best ensemble method for a given stock is selected through a cross-validation evaluation. In [23], a fusion network is proposed to extract text and numerical information for stock price prediction, with the addition of an attention mechanism to improve the overall model performance.

A stacking ensemble approach for predicting stock closing prices is proposed in [24], where a competitive performance is obtained when contrasted with conventional machine learning ensemble models such as random

forest, AdaBoost, and gradient boosting machines. A stacking approach was also explored in [41] with joint consideration of news headlines, multivariate time series data, and multiple base models as predictors. Authors in the work by [25] propose a hybrid forecasting model for stock prices that integrates various deep learning models, specifically, CNN-LSTM [42], GRU-CNN [37], and ensemble models. The work in [26] introduces a hybrid model denoted as PCA-EMD-LSTM, which combines principal component analysis, empirical mode decomposition, and LSTM for predicting stock market trends in Thailand. The hybrid model proposed in [27] utilizes decomposition techniques, multi-factor analysis, and attention-based LSTM to forecast stock market price trends in four major Asian countries. In [28], a hybrid method for analyzing stock markets is introduced, which combines an autoencoder-based feature extraction network with a temporal convolutional model architecture and a temporal clustering optimization algorithm utilizing the KL (Kullback–Leibler) divergence. The approach in [43] employs a CNN model to perform sentiment classification and integrates it to a LSTM analyzing technical indicators, showing that the joint consideration of both aspects leads to improved predictions. A deep learning approach is proposed in [29], where future stock prices are predicted by a blending ensemble learning model that combines two recurrent neural networks followed by a fully connected neural network. The authors in [30] conduct an analysis of the collective sentiment's significance on popular S &P500 stocks and assess its efficacy in investment decision-making. A study in [31] presents a framework based on LSTM and convolutional neural networks to predict the closing prices of Tesla and Apple, utilizing historical data collected over the past two years. Two stock trading decision methods have been applied in [32]: nested reinforcement learning (Nested RL) using three deep reinforcement learning models, and a weighted random selection with confidence (WRSC) strategy. The results show that their approach outperforms baselines, enhancing portfolio management for higher profits at the same risk level.

3 Method

In this section, we describe our method in detail, focusing on its workflow. The proposed method involves a multimodal deep learning approach that combines information from historical stock prices and statistical indicators with news headlines. The model employs long short-term memory (LSTM) networks and bidirectional encoder representations from transformers (BERT) models to capture both quantitative and qualitative information. For the text component, we leverage FinBERT to conduct sentiment

analysis. A fusion layer combines the two modalities and yields a final stock price prediction. A graphical representation of the method is shown in Fig. 1.

3.1 Data gathering, preprocessing, and fusion

Our method aims to comprehensively analyze stock prices, statistical trading indicators, and news headlines. To achieve this goal, we obtain data from two Application Programming Interfaces (APIs). The initial API, referred to as Yahooquery,[1] is employed for retrieving historical stock prices, serving as an unofficial substitute for the obsolete Yahoo Finance API. We input stock tickers in string format, and the API provides all historical data available for a specified stock within a given date range. The data obtained from the API comprises the opening price, high and low values, adjusted closing prices, and the daily observed volume. Subsequently, we employ the TA-lib Python library[2] to incorporate computed statistical indicators into the extracted stock data. These statistical indicators, widely used in technical analysis, encompass the exponential moving average (12-day, 26-day), moving average convergence/divergence (MACD), parabolic SAR, Bollinger bands (upper band, middle band, and lower band), and stochastic (Slow k, Slow d). Before model training, all numerical data undergoes min-max normalization. In our study, we do not perform feature selection to identify a subset of suitable features for each stock. Although some features may be more relevant than others for a given stock under analysis, the adopted deep learning architecture should, in principle, automatically learn feature influence via gradient descent optimization. Specifically, weak features will be characterized by small weights with a vanishing effect in the deeper layers of the network and a tendency to be discarded for prediction. In contrast, relevant features will lead to strong/high activation values that influence the prediction significantly.

As a secondary API, we utilize the end-of-day historical financial data (EODHD) API to fetch news headlines based on a given stock ticker. For instance, a query with the "aapl" ticker as input yields data on news articles, including posting time, titles, article content, URL links, as well as tagged symbols and tickers. To maintain focus on reliable news sources, we exclusively retain articles sourced from Yahoo Finance, discarding those from other origins.
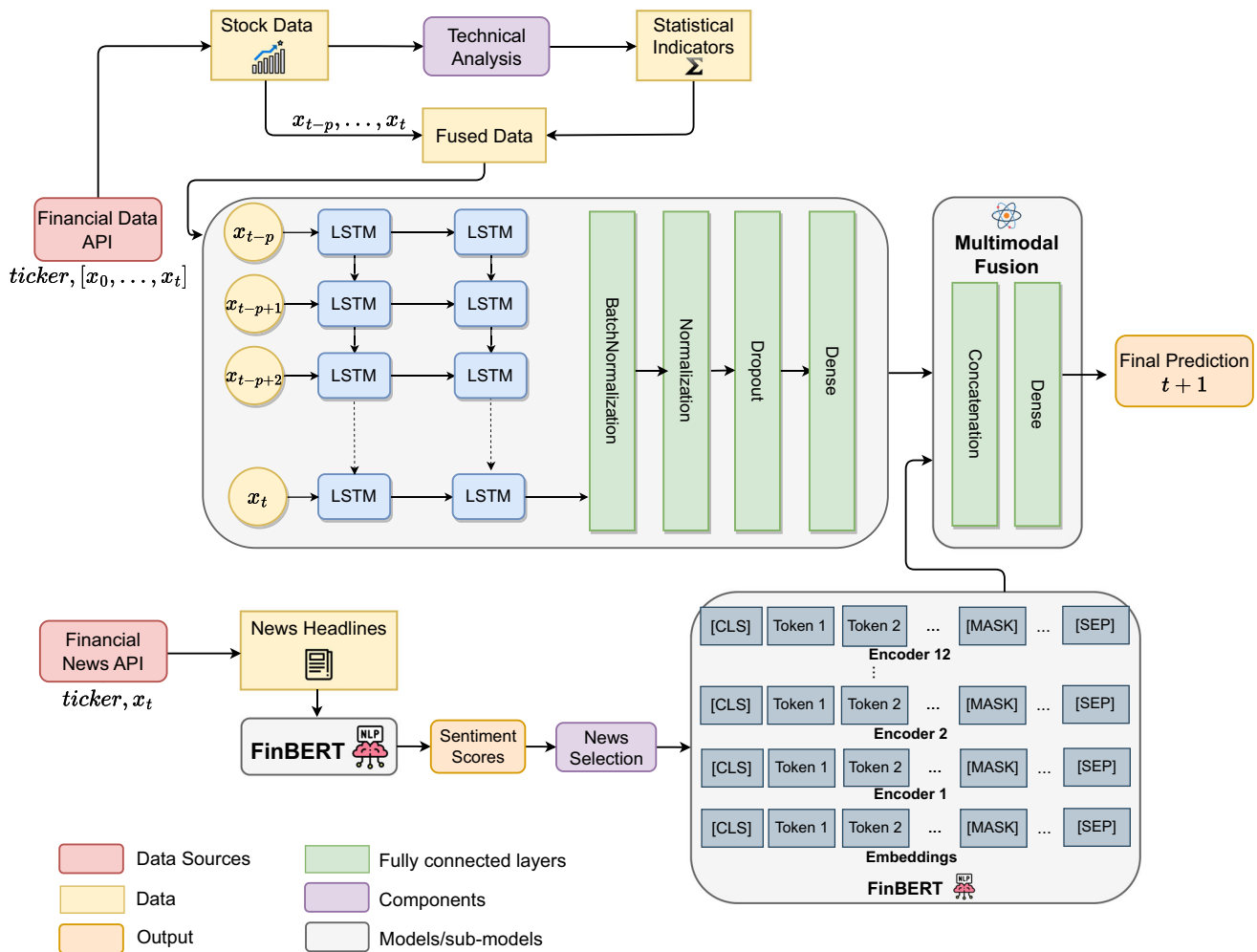
---

[1] https://pypi.org/project/yahooquery/.

[2] https://pypi.org/project/TA-Lib/.

Fig. 1 Overview of the proposed multimodal fusion model for stock market prediction

3.2 Long short-term memory (LSTM)

Long short-term memory (LSTM) neural networks represent a category of recurrent neural networks (RNN) extensively applied in the analysis of time series data, owing to their ability of capturing prolonged dependencies within sequential data [44]. The utility of LSTM models lies in their capacity to discern and forecast patterns in time series data, which makes them valuable for predictive tasks. LSTMs address the challenge of vanishing and exploding gradients encountered in traditional RNNs [45, 46] by introducing memory cells to replace recurrent nodes. A distinguishing feature of a memory cell is its internal state, facilitating the flow of gradients across multiple time steps without vanishing or exploding [44].

Each memory cell comprises multiple nodes referred to as gates. The data from the current time step is fed into these LSTM gates, as well as into the hidden state from the preceding time step. Subsequently, three fully connected layers compute the values associated with the input, forget, and output gates. A sigmoid activation function is applied to these values to yield the final output, constrained within a (0, 1) range.

An input node undergoes computation through a *tanh* activation function. In essence, the gates modulate the significance of the information passed to the model at distinct time steps. The *input gate* gauges the proportion of the input node's value to be added to the current internal state of the cell. The *forget gate* determines whether the prevailing value of the cell should be retained or discarded. Finally, the *output gate* decides if the memory cell should contribute to the output of the ongoing time step.

Assuming the presence of d inputs, h hidden units, and a batch size of n, the input is defined as $\mathbf{X}_t \in \mathbb{R}^{n \times d}$, and the hidden state of the previous time step is defined as $\mathbf{H}_{t-1} \in \mathbb{R}^{n \times h}$. The gates at time step t are defined as follows: the input gate is $\mathbf{I}_t \in \mathbb{R}^{n \times h}$, the forget gate is $\mathbf{F}_t \in \mathbb{R}^{n \times h}$, and the output gate is $\mathbf{O}_t \in \mathbb{R}^{n \times h}$. Formally, they are calculated as:

Table 1 Details on hyperparameter tuning: LSTM branch

Hyperparameter	Description	Range
Layers size	The number of layers in a LSTM model	(2, 4, 6, 8)
Hidden units size	The number of hidden units per layer	(100, 150, 200, 250, 300, 350, 400, 450, 500)
Normalization	Standardizes input for efficient training	(Included, non-included)
BatchNormalization	Standardizes activations in each batch	(Included, non-included)
Dropout rate	Randomly excludes input neurons	(0.1, 0.2, 0.3, 0.4)
Learning rate	Controls step size in optimization	(0.01, 0.003, 0.0001)
Representation size	Length of the LSTM vector representation	(36, 72, 108, 140, 176)

Table 2 Details on hyperparameter tuning: FinBERT branch

Hyperparameter	Description	Range
Layers size	Number of dense layers	(2, 4, 6, 8)
Hidden units size	The number of hidden units per layer	(40, 60, 80, 100, 120, 140, 160)
Normalization	Standardizes input for efficient training	(Included, non-included)
BatchNormalization	Standardizes activations in each batch	(Included, non-included)
Dropout rate	Randomly excludes input neurons	(0.2, 0.3, 0.4)
Learning rate	Controls step size in optimization	(0.01, 0.003, 0.0001)
Representation size	Length of the embedding vector representation	(36, 72, 108)

Table 3 Details on hyperparameter tuning: Multimodal Fusion branch

Hyperparameter	Description	Range
Layers size	The number of layers in the multimodal fusion branch	(2, 4, 6, 8)
Hidden units size	The number of hidden units per layer	(40, 60, 80, 100, 120, 140, 160)
Normalization	Standardizes input for efficient training	(Included)
Learning rate	Controls step size in optimization	(0.0001)

$$\mathbf{I}_t = \sigma(\mathbf{X}_t \mathbf{W}_{xi} + \mathbf{H}_{t-1} \mathbf{W}_{hi} + \mathbf{b}_i), \tag{1}$$

$$\mathbf{F}_t = \sigma(\mathbf{X}_t \mathbf{W}_{xf} + \mathbf{H}_{t-1} \mathbf{W}_{hf} + \mathbf{b}_f), \tag{2}$$

$$\mathbf{O}_t = \sigma(\mathbf{X}_t \mathbf{W}_{xo} + \mathbf{H}_{t-1} \mathbf{W}_{ho} + \mathbf{b}_o), \tag{3}$$

where $\mathbf{W}_{xi}, \mathbf{W}_{xf}, \mathbf{W}_{xo} \in \mathbb{R}^{d \times h}$ are weight parameters and $\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_o \in \mathbb{R}^{1 \times h}$ are bias parameters.

The incorporation of LSTM cells equips the model with the ability to address intricate temporal patterns in multivariate data, enabling the capture of nonlinear and enduring relationships among various features and timestamps. This capability is leveraged to allow the model to discern resilient patterns within historical data, encompassing statistical indicators, and facilitate the extraction of relationships between stock prices and other descriptive features.

3.3 FinBERT

BERT (bidirectional encoder representations from transformers) is a complex deep neural network model for natural language processing (NLP). BERT achieved state-of-the-art results in various NLP tasks such as text classification, question answering, and named entity recognition. BERT uses a transformer architecture that allows capturing long-range dependencies and context in text data, making it highly effective for tasks involving understanding and processing human language. The high accuracy documented in several research works supports the adoption of BERT as a versatile model for many different NLP tasks [47]. Among them, BERT is often used to extract contextual embedding vectors from text, which can be adopted for subsequent downstream tasks. However, the performance of the model is strictly related to the pertinence of the dataset used to train the model. While using pre-trained general-purpose language may be a practical solution to avoid expensive training costs, it may result in a poor representation of topic-specific textual content [48]. To overcome this limitation, we leverage FinBERT [35], a language model specialized for financial data analysis, which obtained the highest scores on FiQA sentiment

scoring and Financial PhraseBank benchmarks, outperforming other popular large language models including GPT-4 [49].

The model architecture consists of multiple stacked transformer layers, which allow the model to capture complex contextual representations. Each layer features a self-attention mechanism, which computes the weighted sum of values (V) based on queries (Q) and keys (K):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

The model adopts multiple attention heads, which can be formalized as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O, \tag{5}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$.

The output of each transformer can be computed as:

$$\text{LayerOutput} = \text{LayerNorm}(x + \text{MultiHead}(x) + \text{FFN}(x)), \tag{6}$$

where FFN is a simple feed-forward neural network, and $LayerNorm(x) = \frac{x-\mu}{\sigma}$ is the layer normalization, with μ and σ being the mean and standard deviation, respectively.

To prevent catastrophic forgetting, FinBERT applies three state-of-the-art techniques: slanted triangular learning rates, discriminative fine-tuning, and gradual unfreezing.

FinBERT takes an initial BERT model trained on BookCorpus and Wikipedia, an fine-tunes it on the TRC2-financial corpus, a subset of Reuters' TRC24, which consists of 1.8M news articles published by Reuters between 2008 and 2010. Subsequently, FinBERT is fine-tuned on Financial Phrasebank corpus consists of 4845 English sentences from financial news found on LexisNexis database, annotated by 16 people with background in finance and business [50].

FinBERT extracts sentiment scores for all news headlines gathered for a specific stock on a given day. It returns a positive, neutral, and negative score for each news. For textual data, we remove stopwords, punctuation marks, square brackets, and lowercase, in order to reduce noise and focus on meaningful words. Initially, we obtain a summary of the day consisting of two values: sum of positive scores, and sum of negative scores. Subsequently, the largest of the two scores determines if the day is overall positive or negative. Based on this information, we select the most representative news headline, i.e., the one with the largest positive or negative score, and extract its embedding vector representation.[3]

We note that sentiment scores extracted by FinBERT are used exclusively for news selection. Separately, we fine-tune FinBERT with the financial news in our dataset, and we replace the output layer with a single-unit dense layer (to predict uptrend/downtrend directly) and optimize it considering different hyperparameter configurations (see Table 2). The best-performing configuration is identified based on accuracy using a validation set.

Afterward, in the proposed model architecture, we remove the classification layer used during optimization and exploit the embedding vector representation of the news for subsequent fusion.

3.4 Multimodal fusion

Our novel multimodal fusion approach is tailored for next-day stock market trend prediction. This branch of the model is responsible for fusing the two data modalities: time series and text. More in detail, the model incorporates time series data processed through LSTM (long short-term memory), a type of recurrent neural network renowned for its effectiveness in handling sequential data, and text embeddings processed through FinBERT, a specialized model specifically tailored for the analysis of financial text data. The primary objective of this approach is to enhance prediction accuracy and robustness by fusing information from different data sources or modalities.

The structural layout of the model is visually depicted in Fig. 1, illustrating how the two data modalities are seamlessly integrated. This model leverages multimodal learning, which empowers the model with a more comprehensive understanding of the underlying patterns and relationships within the data. This, in turn, can result in an improved ability to withstand unexpected market fluctuations and enhance prediction resilience [51]. This observation is substantiated by prior research conducted across various applications.

The fusion of these data modalities is achieved through a specific process involving a concatenation layer and a series of dense layers.

The temporal granularity of data processed by the different model branches is aligned. For each day, the LSTM model processes a single multivariate data instance containing stock prices and technical indicators. Likewise, the FinBERT model leverages the most representative news headline of the day (as explained in the previous subsection). Since the downstream task of interest is next-day stock trend prediction, a daily time granularity is appropriate, and it allows us to train models efficiently

---

[3] Alternative schemes such as concatenation or aggregation of all news and their consideration in the embedding vector representation

Footnote 3 continued
were tried, but yielded sub-par results in our experiments, possibly due to the noise introduced by conflicting information.

considering a large time frame. Both model branches generate a vector embedding which is subsequently provided to the concatenation layer and results in the vertical concatenation of the two vector embeddings.

More in detail, for FinBERT, the preprocessing phase involves tokenizing the input text for BERT, adding special tokens like [CLS] and [SEP]. The embeddings generated by BERT offer contextual representations for each token, capturing nuanced contextual relationships. These BERT embeddings are then merged into the LSTM embeddings by concatenating them with the input sequences, allowing the multimodal model to leverage the rich contextual information from both BERT and LSTM. Notably, the configuration of these layers is customized to suit the dataset's characteristics and the particular prediction task at hand. This strategy for model architecture optimization was proven to be beneficial in [52]. Details of the architecture's optimization and tuning are shown in Table 3, which contains information on the various hyperparameters and configurations considered in the optimization process. We conduct a tuning process leveraging AAPL, TSLA, and MSFT stocks to identify an effective model architecture configuration (layers, number of neurons, etc.).

To this end, LSTM and FinBERT base models are optimized independently, and are then combined in the multimodal architecture, which is further optimized. For hyperparameters used in this process, see Tables 1, 2, and 3, respectively. The selected values for the hyperparameters optimization stage are motivated by works in the literature providing effective heuristics. Specifically, for learning rate, the authors in [53] suggest to start from a default value of 0.01 and experiment with a decreasing factor (negative power of 10), where 10^{-6} is considered as an extremely small value. Dropout and batch normalization have been widely recognized as beneficial as a regularization technique to reduce overfitting in neural networks. As for the dropout rate, works in [54] and [55] have shown that values below 0.5 should be preferred, in order to avoid the removal of too many neurons that would cause an under-learning phenomenon. The number of LSTM and dense layers, as well as the number of neurons in each layer, also play an important factor, where multiple layers generally allow the model to learn higher-level representations. Although their ideal value can be domain specific, a restricted set of multiples of 2 (layers) and 20 (neurons) are usually good candidates [56]. Models are optimized on the training data (first 75% of days), and the best-performing architecture for base models and for the multimodal branch is then selected to conduct actual experiments with all stocks.

Our study's priority is a diverse and comprehensive evaluation of model performance. Considering that all data modalities are required for evaluating the proposed multimodal model, we focus on two representative evaluation periods: uptrend (June to December 2021) and downtrend (January to September 2022), where this condition holds (i.e., financial news is unavailable before 2021). This choice allows us to showcase and discuss model performance in different market conditions. For each day in the evaluation period, models are trained up to the previous day, and predict the stock trend for the current day, leading to a sliding window fine-tuning and evaluation approach. For the LSTM branch, the complete set of hyperparameters optimized in our experiments is shown in Table 1. The complete set of hyperparameters optimized for the Fin-BERT branch is shown in Table 2. To this end, instead of using grid search, we apply the Keras tuner to help us select the best hyperparameter values. As for the multimodal fusion branch, the complete set of hyperparameters is shown in Table 3.

Once the model combines the information from these two data modalities and undergoes optimization, it is ready to make predictions regarding the next-day trends in the stock market. The model's prediction is represented by a single neuron with a sigmoid activation function, a common choice for binary classification tasks. This neuron outputs a value between 0 and 1, serving as an indicator of the likelihood or probability of a specific stock's next-day trend. A value closer to 1 suggests a higher probability of a positive trend, while a value closer to 0 indicates a higher likelihood of a negative trend in the stock's performance.

The dataset used in our work and the implementation of the proposed approach are publicly available at the following online repository: https://github.com/rcorizzo/fin bert-lstm.

4 Experiments

In this section, we provide a description of the experimental setup employed in our study. It encompasses the configuration of hyperparameters used for the various models, and the evaluation metrics used in our evaluation. Subsequently, we discuss the outcomes of our experiments in relation to the accuracy achieved by the different models. Lastly, we focus our attention on portfolio analysis, where we examine the consequences of utilizing model predictions as triggers for buying and selling decisions in real market scenarios.

