

Received March 3, 2022, accepted March 16, 2022, date of publication March 21, 2022, date of current version March 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160797

How to Handle Data Imbalance and Feature Selection Problems in CNN-Based Stock Price Forecasting

ZİNNET DUYGU AKŞEHİR^{ID} AND ERDAL KILIÇ^{ID}

Department of Computer Engineering, Ondokuz Mayıs University, 55139 Samsun, Turkey

Corresponding author: Zinnet Duygu Akşehir (duygu.aksehir@bil.omu.edu.tr)

ABSTRACT Stock market forecasting is a time series problem that aims to predict possible future prices or directions of an index/stock. The stock data contains high uncertainty and is influenced by too many factors; hence it isn't easy to achieve the goal by traditional time series methods. In literature, the convolutional neural networks (CNN) models were used for stock market forecasting and gave successful results. But, data imbalance due to labeling and feature selection problems were seen when considering these models. Hence, this study proposed a new rule-based labeling algorithm and a new feature selection approach to solve the issues. In addition, a CNN-based model, which was presented to predict the next day's trade action of stocks in the Dow30 index, was constructed to check the effectiveness of the data labeling and the feature selection approach. Different image-based input variable sets were created using technical indicators, gold, and oil price data to feed the CNN model. The prediction performance of CNN models was compared with other studies in the literature. The experimental results showed that the CNN prediction model, which uses the proposed feature selection and labeling approaches in this study, performs 3-22% higher accuracy than the CNN-based models taking part in other studies. Also, the labeling approach proposed is more successful than Chen and Huang's data weighting approach to solve the stock data imbalance problem. This algorithm reduced the ratio between labeled data from 15 times to 1.8 times.

INDEX TERMS CNN model, feature selection, labeling, stock prediction.

I. INTRODUCTION

Implementing a forecasting model by analyzing current data and predicting future behavior has become one of the most important issues in the academic and business world, especially in the money/financial markets that provide a return on profits. Money markets consist of financial assets such as stocks, indices, futures contracts, and all these assets are treated as financial time series. Many political, social, and economic factors affect financial assets, such as uncertainty and volatility in markets, political events, general economic situation, movements in other country markets, and investors' expectations. All these factors cause nonlinear behavior on the prices of financial assets. Therefore it is too hard to predict future behavior in financial time series. In this direction, it has been observed that deep learning methods have come to the forefront in recent years, with the frequent use of regression,

decision trees, support vector machines, and artificial neural networks from artificial intelligence techniques.

A. LITERATURE REVIEW

Stock market forecasting is a time series problem that estimates possible future direction or price value based on historical price data. But, the fact that the stock market data is affected by too many factors and is not linear makes it challenging to perform this estimation effectively with traditional time series methods. However, the autoregressive integrated moving averages (ARIMA) model is selected from the time series methods within the scope of stock market/share forecasting, and the success of this model is compared with other methods [1]–[4]. Selvin *et al.* estimated the prices of stocks belonging to three companies listed on the NSE stock market. Recurrent neural networks (RNN), long short-term memory (LSTM), and convolutional neural networks (CNN) were used from deep learning methods to perform the prediction. Besides, ARIMA models were used for time series

The associate editor coordinating the review of this manuscript and approving it for publication was Li He^{ID}.

analysis methods. When they compared the performance of the four models, they found that the deep learning models were more successful than the ARIMA model [4]. In addition to time series methods, machine learning methods such as support vector machines, decision trees, artificial neural networks are also applied to stock price prediction problems [1], [2], [5]–[8]. Vijn *et al.* estimated the next-day closing price of five stocks on the New York stock market using artificial neural networks (ANN) and random forest methods from machine learning techniques. When they evaluated the performance of the prediction models, they saw that the ANN model was more successful [5]. In another study, Parmar *et al.* estimated the future value of a company's stock using LSTM and regression methods. They found that LSTM was more successful among the proposed models [9].

In contrast to statistical and machine learning methods within stock market forecasting, it was observed that deep learning methods have become prominent [10]–[13] in recent years, and LSTM is frequently preferred among these methods [14]–[18]. The stock market data is a time series, and LSTM successfully performs the learning process with the inputs in the time series. Hossain *et al.* proposed a hybrid model to estimate the closing price of the next day's S&P500 index. This proposed hybrid model consists of LSTM and gated recurrent unit (GRU) methods [15]. In another study, Du *et al.* proposed a forecasting model with the LSTM method to predict the next day's Apple stock closing price. They used two different sets of input variables for the proposed model. When comparing the performance of models, they saw that the model created using multiple variables was more successful [16]. In another study, Ji *et al.* proposed a hybrid model (IPSO-LSTM) for estimating the Australian stock market (ASM) index. This hybrid model consists of a new particle swarm optimization (PSO) and LSTM method. They observed that the proposed model is more successful than support-vector regression, LSTM, and PSO-LSTM models [17].