4.1 Setup

In our experiments, we compare our results with popular baselines. For baselines, suitable values for hyperparameter tuning were chosen based on configurations reported as

Table 4 One-day-ahead stock prediction performance in Uptrend (July 1, 2021 to Dec 31, 2021—left) and Downtrend (Jan 1, 2022 to Sep 20, 2022—right) market conditions in terms of Precision, Recall, and F1-Score (micro average) with all analyzed methods and stocks (ATVI, NFLX, SBUX, TSLA)

	Method	Precision	Recall	F1-Score	Precision	Recall	F1-Score
ATVI	ARIMA	0.45	0.45	0.45	0.52	0.52	0.52
	LSTM	0.58	0.58	0.58	0.5	0.49	0.49
	GBTs	0.5	0.5	0.49	0.44	0.45	0.45
	Polarity	0.63	0.63	0.63	0.57	0.49	0.46
	Proposed	0.53	0.53	0.53	0.53	0.55	0.53
	Bi-LSTM	0.46	0.46	0.45	0.5	0.56	0.49
	CNN Seq2Seq	0.41	0.41	0.4	0.38	0.55	0.43
	Attention-CNN-LSTM	0.4	0.41	0.39	0.46	0.54	0.46
	Dilated CNN Seq2Seq	0.39	0.42	0.38	0.44	0.54	0.45
	CNN-LSTM	0.43	0.43	0.42	0.48	0.56	0.47
	GRU-CNN	0.41	0.43	0.39	0.43	0.41	0.28
NFLX	ARIMA	0.41	0.41	0.41	0.51	0.5	0.5
	LSTM	0.51	0.51	0.51	0.44	0.44	0.44
	GBTs	0.54	0.54	0.54	0.5	0.5	0.5
	Polarity	0.57	0.57	0.51	0.51	0.48	0.46
	Proposed	0.47	0.47	0.47	0.5	0.53	0.48
	Bi-LSTM	0.34	0.5	0.38	0.47	0.52	0.45
	CNN Seq2Seq	0.51	0.55	0.45	0.31	0.53	0.39
	Attention-CNN-LSTM	0.31	0.54	0.39	0.31	0.54	0.4
	Dilated CNN Seq2Seq	0.31	0.56	0.4	0.31	0.56	0.4
	CNN-LSTM	0.31	0.56	0.4	0.31	0.54	0.39
	GRU-CNN	0.58	0.57	0.44	0.52	0.47	0.41
SBUX	ARIMA	0.52	0.52	0.52	0.48	0.47	0.47
	LSTM	0.51	0.51	0.51	0.46	0.46	0.46
	GBTs	0.52	0.52	0.52	0.5	0.5	0.5
	Polarity	0.56	0.56	0.51	0.48	0.46	0.45
	Proposed	0.49	0.50	0.49	0.54	0.53	0.53
	Bi-LSTM	0.5	0.5	0.5	0.57	0.56	0.55
	CNN Seq2Seq	0.38	0.46	0.38	0.56	0.56	0.53
	Attention-CNN-LSTM	0.51	0.52	0.51	0.55	0.55	0.51
	Dilated CNN Seq2Seq	0.42	0.46	0.41	0.59	0.59	0.57
	CNN-LSTM	0.43	0.43	0.43	0.28	0.53	0.37
	GRU-CNN	0.54	0.48	0.37	0.48	0.48	0.48
TSLA	ARIMA	0.49	0.47	0.47	0.52	0.53	0.52
	LSTM	0.56	0.57	0.56	0.47	0.47	0.47
	GBTs	0.48	0.49	0.48	0.48	0.48	0.48
	Polarity	0.52	0.53	0.53	0.56	0.55	0.55
	Proposed	0.42	0.50	0.44	0.53	0.53	0.52
	Bi-LSTM	0.55	0.58	0.54	0.48	0.48	0.48
	CNN Seq2Seq	0.56	0.57	0.56	0.51	0.5	0.48
	Attention-CNN-LSTM	0.55	0.58	0.54	0.49	0.48	0.48
	Dilated CNN Seq2Seq	0.56	0.57	0.56	0.49	0.48	0.47
	CNN-LSTM	0.5	0.5	0.5	0.23	0.48	0.31
	GRU-CNN	0.38	0.39	0.27	0.42	0.47	0.34

Table 5 One-day-ahead stock prediction performance in Uptrend (July 1, 2021 to Dec 31, 2021—left) and Downtrend (Jan 1, 2022 to Sep 20, 2022—right) market conditions in terms of Precision, Recall, and F1-Score (micro average) with all analyzed methods and stocks (NVDA, AAPL, AMT, PLD)

	Method	Precision	Recall	F1-Score	Precision	Recall	F1-Score
NVDA	ARIMA	0.44	0.44	0.44	0.49	0.49	0.49
	LSTM	0.45	0.45	0.45	0.5	0.5	0.5
	GBTs	0.45	0.45	0.45	0.49	0.49	0.49
	Polarity	0.57	0.55	0.48	0.58	0.54	0.5
	Proposed	0.4	0.45	0.38	0.51	0.51	0.45
	Bi-LSTM	0.51	0.5	0.5	0.48	0.48	0.47
	CNN Seq2Seq	0.51	0.51	0.39	0.45	0.48	0.42
	Attention-CNN-LSTM	0.44	0.44	0.44	0.43	0.46	0.41
	Dilated CNN Seq2Seq	0.4	0.44	0.38	0.43	0.48	0.39
	CNN-LSTM	0.26	0.51	0.35	0.27	0.52	0.35
	GRU-CNN	0.48	0.5	0.44	0.45	0.47	0.44
AAPL	ARIMA	0.49	0.44	0.4	0.56	0.55	0.55
	LSTM	0.52	0.52	0.52	0.47	0.47	0.47
	GBTs	0.5	0.5	0.5	0.45	0.45	0.45
	Polarity	0.56	0.57	0.55	0.48	0.48	0.47
	Proposed	0.43	0.5	0.44	0.55	0.54	0.51
	Bi-LSTM	0.34	0.58	0.43	0.5	0.5	0.5
	CNN Seq2Seq	0.34	0.58	0.43	0.48	0.5	0.45
	Attention-CNN-LSTM	0.34	0.58	0.43	0.49	0.5	0.46
	Dilated CNN Seq2Seq	0.34	0.58	0.43	0.49	0.51	0.46
	CNN-LSTM	0.33	0.55	0.41	0.51	0.52	0.5
	GRU-CNN	0.47	0.44	0.43	0.5	0.49	0.41
AMT	ARIMA	0.51	0.5	0.5	0.41	0.4	0.4
	LSTM	0.44	0.44	0.44	0.47	0.47	0.47
	GBTs	0.52	0.51	0.51	0.54	0.54	0.54
	Polarity	0.54	0.56	0.49	0.5	0.49	0.4
	Proposed	0.48	0.46	0.46	0.48	0.48	0.48
	Bi-LSTM	0.38	0.47	0.4	0.41	0.46	0.4
	CNN Seq2Seq	0.31	0.56	0.4	0.46	0.46	0.46
	Attention-CNN-LSTM	0.31	0.55	0.4	0.43	0.44	0.43
	Dilated CNN Seq2Seq	0.31	0.56	0.4	0.49	0.49	0.48
	CNN-LSTM	0.4	0.54	0.4	0.45	0.45	0.45
	GRU-CNN	0.31	0.56	0.4	0.51	0.49	0.48
PLD	ARIMA	0.6	0.59	0.59	0.4	0.42	0.41
	LSTM	0.55	0.54	0.54	0.51	0.5	0.51
	GBTs	0.57	0.55	0.55	0.58	0.58	0.58
	Polarity	0.58	0.62	0.57	0.47	0.42	0.3
	Proposed	0.51	0.57	0.53	0.53	0.55	0.53
	Bi-LSTM	0.48	0.56	0.5	0.31	0.53	0.39
	CNN Seq2Seq	0.46	0.58	0.49	0.37	0.49	0.39
	Attention-CNN-LSTM	0.44	0.59	0.48	0.3	0.52	0.38
	Dilated CNN Seq2Seq	0.47	0.59	0.49	0.47	0.51	0.47
	CNN-LSTM	0.45	0.6	0.49	0.31	0.53	0.39
	GRU-CNN	0.55	0.59	0.55	0.6	0.6	0.6

Table 6 One-day-ahead stock prediction performance in Uptrend (July 1, 2021 to Dec 31, 2021—left) and Downtrend (Jan 1, 2022 to Sep 20, 2022—right) market conditions in terms of Precision, Recall, and F1-Score (micro average) with all analyzed methods and stocks (NDAQ, SCHW, BIO, JNJ)

	Method	Precision	Recall	F1-Score	Precision	Recall	F1-Score
NDAQ	ARIMA	0.58	0.59	0.58	0.46	0.46	0.46
	LSTM	0.55	0.55	0.55	0.43	0.43	0.43
	GBTs	0.46	0.45	0.46	0.52	0.52	0.52
	Polarity	0.31	0.53	0.39	0.53	0.47	0.36
	Proposed	0.47	0.54	0.45	0.52	0.5	0.5
	Bi-LSTM	0.5	0.52	0.5	0.45	0.52	0.43
	CNN Seq2Seq	0.49	0.48	0.48	0.3	0.55	0.39
	Attention-CNN-LSTM	0.49	0.52	0.49	0.3	0.55	0.39
	Dilated CNN Seq2Seq	0.48	0.5	0.48	0.3	0.54	0.39
	CNN-LSTM	0.47	0.49	0.47	0.45	0.54	0.4
	GRU-CNN	0.47	0.53	0.45	0.48	0.54	0.41
SCHW	ARIMA	0.47	0.48	0.46	0.49	0.49	0.49
	LSTM	0.4	0.4	0.4	0.49	0.49	0.49
	GBTs	0.53	0.53	0.53	0.55	0.55	0.55
	Polarity	0.54	0.53	0.43	0.56	0.49	0.38
	Proposed	0.49	0.5	0.5	0.47	0.49	0.45
	Bi-LSTM	0.46	0.46	0.45	0.36	0.51	0.37
	CNN Seq2Seq	0.44	0.44	0.44	0.48	0.52	0.43
	Attention-CNN-LSTM	0.42	0.42	0.42	0.28	0.53	0.37
	Dilated CNN Seq2Seq	0.4	0.4	0.4	0.54	0.54	0.47
	CNN-LSTM	0.44	0.47	0.36	0.52	0.53	0.39
	GRU-CNN	0.36	0.47	0.33	0.49	0.52	0.42
BIO	ARIMA	0.46	0.45	0.44	0.47	0.45	0.45
	LSTM	0.52	0.52	0.52	0.52	0.52	0.52
	GBTs	0.5	0.5	0.5	0.55	0.54	0.55
	Polarity	0.32	0.56	0.4	0.47	0.42	0.26
	Proposed	0.50	0.52	0.5	0.43	0.54	0.43
	Bi-LSTM	0.4	0.46	0.41	0.54	0.54	0.54
	CNN Seq2Seq	0.32	0.57	0.41	0.33	0.56	0.42
	Attention-CNN-LSTM	0.36	0.52	0.4	0.34	0.57	0.42
	Dilated CNN Seq2Seq	0.46	0.53	0.45	0.33	0.56	0.42
	CNN-LSTM	0.32	0.57	0.41	0.33	0.56	0.42
	GRU-CNN	0.48	0.46	0.46	0.57	0.56	0.56
JNJ	ARIMA	0.61	0.52	0.46	0.54	0.55	0.54
	LSTM	0.52	0.51	0.51	0.48	0.48	0.48
	GBTs	0.57	0.56	0.56	0.47	0.47	0.47
	Polarity	0.53	0.52	0.52	0.57	0.53	0.49
	Proposed	0.51	0.51	0.5	0.48	0.49	0.47
	Bi-LSTM	0.56	0.56	0.52	0.49	0.48	0.48
	CNN Seq2Seq	0.48	0.48	0.48	0.45	0.51	0.42
	Attention-CNN-LSTM	0.5	0.5	0.5	0.46	0.53	0.41
	Dilated CNN Seq2Seq	0.46	0.46	0.46	0.42	0.45	0.42
	CNN-LSTM	0.48	0.48	0.48	0.44	0.53	0.4
	GRU-CNN	0.53	0.52	0.51	0.3	0.54	0.38

effective in the original papers or following the above-discussed rationale used for our proposed model, performing grid search using solely historical training data. For conciseness, in the following, we report the final best-performing configurations adopted in the experimental analysis:

Table 7 Average performance of different methods across multiple stocks (F1-Score) and statistical analysis with Wilcoxon signed rank tests (*p*-value) comparing all pairwise combinations of methods

	Downtrend					
	ARIMA	LSTM	GBTs	Polarity	Bi-LSTM	CNN Seq2Seq
F1-Score	0.4833	0.4775	0.5067	0.4233	0.4625	0.4342
p-Value	0.8174	0.3123	0.3556	0.0377*	0.2602	0.0035*
	Attention	Dilated				
	CNN-LSTM	CNN Seq2Seq	CNN-LSTM	GRU-CNN	Proposed	
F1-Score	0.4267	0.4492	0.4033	0.4342	0.4900	
p-Value	0.0022*	0.0226*	0.0005*	0.0327*		
	Uptrend					
	ARIMA	LSTM	GBTs	Polarity	Bi-LSTM	CNN Seq2Seq
F1-Score	0.4767	0.5075	0.5075	0.5008	0.4650	0.4425
p-Value	0.7508	0.0606	0.0606	0.2366	0.7508	0.0783
	Attention	Dilated				
	CNN-LSTM	CNN Seq2Seq	CNN-LSTM	GRU-CNN	Proposed	
F1-Score	0.4492	0.4367	0.4267	0.4200	0.4742	
p-Value	0.2253	0.0531*	0.0226*	0.0433*		

*Comparison is statistically significant (*p*-value <0.05)

- **ARIMA** [57]: ARIMA, an abbreviation for autoregressive integrated moving average, is a forecasting approach characterized by three constituent components, each of which is controlled by a specific parameter. These components include the count of autoregressive terms denoted as (p), the quantity of nonseasonal differencing steps required to achieve stationarity marked as (d), and the number of lagged forecast errors integrated into the prediction equation represented by (q). The comprehensive ARIMA model can be formally expressed as follows:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (7)$$

where y'_t represents a differenced time series, subject to differencing operations multiple times. The predicted values on the right-hand side encompass both past y'_t values and previous prediction errors. We use this widely known autoregressive method to forecast the closing price of each stock for the following day, using historical price data. The key aspect of this approach is to identify price trends by relying solely on the target variable of interest. The prediction is then transformed into a binary format, i.e., either an uptrend or a downtrend, by comparing the predicted value (next day) with the most recently observed closing price of the stock

(current day). To automatically determine the best configuration for the parameters (p, d, q) based on historical training data, we use the Auto-ARIMA implementation provided in "pmdarima."[4]

- **GBTs** [58]: For all other parameters, we adhere to the default configuration recommended by the method's documentation in scikit-learn. gradient boosted trees (GBTs) are a competitive ensemble method within machine learning algorithms. In essence, GBTs blend "weak" machine learning models, such as decision trees, into a more resilient and precise ensemble machine learning model. The fundamental principle of gradient boosting involves iteratively enhancing the model's predictions by training it to minimize the residual error of previous base models. Concretely, an incremental training strategy is employed, where one tree is acquired in each iteration. For every individual example, its prediction is generated by each model, and the final prediction is computed as the summation of the scores provided by all models. To be more precise, the objective function governing the optimization of tree boosting can be formally expressed as follows:

---

[4] https://pypi.org/project/pmdarima/.

Table 8 Simulated portfolio gains (ATVI stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	− 2.07	− 10.34	− 14.75	− 9.06	1.36	6.82	9.17	5.78
CNN-LSTM	− 2.78	− 13.52	− 17.48	− 11.26	1.36	6.81	13.49	7.22
Dilated CNN Seq2Seq	− 1.54	− 7.70	− 20.08	− 9.77	− 0.59	− 2.95	− 4.27	− 2.60
GBTs	− 1.25	− 6.24	− 12.47	− 6.65	− 0.30	− 1.49	− 0.37	− 0.72
GRU-CNN	− 2.04	− 10.18	− 12.28	− 8.17	− 1.93	1.90	8.31	2.76
LSTM	− 1.56	− 7.81	− 15.61	− 8.33	0.86	4.30	8.60	4.58
Bi-LSTM	− 3.45	− 17.26	− 29.15	− 16.62	− 0.23	− 1.15	− 1.98	− 1.12
Polarity	− 0.82	− 4.10	− 8.72	− 4.55	− 1.62	0.41	5.64	1.48
Attention-CNN-LSTM	− 3.55	− 17.75	− 29.86	− 17.05	0.31	1.56	2.97	1.62
CNN Seq2Seq	− 1.92	− 9.62	− 25.17	− 12.24	− 0.01	− 0.07	− 0.14	− 0.07
Proposed	− 1.89	− 9.47	− 11.52	− 7.63	0.77	4.25	10.97	5.33
Buy and hold	NA			− 27.99				12.61

The last column highlights average gains for each model computed across all Max Shares configurations

Table 9 Simulated portfolio gains (NFLX stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	1.48	6.06	10.90	6.15	− 44.54	− 43.04	− 40.48	− 42.69
CNN-LSTM	14.39	13.36	13.82	13.86	− 1.50	− 7.04	− 8.71	− 5.75
Dilated CNN Seq2Seq	14.39	13.36	13.82	13.86	− 0.06	− 0.31	− 0.62	− 0.33
GBTs	12.25	11.81	7.98	10.68	3.55	− 0.79	− 12.09	− 3.11
GRU-CNN	14.18	12.29	11.91	12.79	− 43.93	− 53.62	− 51.49	− 49.68
LSTM	11.48	6.65	4.33	7.49	− 5.70	− 30.31	− 31.95	− 22.65
Bi-LSTM	10.67	6.15	7.30	8.04	2.93	− 10.28	− 13.48	− 6.95
Polarity	16.09	12.98	13.57	14.22	− 48.35	− 55.46	− 52.04	− 51.95
Attention-CNN-LSTM	14.04	11.63	10.55	12.08	− 1.22	− 6.09	− 7.41	− 4.91
CNN Seq2Seq	15.02	13.82	13.62	14.15	− 1.88	− 7.59	− 8.77	− 6.08
Proposed	17.00	21.23	20.05	19.42	− 0.56	1.63	0.94	0.67
Buy and hold	NA			14.14				− 56.6

The last column highlights average gains for each model computed across all Max Shares configurations

Table 10 Simulated portfolio gains (SBUX stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	0.77	1.33	1.67	1.25	0.46	− 14.74	− 14.76	− 9.68
CNN-LSTM	− 2.22	− 6.42	− 7.31	− 5.32	− 0.02	− 0.12	− 0.24	− 0.13
Dilated CNN Seq2Seq	0.12	− 3.73	− 1.43	− 1.68	2.45	7.62	10.16	6.75
GBTs	− 0.59	− 2.95	− 7.93	− 3.82	1.13	5.65	8.29	5.02
GRU-CNN	0.01	0.03	0.06	0.03	− 0.92	− 4.05	− 1.24	− 2.07
LSTM	− 0.39	− 1.96	− 4.09	− 2.14	0.56	2.79	4.32	2.56
Bi-LSTM	0.75	1.27	2.89	1.64	4.84	2.05	− 1.81	1.69
Polarity	1.38	− 1.71	1.29	0.32	1.75	− 7.50	− 9.49	− 5.08
Attention-CNN-LSTM	− 0.03	2.04	2.92	1.64	1.54	5.74	5.71	4.33
CNN Seq2Seq	0.56	− 2.74	− 0.74	− 0.97	2.06	6.20	7.81	5.36
Proposed	0.13	− 0.24	3.20	1.03	3.83	10.50	7.00	7.11
Buy and hold	NA			2.49				− 20.86

The last column highlights average gains for each model computed across all Max Shares configurations

Table 11 Simulated portfolio gains (TSLA stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	7.12	14.25	17.91	13.09	1.95	− 22.32	− 21.45	− 13.94
CNN-LSTM	18.83	39.66	41.31	33.26	− 0.17	− 0.84	− 1.67	− 0.89
Dilated CNN Seq2Seq	35.41	44.91	43.53	41.28	− 14.49	− 21.80	− 18.07	− 18.12
GBTs	21.57	30.04	33.29	28.30	− 2.74	5.05	− 17.50	− 5.06
GRU-CNN	0.71	3.53	7.06	3.77	− 0.47	− 2.34	− 4.69	− 2.50
LSTM	19.59	51.51	50.52	40.54	− 4.79	− 16.78	− 10.55	− 10.71
Bi-LSTM	45.77	55.97	53.21	51.65	− 21.82	− 20.82	− 12.62	− 18.42
Polarity	11.03	35.75	32.66	26.48	− 9.98	− 11.73	− 8.49	− 10.07
Attention-CNN-LSTM	46.33	54.28	52.19	50.93	− 24.51	− 33.23	− 28.59	− 28.78
CNN Seq2Seq	35.59	45.80	46.43	42.60	1.45	− 3.97	− 1.44	− 1.32
Proposed	37.64	48.83	45.58	44.02	13.60	19.18	5.81	12.86
Buy and hold	NA			57.55				− 22.71

The last column highlights average gains for each model computed across all Max Shares configurations

Table 12 Simulated portfolio gains (NVDA stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	2.87	17.58	29.39	16.61	− 9.02	− 19.77	− 23.35	− 17.38
CNN-LSTM	45.91	48.26	47.32	47.16	− 0.08	− 0.42	− 0.83	− 0.44
Dilated CNN Seq2Seq	41.40	39.55	29.65	36.87	− 7.12	− 25.59	− 24.27	− 18.99
GBTs	4.72	25.70	32.34	20.92	− 11.73	− 31.87	− 41.13	− 28.25
GRU-CNN	39.50	47.82	47.08	44.80	− 8.36	− 20.78	− 20.22	− 16.45
LSTM	4.61	28.23	34.03	22.29	− 6.61	− 30.30	− 29.54	− 22.15
Bi-LSTM	3.21	22.48	29.66	18.45	− 34.49	− 35.15	− 36.20	− 35.28
Polarity	41.57	43.42	34.22	39.74	− 47.65	− 55.96	− 59.06	− 54.22
Attention-CNN-LSTM	3.04	17.56	19.42	13.34	− 22.17	− 14.58	− 12.57	− 16.44
CNN Seq2Seq	45.94	48.41	47.63	47.33	− 8.84	− 28.10	− 27.34	− 21.43
Proposed	41.55	48.11	49.41	46.36	− 1.89	− 20.23	− 27.36	− 16.49
Buy and hold	NA			57.55				− 55.24

The last column highlights average gains for each model computed across all Max Shares configurations

$$\mathrm{obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \quad (8)$$

$$= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \mathrm{c}, \quad (9)$$

where in l represents a differentiable convex loss function quantifying the disparity between the prediction, denoted as $\hat{y}_i$, and the target value, y_i. Additionally, Ω is used to impose a penalty on the model's complexity through a regularization term, which serves the purpose of mitigating overfitting. Within our methodology, this model is harnessed for the purpose of

identifying underlying patterns within the multi-dimensional feature space and uncovering nonlinear associations among price data, statistical indicators, and future price trends. Specifically, we adopt the implementation of Gradient Boosted Trees provided by scikit-learn, accessible at the following.[5] We proceed to train the model using the ensuing hyperparameter configuration: $\{n\_estimators = 50, learning\_rate = 1.0, max\_depth = 10\}$.