In recent years, CNN method, which is widely used in image recognition due to its particularly significant pattern recognition ability, has also been applied in the field of financial forecasting, and its scope expanded [19]–[24]. Gunduz *et al.* predicted the daily movement direction of the three stocks in BIST30 with the CNN method. They investigated the effect on the forecasting model of the technical indicators calculated from gold and dollar prices. As a result of experiments, it was shown that the use of dollar-gold attributes in addition to price attributes improves classification performance for all three stocks [19]. Alhazbi *et al.* investigated whether external factors such as oil price affected the stock market direction and predicted the daily movement of the Qatar Stock Exchange using the CNN method. They stated that adding these external factors to the stock market data increases the model performance [20]. In another study, Sezer and Ozbayoglu proposed a new prediction model called CNN-TA that uses 2-D CNN based on image processing in their research. They aimed to predict the buy/sell/hold

position of Dow30 stocks' prices and Exchange-Traded Funds (ETF) prices for the next day with the proposed model. The financial data were converted into 15×15 images for the input variables of the CNN-TA model. In addition, a sliding window training-test approach is adopted for the proposed model. They evaluated the performance of the CNN-TA model for the last ten years' data [21]. In another CNN-based study using the labeling algorithm in [21], Chen and Huang [22] proposed two different labeling algorithms to predict the future S&P500 index. Technical indicators, gold prices, gold price volatility index, crude oil, and crude oil price volatility index were determined as input features of the prediction model. They compared the performance of their proposed models, called CNN8 and LSTM8, with the CNN-TA [21] model in terms of accuracy, precision, recall, and F1 score metrics, and these models outperformed the CNN-TA model. Also, it has been concluded that considering oil prices, oil volatility index, gold prices, and gold volatility indices instead of considering only technical indicators as input features of the model positively affect the stock market forecasting problem. Sim *et al.* [23] pointed out that using many technical indicators did not positively affect the prediction models. Another model called TI-CNN based on CNN and technical indicators have been proposed by Chandar [24] to estimate the stock trade action. He calculated ten technical indicators using daily stock prices for the proposed model and used Gramian Angular Field (GAF) method to obtain images from these technical indicators. The performance of the proposed model was compared with three earlier studies [21], [25], [26] in the literature. As a result of the comparisons, he saw that the TI-CNN model was better than the other models.

It is seen that the determination of the input variables of the model in stock price estimation studies, in other words, the feature selection process, directly affects the model performance. Within this context, most of the studies use the technical indicators and stocks historical data. However, recent studies argue that using these indicators does not affect the model since technical indicators show similar behavior to the closing prices of stocks [22], [23]. In this context, it has been proposed that using external factors such as gold and oil price may positively affect the forecasting model's performance.

Another factor affecting the prediction model's performance is the imbalance dataset. This problem is encountered in many fields such as image classification [27], speech recognition [28], and natural language processing [29]. This problem also occurs in stock price prediction studies [22], [30], [31]. There are various approaches to handle the imbalance problem in the literature. Some studies use complex methods, others use more straightforward approaches such as copying [32] and weighting data [22] to solve this problem.

In this study, we proposed a CNN-based model to predict the next day's trade actions of the stocks in the Dow30 index. While developing this model, a rule-based labeling method was proposed to solve the stock data imbalance problem.

We also analyzed and decided the input parameters that positively affect the performance of the stock price prediction model to provide the basis for the feature selection approach.

B. MOTIVATION AND CONTRIBUTIONS

The prediction models in 2D-CNN-based stock price/market forecasting studies achieved successful results in recent years. However, there are still some problems in issues such as data imbalance and feature selection that need to be corrected. Therefore, the motivation of this study is to identify these problems and find solutions. The contributions of this research are as follows:

- It was seen that using only technical indicators for the prediction model did not have a positive effect on the model performance. Within this scope, external factors such as gold and oil prices were used for the proposed model and created sets of different input variables to examine the effect on the model.
- Each trading data was labeled as buy/sell/hold using stock price data. Then, we created images of each trading day for different sets of input variables. To solve the data imbalance problem with a simpler (straightforward) approach, we examined the labels of the stock's last five days. We proposed a new labeling algorithm based on some rules to predict the next day's trade action within this scope.
- We referenced the input variables used in [21]. In this study, we investigated the necessity of the determined period values (6-20 days) for the 15 technical indicators. Considering other studies in literature, experiments conducted, and practical applications, it was observed that the use of 7, 14, and 21-day period values positively affected the model's success.

As a result of these contributions, we observed that the proposed stock's price prediction model, outperforms other studies [21], [22], [24] in the literature.

C. ORGANIZATION

The remaining parts of the study are organized as follows: In Section II, the proposed CNN model and how the input dataset for this model is generated are detailed. In this context, information about the dataset, labeling algorithm, feature selection, and image creation are discussed. The experimental results from the study are detailed in Section III. The last section consists of the conclusion and future works.

II. METHOD

In this study, we proposed two approaches to solve the data imbalance problem and determine more meaningful features for the model. Accordingly, we observed the performances of the approaches on a CNN-based model for forecasting the next day's trade action (buy, sell, hold) of Dow30 stocks.

A. DATASET

The daily price data of Dow30 stocks were obtained between 01/01/2008 to 15/12/2021 from finance.yahoo.com.

This dataset includes the open, close, adjusted close, high, low, and volume values of these stocks. The adjusted closing price data in the dataset was used to label the images as buy, sell or hold, and the other price data (close, high, low, volume) were used to calculate the technical indicator values. In addition, the daily closing prices of gold and oil in the determined period were obtained from the same website. We used the sliding window, and the traditional (standard) approaches for training and test processes in the study. In the sliding window approach, we chose the first five-year period for training and the following year for testing, i.e., the training period: 2008-2012 and the testing period: 2013. Then we shifted the training and testing periods by one year, i.e., training period: 2009-2013 and the testing period: 2014. In the traditional approach, 70% of the obtained stock data was used for training and the remaining for the test process. indicator values.

B. LABELING

After creating the dataset for the model, each trading day was labeled as "Buy," "Sell," and "Hold" with the labeling algorithm proposed by Sezer and Ozbayoglu. In this proposed algorithm, the 11-day sliding window approach is used. Labeling was realized according to the midpoint of the 11-day window size, in other words, the closing price of the 6th day. If the 6th day's closing price was the highest in this 11-day window, it was labeled "Sell". If the 6th day's closing price was the lowest, it was marked "Buy". Otherwise, it was marked as "Hold".

We labeled the used datasets in this study for both approaches with the labeling algorithm proposed in [21]. Then, we examined the labels of the datasets to be used in the sliding window approach. We found that the number of data labeled as Hold in the dataset was approximately 15 times those labeled as Buy and Sell. We proposed a new rule-based labeling algorithm based on the labeling algorithm given in [21] to solve the data imbalance problem. The proposed rule-based labeling algorithm is detailed following and given in Algorithm 1:

- Firstly, the labeling algorithm proposed in [21] was applied to the dataset. Thus, each trading day in the dataset is labeled as "Buy", "Sell" or "Hold".
- Then, these labeled data are combined according to the past five trading days. The obtained label values resulting from the merging were evaluated and relabeled the data for the past five days within the following rules:
 - If the count of "Buy" is 1 and "Sell" is 0 in the label values of the past five days, this data is labeled as "Buy".
 - If the count of "Sell" is 1 and "Buy" is 0 in the label values of the past five days, this data is labeled as "Sell".
 - If the label values of the past five days contain ['Buy', 'Hold', 'Hold', 'Sell'] or ['Buy', 'Hold', 'Hold', 'Hold', 'Sell'], this data is labeled as "Sell".

Algorithm 1 Proposed Rule Based Labeling Algorithm

```

1: calculate label (Buy/Sell/Hold) using labeling algorithm
   of [21] study
2: sublist1 = ['Buy', 'Hold', 'Hold', 'Sell']
3: sublist2 = ['Buy', 'Hold', 'Hold', 'Hold', 'Sell']
4: sublist3 = ['Sell', 'Hold', 'Hold', 'Buy']
5: sublist4 = ['Sell', 'Hold', 'Hold', 'Hold', 'Buy']
6: create labels[] mergering label of last five day's
7: if ((labels.count('Buy') == 1) & (labels.count('Sell')
   == 0)) then
8:   result = 'Buy'
9: else if ((labels.count('Sell') == 1) &
   (labels.count('Buy') == 0)) then
10:  result = 'Sell'
11: else if ((is_sublist(sublist1, labels) | is_sublist(sublist2,
   labels)) then
12:  result = 'Sell'
13: else if ((is_sublist(sublist3, labels) | is_sublist(sublist4,
   labels)) then
14:  result = 'Buy'
15: else
16:  result = 'Hold'
17: end if
    
```

- If the label values of the past five days contain ['Sell', 'Hold', 'Hold', 'Buy'] or ['Sell', 'Hold', 'Hold', 'Hold', 'Buy'], this data is labeled as "Buy".
- In other cases, the data is labeled as "Hold".

The performance of the proposed labeling algorithm is discussed in Section III.

C. FEATURE SELECTION

It is essential to determine the input variables that the model will use to predict the next day's trade action of stock prices. For this purpose, it is necessary to analyze the factors affecting stock prices. Most of the proposed forecasting models use technical indicators in addition to stock prices. However, in some of the studies in the literature [22], [23], it has been stated that these indicators do not positively affect the performance of the forecasting model because some of the technical indicators behave as stock's closing prices. In addition, it has been stated that external factors such as gold and oil price may be more effective on a prediction model performance. We performed various experiments and analyzed practical applications within this scope to create different input variable sets. The methods of their creation are given following:

- Firstly, 15 technical indicator (RSI, Williams's % R, WMA, EMA, SMA, HMA, Triple EMA, CCI, CMO, MACD, PPO, ROC, CMFI, DMI and PSI) feature values for different intervals of 6-20 days, used in the CNN-TA model [21], were handled. Then, gold and oil price data were added to these features.

TABLE 1. Image dimensions and properties.

Dimension	Properties
5x11	11 attributes were determined after the attribute selection process and combined the features of the last five days.
5x13	Gold and oil prices, in addition to the 11 attributes, were determined after the attribute selection process and combined the features of the last five days.
15x15	7, 14, and 21-day period values for 15 technical indicators and combined the features of the last five days.
15x17	Gold and oil prices, in addition to 7, 14, and 21-day period values for 15 technical indicators and combined the features of the last five days.
75x15	6-20 period values for 15 technical indicators and combined the features of the last five days.
75x17	Gold and oil prices, in addition to 6-20 period values for 15 technical indicators and combined the features of the last five days.