- **LSTM** [36]: The model is optimized via gradient descent using the Adam optimizer and the binary cross-

---

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.

Table 13 Simulated portfolio gains (AAPL stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	0.03	0.15	0.30	0.16	− 0.98	− 11.18	− 17.09	− 9.75
CNN-LSTM	22.19	22.81	22.92	22.64	− 4.35	0.71	2.85	− 0.27
Dilated CNN Seq2Seq	22.41	24.11	25.79	24.10	− 3.57	1.72	5.17	1.11
GBTs	2.65	13.10	12.65	9.47	− 0.98	− 3.24	− 4.86	− 3.03
GRU-CNN	− 0.79	− 1.43	0.14	− 0.70	− 3.53	− 11.07	− 11.05	− 8.55
LSTM	5.33	21.32	21.60	16.08	− 1.62	− 7.03	− 0.23	− 2.96
Bi-LSTM	22.41	24.11	25.79	24.10	− 3.85	0.42	− 4.11	− 2.51
Polarity	15.62	24.17	25.30	21.70	− 9.56	− 11.82	− 18.32	− 13.23
Attention-CNN-LSTM	22.41	24.11	25.79	24.10	− 4.16	− 1.04	1.69	− 1.17
CNN Seq2Seq	22.41	24.11	25.79	24.10	− 3.43	− 0.11	1.00	− 0.85
Proposed	19.75	19.93	21.81	20.50	3.39	18.32	18.75	13.49
Buy and hold	NA			29.47				− 1.49

The last column highlights average gains for each model computed across all Max Shares configurations

entropy loss. The base LSTM model considered consists of two LSTM layers (500 and 100 units, respectively), a dropout layer, and a dense layer with a single neuron and sigmoid activation function to predict next-day trends.

- **Polarity** [41]: In this work, we utilize sentiment analysis predictions acquired through the EODHD Financial Data API.[6] The sentiment analysis encompasses four distinct categories: polarity, negative, neutral, and positive, with each score ranging from -1 to 1. Upon obtaining sentiment scores for all news articles gathered about a specific stock on a given day, we proceed to consolidate them and derive a binary label serving as the global indicator for that stock on that particular day. Specifically, if the cumulative polarity score threshold is above 0.4, we assign the label "1"; otherwise, it is designated as "0." The underlying rationale for this binary classification is rooted in the idea that positive media coverage could potentially signify an imminent uptrend. In contrast, negative media coverage implies uncertainty and skepticism, which may lead to a downtrend in the stock's performance.

- **Bi-LSTM** [38]: The model is optimized via gradient descent using the Adam optimizer and the binary cross-entropy loss. The base LSTM model considered consists of two LSTM layers (500 units). Following up, a bidirectional LSTM layer with (128 units) and a dropout rate of 0.8 is employed to enhance the model's ability to learn complex patterns bidirectionally, while dropout regularization helps mitigate overfitting.

Finally, a dense layer with a single unit and sigmoid activation is added to produce output predictions.

- **CNN Seq2Seq** [42]: The model is a sequence-to-sequence architecture for time series data. The encoder starts with a 1D convolutional layer (128 filters, ReLU activation), followed by max pooling. An LSTM layer (128 units) captures temporal dependencies with 0.8 dropout regularization. A RepeatVector layer prepares encoded representation. In the decoder, dilated convolutional layers replace standard ones. A dilated convolutional layer (128 filters, ReLU activation, dilation rate 2) followed by max pooling. An LSTM layer (128 units) decodes temporal information with a dropout rate of 0.8. Finally, a time-distributed dense layer (single unit, sigmoid activation) generates output predictions for each time step.

- **Attention-CNN-LSTM** [39]: The model architecture includes an encoder–decoder structure for time series data. The encoder starts with a 1D convolutional layer (128 filters, ReLU activation) for feature extraction, followed by max pooling. An LSTM layer (128 units) captures temporal dependencies with 0.8 dropout regularization. An attention mechanism enhances performance by focusing on relevant information, concatenated with encoder output for enriched representation. The decoder includes a similar convolutional layer, max pooling, and LSTM for decoding. Dropout regularization is applied again. Finally, a time-distributed dense layer (sigmoid activation) generates output predictions for each time step. This architecture integrates convolutional and LSTM layers, dropout regularization, and attention for effective time series

---

[6] https://eodhistoricaldata.com/.

Table 14 Simulated portfolio gains (AMT stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	2.22	− 3.20	− 3.68	− 1.55	1.41	− 9.89	− 12.46	− 6.98
CNN-LSTM	2.12	3.09	3.60	2.94	− 8.80	− 13.24	− 17.53	− 13.19
Dilated CNN Seq2Seq	2.65	4.54	5.41	4.20	− 8.96	− 12.08	− 13.05	− 11.36
GBTs	− 1.33	− 4.91	4.27	− 0.66	− 2.33	− 5.89	− 7.42	− 5.21
GRU-CNN	2.65	4.54	5.41	4.20	− 2.53	− 7.99	− 11.41	− 7.31
LSTM	0.31	− 0.17	− 4.47	− 1.45	− 3.16	− 5.68	− 4.50	− 4.45
Bi-LSTM	1.71	− 0.82	− 2.32	− 0.48	− 7.68	− 20.91	− 20.52	− 16.37
Polarity	0.71	4.61	6.51	3.94	− 0.17	− 8.26	− 9.64	− 6.02
Attention-CNN-LSTM	2.52	4.37	5.08	3.99	− 12.57	− 22.35	− 23.28	− 19.40
CNN Seq2Seq	2.65	4.54	5.41	4.20	− 10.08	− 15.71	− 18.84	− 14.88
Proposed	− 0.97	− 3.25	1.95	− 0.76	− 11.14	− 16.34	− 18.24	− 15.24
Buy and hold	NA			7.49				− 13.22

The last column highlights average gains for each model computed across all Max Shares configurations

Table 15 Simulated portfolio gains (PLD stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	5.87	10.71	13.97	10.19	− 0.44	− 3.64	− 12.01	− 5.36
CNN-LSTM	27.00	31.44	31.92	30.12	− 0.28	− 1.39	− 2.77	− 1.48
Dilated CNN Seq2Seq	26.43	30.99	31.56	29.66	− 6.12	− 15.50	− 15.43	− 12.35
GBTs	1.77	5.87	11.90	6.51	− 0.66	− 3.31	− 3.74	− 2.57
GRU-CNN	23.20	29.02	28.79	27.00	− 1.72	− 8.62	− 14.30	− 8.21
LSTM	6.39	20.11	16.38	14.29	− 2.00	− 9.99	− 16.05	− 9.34
Bi-LSTM	21.87	22.14	17.73	20.58	− 0.28	− 1.39	− 2.77	− 1.48
Polarity	24.48	33.66	36.00	31.38	− 26.32	− 26.98	− 28.01	− 27.10
Attention-CNN-LSTM	26.83	31.35	31.95	30.05	− 0.24	− 1.21	− 2.58	− 1.34
CNN Seq2Seq	26.26	30.84	31.36	29.49	− 0.91	− 4.56	− 5.73	− 3.73
Proposed	18.21	24.45	25.55	22.74	− 1.56	− 7.79	− 7.40	− 5.58
Buy and hold	NA			39.83				− 31.26

The last column highlights average gains for each model computed across all Max Shares configurations

processing and prediction.

- **Dilated CNN Seq2Seq** [40]: The adopted model is a sequence-to-sequence architecture for time series data. The encoder consists of a 1D convolutional layer with 128 filters and ReLU activation, followed by max pooling. An LSTM layer with 128 units and dropout rate of 0.8 captures temporal dependencies, and a RepeatVector layer prepares the encoded representation for each time step. In the decoder, dilated convolutional layers replace standard convolutional layers. The first layer has 128 filters, ReLU activation, and a dilation rate of 2 for broader contextual information. Max pooling retains relevant features. An LSTM layer with 128 units decodes temporal information, with dropout rate of 0.8. Finally, a time-distributed dense layer with a single unit and sigmoid activation generates output predictions for each time step.

- **CNN-LSTM** [42]: CNNs are effective for learning from time series data, with 1D convolutional layers filtering noise and extracting features. Causal convolution ensures influence only from previous time steps. RNNs excel in sequential learning tasks. We compare our model with CNN-LSTM, combining 1D CNN with LSTM, featuring convolutional layer, LSTM, batch normalization, dropout, and dense layer. Various model variants explored for optimal parameters: hidden layers (1 and 2), neurons (64 and 128), batch sizes (32 and 64), and dropout rates (0.2 and 0.5). Best-performing CNN-

Table 16 Simulated portfolio gains (NDAQ stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	1.93	8.36	9.56	6.62	0.47	2.35	− 6.68	− 1.29
CNN-LSTM	2.22	7.45	7.34	5.67	− 0.07	− 0.37	− 0.73	− 0.39
Dilated CNN Seq2Seq	2.52	7.96	7.52	6.00	− 0.02	− 0.09	− 0.18	− 0.10
GBTs	0.90	4.48	7.12	4.17	− 0.43	− 2.13	− 4.27	− 2.28
GRU-CNN	4.27	11.67	13.88	9.94	− 0.05	− 0.27	− 0.54	− 0.29
LSTM	0.73	3.63	7.06	3.80	− 0.30	− 1.49	− 2.98	− 1.59
Bi-LSTM	2.75	8.21	9.01	6.66	− 0.79	− 3.95	− 7.90	− 4.21
Polarity	5.20	13.72	16.27	11.73	6.35	0.49	− 2.75	1.36
Attention-CNN-LSTM	2.93	8.33	9.01	6.76	− 0.02	− 0.08	− 0.17	− 0.09
CNN Seq2Seq	1.57	5.58	4.24	3.80	− 0.02	− 0.08	− 0.17	− 0.09
Proposed	3.55	9.24	10.91	7.90	1.94	6.92	2.90	3.92
Buy and hold	NA			18.75				− 10

The last column highlights average gains for each model computed across all Max Shares configurations

LSTM has a convolutional layer with 32 filters of size 3, causal padding, ReLU activation, LSTM with 128 units, tanh activation, followed by batch normalization, dropout (rate 0.2), and dense layer with ReLU activation.

- **GRU-CNN** [37]: The GRU-CNN model combines GRU and 1D CNN, offering simpler training and improved performance. Parameters are similar to CNN-LSTM. A key difference stands in the arrangement of RNN and CNN layers. GRU-CNN is composed of a GRU layer (128 units, tanh activation), a 1D convolutional layer (32 filters, size 3, stride 1, causal padding, ReLU activation), global max pooling, batch normalization, dense layer (10 units, ReLU activation), dropout (rate 0.2), and a dense layer (prediction window size, ReLU activation). The GRU layer returns a sequence, whereas global max pooling retains significant features and reduces dimensionality.

- **FinBERT-LSTM (Proposed)**: The best-performing model architecture was obtained with two LSTM layers with 500 and 450 hidden units, interleaved by a normalization layer, followed by a batch normalization layer, and a dense representation layer of size 72. The best-performing FinBERT model following its standard architecture was obtained with a normalization layer followed by two dense layers of size 80 and 36, each of which was interleaved by a normalization layer. The best-performing architecture following to the fusion of the two modalities was obtained with two dense layers of size 40 and 60, interleaved by normalization layers.

For a quantitative evaluation of our results, we adopt conventional classification metrics such as Precision, Recall (R), and F1-Score, defined as:

$$\text{Precision} = \frac{T_p}{T_p + F_p}; \quad \text{Recall} = \frac{T_p}{T_p + F_n}; \quad F1 - \text{Score}$$
$$= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where T_p is the number of true positives, and F_p is the number of false negatives. Specifically, we adopt micro-averaged variants of Precision, Recall, and F-Measure, i.e., metrics are computed globally by considering each element of the label indicator matrix as a label.

4.2 Model accuracy

In our analysis of different models' performance on stock trend prediction, we aim to provide a broad evaluation and showcase model performance on a general set of real-world stocks with great diversity. To this end, we employ data from 12 real-world stocks in different sectors: Communication Services (ATVI, NFLX), Consumer Discretionary (SBUX, TSLA), Information Technology (NVDA, AAPL), Real Estate (AMT, PLD), Financials (NDAQ, SCHW), Healthcare (JNJ, BIO). One relevant perspective for results analysis is the model's classification accuracy on one-day-ahead stock trend prediction. In the following, we describe summarized results by financial sector. A more fine-grained discussion is provided in the Appendix.

4.2.1 Communication services

For ATVI (Activision Blizzard), the best-performing model in uptrend market conditions is Polarity with an F1-Score of 0.63. In downtrend conditions, the proposed model performs best with an F1-Score of 0.53. The worst-performing models are GRU-CNN in downtrend (0.28) and

Table 17 Simulated portfolio gains (SCHW stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	4.23	16.16	17.86	12.75	− 0.89	− 5.94	− 11.23	− 6.02
CNN-LSTM	0.00	0.01	0.03	0.01	− 0.11	− 0.54	− 1.08	− 0.58
Dilated CNN Seq2Seq	1.63	8.13	2.15	3.97	− 0.09	− 0.46	− 0.08	− 0.21
GBTs	0.33	1.67	1.96	1.32	− 0.11	− 0.55	0.58	− 0.03
GRU-CNN	− 0.02	− 0.10	− 0.21	− 0.11	− 0.03	− 0.14	− 0.27	− 0.14
LSTM	0.73	3.66	7.32	3.91	− 0.75	− 3.74	− 7.49	− 3.99
Bi-LSTM	1.02	5.08	2.52	2.87	0.00	0.02	0.04	0.02
Polarity	8.16	19.93	22.79	16.96	− 6.24	− 21.29	− 21.85	− 16.46
Attention-CNN-LSTM	3.25	12.51	6.20	7.32	0.02	0.10	0.20	0.11
CNN Seq2Seq	1.32	2.98	1.65	1.98	− 0.15	− 0.76	− 1.52	− 0.81
Proposed	0.41	2.03	3.65	2.03	− 0.95	− 4.74	− 9.53	− 5.07
Buy and hold	NA			14.73				− 13.98

The last column highlights average gains for each model computed across all Max Shares configurations

Table 18 Simulated portfolio gains (JNJ stock) in uptrend (left) and downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	0.08	0.38	0.76	0.40	− 1.25	− 5.13	− 5.17	− 3.85
CNN-LSTM	− 3.54	− 3.03	− 3.16	− 3.25	− 0.42	− 2.10	− 3.69	− 2.07
Dilated CNN Seq2Seq	− 4.23	− 4.18	− 4.97	− 4.46	− 1.02	− 6.81	− 11.04	− 6.29
GBTs	− 0.59	− 1.70	1.71	− 0.19	− 0.50	− 2.52	− 6.09	− 3.04
GRU-CNN	− 3.72	− 2.40	− 1.96	− 2.69	− 0.00	− 0.02	− 0.05	− 0.02
LSTM	− 1.36	− 6.34	− 4.74	− 4.14	− 0.29	− 1.46	− 2.03	− 1.26
Bi-LSTM	0.18	1.29	2.18	1.21	− 5.86	− 7.15	− 11.20	− 8.07
Polarity	− 0.32	− 1.62	− 3.16	− 1.70	− 3.55	− 0.71	− 0.87	− 1.71
Attention-CNN-LSTM	− 4.85	− 4.72	− 4.13	− 4.57	− 0.42	− 2.10	− 3.68	− 2.07
CNN Seq2Seq	− 4.02	− 3.81	− 4.83	− 4.22	− 0.69	− 3.46	− 6.68	− 3.61
Proposed	− 2.90	− 4.17	− 0.22	− 2.43	− 1.19	− 1.06	− 1.99	− 1.41
Buy and hold	NA			4.53				− 3.05

The last column highlights average gains for each model computed across all Max Shares configurations

Dilated CNN Seq2Seq in uptrend conditions (0.38). For NFLX (Netflix) in uptrend conditions, the best model is GBTs with an F1-Score of 0.54. In downtrend predictions, GBTs and ARIMA perform best with an F1-Score of 0.50. The worst models are Bi-LSTM (0.38) for uptrends and CNN-LSTM, CNN Seq2Seq (0.39) for downtrends.

4.2.2 Consumer discretionary

For SBUX (Starbucks), the best models for uptrends are ARIMA and GBTs with an F1-Score of 0.52. For downtrends, the best is Dilated CNN Seq2Seq (0.57). The worst models are GRU-CNN (0.37) for uptrends and CNN-LSTM (0.37) for downtrends. For TSLA (Tesla), the best uptrend models are Dilated CNN Seq2Seq, CNN Seq2Seq, and LSTM with an F1-Score of 0.56. For downtrends, Polarity performs best (0.55). The worst models are GRU-CNN (0.27) in uptrend and CNN-LSTM (0.31) in downtrend conditions.

4.2.3 Information technology

For NVDA (Nvidia), the best-performing uptrend model is Bi-LSTM with an F1-Score of 0.50. For downtrends, LSTM and Polarity lead with an F1-Score of 0.50. The worst-performing model in both conditions is CNN-LSTM

Table 19 Simulated portfolio gains (BIO stock) in Uptrend (left) and Downtrend (right): absolute (USD) and relative (percentage) with respect to the initial investment, with all models and different Max Shares configurations (1, 5, 10)

Method	Gain % (1)	Gain % (5)	Gain % (10)	Avg %	Gain % (1)	Gain % (5)	Gain % (10)	Avg %
ARIMA	1.90	1.74	− 0.38	1.09	− 12.42	− 14.46	− 20.51	− 15.80
CNN-LSTM	11.53	13.02	13.50	12.69	− 2.02	− 5.89	− 7.91	− 5.27
Dilated CNN Seq2Seq	12.98	12.29	13.07	12.78	− 1.43	− 6.27	− 8.35	− 5.35
GBTs	1.95	14.83	14.74	10.51	− 22.60	− 14.74	− 10.89	− 16.08
GRU-CNN	− 3.52	− 2.30	− 1.45	− 2.42	− 10.27	− 8.96	− 6.51	− 8.58
LSTM	8.22	4.81	− 4.25	2.93	− 15.85	− 26.76	− 24.66	− 22.42
Bi-LSTM	7.92	− 1.72	− 0.58	1.87	− 16.16	− 22.18	− 23.17	− 20.50
Polarity	11.45	12.72	13.50	12.56	− 29.48	− 36.37	− 37.13	− 34.33
Attention-CNN-LSTM	8.38	5.51	6.18	6.69	− 1.21	− 5.74	− 6.36	− 4.44
CNN Seq2Seq	11.53	13.02	13.50	12.69	− 1.43	− 6.27	− 8.35	− 5.35
Proposed	7.19	10.89	13.07	10.38	− 4.05	− 1.85	− 5.35	− 3.75
Buy and hold	NA			15.14				− 37.69

The last column highlights average gains for each model computed across all Max Shares configurations

(0.35 for uptrend and downtrend). For AAPL (Apple), Polarity is the best for uptrend with an F1-Score of 0.55. ARIMA leads in downtrend predictions with an F1-Score of 0.55. The worst models are ARIMA (0.40) for uptrends and GRU-CNN (0.41) for downtrends.

4.2.4 Real estate

For AMT (American Tower), the best uptrend model is GBTs with an F1-Score of 0.51. For downtrends, GBTs also lead with 0.54. The worst models are several (F1-Score of 0.40) for uptrends, and Bi-LSTM (0.40) for
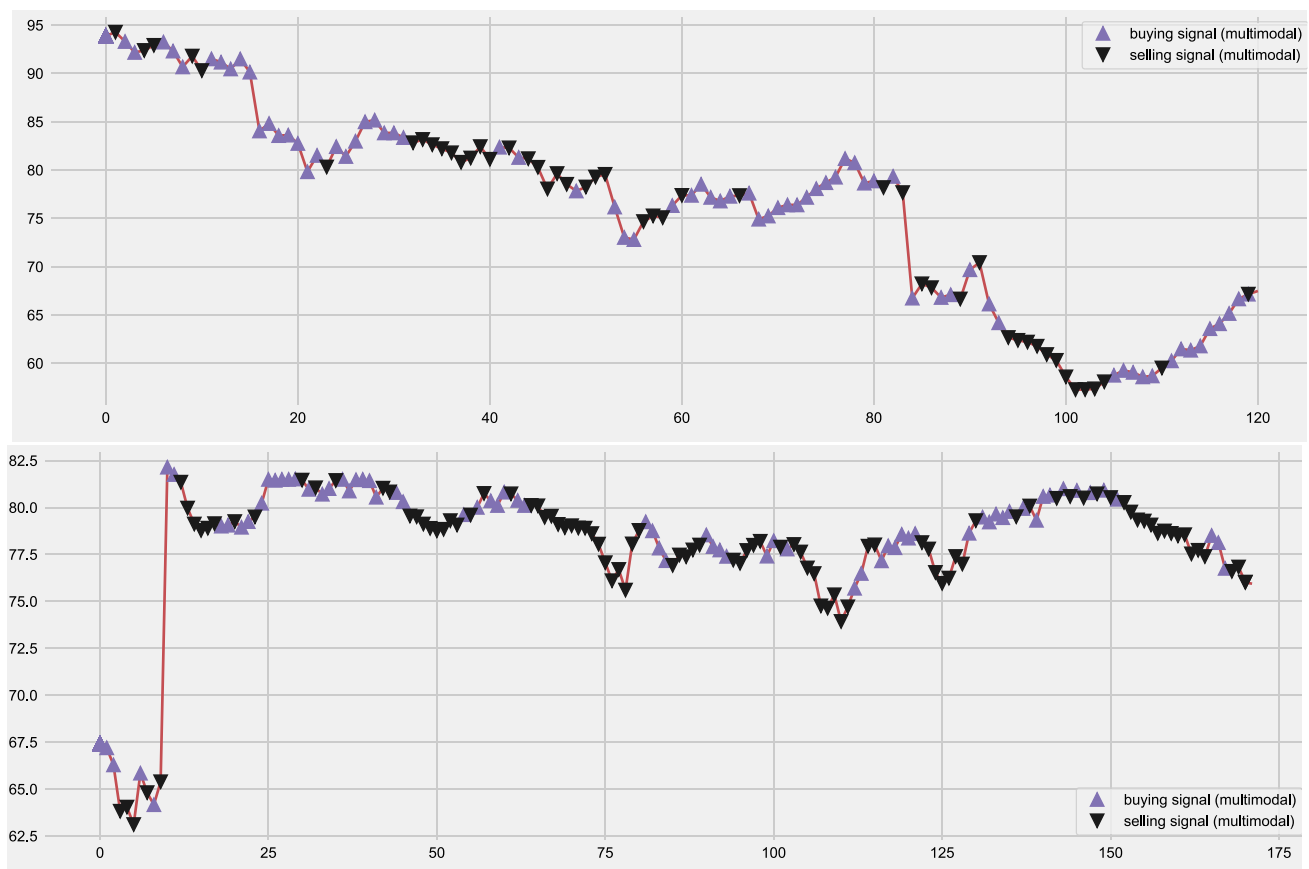
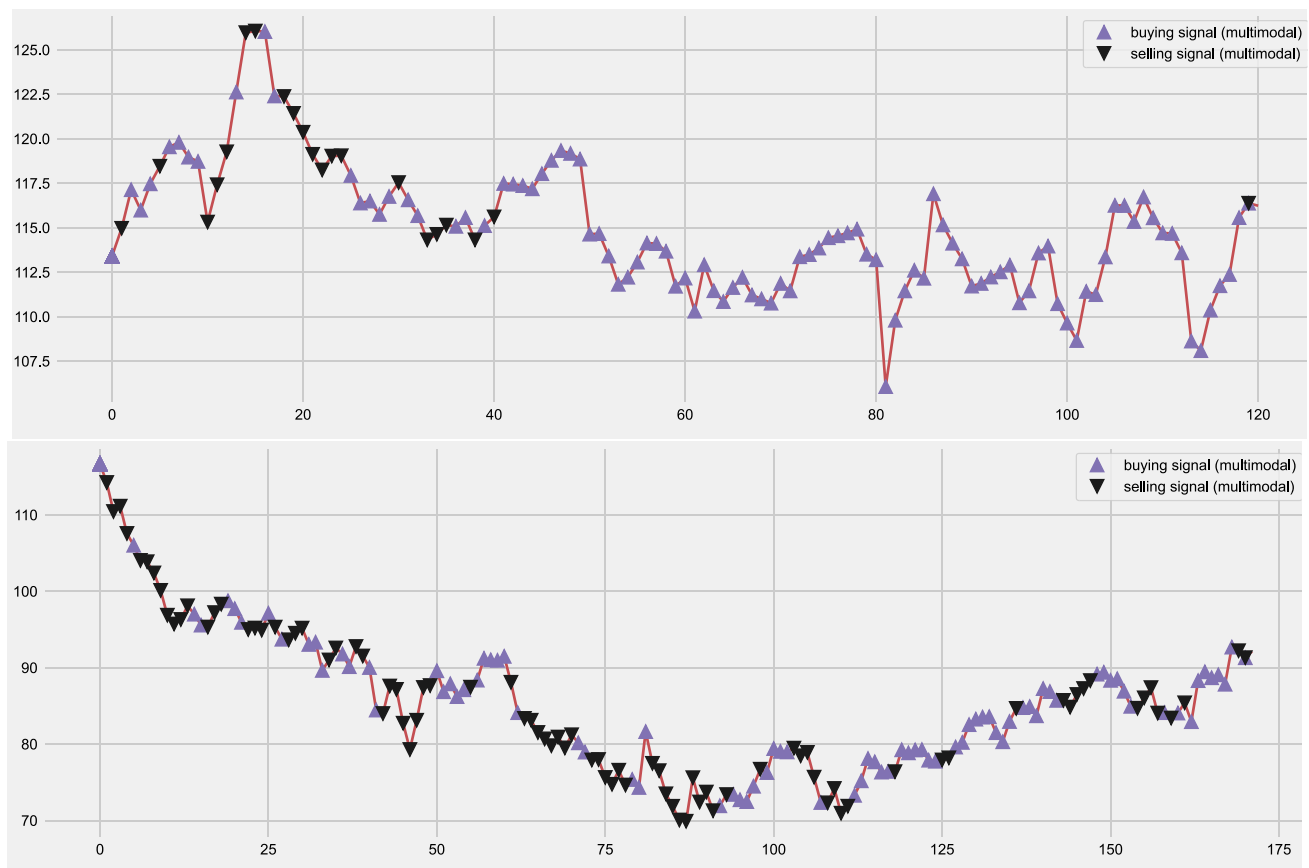**Fig. 2** Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (ATVI stock)
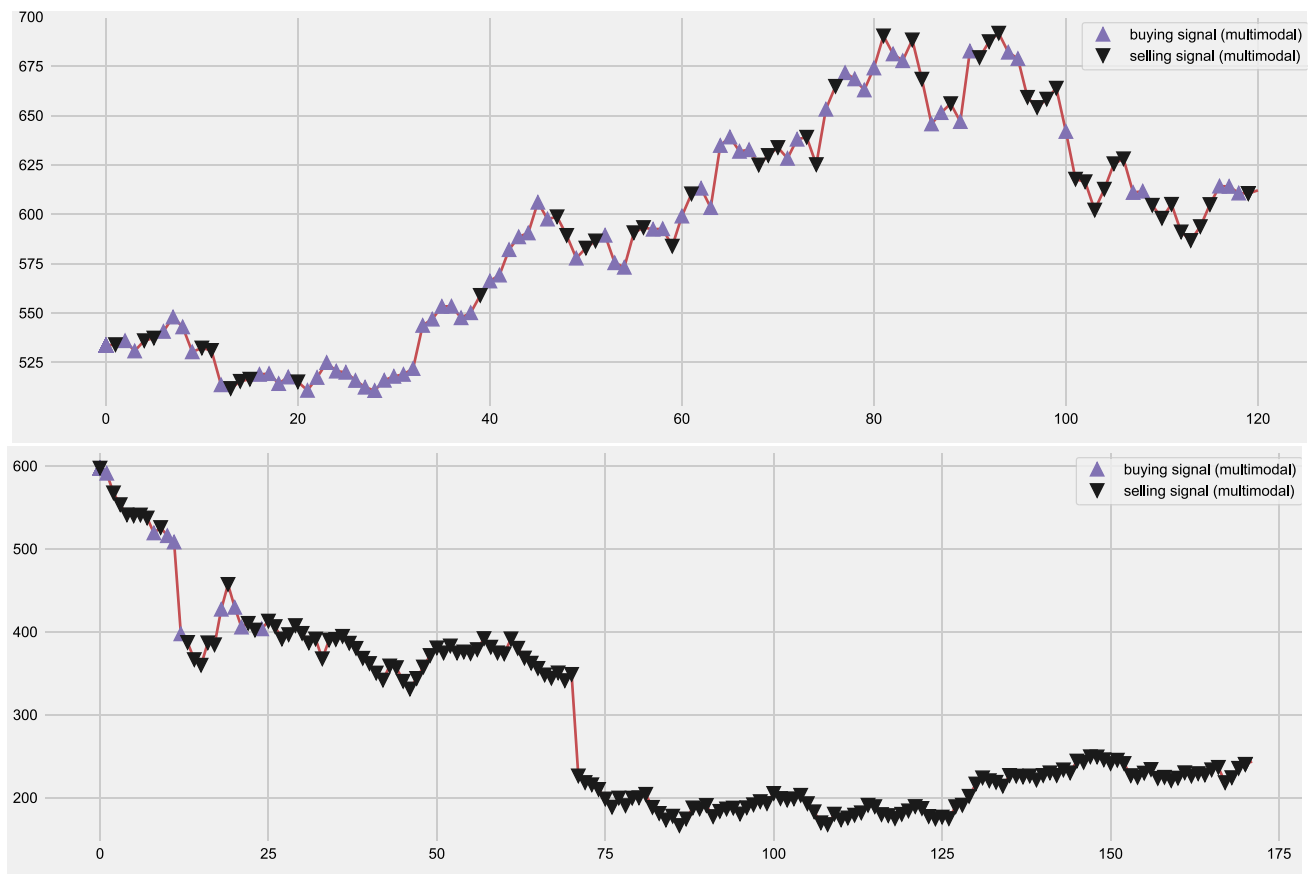
Fig. 3 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (NFLX stock)

downtrends. For PLD (Prologis) ARIMA is the best model for uptrends with an F1-Score of 0.59. For downtrends, GRU-CNN performs best (0.60). The worst models are Attention-CNN-LSTM (0.48) for uptrends and Polarity (0.30) for downtrends.

4.2.5 Financials

For NDAQ (Nasdaq) ARIMA is the best for uptrends with an F1-Score of 0.58. GBTs perform best for downtrends (0.52). The worst model is Polarity (0.39 for uptrends and 0.36 for downtrends). For SCHW (Charles Schwab) GBTs is the best model for uptrends with an F1-Score of 0.55. For downtrends, GBTs lead with 0.53. The worst models are CNN-LSTM (0.36) for uptrends and Attention-CNN-LSTM, Bi-LSTM (0.37) for downtrends.

4.2.6 Health

For JNJ (Johnson & Johnson), the best uptrend model is GBTs with an F1-Score of 0.56. ARIMA leads in downtrend predictions with an F1-Score of 0.54. The worst models are Polarity (0.49) for uptrends and GRU-CNN (0.38) for downtrends. For BIO (Bio-Rad Laboratories), the best-performing model for uptrends is GRU-CNN with an F1-Score of 0.56. For downtrends, GRU-CNN again leads with 0.56. The worst models are Attention-CNN-LSTM, Polarity (0.40) for uptrends and Polarity (0.26= for downtrends.

Analyzing the performance of different models for each stock provides insights into their effectiveness in predicting stock trends. Considering base models performance with different stocks, GBTs emerge as the most robust approach across all stocks, followed by LSTM, Polarity, and ARIMA. We also observe that Polarity performs particularly well in predicting TSLA stock trends, while LSTM
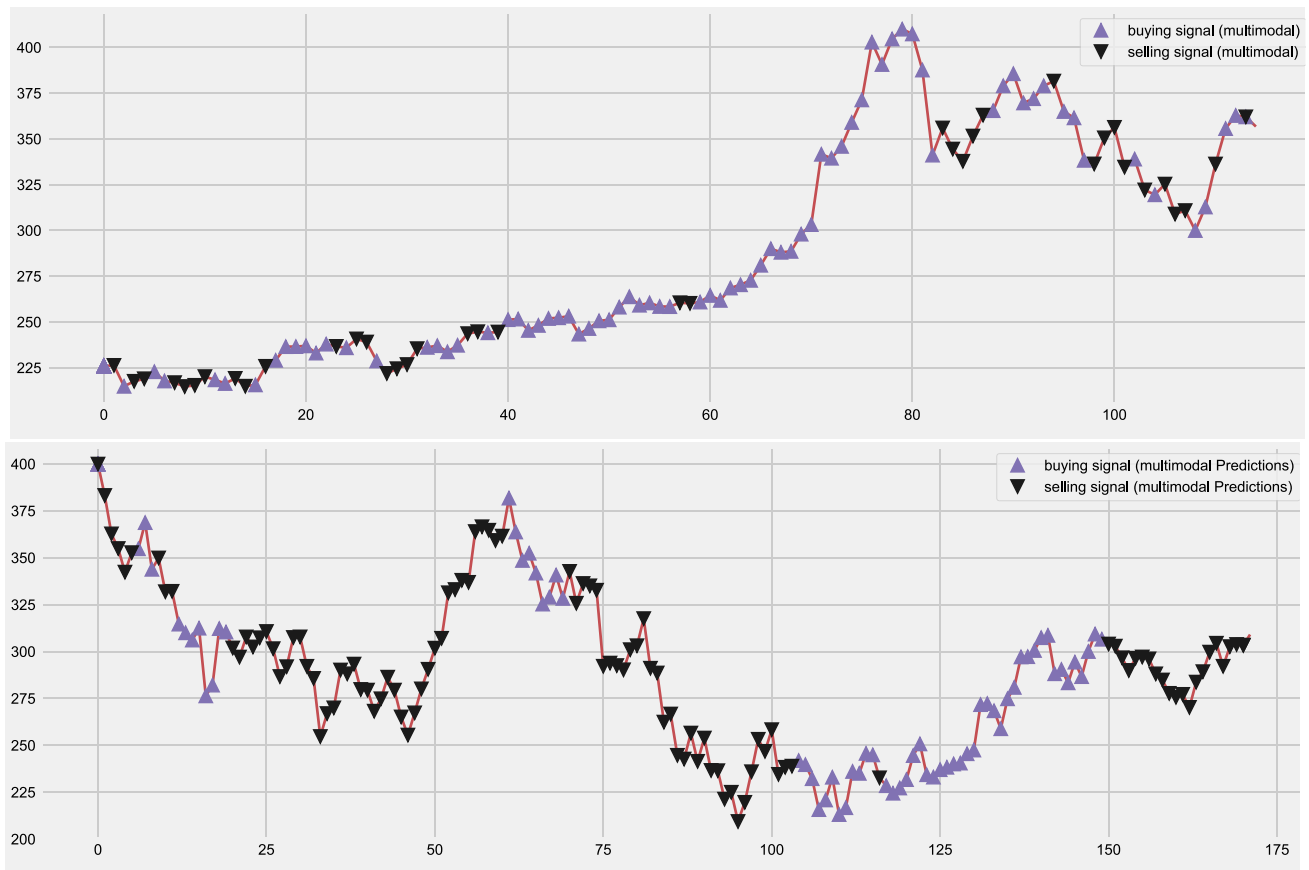
Fig. 4 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (SBUX stock)

shows competitive performance in AAPL predictions. However, experimental results in Tables 4, 5 and 6 show that different models exhibit varying performance in capturing uptrends and downtrends in stock prices.

In uptrend market conditions, GBTs consistently achieves the highest F1-Scores across multiple stocks, indicating its effectiveness in capturing upward price movements. ARIMA and Proposed also demonstrate a competitive performance in predicting uptrends for certain stocks. However, convolutional neural network-based models such as CNN-LSTM and GRU-CNN consistently underperform other approaches across various stocks and market conditions.

In downtrend market conditions, however, the proposed model outperforms other models, as it consistently achieves the highest F1-Scores across various stocks, showcasing strength in capturing downward price movements. Overall, considering both uptrend and downtrend predictions, the proposed model emerges as the top-performing approach to capture the overall stock market trend. It excels in predicting both upward and downward price movements, representing a robust choice for stock trend analysis. GBTs, ARIMA, Polarity, LSTM, BI-LSTM, CNN Seq2Seq, Attention-CNN-LSTM, Dilated CNN Seq2Seq,

CNN-LSTM, and GRU-CNN can, in some cases, also provide a competitive performance, although their level of reliability significantly varies across different stocks.

One possible explanation for the generally low performance of CNN-based models is the noisy/highly fluctuating nature of financial data, as well as the lack of spatial dependencies in the data, which makes CNN-based models ineffective. On the other hand, models that focus on temporal dependencies (LSTM, ARIMA), models that are more robust to noise (GBTs), and models that consider a holistic combination of data modalities (Proposed) appear more effective in predicting stock trends. Another relevant observation from a bias-variance perspective is that ARIMA models are typically characterized by low variance, while GBTs are flexible enough to capture complex patterns without overfitting excessively, leading to more accurate predictions than CNN-based models.

It is important to note that the performance of these models may be influenced by changing macroeconomic conditions and concept drift in the considered time frames. The upward trajectory of stock prices in June 2021 can be attributed to a confluence of economic factors. Firstly, the global economy was in a recovery phase after the severe economic downturn triggered by the COVID-19 pandemic.

Fig. 5 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (TSLA stock)

Governments and central banks across the world had implemented a series of monetary and fiscal policies to stimulate economic growth, which boosted investor confidence. Secondly, companies began releasing favorable earnings reports during this period. Positive financial performance exceeding market expectations can act as a catalyst for stock price appreciation as it instills investor optimism. Thirdly, the persistently low interest rates in 2021 made stocks an attractive investment option due to the relatively higher potential returns when compared to low-yield fixed-income securities. Furthermore, sectors such as technology and growth stocks exhibited considerable appeal to investors, with their higher growth potential contributing to increased stock prices. Lastly, speculative trading activities, epitomized by events like the GameStop and AMC short squeezes, generated significant retail investor participation and volatility, influencing stock price movements.

The decline in stock prices during 2022 can be attributed to a combination of economic and market factors. First and foremost, economic conditions, particularly the concerns over rising inflation and interest rates, were central to the reduced attractiveness of stocks. The expectation of increasing inflation and interest rates led investors to consider alternative investments that offered better protection against eroding purchasing power and higher fixed returns. Secondly, geopolitical events, such as trade disputes and international conflicts, introduced uncertainty into the market, undermining investor sentiment. Geopolitical tensions can result in market volatility, which has a detrimental impact on stock prices. Additionally, corporate earnings played a pivotal role in driving down stock prices. Companies reporting weaker-than-expected earnings, often compounded by supply chain disruptions and increased production costs, faced downward pressure on their stock prices. Moreover, central bank policies, including potential interest rate hikes, can lead to stock market declines, as higher borrowing costs and reduced liquidity negatively affect equity valuations. Finally, market sentiment, determined by factors like fear, uncertainty, and pessimism, significantly influenced stock market performance in 2022, contributing to the overall decline in stock prices.
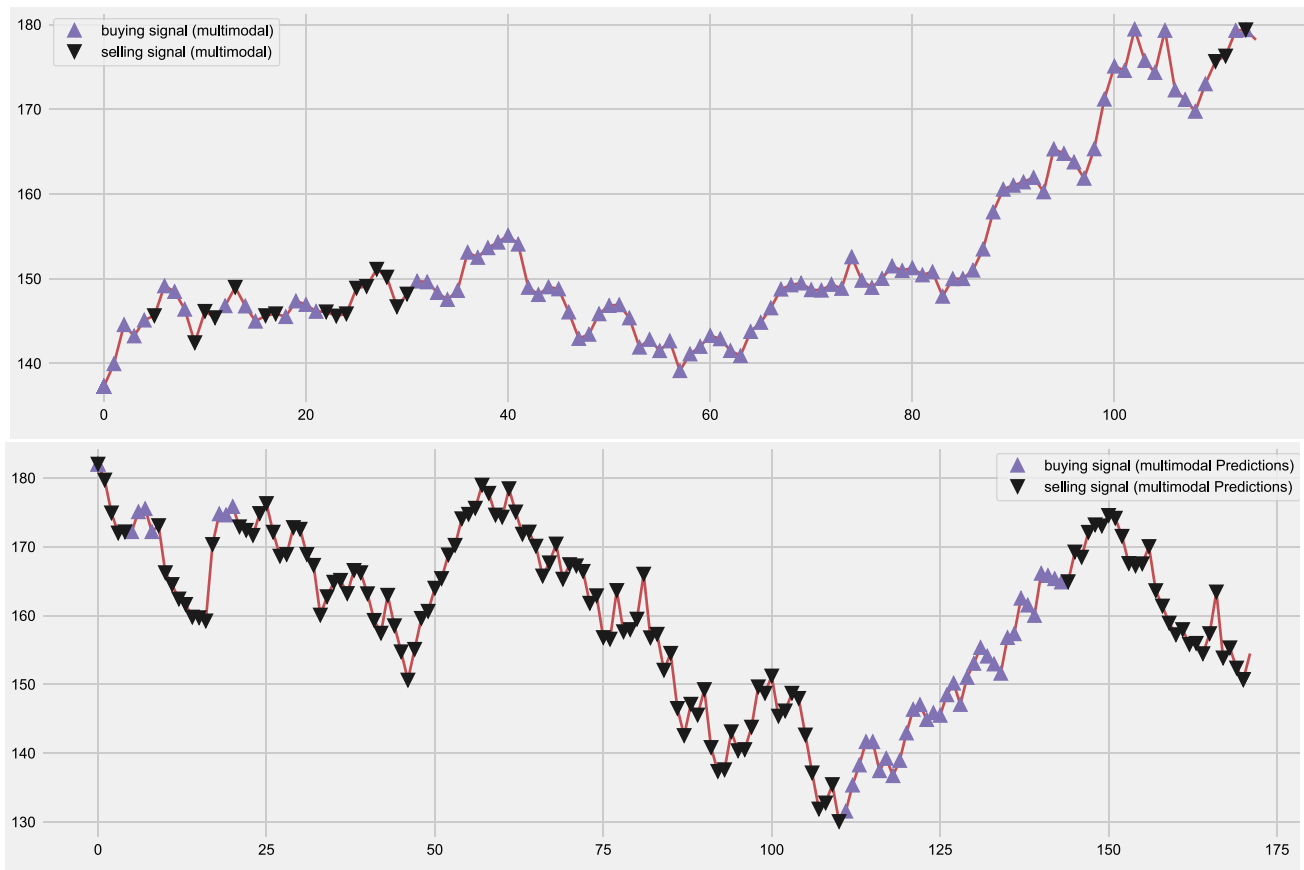
Fig. 6 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (NVDA stock)

To validate the statistical significance of our results, we adopt Wilcoxon signed rank tests on all pairwise combinations of methods across multiple executions, obtained considering the average F1-Score with different stocks. Based on results reported in Table 7, we can infer that in downtrend market conditions, the proposed method outperforms 9 out of 10 baselines in terms of F1-Score, and 6 of 10 comparisons are statistically significant. In the single case where one baseline (GBTs) outperform the proposed method, this comparison is not statistically significant (p-value of 0.3556). In uptrend market conditions, the proposed method outperforms 6 out of 10 baselines in terms of F1-Score, and 3 of 6 comparisons are statistically significant. In the 4 other cases where baselines (ARIMA, LSTM, GBTs, and Polarity) outperform the proposed method, all cases are not statistically significant (p-values: 0.7508, 0.0606, 0.0606, and 0.2366).