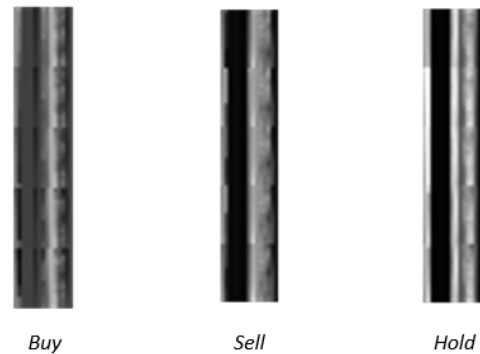


FIGURE 1. Labeled sample images of the 75 × 15 dimensions.

- Fifteen different period values for 15 technical indicators in CNN-TA were investigated to decide which period is meaningful for the proposed forecasting model. As a result, the use of only 7, 14, and 21-day period values of these technical indicators were chosen. Then, gold and oil price data were added to these features.
- The correlation-based feature selection method was applied by considering 15 different periods of each technical indicator used in the CNN-TA model. As a result, 11 feature values were found to be more significant. Then, gold and oil price data were added to these features.

D. IMAGE CREATION

According to the new labeling algorithm, various experiments were done to determine the stock's next day's trading action. As a result of these experiments, we observed that it is more successful in considering the last five days' features. These features include technical indicators and with or without gold and oil daily closing prices obtained from Yahoo Finance. We constructed a matrix in which columns represent features and rows represent daily price by selecting period.

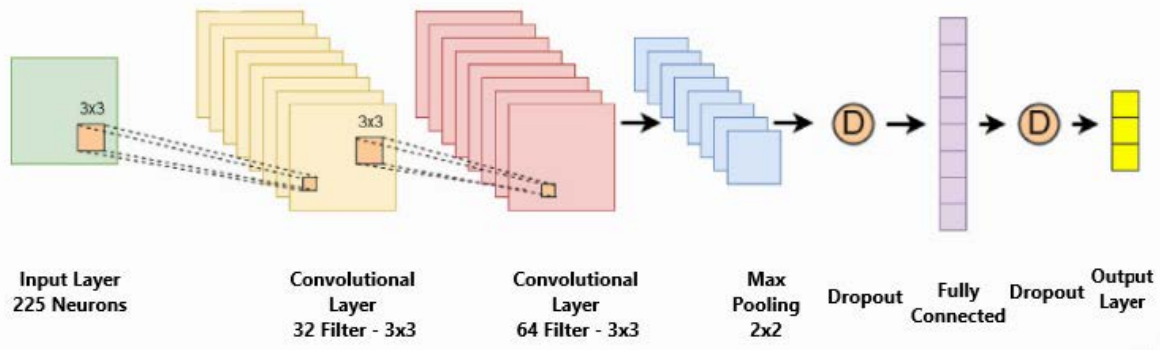


FIGURE 2. CNN structure [21].

For example, to build a 15×15 dimension matrix, first 15 technical indicators values were calculated for 7, 14, and 21-day periods. By using these calculated values, a 3×15 matrix was obtained for one trading day. When merging calculated these values for the last five days, this matrix dimension became 15×15 . Then this matrix was converted to an image. With the same thinking, all images can be created. These created images dimensions and properties are given in Table 1.

Then, created images labeled according to Algorithm 1 to determine the stock's next day's trade action value. Also, Figure 1 shows sample images with a dimension of 75×15 obtained according to the trade action label. The stocks' trade action of the next day was estimated using the images and the CNN model detailed in the following subsection.

E. CNN MODEL

In this study, to test the performance of labeling and feature selection approaches the CNN architecture proposed by Sezer and Ozbayoglu was used. This model architecture was given in Figure 2 and adapted according to our created different input image sets. The proposed CNN architecture consists of 8 layers: input layer, two convolutional layers, a max pooling, two dropouts, fully connected layer, and output layer. For the first two convolution layers where the convolution operation is performed, filter size 3×3 and the number of filters were selected as 32 and 64, respectively. The convolution operation on two-dimensional images is performed using Equation 1. K and I denote the kernel and input image in the equation, respectively. After these layers, the max-pooling layer with a 2×2 filter size was used to select the highest values in the convolutional matrix. In addition, two dropout layers (0.25, 0.5) were added to prevent overfitting. By using these layers, a deep neural network architecture is built in this study. Equation 2 provides the details about the neural network architecture. W , x , and b denote the weights, input and bias in the equation, respectively. The fully connected layer was used to flatten and connect the neurons to the next layer.

TABLE 2. Sample of confusion matrix.