4.3 Portfolio analysis

In our experiments, we begin with a budget of 10,000 USD, and allocate the entire amount to purchase as many shares as possible at the prevailing market price on that day. The parameter "Max Shares" with values of 1, 5, or 10 is used to determine the number of stocks to be bought or sold each day based on the trend predictions generated by various models. When a model predicts a downtrend for the following day, we sell the desired number of shares at the current market price to mitigate potential losses. This action increases the available USD balance in the portfolio, which can then be reinvested in purchasing additional shares. On the other hand, if a model predicts an uptrend and the available USD balance is positive, we utilize the available balance to buy a number of shares that is either equal to or less than the desired number.

In cases where a model predicts an uptrend, but there is no available USD balance, we choose to hold onto all the previously purchased shares. At the conclusion of the simulation, all shares held in the portfolio are sold at their

Fig. 7 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (AAPL stock)

respective market prices. The absolute value of the portfolio in USD at this point is used to calculate the gains or losses relative to the initial budget of 10,000 USD.

In our simulation, we do not consider trading fees as some brokers, such as Charles Schwab,[7] do not apply them to online transactions and profit from different revenue streams. A custom trading fee can be easily considered using our public code implementation[8]. In our experiments, we perform two separate trading simulations in the two time frames covered in our performance analysis: uptrend (July 1, 2021 to Dec 31, 2021) and downtrend (Jan 1, 2022 to Sep 20, 2022). The findings of our portfolio analysis are outlined in Tables 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19, while Figs. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 offer time series visualizations depicting the closing prices of each stock over time, alongside the buy and sell signals generated by all models. To maintain conciseness in our discussion, we present the plot for the most effective model, as determined by the portfolio results.

Examining the time series plots enables us to gain valuable insights into the market conditions of the various

stocks on the overall time frame of our analysis (both uptrend and downtrend). For ATVI, Fig. 2 reveals that the stock started at $95, observed a downtrend up to the middle of the evaluation time frame (with a bottom at $55), followed by an uptrend up to $82.5, after which the price retraced to about $75, which is lower than the first day of the evaluation. Moving to NFLX, Fig. 3 reveals that this stock observed a significant downtrend throughout the entire evaluation time frame, starting with a value of over $525 per share and ending up with $210. As for SBUX, a similar uncertainty pattern can be observed in Fig. 4, with a starting price of $112.5, a fast uptrend phase up to $126, followed by a strong downtrend, which led to a bottom price of $70, with a subsequent sideways phase. The stock recouped its value in the final uptrend phase, ending with a value of $92, higher than the first evaluation day. For TSLA, Fig. 5 reveals that the stock started at $225 and observed a sideways phase followed by an uptrend phase, reaching a maximum price above $400. Subsequently, the stock observed a retracement phase in the channel $300 − $375 followed by a downtrend with low peaks at $250 and $200, recouping to $300 toward the end of the evaluation time frame. In the case of NVDA (Fig. 6), the stock exhibited an initial value of $200 and experienced an
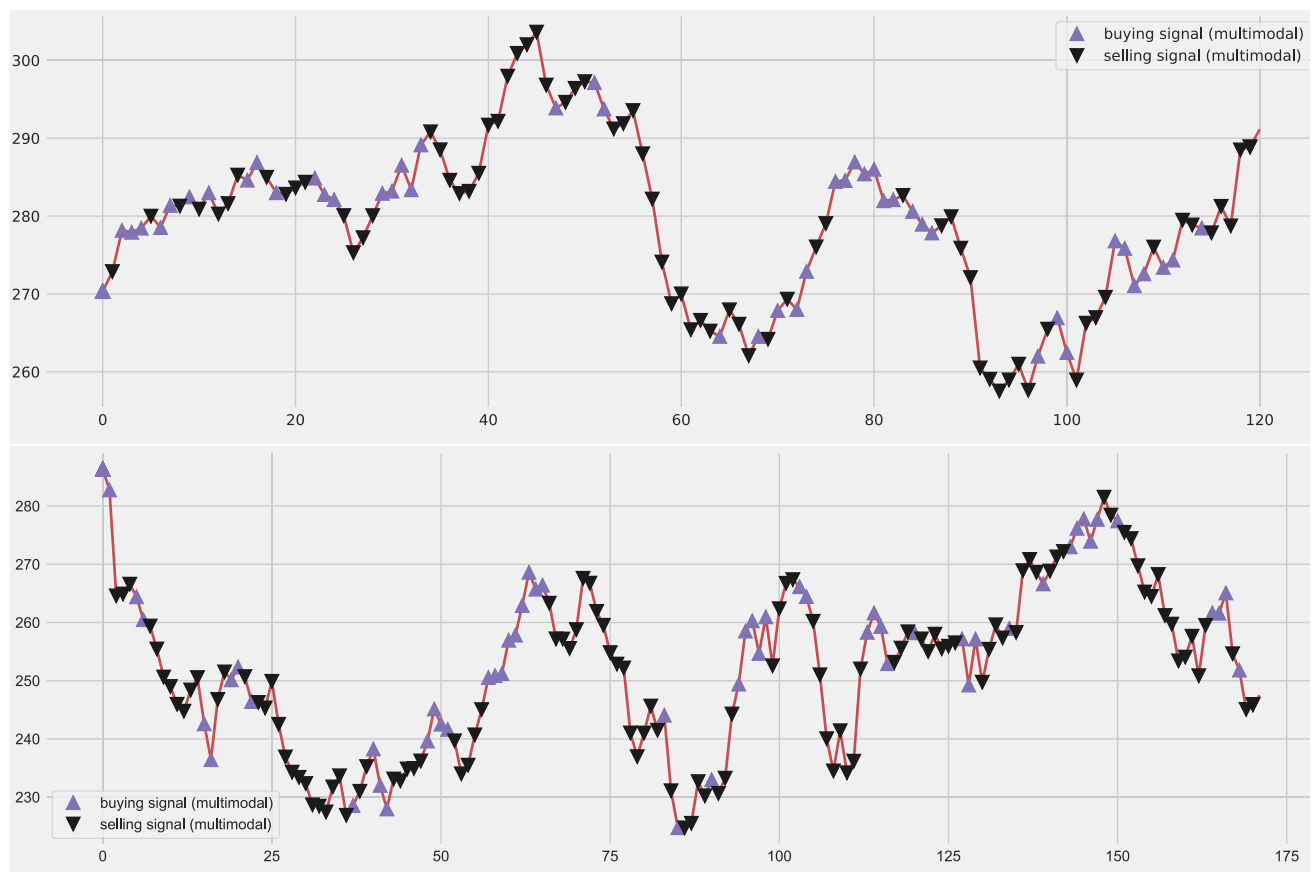
Fig. 8 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (AMT stock)

uptrend, reaching a peak of $340 around the midpoint of the evaluation period. Afterward, it encountered fluctuations and declined to $280 before recovering to $300, which was higher than the value on the first evaluation day. Analyzing AAPL (Fig. 7), it displayed a substantial uptrend throughout the entire evaluation duration, starting at approximately $135 per share and concluding at $180. Following this, there were fluctuations leading to a dip to $130, with a subsequent recovery to $150, which was slightly higher than the value on the initial evaluation day. Regarding AMT, a similar pattern of uncertainty was observed (Fig. 8), with an initial price of $270, followed by a rapid uptrend phase up to $305. However, this was followed by a strong downtrend, resulting in a bottom price of $230, leading to a subsequent sideways phase. Toward the end of the evaluation period, the stock recovered its value during the final uptrend phase, ending with a value of $245, higher than its value on the first evaluation day. In the case of PLD (Fig. 9), the stock displayed an initial value of $120 and experienced an uptrend, reaching a peak of $172 around the midpoint of the evaluation period. Subsequently, it encountered fluctuations and declined to $110 before recovering to $112, a lower value than that observed

on the first evaluation day. In the case of NDAQ (Fig. 10), the stock showed a significant uptrend during the entire evaluation period, commencing at around $58 per share and reaching $73. However, there were fluctuations that caused a dip to $47.5, followed by a recovery to $61, a value similar to that of the first evaluation day. Examining SCHW (Fig. 11), the stock initiated at $74 and underwent an uptrend, peaking at $95 around the midpoint of the evaluation period. Following this, there were fluctuations leading to a decline to $60 before recovering to $74, a value similar to that on the first evaluation day. Similarly, JNJ exhibited a similar pattern of uncertainty (Fig. 12), with an initial price of $165, followed by a rapid uptrend phase up to $180. However, a significant downtrend ensued, resulting in a bottom price of $155, followed by a subsequent sideways phase. Toward the end of the evaluation period, the stock recovered its value during the final uptrend phase, ending with a value of $169, higher than its value on the first evaluation day. Lastly, examining BIO (Fig. 13), the stock displayed a significant uptrend throughout the entire evaluation period, starting at approximately $650 per share and ending at $825. However, there were fluctuations that caused a dip to $710,
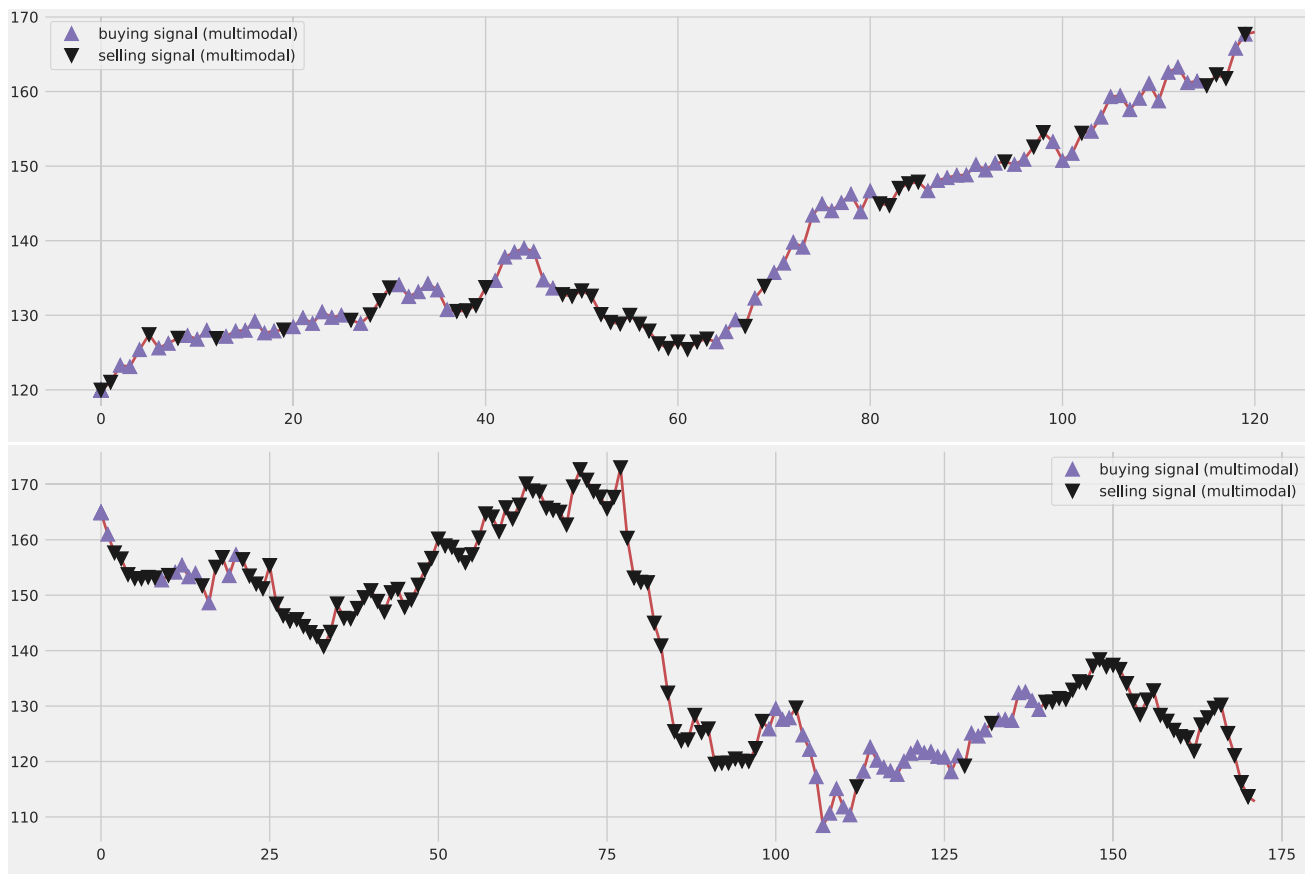
Fig. 9 Stock prices with buy and sell signals extracted by all models during the evaluation period from July 1, 2021, to Sep 20, 2022 (PLD stock)

followed by a retracement to $450, a value significantly lower than its value on the first evaluation day.

In analyzing model profitability, the findings presented in Tables 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19 indicate that the proposed multimodal approaches outperform the individual base models. This observation is particularly significant as it demonstrates that the accuracy rankings of different models, as discussed in the previous section, may not necessarily align with their profitability. A notable example of this phenomenon is evident in the case of AAPL (see Table 13), where the proposed model achieves a lower F1-Score compared to the best-performing ARIMA base model (as seen in Table 4). Despite this, the proposed model manages to achieve the largest gains, averaging 13.49% in the downtrend time frame (see

Table 13), which is notably higher than the −9.75% obtained with ARIMA.

The observed phenomenon can be explained by the tendency of binary performance metrics, like F1-Scores, to assign equal significance to the accurate prediction of trends across all days. However, these metrics may not adequately address the fluctuations in price changes occurring within a 24-hour time frame. As a result, a model that effectively predicts trends on days characterized by significant price volatility might identify advantageous opportunities for buying and selling. Conversely, a model that accurately forecasts trends for a larger number of days when prices are relatively stable may not generate a comparable level of profitability. This underscores the importance of incorporating additional considerations, such

Fig. 10 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (NDAQ stock)

as price variability, when evaluating the performance and profitability of models in the context of predicting stock trends.

The superiority of the proposed multimodal approach is also evident in its performance on uptrends in NFLX and downtrends in NFLX, SBUX, TSLA, AAPL, NDAQ, and BIO. It is worth noting that our model is particularly conservative in a downtrend phase, resulting in better capital preservation when compared with a buy-and-hold strategy. In the following, we focus on significant examples where this phenomenon is observed.

For NFLX (see Table 9), the proposed multimodal approach achieves an average gain of 19.42% in USD, surpassing the second-ranked Polarity model's performance of 14.22% in uptrend market conditions. In the case of SBUX (see Table 10), the proposed multimodal approach achieves an average gain of 7.11% in USD in downtrend market conditions, surpassing the second-ranked Dilated CNN Seq2Seq model's performance of 6.75%. In the case of TSLA (see Table 11), the proposed model stands out in downtrend market conditions as the only model with a positive average gain percentage,

exceeding CNN-LSTM, the second-best model in TSLA, by an impressive 13.75% in USD. In the case of AAPL (see Table 13), the proposed model emerges as the outperforming model demonstrating a favorable average gain percentage. It outperforms the second-ranked model, Dilated CNN Seq2Seq, in the AAPL context, by a notable margin of 12.38% in USD. Similarly, in the case of NDAQ (see Table 16), the proposed multimodal approach achieves an average gain of 3.92% in USD, outperforming the second-ranked Polarity, which obtains an average gain of 1.36%. For the case of BIO (see Table 19), the proposed multimodal approach achieves an average gain of −3.75% in USD, outperforming the second-ranked Attention-CNN-LSTM, which obtains an average gain of −4.44%.

In our analysis, it is noteworthy that in the uptrend time frame, our proposed model outperforms the buy-and-hold strategy with two stocks (ATVI, NFLX). Furthermore, our model identifies superior performance relative to the buy-and-hold strategy for nine stocks (NFLX, SBUX, NVDA, AAPL, PLD, NDAQ, SCHW, JNJ, and BIO), despite them being in a downtrend.
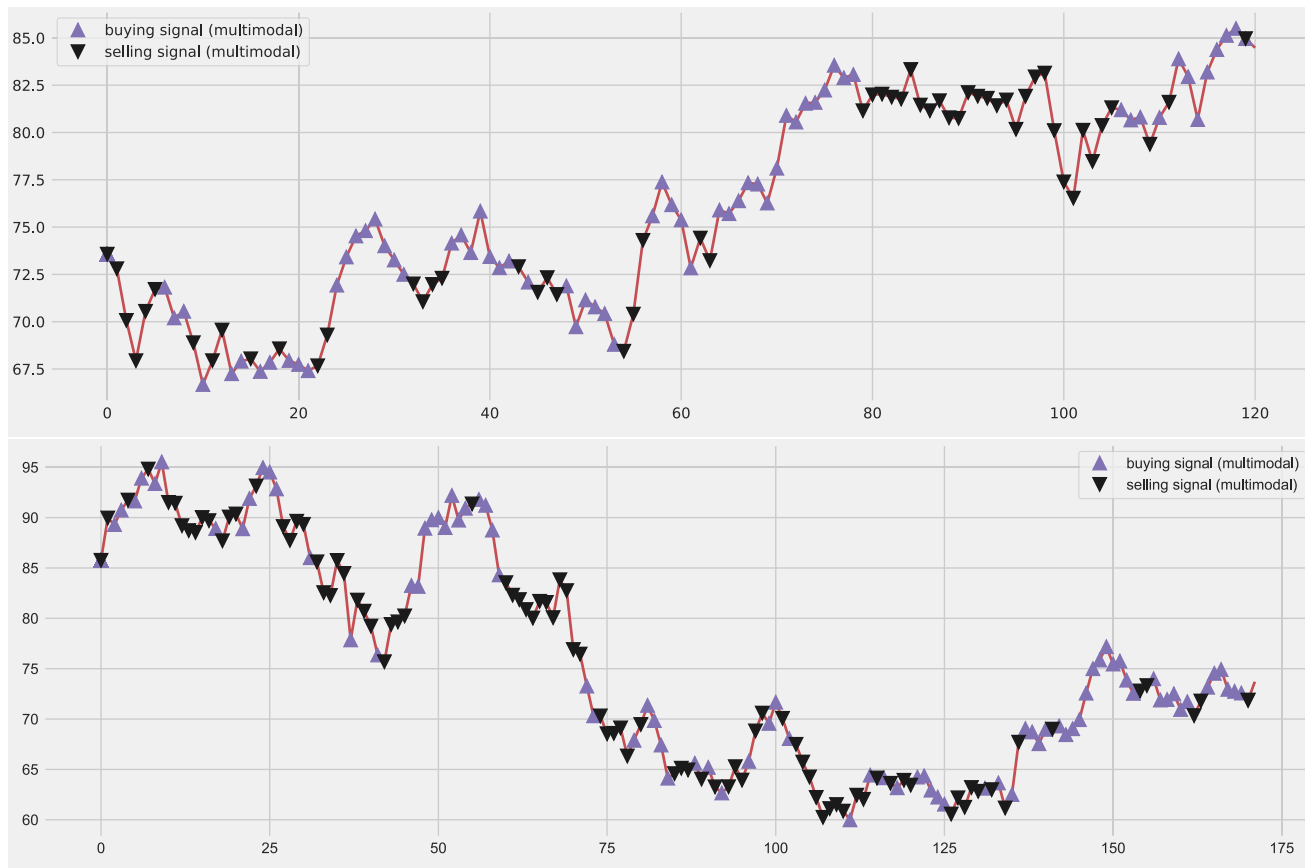
Fig. 11 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (SCHW stock)

From a theoretical viewpoint, the observed results reveal that analyzing multiple perspectives (such as stock prices, technical indicators, sentiment in news headlines) through multimodal data fusion yields a model that requires stronger evidence to predict changes in stock market trends. On the contrary, other baseline models that are mostly focused on temporal aspects of data, appear more sensitive. This consideration explains why such baseline models may have an advantage in an uptrend phase, whereas our proposed model appears more conservative. As a consequence, our multimodal fusion model presents a more limited exposure in an uptrend phase, but can provide better capital preservation capabilities than others in a downtrend phase.

In general, the superior performance observed for the proposed model can be attributed to its ability to integrate diverse data sources, extract complex and high-level features, capture nonlinear relationships, and learn in an end-to-end manner, leading to reduced bias and improved generalizability. These advantages enable the proposed multimodal model to provide more accurate and robust predictions compared to simpler models that might only utilize a single type of data or simpler feature extraction techniques.

Another interesting perspective is provided by analyzing the impact of the "Max Shares" parameter in our experiments. In the analysis of various stocks during both uptrends and downtrends, several trends emerge, although these patterns are not consistent across all cases. During uptrends, there is a tendency for multiple stocks, such as ATVI, NFLX, and SBUX, to perform better with a "Max Shares" value set to 1, while AAPL, AMT, PLD, NDAQ, JNJ, and BIO achieve best with a "Max Shares" value set to 10.