		Predicted Class	
		Negative	Positive
Actual Class	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Finally, the softmax activation function, given in Equation 3 (y denotes output), was selected as the activation function in the output layer to perform the classification process [21].

$$S(i, j) = (I * K)(i, j) \\ = \sum_m \sum_n I(m, n)K(i - m, j - n) \quad (1)$$

$$e_i = \sum_j W_{i,j}x_j + b_i \quad (2)$$

$$y = \text{softmax}(e) \quad (3)$$

F. PERFORMANCE MEASURE

The confusion matrix given as an example in Table 2 is frequently used to evaluate the model's performance within the scope of classification problems. The information included in this matrix refers to follows [33]:

- **True Positive (TP)**: It is the number of samples that are actual class value is "positive" and labeled the "positive" class as a result of the classification.
- **False Positive (FP)**: It is the number of samples that are actual class value is "negative" and labeled the "positive" class as a result of the classification.
- **True Negative (TN)**: It is the number of samples that are actual class value is "negative" and labeled the "negative" class as a result of the classification.
- **False Negative (FN)**: It is the number of samples that are actual class value is "positive" and labeled the "negative" class as a result of the classification.

By using these TP, FP, FN, and TN values in the matrix, the accuracy, recall, precision, F1 score metrics given in Equations 4-7 are calculated, respectively. It is not sufficient

TABLE 3. Results of the CNN model using the sliding window training-test approach.

Test Dataset	Performance Metrics		Dimensions of Input Images					
			5x11	5x13	15x15	15x17	75x15	75x17
Ten years data	Accuracy (%)		76.74	76.82	80.23	78.85	80.23	78.62
	F1 score	Buy	0.78	0.77	0.81	0.80	0.81	0.80
		Hold	0.77	0.78	0.81	0.78	0.81	0.79
		Sell	0.24	0.73	0.78	0.78	0.78	0.76
AAPL	Accuracy (%)		75	75.20	77.24	75.81	77.03	78.46
	F1 score	Buy	0.80	0.79	0.73	0.75	0.73	0.74
		Hold	0.76	0.79	0.81	0.78	0.81	0.82
		Sell	0.68	0.63	0.74	0.73	0.73	0.76
TRV	Accuracy (%)		73.78	72.76	75.61	74.59	77.03	79.67
	F1 score	Buy	0.79	0.77	0.79	0.78	0.80	0.83
		Hold	0.72	0.71	0.72	0.69	0.75	0.76
		Sell	0.75	0.71	0.76	0.76	0.76	0.81
INTC	Accuracy (%)		75.20	73.17	74.80	73.58	80.89	76.42
	F1 score	Buy	0.81	0.74	0.76	0.73	0.82	0.76
		Hold	0.75	0.75	0.74	0.73	0.81	0.77
		Sell	0.70	0.70	0.74	0.74	0.78	0.76
GS	Accuracy (%)		73.17	72.36	75.81	71.54	77.03	72.15
	F1 score	Buy	0.74	0.67	0.79	0.77	0.81	0.77
		Hold	0.74	0.76	0.78	0.72	0.78	0.72
		Sell	0.71	0.69	0.70	0.67	0.73	0.68
WBA	Accuracy (%)		76.63	74.39	77.06	78.66	80.49	78.05
	F1 score	Buy	0.73	0.78	0.72	0.78	0.81	0.76
		Hold	0.75	0.73	0.77	0.76	0.80	0.76
		Sell	0.82	0.72	0.82	0.83	0.80	0.83
CSCO	Accuracy (%)		74.5	73.58	78.05	77.85	79.47	76.02
	F1 score	Buy	0.75	0.75	0.82	0.79	0.80	0.72
		Hold	0.72	0.74	0.78	0.76	0.80	0.75
		Sell	0.78	0.71	0.74	0.33	0.79	0.76
NKE	Accuracy (%)		72.36	73.78	74.39	74.19	76.22	74.80
	F1 score	Buy	0.73	0.73	0.78	0.76	0.76	0.77
		Hold	0.72	0.75	0.75	0.73	0.75	0.74
		Sell	0.73	0.73	0.71	0.74	0.79	0.74
MMM	Accuracy (%)		72.56	69.11	74.80	75.20	74.39	77.24
	F1 score	Buy	0.74	0.71	0.74	0.75	0.74	0.78
		Hold	0.71	0.70	0.74	0.74	0.73	0.77
		Sell	0.73	0.65	0.76	0.78	0.77	0.76
KO	Accuracy (%)		73.17	71.95	75.61	75.41	79.47	78.25
	F1 score	Buy	0.76	0.76	0.77	0.79	0.80	0.79
		Hold	0.69	0.71	0.75	0.73	0.78	0.78
		Sell	0.75	0.69	0.76	0.75	0.80	0.79
AXP	Accuracy (%)		70.12	72.36	79.07	74.19	78.86	75.81
	F1 score	Buy	0.77	0.78	0.79	0.78	0.83	0.83
		Hold	0.67	0.74	0.80	0.71	0.78	0.73
		Sell	0.68	0.65	0.78	0.75	0.77	0.74

TABLE 4. Comparison with other studies.

		Proposed model	CNN-TA [21]	CNN8 [22]
Accuracy (%)		80.23	58	78
Precision	Buy	0.78	0.22	0.30
	Hold	0.83	0.95	0.98
	Sell	0.78	0.18	0.29
Recall	Buy	0.84	0.80	0.92
	Hold	0.80	0.55	0.77
	Sell	0.77	0.81	0.77
F1 score	Buy	0.81	0.34	0.45
	Hold	0.81	0.70	0.86
	Sell	0.78	0.29	0.43

to compare models only in terms of accuracy metric. For this reason, the F1 score metric, which is calculated with the harmonic average of the recall and precision, should be used.

$$accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{4}$$

$$recall = \frac{TP}{(TP + FN)} \tag{5}$$

TABLE 5. Comparison with TI-CNN model.

		Proposed model	TI-CNN [24]
Accuracy (%)		80.23	77.75
AAPL	Accuracy (%)	77.03	65
	F1 score	0.76	0.68
GS	Accuracy (%)	77.03	78
	F1 score	0.77	0.74

$$precision = \frac{TP}{(TP + FP)} \tag{6}$$

$$F1\ score = \frac{2 \cdot precision \cdot recall}{(precision + recall)} \tag{7}$$

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

Datasets with six different image dimensions were created for the stocks in the Dow30 index with the determined features and the proposed labeling algorithm. Then, the trade action

TABLE 6. Results of the CNN model using the traditional training-test approach.

Test Dataset	Performance Metrics		Dimensions of Input Images					
			5x11	5x13	15x15	15x17	75x15	75x17
Ten years data	Accuracy (%)		69.61	69.37	71.26	73.51	72.66	73.48
	F1 score	Buy	0.69	0.69	0.71	0.75	0.72	0.75
		Hold	0.72	0.73	0.74	0.77	0.76	0.76
AAPL	Accuracy (%)		75.81	74.39	81.10	81.5	85.37	87.80
	F1 score	Buy	0.78	0.73	0.82	0.89	0.88	0.89
		Hold	0.79	0.79	0.85	0.84	0.89	0.89
TRV	Accuracy (%)		67.89	65.04	66.87	62.20	64.63	61.59
	F1 score	Buy	0.71	0.68	0.68	0.63	0.66	0.62
		Hold	0.67	0.68	0.67	0.68	0.66	0.64
INTC	Accuracy (%)		68.29	72.15	79.27	77.85	83.54	82.32
	F1 score	Buy	0.69	0.71	0.79	0.83	0.83	0.81
		Hold	0.69	0.76	0.81	0.80	0.85	0.83
GS	Accuracy (%)		75.41	75.61	79.88	83.74	82.11	80.69
	F1 score	Buy	0.70	0.69	0.77	0.85	0.76	0.77
		Hold	0.80	0.79	0.84	0.87	0.86	0.84
WBA	Accuracy (%)		70.73	66.87	64.23	60.16	62.80	61.99
	F1 score	Buy	0.69	0.68	0.60	0.56	0.59	0.59
		Hold	0.71	0.70	0.64	0.62	0.65	0.65
CSCO	Accuracy (%)		71.34	69.51	73.98	77.44	80.08	79.27
	F1 score	Buy	0.74	0.70	0.76	0.83	0.82	0.81
		Hold	0.72	0.72	0.75	0.79	0.82	0.81
NKE	Accuracy (%)		74.19	70.73	79.88	78.89	81.50	80.08
	F1 score	Buy	0.77	0.75	0.82	0.80	0.83	0.80
		Hold	0.77	0.75	0.81	0.82	0.83	0.82
MMM	Accuracy (%)		66.87	71.90	72.76	74.39	77.44	80.49
	F1 score	Buy	0.68	0.74	0.71	0.83	0.79	0.83
		Hold	0.69	0.76	0.75	0.78	0.79	0.81
KO	Accuracy (%)		72.56	73.17	75.20	77.64	76.22	77.24
	F1 score	Buy	0.74	0.79	0.78	0.85	0.79	0.80
		Hold	0.74	0.76	0.76	0.80	0.77	0.79
AXP	Accuracy (%)		73.98	73.58	81.71	81.10	83.74	83.74
	F1 score	Buy	0.76	0.76	0.87	0.90	0.88	0.88
		Hold	0.74	0.76	0.82	0.83	0.84	0.83
AXP	F1 score		0.71	0.66	0.77	0.67	0.78	0.80

of the stocks the next day’s closing price was estimated with the CNN model, which uses both sliding window and traditional training-test approaches. In addition, test datasets were created containing the last ten-year data of the stocks and the previous one-year data of the determined ten stocks (AAPL, TRV, INTC, GS, WBA, CSCO, NKE, MMM, KO, AXP). The performance of the CNN model was evaluated using these datasets and compared with other studies in the literature [21], [22], [24].

The obtained results are discussed in the next two subsections. In the first subsection, the evaluation of the labeling algorithm is evaluated. In the second subsection, the performance of the CNN model using the feature selection method and labeling algorithm is discussed.

A. EVALUATION OF THE LABELING ALGORITHM

When the proposed labeling algorithm in [21] is used for the five-year training dataset of Dow30 stocks within the scope

of the sliding window training testing approach, the number of data labeled as buy, sell and hold is approximately 2200, 2180, and 30600, respectively. It was seen that the number of data labeled as Hold is approximately 15 times those labeled as Buy and Sell. When the rule-based labeling algorithm in this study was used for the same dataset, the number of data labeled as buy, sell, hold is approximately 9300, 9200, and 16400, respectively. These two results show that the proposed labeling algorithm and feature selection approaches reduced the data imbalance from 15 times to 1.8 times. Hence the proposed approaches are very effective. In addition, the data weighting approach used in the study [22] to solve the data imbalance problem in [21] does not increase the number of data but improves the classification performance.