However, during downtrends, the choice of "Max Shares" value varies widely among different stocks, with some favoring a value of 10 (e.g., ATVI, AAPL, SCHW) and others performing better with a value of 1 (e.g., NFLX, TSLA, NVDA, AMT, PLD, NDAQ, SCHW, JNJ, BIO). The performance of specific models, including CNN-LSTM, GRU-LSTM, LSTM, Polarity, ARIMA, and GBTs, exhibits inconsistency, demonstrating their varied efficacy
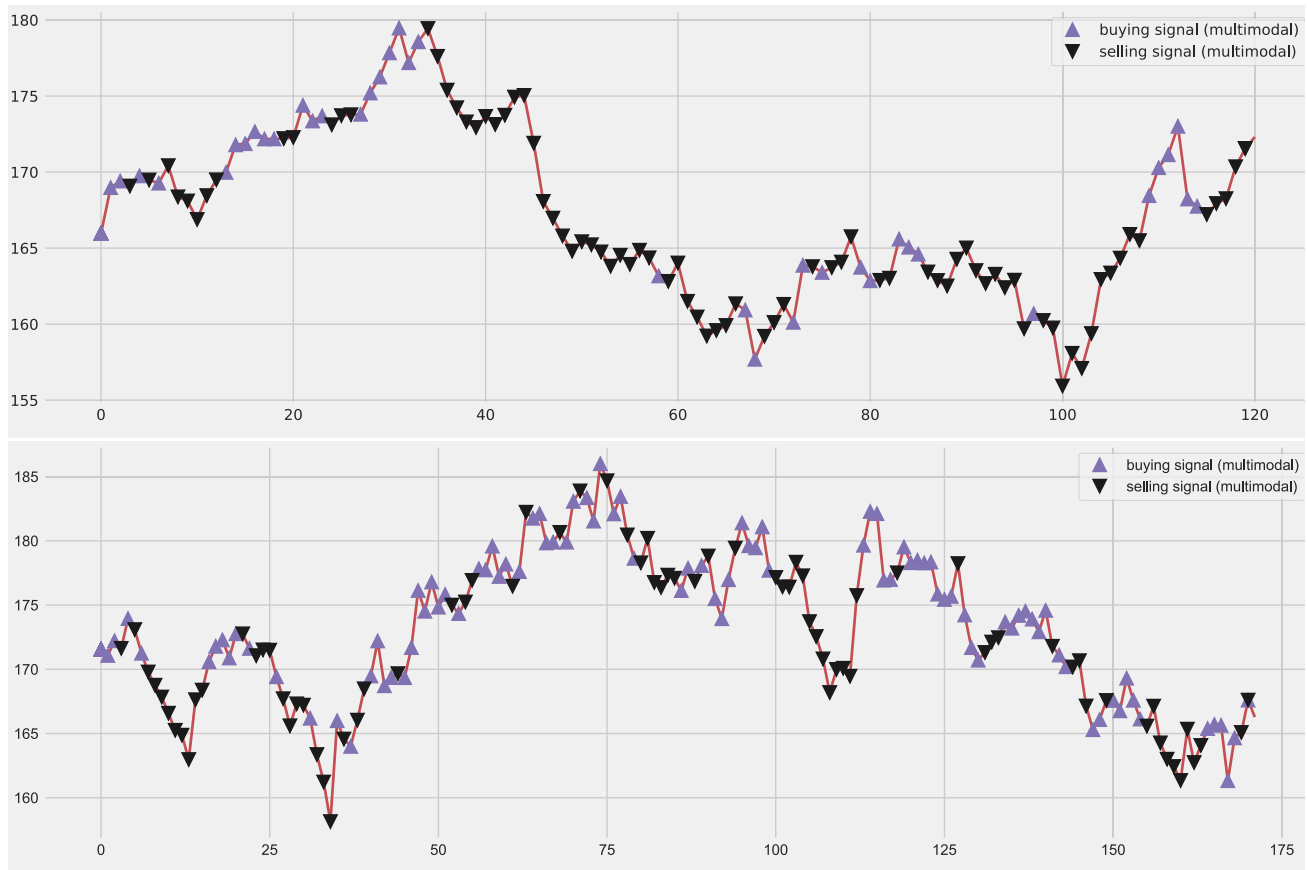
Fig. 12 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (JNJ stock)

depending on the specific stock and market conditions. These patterns suggest that the relationship between "Max Shares" and model performance is stock-specific and influenced by unique price behaviors, highlighting the need for individualized analysis and tailored strategies for each stock.

5 Conclusion

In this article, we propose a novel multimodal deep learning method for financial time series forecasting. The primary challenge of next-day trend prediction in the financial sector with joint exploitation of text and time series data is addressed by our method. To this end, our model consists of a BERT-based model branch fine-tuned on financial news, and a LSTM branch to capture useful temporal patterns. Our extensive experiments using real-world stock market datasets periods showed that our proposed method is competitive with respect to popular baselines, in both uptrend and downtrend market conditions.

Our portfolio analysis showed that our method could be fruitfully adopted in a trading scenario, yielding positive gains in an uptrend phase, as well as capital preservation in a downtrend phase, outperforming other baselines as well as a buy-and-hold strategy. Possible limitations of our work may include considering a single source for news headlines and the non-exploitation of correlations between stocks for
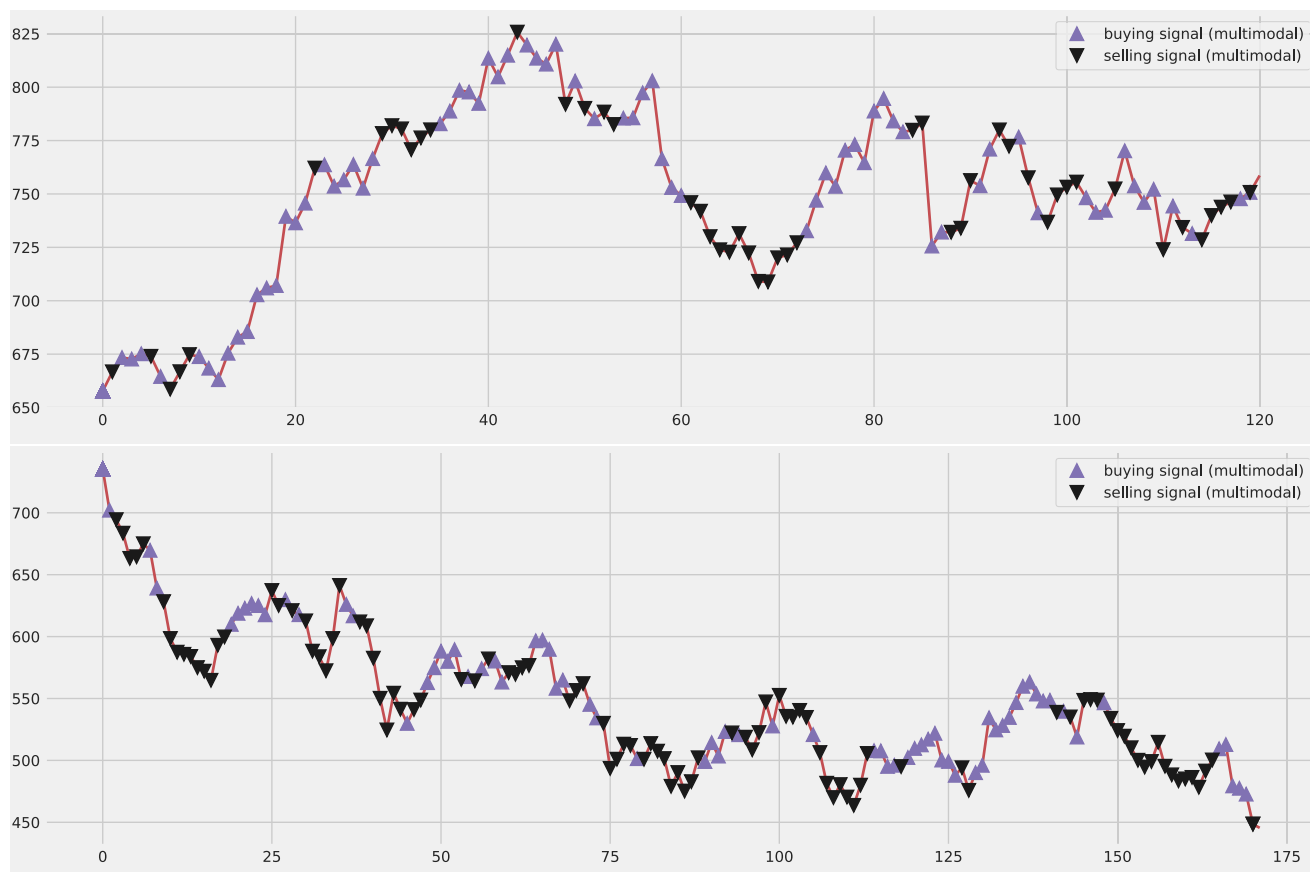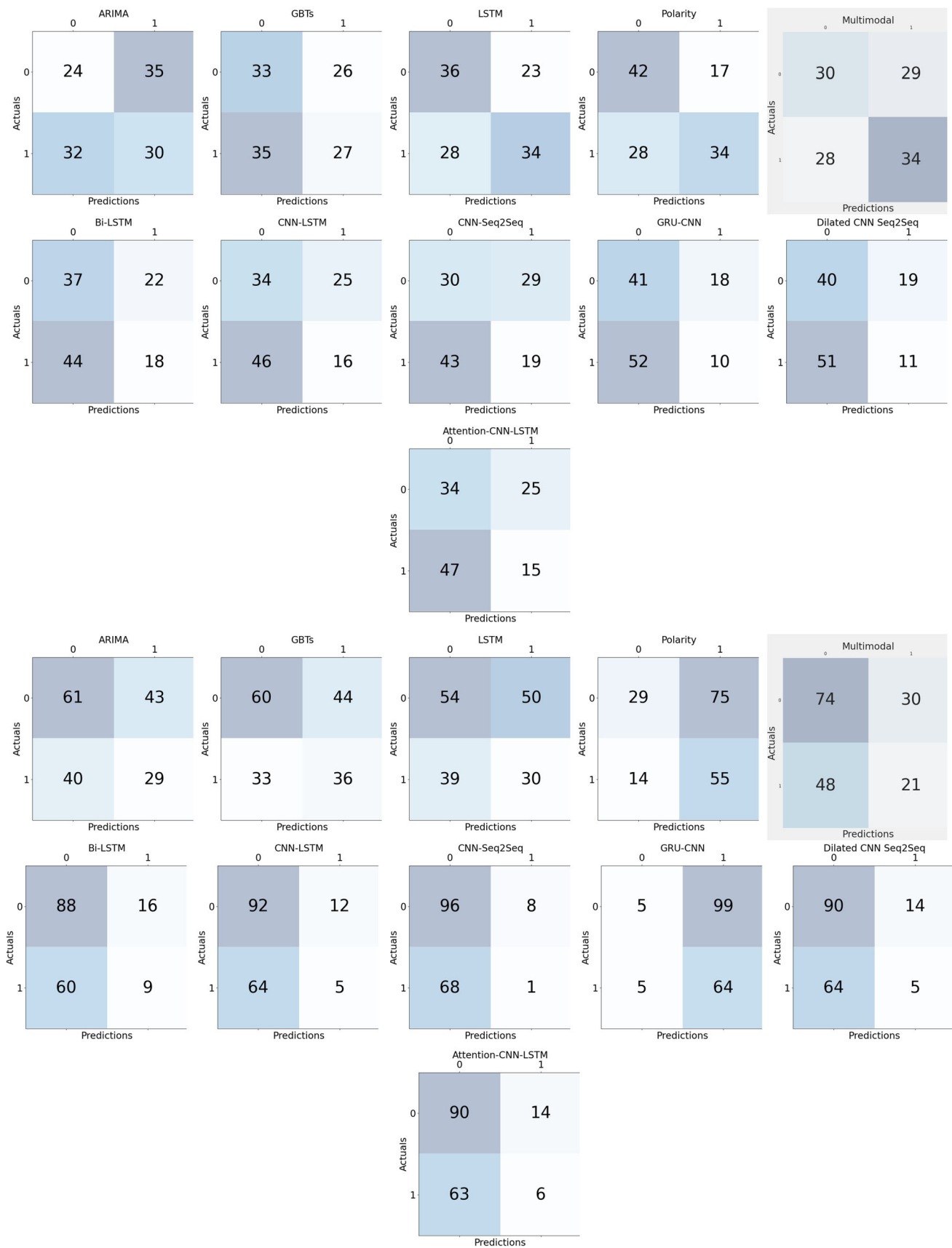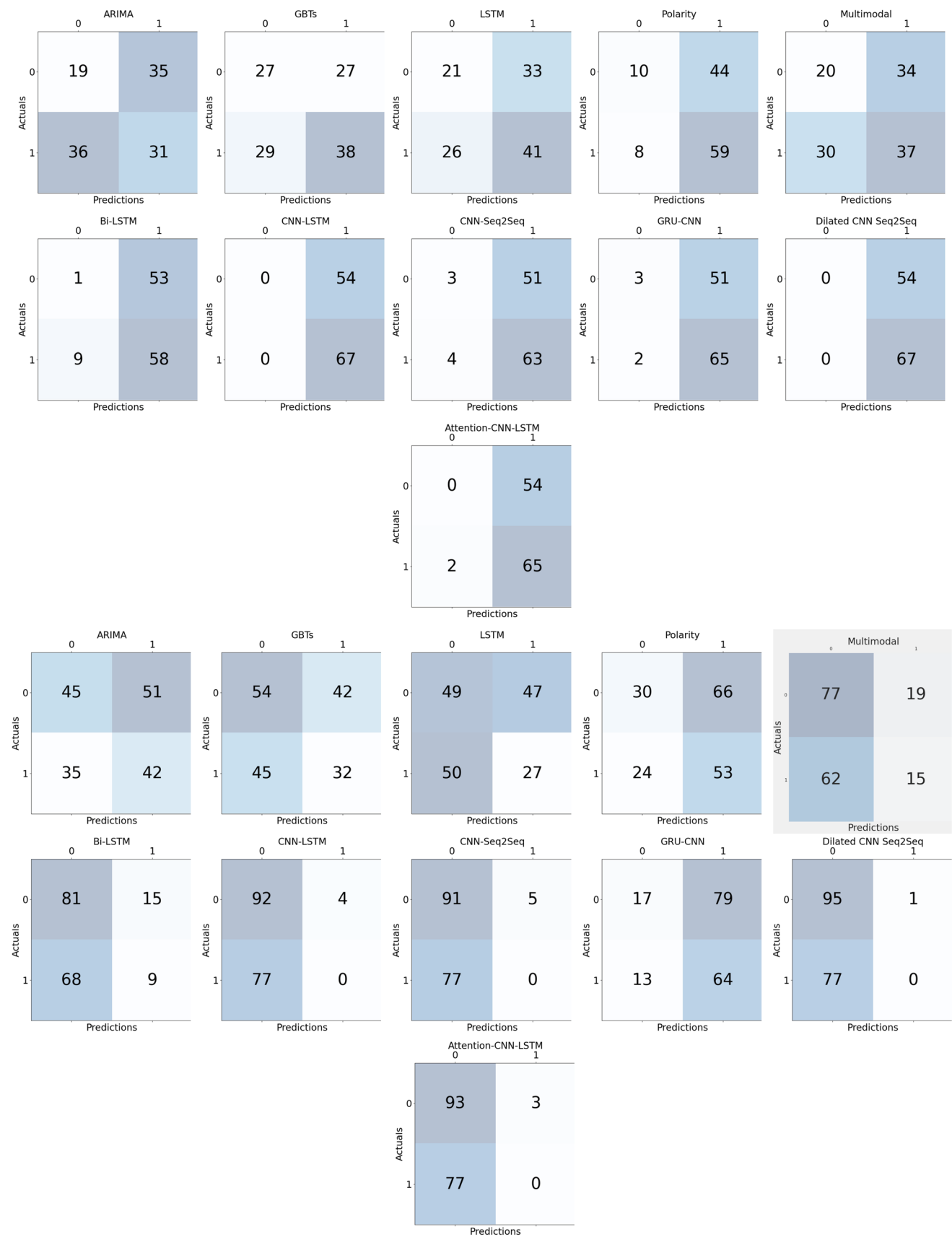
Fig. 13 Stock prices with buy and sell signals during the evaluation period from July 1, 2021, to Sep 20, 2022 (BIO stock)

the prediction of a single stock. In future research, we will study the exploitation of additional data modalities for next-day trend stock market prediction. Moreover, we will investigate the adoption of deep learning methods such as attention and graph convolution for this analytical task.

Fig. 14 Confusion matrices obtained by all methods on the next-day ▶ trend prediction task in the evaluation period from July 1, 2021 to Dec 31, 2021 (Uptrend) and Jan 1, 2022 to Sep 20, 2022 (Downtrend) (ATVI stock)

ARIMA

	0	1
0	19	35
1	36	31

Actuals / Predictions

GBTs

	0	1
0	27	27
1	29	38

Actuals / Predictions

LSTM

	0	1
0	21	33
1	26	41

Actuals / Predictions

Polarity

	0	1
0	10	44
1	8	59

Actuals / Predictions

Multimodal

	0	1
0	20	34
1	30	37

Actuals / Predictions

Bi-LSTM

	0	1
0	1	53
1	9	58

Actuals / Predictions

CNN-LSTM

	0	1
0	0	54
1	0	67

Actuals / Predictions

CNN-Seq2Seq

	0	1
0	3	51
1	4	63

Actuals / Predictions

GRU-CNN

	0	1
0	3	51
1	2	65

Actuals / Predictions

Dilated CNN Seq2Seq

	0	1
0	0	54
1	0	67

Actuals / Predictions

Attention-CNN-LSTM

	0	1
0	0	54
1	2	65

Actuals / Predictions

ARIMA

	0	1
0	45	51
1	35	42

Actuals / Predictions

GBTs

	0	1
0	54	42
1	45	32

Actuals / Predictions

LSTM

	0	1
0	49	47
1	50	27

Actuals / Predictions

Polarity

	0	1
0	30	66
1	24	53

Actuals / Predictions

Multimodal

	0	1
0	77	19
1	62	15

Actuals / Predictions

Bi-LSTM

	0	1
0	81	15
1	68	9

Actuals / Predictions

CNN-LSTM

	0	1
0	92	4
1	77	0

Actuals / Predictions

CNN-Seq2Seq

	0	1
0	91	5
1	77	0

Actuals / Predictions

GRU-CNN

	0	1
0	17	79
1	13	64

Actuals / Predictions

Dilated CNN Seq2Seq

	0	1
0	95	1
1	77	0

Actuals / Predictions

Attention-CNN-LSTM

	0	1
0	93	3
1	77	0

Actuals / Predictions

◄**Fig. 15** Confusion matrices obtained by all methods on the next-day trend prediction task in the evaluation period from July 1, 2021 to Dec 31, 2021 (Uptrend) and Jan 1, 2022 to Sep 20, 2022 (downtrend) (NFLX stock)

Appendix

Model accuracy

For ATVI, the best-performing model in uptrend market conditions is Polarity with an F1-Score of 0.63, followed by LSTM (0.58), Proposed (0.53), GBTs (0.49), ARIMA (0.45), Bi-LSTM (0.45), CNN-LSTM (0.42), CNN Seq2-Seq (0.4), Attention-CNN-LSTM (0.39), GRU-CNN (0.39), and Dilated CNN Seq2Seq (0.38). The worst-performing model in uptrend is Dilated CNN Seq2Seq with an F1-Score of 0.38. In downtrend market conditions, the best model is Proposed with an F1-Score of 0.53, followed by ARIMA (0.52), LSTM (0.49), Bi-LSTM (0.49), CNN-LSTM (0.47), Polarity (0.46), Attention-CNN-LSTM (0.46), GBTs (0.45), Dilated CNN Seq2Seq (0.45), CNN Seq2Seq (0.43), and GRU-CNN (0.28). The worst-performing model in downtrend is GRU-CNN with an F1-Score of 0.28.

Moving to NFLX, the best model for uptrends is GBTs with an F1-Score of 0.54, followed by LSTM (0.51), Polarity (0.51), Proposed (0.47), CNN Seq2Seq (0.45), GRU-CNN (0.44), ARIMA (0.41), Dilated CNN Seq2Seq (0.4), CNN-LSTM (0.4), Attention-CNN-LSTM (0.39), and Bi-LSTM (0.38). The worst-performing model for uptrends is Bi-LSTM with an F1-Score of 0.38. For downtrend prediction, the best models are GBTs and ARIMA with an F1-Score of 0.5, followed by Proposed (0.48), Polarity (0.46), Bi-LSTM (0.45), LSTM (0.44), GRU-CNN (0.41), Dilated CNN Seq2Seq (0.4), Attention-CNN-LSTM (0.4), CNN-LSTM (0.39), and CNN Seq2Seq (0.39). The worst-performing models for downtrends are CNN-LSTM and CNN Seq2Seq with an F1-Score of 0.39.

For SBUX, the best-performing models for uptrend prediction are ARIMA and GBTs with an F1-Score of 0.52, followed by LSTM (0.51), Polarity (0.51), Attention-CNN-LSTM (0.51), Bi-LSTM (0.5), Proposed (0.49), CNN-LSTM (0.43), Dilated CNN Seq2Seq (0.41), CNN Seq2Seq (0.38), and GRU-CNN (0.37). The worst-performing model for uptrends is GRU-CNN with an F1-Score of 0.37. For downtrend prediction, the best model is Dilated CNN Seq2Seq with an F1-Score of 0.57, followed by Bi-LSTM (0.55), Proposed (0.53), CNN Seq2Seq (0.53), Attention-CNN-LSTM (0.51), GBTs (0.5), GRU-CNN (0.48), ARIMA (0.47), LSTM (0.46), Polarity (0.45), and CNN-LSTM (0.37). The worst-performing model for downtrends is CNN-LSTM with an F1-Score of 0.37.

For TSLA, the best-performing models in uptrend market conditions are Dilated CNN Seq2Seq, CNN Seq2-Seq, and LSTM with an F1-Score of 0.56, followed by Bi-LSTM (0.54), Attention-CNN-LSTM (0.54), Polarity (0.53), CNN-LSTM (0.5), GBTs (0.48), ARIMA (0.47), Proposed (0.44), and GRU-CNN (0.27). The worst-performing model in uptrend is GRU-CNN with an F1-Score of 0.27. In downtrend market conditions, the best model is Polarity with an F1-Score of 0.55, followed by Proposed (0.52), ARIMA (0.52), Attention-CNN-LSTM (0.48), Bi-LSTM (0.48), CNN Seq2Seq (0.48), GBTs (0.48), Dilated CNN Seq2Seq (0.47), LSTM (0.47), GRU-CNN (0.34), and CNN-LSTM (0.31). The worst-performing model in downtrend is CNN-LSTM with an F1-Score of 0.31.