B. EVALUATION OF THE CNN MODEL PERFORMANCE

The obtained results for datasets with six different image dimensions of the CNN model using the sliding window

training-test approach were given in Table 3. When evaluating the F1 score and accuracy metrics values for this approach, it was seen that higher values were obtained on 75×15 dimensions images. Therefore, the CNN model performed more successfully on these image dataset. In addition, this successful model was compared CNN-TA and CNN8, and the results were given in Table 4. As a result of the comparisons, the proposed prediction model outperforms the CNN-TA model in terms of performance metrics due to the proposed new labeling algorithm and feature selection approach. While the success of the CNN-TA model was 58%, the model's success was improved by approximately 22%, in the direction of the approaches suggested in this study, was reached an accuracy of 80.23%. In addition, recall and F1 score metrics, especially for Buy and Sell, were improved by approximately 50%. The CNN8 model was also compared with the prediction model in this study. As a result of the comparisons, it was seen that the proposed rule-based approach in this study improved the model performance more than CNN8 in terms of accuracy, precision, recall and F1 score metrics. The proposed model was also compared with the TI-CNN model [24]. The study also used the training-test approach and the labeling algorithm in [21]. This model estimated the next day's trade action of some stocks in NASDAQ and NYSE. The comparison results were given in Table 5. The results showed that the proposed model was more successful than TI-CNN model. In addition, the model outperformed TI-CNN in terms of accuracy and F1 score metrics for AAPL stock and F1 score for GS. Table 6 shown the traditional training-test approach results for datasets with six different image dimensions. It was seen that the CNN model performed more successful predictions on 75×15 and 75×17 dimensions images. In addition, it was observed that the gold and oil prices positively affected the prediction performance.

Table 6 shown the traditional training-test approach results for datasets with six different image dimensions. It was seen that the CNN model performed more successful predictions on 75×15 and 75×17 dimensions images. In addition, it was observed that the gold and oil prices positively affected the prediction performance.

In our stock trading system, most of the "Buy", "Sell", and "Hold" points are captured correctly by the proposed model. The main reason for this is that the number of the "Buy", "Sell", "Hold" points are nearly equal; hence the deep neural network catches the entry and exit points correctly. In other words, to be able to catch most of the "Buy" and "Sell" points (recall), the model generates true alarms for existent entry and exit points (precision). As a result, the standard annualized returns of the stock trading algorithm will be higher. Besides, compared with the other stock trading models in the literature, it will be better than the majority.

IV. CONCLUSION AND FUTURE WORKS

In this study, a new rule-based labeling algorithm was proposed to solve the imbalance problem in the dataset used in the stock price prediction. In addition, a feature selection

approach was proposed by investigating what should be considered when determining the input variables to be used in the CNN-based stock prediction model. To see the effect of the proposed labeling algorithm and feature selection approaches on the prediction model, a 2D-CNN-based deep neural network was used. While realizing this, both sliding window and traditional training-test approaches were used. When the obtained results are compared with the results of other prediction models in the literature, it has been observed that the proposed approaches had a positive effect on the model's prediction performance. These results prove that the proposed approaches are applicable within the scope of stock price prediction.

As future works, hybrid methods can increase the success of developed CNN-based approaches to predict stock prices. In particular, attention-based deep learning and deep q learning approaches may be applied to this field.

REFERENCES

- [1] B. Unal and C. Aladag, "Stock exchange prediction via long short-term memory networks," in *Proceedings Book*. Van, Turkey: Van Yuzuncu Yil University, 2019, p. 246.
- [2] S. Tekin and E. Canakoglu, "Analysis of price models in Istanbul stock exchange," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.
- [3] A. L. P. Selçuk, Ö. E. Yiğit, and Ö. Z. Ersoy, "Prediction of bist price indices: A comparative study between traditional and deep learning methods," *Sigma J. Eng. Natural Sci.*, vol. 38, no. 4, pp. 1693–1704, 2020.
- [4] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 1643–1647.
- [5] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock closing price prediction using machine learning techniques," *Proc. Comput. Sci.*, vol. 167, pp. 599–606, Jan. 2020.
- [6] K. J. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, nos. 1–2, pp. 307–319, Sep. 2003.
- [7] E. Altay and M. H. Satman, "Stock market forecasting: Artificial neural network and linear regression comparison in an emerging market," *J. Financial Manage. Anal.*, vol. 18, no. 2, p. 18, 2005.
- [8] R. Y. M. Li, S. Fong, and K. W. S. Chong, "Forecasting the REITs and stock indices: Group method of data handling neural network approach," *Pacific Rim Property Res. J.*, vol. 23, no. 2, pp. 123–160, May 2017.
- [9] I. Parmar, N. Agarwal, S. Saxena, R. Arora, S. Gupta, H. Dhiman, and L. Chouhan, "Stock market prediction using machine learning," in *Proc. 1st Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, Dec. 2018, pp. 574–576.
- [10] R. Halil and M. Demirci, "Predicting the Turkish stock market bist 30 index using deep learning," *Int. J. Eng. Res. Develop.*, vol. 11, no. 1, pp. 253–265, 2019.
- [11] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications: A survey," *Appl. Soft Comput.*, vol. 93, Aug. 2020, Art. no. 106384.
- [12] Z. Hu, Y. Zhao, and M. Khushi, "A survey of forex and stock price prediction using deep learning," *Appl. Syst. Innov.*, vol. 4, no. 1, p. 9, Feb. 2021.
- [13] W. Jiang, "Applications of deep learning in stock market prediction: Recent progress," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115537.
- [14] T. Gao, Y. Chai, and Y. Liu, "Applying long short term memory neural networks for predicting stock closing price," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2017, pp. 575–578.
- [15] M. A. Hossain, R. Karim, R. Thulasiram, N. D. B. Bruce, and Y. Wang, "Hybrid deep learning model for stock price prediction," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1837–1844.
- [16] J. Du, Q. Liu, K. Chen, and J. Wang, "Forecasting stock prices in two ways based on LSTM neural network," in *Proc. IEEE 3rd Inf. Technol., Netw., Electron. Automat. Control Conf. (ITNEC)*, Mar. 2019, pp. 1083–1086.

- [17] Y. Ji, A. W.-C. Liew, and L. Yang, "A novel improved particle swarm optimization with long-short term memory hybrid model for stock indices forecast," *IEEE Access*, vol. 9, pp. 23660–23671, 2021.
- [18] D. M. Q. Nelson, A. C. M. Pereira, and R. A. de Oliveira, "Stock market's price movement prediction with LSTM neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1419–1426.
- [19] H. Gunduz, Z. Cataltepe, and Y. Yaslan, "Stock market direction prediction using deep neural networks," in *Proc. 25th Signal Process. Commun. Appl. Conf. (SIU)*, May 2017, pp. 1–4.
- [20] S. Alhazbi, A. B. Said, and A. Al-Maadid, "Using deep learning to predict stock movements direction in emerging markets: The case of Qatar stock exchange," in *Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIoT)*, Feb. 2020, pp. 440–444.
- [21] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach," *Appl. Soft Comput.*, vol. 70, pp. 525–538, Apr. 2018.
- [22] Y.-C. Chen and W.-C. Huang, "Constructing a stock-price forecast CNN model with gold and crude oil indicators," *Appl. Soft Comput.*, vol. 112, Nov. 2021, Art. no. 107760.
- [23] H. S. Sim, H. I. Kim, and J. J. Ahn, "Is deep learning for image recognition applicable to stock market prediction?" *Complexity*, vol. 2019, pp. 1–10, Feb. 2019.
- [24] S. K. Chandar, "Convolutional neural network for stock trading using technical indicators," *Automated Softw. Eng.*, vol. 29, no. 1, pp. 1–14, May 2022.
- [25] S. Chen and H. He, "Stock prediction using convolutional neural network," in *Proc. IOP Conf. Mater. Sci. Eng.*, vol. 435, no. 1, Oct. 2018, Art. no. 012026.
- [26] B. Xu, D. Zhang, S. Zhang, H. Li, and H. Lin, "Stock market trend prediction using recurrent convolutional neural networks," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.* Hohhot, China: Springer, 2018, pp. 166–177.
- [27] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
- [28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.
- [29] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," 2018, *arXiv:1805.06201*.
- [30] S. W. Lee and H. Y. Kim, "Stock market forecasting with super-high dimensional time-series data using ConvLSTM, trend sampling, and specialized data augmentation," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113704.
- [31] X. Teng, T. Wang, X. Zhang, L. Lan, and Z. Luo, "Enhancing stock price trend prediction via a time-sensitive data augmentation method," *Complexity*, vol. 2020, pp. 1–8, Feb. 2020.
- [32] O. B. Sezer, A. M. Ozbayoglu, and E. Dogdu, "An artificial neural network-based stock trading system using technical analysis and big data framework," in *Proc. SouthEast Conf.*, Apr. 2017, pp. 223–226.
- [33] D. O. Sahin, S. Akleylek, and E. Kiliç, "LinRegDroid: Detection of Android malware using multiple linear regression models-based classifiers," *IEEE Access*, vol. 10, pp. 14246–14259, 2022.



ZİNNET DUYGU AKŞEHİR received the bachelor's degree in computer engineering from Karadeniz Technical University, Trabzon, in 2017, and the master's degree in computer engineering from Ondokuz Mayıs University, Samsun, in 2020. She is currently pursuing the Ph.D. degree in computational sciences. Her research interests include machine learning and data mining.



ERDAL KILIÇ received the bachelor's and master's degrees in electrical and electronic engineering from Karadeniz Technical University, Trabzon, in 1991 and 1996, respectively, and the Ph.D. degree in electrical and electronic engineering from Middle East Technical University, Ankara, in 2005. He is currently a Full Professor with the Department of Computer Engineering, Ondokuz Mayıs University. His research interests include neural networks, machine learning, and data mining.

...