For NVDA, the best-performing model in uptrend market conditions is Bi-LSTM with an F1-Score of 0.5, followed by Polarity (0.48), GBTs (0.45), LSTM (0.45), GRU-CNN (0.44), ARIMA (0.44), Attention-CNN-LSTM (0.44), CNN Seq2Seq (0.39), Proposed (0.38), Dilated CNN Seq2Seq (0.38), and CNN-LSTM (0.35). The worst-performing model in uptrend is CNN-LSTM with an F1-Score of 0.35. In downtrend market conditions, the best models are LSTM and Polarity with an F1-Score of 0.5, followed by GBTs (0.49), ARIMA (0.49), Bi-LSTM (0.47), Proposed (0.45), GRU-CNN (0.44), CNN Seq2Seq (0.42), Attention-CNN-LSTM (0.41), Dilated CNN Seq2Seq (0.39), and CNN-LSTM (0.35). The worst-performing model in downtrend is CNN-LSTM with an F1-Score of 0.35.

In AAPL, the best-performing model in uptrend market conditions is Polarity with an F1-Score of 0.55, followed by LSTM (0.52), GBTs (0.5), Proposed (0.44), Bi-LSTM (0.43), CNN Seq2Seq (0.43), Attention-CNN-LSTM
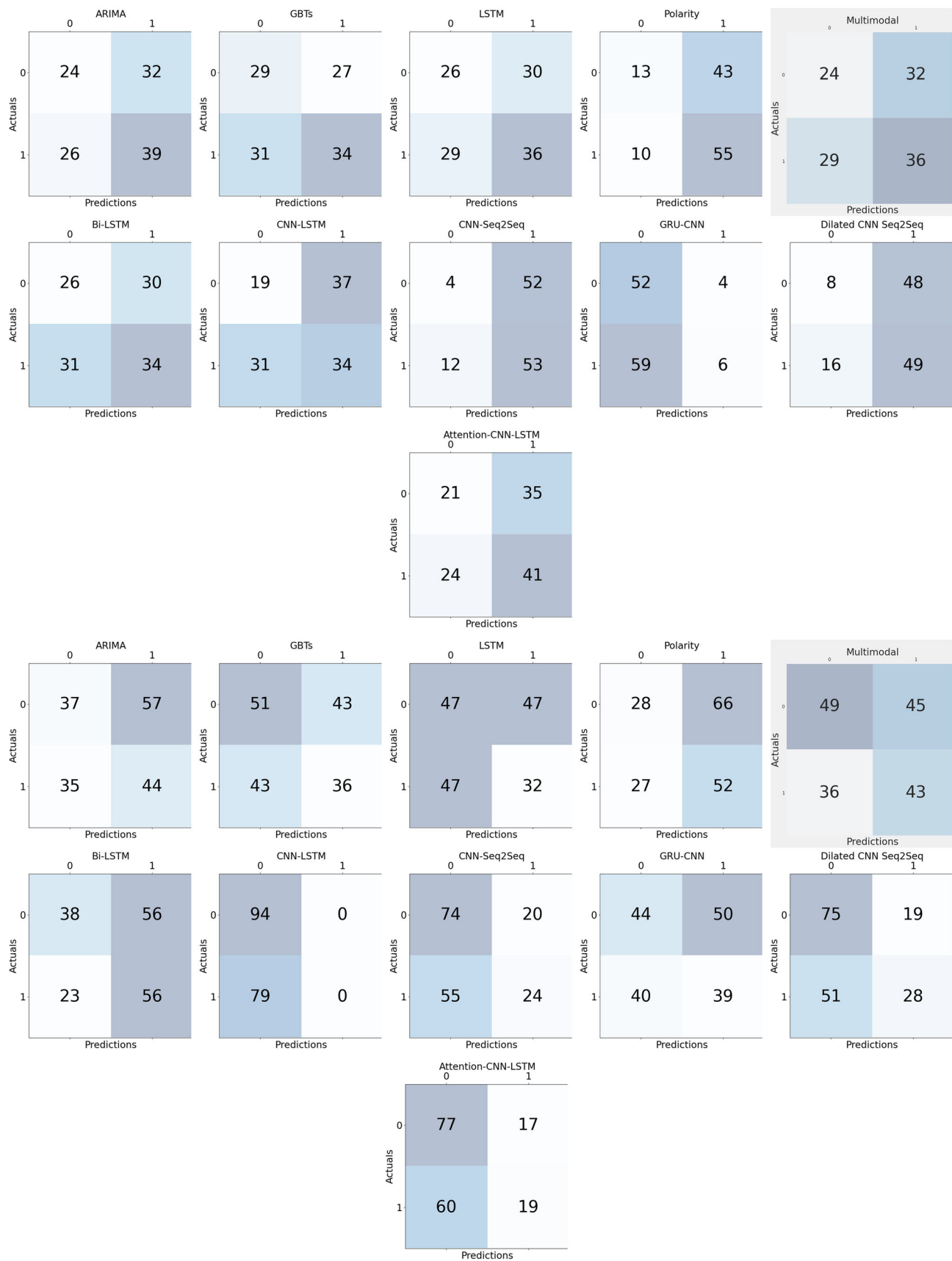
◄**Fig. 16** Confusion matrices obtained by all methods on the next-day trend prediction task in the evaluation period from July 1, 2021 to Dec 31, 2021 (Uptrend) and Jan 1, 2022 to Sep 20, 2022 (downtrend) (SBUX stock)

(0.43), Dilated CNN Seq2Seq (0.43), GRU-CNN (0.43), CNN-LSTM (0.41), and ARIMA (0.4). The worst-performing model in uptrend is ARIMA with an F1-Score of 0.4. In downtrend market conditions, the best model is ARIMA with an F1-Score of 0.55, followed by Proposed (0.51), Bi-LSTM (0.5), CNN-LSTM (0.5), Polarity (0.47), LSTM (0.47), CNN Seq2Seq (0.45), Attention-CNN-LSTM (0.46), Dilated CNN Seq2Seq (0.46), GBTs (0.45), and GRU-CNN (0.41). The worst-performing model in downtrend is GRU-CNN with an F1-Score of 0.41.

For AMT, the best-performing model in uptrend market conditions is GBTs with an F1-Score of 0.51, followed by ARIMA (0.5), Polarity (0.49), Proposed (0.46), LSTM (0.44), Bi-LSTM (0.4), CNN Seq2Seq (0.4), Attention-CNN-LSTM (0.4), Dilated CNN Seq2Seq (0.4), CNN-LSTM (0.4), and GRU-CNN (0.4). The worst-performing models in uptrend are Bi-LSTM, CNN Seq2Seq, Attention-CNN-LSTM, Dilated CNN Seq2Seq, and CNN-LSTM with an F1-Score of 0.4. In downtrend market conditions, the best model is GBTs with an F1-Score of 0.54, followed by GRU-CNN (0.48), Proposed (0.48), LSTM (0.47), Dilated CNN Seq2Seq (0.46), CNN Seq2Seq (0.46), CNN-LSTM (0.45), ARIMA (0.4), Polarity (0.4), Attention-CNN-LSTM (0.43), and Bi-LSTM (0.4). The worst-performing model in downtrend is Bi-LSTM with an F1-Score of 0.4.

Moving to PLD, the best model for uptrends is ARIMA with an F1-Score of 0.59, followed by Polarity (0.57), GBTs (0.55), GRU-CNN (0.55), LSTM (0.54), Proposed (0.53), Bi-LSTM (0.5), Dilated CNN Seq2Seq (0.49), CNN Seq2Seq (0.49), CNN-LSTM (0.49), and Attention-CNN-LSTM (0.48). The worst-performing model for uptrends is Attention-CNN-LSTM with an F1-Score of 0.48. For downtrend prediction, the best model is GRU-CNN with an F1-Score of 0.6, followed by GBTs (0.58), Proposed (0.53), LSTM (0.51), Dilated CNN Seq2Seq (0.47), ARIMA (0.41), CNN Seq2Seq (0.39), Bi-LSTM (0.39), CNN-LSTM (0.39), Attention-CNN-LSTM (0.38), and

Polarity (0.3). The worst-performing model for downtrends is Polarity with an F1-Score of 0.3.

In NDAQ, the best-performing model in uptrend market conditions is ARIMA with an F1-Score of 0.58, followed by LSTM (0.55), Bi-LSTM (0.5), Attention-CNN-LSTM (0.49), Dilated CNN Seq2Seq (0.48), CNN Seq2Seq (0.48), CNN-LSTM (0.47), GBTs (0.46), Proposed (0.45), GRU-CNN (0.45), and Polarity (0.39). The worst-performing model in uptrend is Polarity with an F1-Score of 0.39. In downtrend market conditions, the best model is GBTs with an F1-Score of 0.52, followed by Proposed (0.5), ARIMA (0.46), LSTM (0.43), Bi-LSTM (0.43), GRU-CNN (0.42), CNN-LSTM (0.4), CNN Seq2Seq (0.39), Attention-CNN-LSTM (0.39), Dilated CNN Seq2Seq (0.39), and Polarity (0.36). The worst-performing model for downtrends is Polarity with an F1-Score of 0.36.

In SCHW, the best-performing model for uptrend prediction is GBTs with an F1-Score of 0.55, followed by Proposed (0.5), LSTM (0.49), ARIMA (0.46), Bi-LSTM (0.45), CNN Seq2Seq (0.44), Polarity (0.43), GRU-CNN (0.42), Attention-CNN-LSTM (0.42), Dilated CNN Seq2-Seq (0.4), and CNN-LSTM (0.36). The worst-performing model for uptrends is CNN-LSTM with an F1-Score of 0.36. For downtrend prediction, the best model is GBTs with an F1-Score of 0.53, followed by LSTM (0.49), Dilated CNN Seq2Seq (0.47), ARIMA (0.46), Proposed (0.45), CNN Seq2Seq (0.43), GRU-CNN (0.42), CNN-LSTM (0.39), Polarity (0.38), Attention-CNN-LSTM (0.37), and Bi-LSTM (0.37). The worst-performing models for downtrends are Attention-CNN-LSTM and Bi-LSTM with an F1-Score of 0.37.

Moving to BIO, the best-performing model in uptrend market conditions is GRU-CNN with an F1-Score of 0.56, followed by GBTs (0.55), LSTM (0.52), Proposed (0.5), Dilated CNN Seq2Seq (0.45), ARIMA (0.44), Bi-LSTM (0.41), CNN-LSTM (0.41), CNN Seq2Seq (0.41), Attention-CNN-LSTM (0.4), and Polarity (0.4). The worst-performing models in uptrend are Attention-CNN-LSTM and Polarity with an F1-Score of 0.4. In downtrend market conditions, the best model is GRU-CNN with an F1-Score of 0.56, followed by GBTs (0.55), Bi-LSTM (0.54), LSTM (0.52), ARIMA (0.45), Proposed (0.43), Dilated CNN Seq2Seq (0.42), CNN Seq2Seq (0.42), CNN-LSTM (0.42),
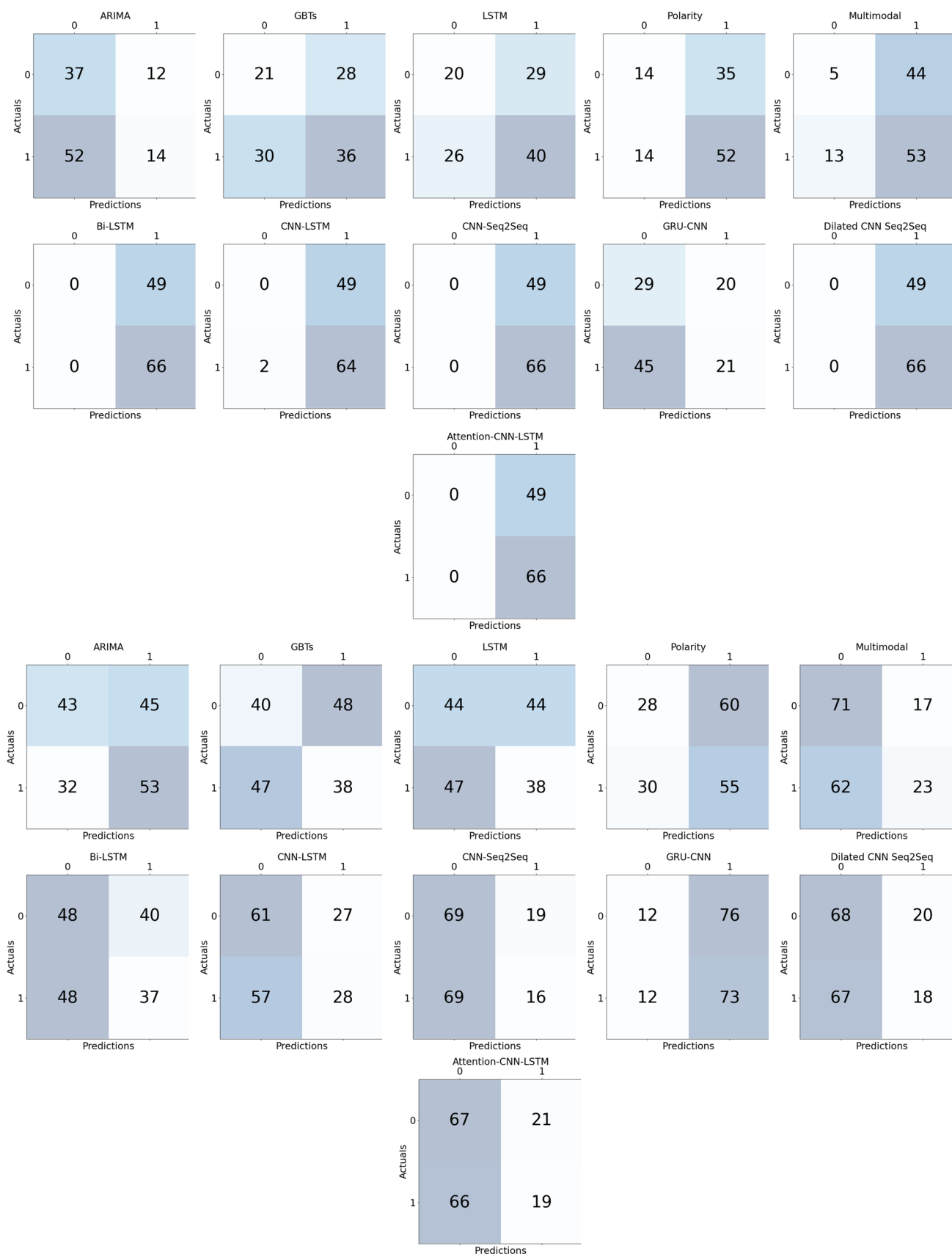
◄ **Fig. 17** Confusion matrices with all methods on the next-day trend prediction task in the evaluation period from July 1, 2021 to Dec 31, 2021 (Uptrend) and Jan 1, 2022 to Sep 20, 2022 (downtrend) (AAPL stock)

Attention-CNN-LSTM (0.42), and Polarity (0.26). The worst-performing model in downtrend is Polarity with an F1-Score of 0.26.

Lastly, for JNJ, the best-performing model in uptrend market conditions is GBTs with an F1-Score of 0.56, followed by Bi-LSTM (0.52) and Polarity (0.52), LSTM (0.51), GRU-CNN (0.51), Proposed (0.5), Attention-CNN-LSTM (0.5), CNN-LSTM (0.48), CNN Seq2Seq (0.48), Dilated CNN Seq2Seq (0.46), ARIMA (0.46). The worst-performing model in uptrend is Polarity with an F1-Score of 0.49. In downtrend market conditions, the best model is ARIMA with an F1-Score of 0.54, followed by Polarity (0.49), LSTM (0.48), Bi-LSTM (0.48), Proposed (0.47), GBTs (0.47), CNN Seq2Seq (0.42), Dilated CNN Seq2Seq (0.42), Attention-CNN-LSTM (0.41), CNN-LSTM (0.4), and GRU-CNN (0.38). The worst-performing model in downtrend is GRU-CNN with an F1-Score of 0.38.

Confusion matrices

It is important to note that relying solely on classification metrics may provide a limited perspective of the results. A closer look at the confusion matrices for a selection of stocks in communication services (ATVI-Fig. 14 and NFLX-Fig. 15), consumer services (SBUX-Fig. 16), information technology (AAPL-Fig. 17), real estate (AMT-Fig. 18), healthcare (JNJ-Fig. 19) reveals the diversity of

model predictions with uptrend (top) and downtrend (bottom).

Taking ATVI as an example (see Fig. 14) we observe that correct predictions for days of uptrend and downtrend (as shown by the number of true negatives—TN and true positives—TP, respectively) are rather balanced for some of the models in uptrend, such as ARIMA (24, 30), LSTM (36, 34), GBTs (33, 27), Polarity (42, 34) and imbalanced for others, such as Bi-LSTM (37, 18), CNN-LSTM (34, 16), CNN Seq2Seq (30, 19), GRU-CNN (41, 10), Dilated CNN Seq2Seq (40, 11), and Attention-CNN-LSTM (34, 15). In downtrend, we observe that some models are relatively imbalanced, like LSTM (54, 30), GBTs (60, 36), Polarity (29, 55), and ARIMA (61, 29), and the proposed multimodal model (74, 21), while others are extremely imbalanced, like Bi-LSTM (88, 9), CNN-LSTM (92, 5), CNN Seq2Seq (96, 1), GRU-CNN (5, 64), Dilated CNN Seq2Seq (90, 5), and Attention-CNN-LSTM (90, 6).

Shifting the focus to NFLX, GBTs presents the most balanced performance in terms of correctly predicted days of the uptrend (27–38), followed by LSTM (21–41) and ARIMA (19–31). Results for the other models are otherwise quite imbalanced. In downtrend, in decreasing order of imbalance we can find CNN Seq2Seq (91, 0), together with Dilated CNN Seq2Seq (90, 0) and attention-CNN-LSTM (93, 0), followed by CNN-LSTM (92, 5), Bi-LSTM (81, 9), GRU-CNN (17, 64), our multimodal approach (77–15), LSTM (49–27), and Polarity (30–53), GBTs (54–32), and ARIMA (45–42), which presents the most balanced performance.

Overall, some of the models show imbalance in their predictions, as observed with the proposed multimodal approach in uptrend for AAPL, and with ARIMA in
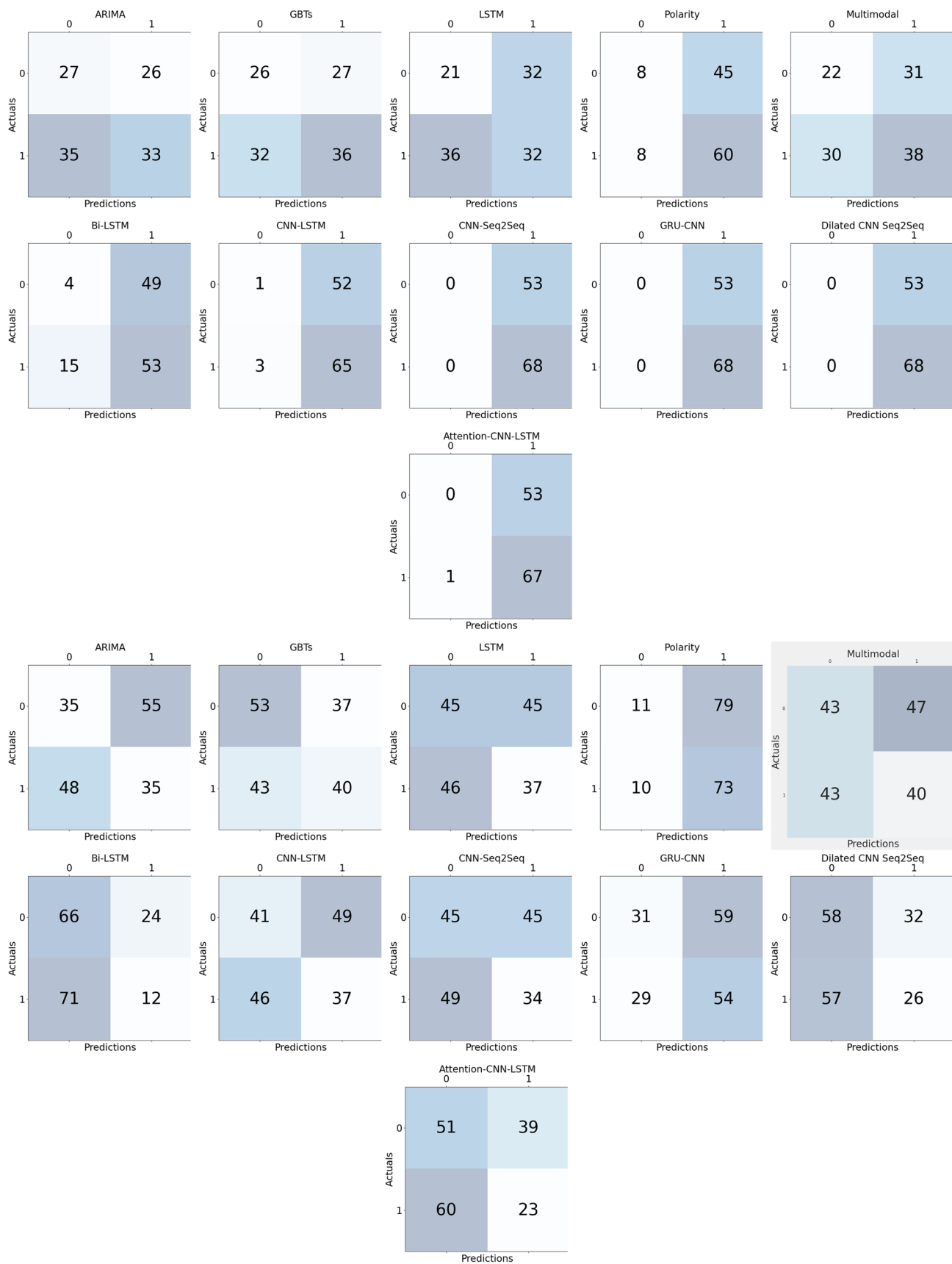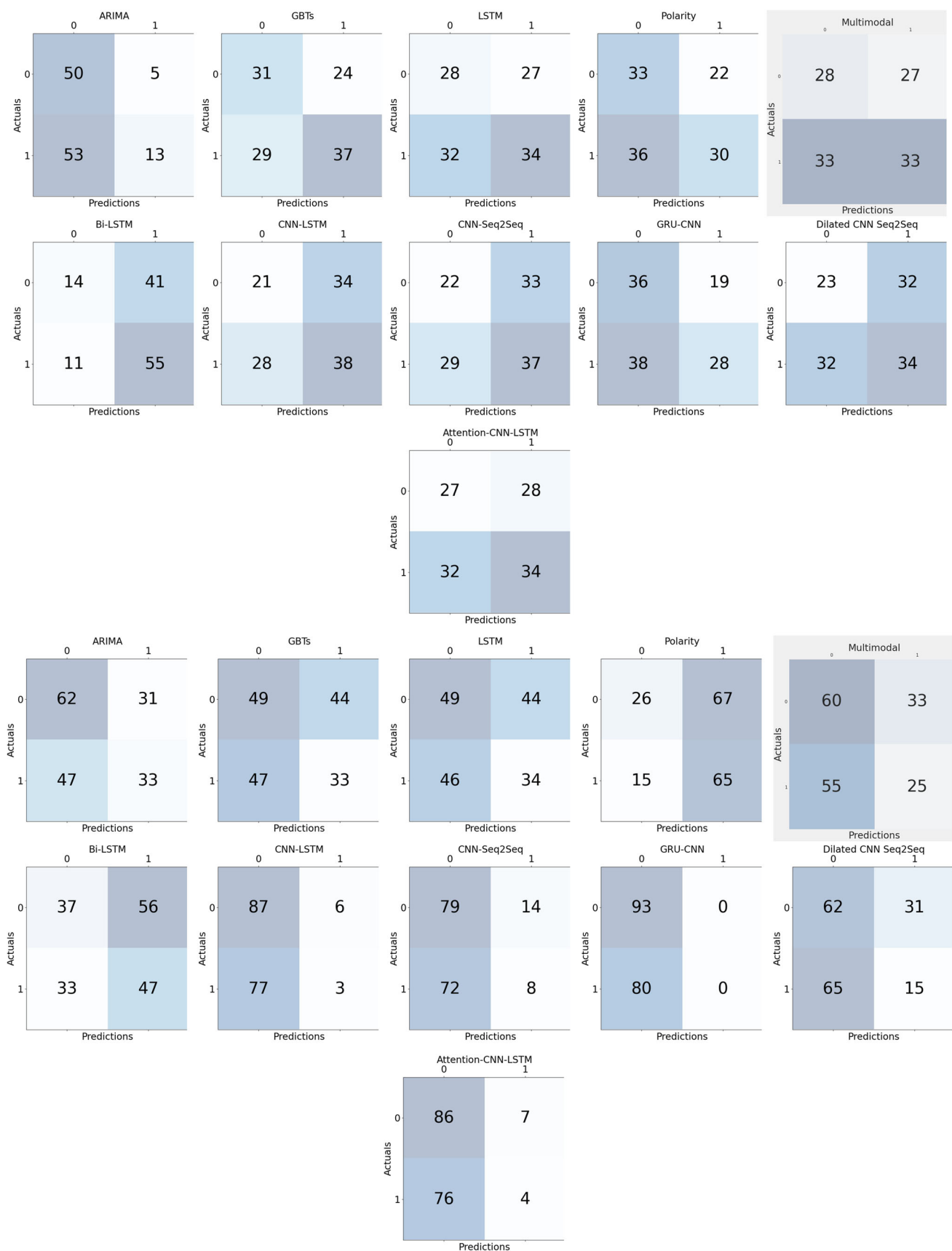
◄**Fig. 18** Confusion matrices obtained by all methods on the next-day trend prediction task in the evaluation period from July 1, 2021 to Dec 31, 2021 (Uptrend) and Jan 1, 2022 to Sep 20, 2022 (downtrend) (AMT stock)

uptrend for JNJ, as well as for the majority of the stocks with the Polarity approach in both uptrend and downtrend (AMT, PLD, NVDA, NDAQ, SCHW, JNJ, and BIO).

Furthermore, it is noteworthy that a majority of stocks analyzed using Bi-LSTM, CNN-LSTM, CNN Seq2Seq, GRU-CNN, Dilated CNN Seq2Seq, and Attention-CNN-LSTM methodologies exhibit imbalances in both uptrends and downtrends. Specifically, stocks such as NFLX, NVDA, AAPL, AMT, PLD, NDAQ, SCHW, and BIO demonstrate imbalances during uptrends, whereas ATVI, NFLX, AAPL, PLD, NDAQ, SCHW, JNJ, and BIO display imbalances during downtrends.

Portfolio analysis: impact of max shares

In the context of an upward trend in the stock price of ATVI, it is observed that all models exhibit superior performance when "Max Shares" is set to 1. In contrast, during a downward trend in ATVI, the models ARIMA, CNN-LSTM, GRU-CNN, LSTM, Polarity, Attention-CNN-LSTM, and the Proposed model demonstrate enhanced performance when "Max Shares" is set to 10, while Dilated CNN Seq2Seq, GBTs, Bi-LSTM, and CNN Seq2Seq yield similar results at a "Max Shares" value of 1 (see Table 8). For the stock price of NFLX in an uptrend, the models CNN-LSTM, Dilated CNN Seq2Seq, GBTs, GRU-CNN, LSTM, Bi-LSTM, Polarity, Attention-CNN-LSTM, and CNN Seq2Seq are found to perform better with a "Max Shares" value of 1, while ARIMA excels with a "Max Shares" value of 10, and the Proposed model performs best at 5. Conversely, during a downtrend in NFLX, only the ARIMA model performs better with a "Max

Shares" value of 10 (although even in this best case it incurs in catastrophic losses of -40.48%), whereas all other models perform best with a "Max Shares" value of 1 (see Table 9). In the case of SBUX during an uptrend, the models CNN-LSTM, Dilated CNN Seq2Seq, GBTs, LSTM, Polarity, and CNN Seq2Seq exhibit a better performance with a "Max Shares" value of 1, while ARIMA, GRU-CNN, Bi-LSTM, Attention-CNN-LSTM, and the Proposed models perform best with a "Max Shares" value of 10. In a downtrend for SBUX, the Dilated CNN Seq2-Seq, GBTs, LSTM, and CNN Seq2Seq models outperform with a "Max Shares" value of 10, while ARIMA, CNN-LSTM, GRU-CNN, Bi-LSTM, and Polarity yield better results with a "Max Shares" value of 1, and the Proposed and Attention-CNN-LSTM models excel at a "Max Shares" value of 5 (see Table 10). For the TSLA stock in an uptrend, the Dilated CNN Seq2Seq, LSTM, Bi-LSTM, Polarity, Attention-CNN-LSTM, and the Proposed models are found to perform better when the "Max Shares" value is set to 5, while ARIMA, CNN-LSTM, GBTs, GRU-CNN, and CNN Seq2Seq yield superior results at a "Max Shares" value of 10. During a TSLA downtrend, models like ARIMA, CNN-LSTM, Dilated CNN Seq2Seq, GRU-CNN, LSTM, Attention-CNN-LSTM, and CNN Seq2Seq excel with a "Max Shares" value of 1, while Bi-LSTM and Polarity perform best with a "Max Shares" value of 10, and GBTs and the Proposed model provide the highest performance at a "Max Shares" value of 5 (see Table 11). In the context of an uptrend for NVDA, models such as ARIMA, GBTs, LSTM, Bi-LSTM, Attention-CNN-LSTM, and Proposed exhibit superior performance when the "Max Shares" value is set to 10, whereas CNN-LSTM, GRU-CNN, Polarity, and CNN Seq2Seq provide the highest performance at a "Max Shares" value of 5, and Dilated CNN Seq2Seq perform best at a "Max Shares" value of 1. During a downtrend in NVDA, all models perform best with a "Max Shares" value of 1 (see Table 12). In the case

ARIMA (Predictions 0, 1; Actuals 0, 1): 50, 5 / 53, 13

GBTs: 31, 24 / 29, 37

LSTM: 28, 27 / 32, 34

Polarity: 33, 22 / 36, 30

Multimodal: 28, 27 / 33, 33

Bi-LSTM: 14, 41 / 11, 55

CNN-LSTM: 21, 34 / 28, 38

CNN-Seq2Seq: 22, 33 / 29, 37

GRU-CNN: 36, 19 / 38, 28

Dilated CNN Seq2Seq: 23, 32 / 32, 34

Attention-CNN-LSTM: 27, 28 / 32, 34

ARIMA: 62, 31 / 47, 33

GBTs: 49, 44 / 47, 33

LSTM: 49, 44 / 46, 34

Polarity: 26, 67 / 15, 65

Multimodal: 60, 33 / 55, 25

Bi-LSTM: 37, 56 / 33, 47

CNN-LSTM: 87, 6 / 77, 3

CNN-Seq2Seq: 79, 14 / 72, 8

GRU-CNN: 93, 0 / 80, 0

Dilated CNN Seq2Seq: 62, 31 / 65, 15

Attention-CNN-LSTM: 86, 7 / 76, 4

◀**Fig. 19** Confusion matrices obtained by all methods on the next-day trend prediction task in the evaluation period from July 1, 2021 to Dec 31, 2021 (Uptrend) and Jan 1, 2022 to Sep 20, 2022 (downtrend) (JNJ stock)

of AAPL during an uptrend, only the GBTs model performs better with a "Max Shares" value of 5, while the other models perform best at a "Max Shares" value of 10. In a downtrend for AAPL, we observe that CNN-LSTM, Dilated CNN Seq2Seq, LSTM, Attention-CNN-LSTM, CNN Seq2Seq, and Proposed models exhibit superior performance at a "Max Shares" value of 10, and Bi-LSTM performs best at 5, while the rest of the models perform best with a "Max Shares" value of 1 (see Table 13). For AMT in an uptrend, the LSTM, ARIMA, and Bi-LSTM models perform better with a "Max Shares" value of 1, while the rest of the models perform best with a "Max Shares" value of 10. In a downtrend for AMT, all models perform best at a "Max Shares" value of 1 (see Table 14). During an uptrend in PLD, the GRU-CNN, LSTM, and Bi-LSTM models exhibit a superior performance with a "Max Shares" value of 5, while the rest obtain the best performance at a "Max Shares" value of 10. In a PLD downtrend, all of the models perform best at a "Max Shares" value of 1 (see Table 15). In the context of an uptrend in NDAQ, it can be observed that the CNN-LSTM, Dilated CNN Seq2Seq, and CNN Seq2Seq models perform best with "Max Shares" equal to 5, while the rest of the models present a better performance with a "Max Shares" value of 10. During a downtrend in NDAQ, only ARIMA and the Proposed model exhibit superior performance with a "Max Shares" value of 5, while the rest perform best at a "Max Shares" value of 1 (see Table 16). For SCHW in an uptrend, most of the models perform better with a "Max Shares" value of 10, while Dilated CNN Seq2Seq, Bi-LSTM, Attention-CNN-LSTM, CNN Seq2Seq performs best at 5, and GRU-CNN obtains the best result at a "Max Shares" value of 1. In a SCHW downtrend, only Dilated CNN Seq2Seq, GBTs, Bi-LSTM, and Attention-CNN-LSTM demonstrate superior performance with a "Max Shares" value of 10, while the rest yield the highest performance at a "Max Shares" value of 1 (see Table 17). During an upward trend in JNJ, the LSTM and Polarity models demonstrate optimal performance when the "Max Shares" parameter is set to 1. Conversely, the CNN-LSTM, Dilated CNN Seq2Seq, and CNN Seq2Seq models achieve their peak performance at a "Max Shares" value of 5. As for the ARIMA, GBTs, GRU-CNN, Bi-LSTM, Attention-CNN-LSTM, and Proposed models, they yield the best results when "Max Shares" is adjusted to 10. In a downtrend scenario for JNJ, only Polarity and the Proposed model exhibit superior performance when the "Max

Shares" value is set to 5. Conversely, the other models perform optimally when "Max Shares" is set to 1. (see Table 18). In the scenario of BIO during an upward trend, both the LSTM and ARIMA models exhibit superior performance when the "Max Shares" parameter is set to 1. Conversely, Polarity and the Proposed model deliver optimal outcomes when "Max Shares" is adjusted to 10, whereas GBTs show the highest effectiveness with a "Max Shares" setting of 5. When facing a BIO downtrend, GBTs and the GRU-CNN showcase enhanced performance with a "Max Shares" value of 10, while the Proposed model demonstrates superior results at a "Max Shares" value of 5. Meanwhile, the remaining models demonstrate their best performance when "Max Shares" is set to 1. (see Table 19).

Declarations

References

1. Cowles A 3rd (1933) Can stock market forecasters forecast? Econometrica: J Econ Soc 1(3):309–324. https://doi.org/10.2307/1907042
2. Cao CQ, Tsay RS (1992) Nonlinear time-series analysis of stock volatilities. J Appl Economet 7(S1):165–185
3. Olorunnimbe K, Viktor H (2023) Deep learning in the stock market-a systematic survey of practice, backtesting, and applications. Artif Intell Rev 56(3):2057–2109
4. Kumar U, Jain V (2010) Arima forecasting of ambient air pollutants (o 3, no, no 2 and co). Stoch Env Res Risk Assess 24(5):751–760

5. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M (2020) Application of the Arima model on the Covid-2019 epidemic dataset. Data Brief 29:105340

6. Taylor SJ, Letham B (2018) Forecasting at scale. Am Stat 72(1):37–45

7. Stock JH, Watson MW (2001) Vector autoregressions. J Econ Perspect 15(4):101–115

8. Zhao Y, Ye L, Pinson P, Tang Y, Lu P (2018) Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting. IEEE Trans Power Syst 33(5):5029–5040

9. Salisu AA, Gupta R, Ogbonna AE (2022) A moving average heterogeneous autoregressive model for forecasting the realized volatility of the us stock market: evidence from over a century of data. Int J Financ Econ 27(1):384–400

10. Al-Shiab M (2006) The predictability of the Amman stock exchange using the univariate autoregressive integrated moving average (Arima) model. J Econ Adm Sci 22(2):17–35

11. Billah MM, Sultana A, Bhuiyan F, Kaosar MG (2024) Stock price prediction: comparison of different moving average techniques using deep learning model. Neural Comput Appl 36(11):5861–5871

12. Arsov M, Zdravevski E, Lameski P, Corizzo R, Koteli N, Gramatikov S, Mitreski K, Trajkovik V, Marín ST (2021) Multi-horizon air pollution forecasting with deep neural networks. Sensors 21(4):14248220

13. Bhandari HN, Rimal B, Pokhrel NR, Rimal R, Dahal KR, Khatri RK (2022) Predicting stock market index using LSTM. Mach Learn Appl 9:100320

14. Corizzo R, Yepez-Lopez R, Gilbert S, Japkowicz N (2022) LSTM-based pulmonary air leak forecasting for chest tube management. In: 2022 IEEE international conference on big data (Big Data), IEEE, pp 5217–5222

15. Corizzo R, Ceci M, Fanaee-T H, Gama J (2021) Multi-aspect renewable energy forecasting. Inf Sci 546:701–722

16. Shah D, Campbell W, Zulkernine FH (2018) A comparative study of LSTM and DNN for stock market forecasting. In: 2018 IEEE international conference on big data (big Data), IEEE, pp 4148–4155

17. Mahjoub S, Chrifi-Alaoui L, Marhic B, Delahoche L (2022) Predicting energy consumption using LSTM, multi-layer GRU and drop-GRU neural networks. Sensors 22(11):4062

18. Banik S, Sharma N, Mangla M, Mohanty SN, Shitharth S (2022) Lstm based decision support system for swing trading in stock market. Knowl-Based Syst 239:107994

19. Althelaya KA, El-Alfy E-SM, Mohammed S (2018) Evaluation of bidirectional LSTM for short-and long-term stock market prediction. In: 2018 9th international conference on information and communication systems (ICICS), IEEE, pp 151–156

20. Shaban WM, Ashraf E, Slama AE (2024) Smp-dl: a novel stock market prediction approach based on deep learning for effective trend forecasting. Neural Comput Appl 36(4):1849–1873

21. Vaziri J, Farid D, Nazemi Ardakani M, Hosseini Bamakan SM, Shahlaei M (2023) A time-varying stock portfolio selection model based on optimized PSO-BILSTM and multi-objective mathematical programming under budget constraints. Neural Comput Appl 35(25):18445–18470

22. Weng B, Lu L, Wang X, Megahed FM, Martinez W (2018) Predicting short-term stock prices using ensemble methods and online data sources. Expert Syst Appl 112:258–273

23. Lin C-T, Wang Y-K, Huang P-L, Shi Y, Chang Y-C (2022) Spatial-temporal attention-based convolutional network with text and numerical information for stock price prediction. Neural Comput Appl 34(17):14387–14395

24. Akter MS, Shahriar H, Chowdhury R, Mahdy M (2022) Forecasting the risk factor of frontier markets: a novel stacking ensemble of neural network approach. Future Internet 14(9):252

25. Song H, Choi H (2023) Forecasting stock market indices using the recurrent neural network based hybrid models: CNN-LSTM, GRU-CNN, and ensemble models. Appl Sci 13(7):4644

26. Srijiranon K, Lertratanakham Y, Tanantong T (2022) A hybrid framework using PCA, EMD and LSTM methods for stock market price prediction with sentiment analysis. Appl Sci 12(21):10823

27. Wang J, Cui Q, Sun X, He M (2022) Asian stock markets closing index forecast based on secondary decomposition, multi-factor analysis and attention-based LSTM model. Eng Appl Artif Intell 113:104908

28. Cen Y, Luo M, Cen G, Zhao C, Cheng Z (2022) Financial market correlation analysis and stock selection application based on TCN-deep clustering. Future Internet 14(11):331

29. Li Y, Pan Y (2022) A novel ensemble deep learning model for stock prediction based on stock prices and news. Int J Data Sci Anal 13(2):139–149

30. Hasselgren B, Chrysoulas C, Pitropakis N, Buchanan WJ (2023) Using social media & sentiment analysis to make investment decisions. Future Internet 15(1):5

31. Aldhyani TH, Alzahrani A (2022) Framework for predicting and modeling stock market prices based on deep learning algorithms. Electronics 11(19):3149

32. Yu X, Wu W, Liao X, Han Y (2023) Dynamic stock-decision ensemble strategy based on deep reinforcement learning. Appl Intell 53(2):2452–2470

33. Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news: The Azfin text system. ACM Trans Inform Syst 27(2):1–19

34. Corizzo R, Rosen J (2023) Stock market prediction with time series data and news headlines: a stacking ensemble approach. J Intell Inf Syst 62(1):27–56

35. Liu Z, Huang D, Huang K, Li Z, Zhao J (2021) Finbert: A pre-trained financial language representation model for financial text mining. In: Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence, pp 4513–4519

36. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

37. Jaiswal R, Singh B (2022) A hybrid convolutional recurrent (CNN-GRU) model for stock price prediction. In: 2022 IEEE 11th international conference on communication systems and network technologies (CSNT), pp 299–304. IEEE

38. Graves A, Mohamed A-R, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, pp 6645–6649

39. Jialin L, Shanwen Q, Zhikai Z, Keyao L, Jiayong M, Toe TT (2022) Cnn-lstm model stock forecasting based on an integrated attention mechanism. In: 2022 3rd international conference on pattern recognition and machine learning (PRML), IEEE, pp 403–408

40. Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499

41. Corizzo R, Rosen J (2024) Stock market prediction with time series data and news headlines: a stacking ensemble approach. J Intell Inf Syst 62(1):27–56

42. Livieris IE, Pintelas E, Pintelas P (2020) A CNN-LSTM model for gold price time-series forecasting. Neural Comput Appl 32:17351–17360

43. Jing N, Wu Z, Wang H (2021) A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Syst Appl 178:115019

44. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 31(7):1235–1270

45. Hochreiter S, Bengio Y, Frasconi P et al. (2001) Gradient flow in recurrent nets: the difficulty of learning long-term dependencies

46. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166

47. Lu Z, Du P, Nie J-Y (2020) VGCN-BERT: augmenting BERT with graph embedding for text classification. In: Advances in information retrieval: 42nd European conference on IR research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42, Springer, pp 369–382

48. Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 328–339

49. Li X, Chan S, Zhu X, Pei Y, Ma Z, Liu X, Shah S (2023) Are CHATGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In: Proceedings of the 2023 conference on empirical methods in natural language processing: industry track, pp 408–422

50. Araci D (2019) Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063

51. Rahate A, Walambe R, Ramanna S, Kotecha K (2022) Multi-modal co-learning: Challenges, applications with datasets, recent advances and future directions. Inf Fus 81:203–239

52. Noreen N, Palaniappan S, Qayyum A, Ahmad I, Imran M, Shoaib M (2020) A deep learning model based on concatenation approach for the diagnosis of brain tumor. IEEE Access 8:55135–55144

53. Bengio Y (2012) Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade: Second Edition, Springer, pp 437–478

54. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

55. Altieri M, Corizzo R, Ceci M (2024) Gap-LSTM: Graph-based autocorrelation preserving networks for geo-distributed forecasting. IEEE Trans Neural Netw Learn Syst

56. Salman AG, Heryadi Y, Abdurahman E, Suparta W (2018) Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting. Proc Comput Sci 135:89–98

57. Pankratz A (2009) Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. John Wiley & Sons, Hoboken

58. Chen T (2014) Introduction to boosted trees. Univ Wash Comput Sci 22(115):14–40